2014

A Capability Maturity Model for Research Data Management.

Jian Qin
Syracuse University

Kevin Crowston
Syracuse University

Arden Kirkland

Recommended Citation

Qin, J., Crowston, K., & Kirkland, A. (2014). A Capability Maturity Model for Research Data Management. Syracuse, NY: School of Information Studies, Syracuse University.

A Capability Maturity Model for Research Data Management.

Description/Abstract

Objective: To support the assessment and improvement of research data management (RDM) practices to increase its reliability, this paper describes the development of a capability maturity model (CMM) for RDM. Improved RDM is now a critical need, but low awareness of – or lack of – data management is still common among research projects.

Methods: A CMM includes four key elements: key practices, key process areas, maturity levels, and generic processes. These elements were determined for RDM by a review and synthesis of the published literature on and best practices for RDM.

Results: The RDM CMM includes five chapters describing five key process areas for research data management: 1) data management in general; 2) data acquisition, processing, and quality assurance; 3) data description and representation; 4) data dissemination; and 5) repository services and preservation. In each chapter, key data management practices are organized into four groups according to the CMM's generic processes: commitment to perform, ability to perform, tasks performed, and process assessment (combining the original measurement and verification). For each area of practice, the document provides a rubric to help projects or organizations assess their level of maturity in RDM.

Conclusions: By helping organizations identify areas of strength and weakness, the RDM CMM provides guidance on where effort is needed to improve the practice of RDM.

Keywords

Capability Maturity Model, Research Data Management, Process Assessment, Performance Assessment

Disciplines

Library and Information Science

Creative Commons License

A CAPABILITY MATURITY MODEL FOR RESEARCH DATA MANAGEMENT

Jian Qin, Kevin Crowston, Arden Kirkland

SCHOOL OF INFORMATION STUDIES, SYRACUSE UNIVERSITY

CMM for RDM A Capability Maturity Model for Research Data Management

Jian Qin Kevin Crowston Arden Kirkland

School of Information Studies
Syracuse University
Syracuse, NY 13244
United States

Abstract

The broad goals of this project are to document, foster and promulgate best practices in research data management (RDM), practices that support research transparency and the replication of scientific results. We do so in order to cultivate a new generation of researchers and data managers who are both the best practice beneficiaries and contributors. Furthermore, as more organizations invest in RDM, it has become increasingly important for administrators, researchers, and managers to be able to evaluate RDM process for sustainability, efficiency, and effectiveness, which requires a baseline for comparison.

Objective: To support the assessment and improvement of research data management (RDM) practices to increase its reliability, this paper describes the development of a capability maturity model (CMM) for RDM. Improved RDM is now a critical need, but low awareness of – or lack of – data management is still common among research projects.

Methods: A CMM includes four key elements: key practices, key process areas, maturity levels, and generic processes. These elements were determined for RDM by a review and synthesis of the published literature on and best practices for RDM.

Results: The RDM CMM includes five chapters describing five key process areas for research data management: 1) data management in general; 2) data acquisition, processing, and quality assurance; 3) data description and representation; 4) data dissemination; and 5) repository services and preservation. In each chapter, key data management practices are organized into four groups according to the CMM's generic processes: commitment to perform, ability to perform, tasks performed, and process assessment (combining the original measurement and verification). For each area of practice, the document provides a rubric to help projects or organizations assess their level of maturity in RDM.

Conclusions: By helping organizations identify areas of strength and weakness, the RDM CMM provides guidance on where effort is needed to improve the practice of RDM.

Acknowledgement

This project is supported by a grant from the Interuniversity Consortium for Political and Social Research, which is part of a larger grant from the Alfred Sloan Foundation.

Cite as:

Qin, J., Crowston, K., & Kirkland, A. (2014). A Capability Maturity Model for Research Data Management. Syracuse, NY: School of Information Studies, Syracuse University.

Table of Contents

0. Introduction

Research in science, social science, and the humanities is increasingly data-intensive, highly collaborative, and highly computational at a large scale. The tools, content and social attitudes for supporting multidisciplinary collaborative research require "new methods for gathering and representing data, for improved computational support and for growth of the online community" (Murray-Rust, 2008). As a result, improved research data management (RDM) is now a critical need, with action needed across the data lifecycle: from data capture, analysis and visualization (Gray, 2007), through curation, sharing and preservation, to support for further discovery and reuse. To enable assessment and improvement of RDM practices that increase the reliability of RDM, this document presents a capability maturity model (CMM) for RDM.

Currently, RDM practices vary greatly depending on the scale, discipline, funding and type of projects. "Big science" research fields—such as astrophysics, geosciences, climate science and system biology—generally have established well-defined RDM policies and practices, with supporting data repositories for data curation, discovery and reuse. RDM in these disciplines often has significant funding support for the necessary personnel and technology infrastructure. By contrast, in most "small science" or humanities research (i.e., projects typically involving a single PI and a few students), RDM is less well developed. However, even in these fields, RDM practices are still critical: the data generated by these projects may be small on an individual level, but they can nevertheless add up to a large volume collectively (Carlson, 2006) and in aggregation can have more complexity and heterogeneity than those generated from big research projects.

The importance of RDM has been raised to a new level, as demonstrated by US National Science Foundation's renewed mandate that proposals include a data management plan. However, low awareness of—or indeed lack of—data management is still common among research projects, especially small science projects. This lack of awareness is affected by factors such as the type and quantity of data produced, the heritage and practices of research communities and size of research teams (Key Perspectives, 2010). Further complicating the discussion of practices, RDM is an interdisciplinary field: communities of practice involve researchers, information technology professionals, librarians and graduate students, each bringing their domain-specific culture and practices to bear on RDM. But as yet, the field lacks a conceptual model upon which practices, policies and performance and impact assessment can be based. Research projects need more concrete guidance to analyze and assess the processes of RDM. The goal of this document is to present the first steps towards development of such a model, in the form of a Capability Maturity Model (CMM) for RDM.

References

Carlson, S. (2006). Lost in a sea of science data. *The Chronicle of Higher Education*, 52: A35. Retrieved from http://chronicle.com/weekly/v52/i42/42a03501.htm

Gray, J. (2007). Jim Gray on eScience: A transformed scientific method. In: T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*, pp. 5-12. Redmond, WA: Microsoft Research. Retrieved from http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf

Key Perspectives. (2010). Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. *SCARP Synthesis Study, Digital Curation Centre*. Retrieved from http://www.dcc.ac.uk/scarp

Murray-Rust, P. (2008). Chemistry for everyone. *Nature*, 451, 648-651. Retrieved from http://www.nature.com/nature/journal/v451/n7179/full/451648a.html

0.1 Research Lifecycle and Data Management Lifecycle

Lifecycle is a term frequently used in our technology-driven society. Examples include information systems lifecycle, information transfer lifecycle, and many other variations depending on the domain for which the term lifecycle is used. In the research data management domain, this term is used in several contexts: research lifecycle, data lifecycle, data curation lifecycle, and data management lifecycle. Each version has a different emphasis but they are often related or overlap in one way or the other. A research lifecycle generally includes study concept and design, data collection, data processing, data access and dissemination, and analysis. As a research project progresses along the stages, different data will be collected, processed, calibrated, transformed, segmented or merged. Data at these stages go through one state to the next after certain processing or condition is performed on them. Some of these data are in the active state and may be changed frequently, while others such as raw data and analysis-ready datasets will be tagged with metadata for discovery and reuse. At each stage of this lifecycle, the context and type of research (Figure 1) can directly affect the types of data generated and requirements for how the data will be processed, stored, managed, and preserved.

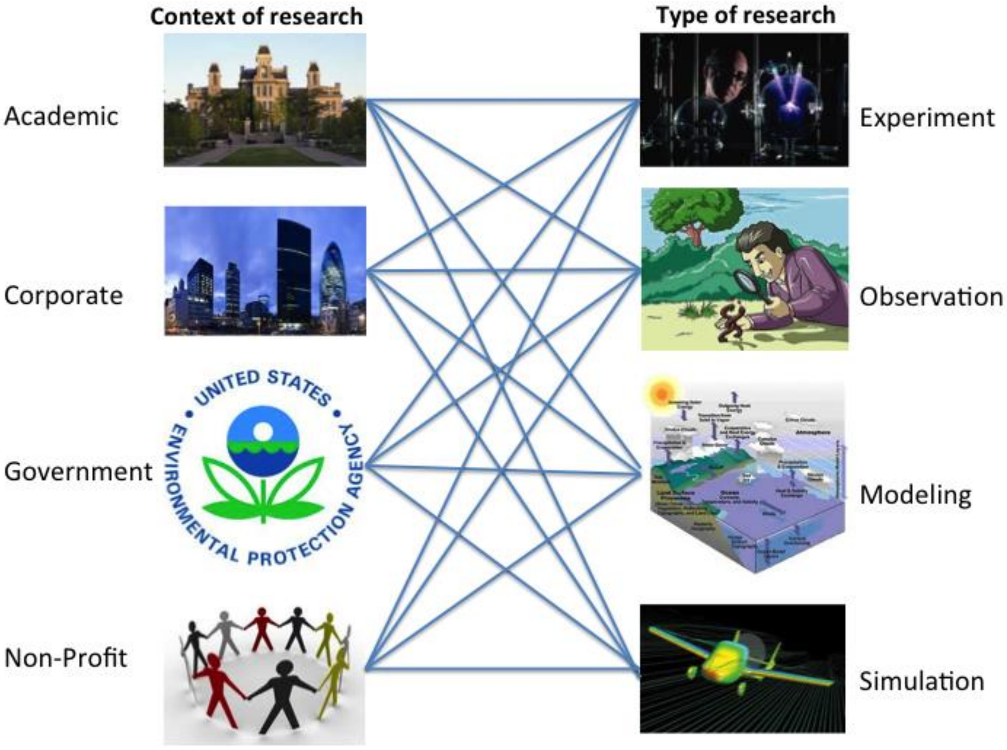

Figure 1. The contexts and types of research as well as their relations

For example, in the United States, national research centers such as the National Center for Atmospheric Research (NCAR, http://ncar.ucar.edu/) and the National Oceanic and Atmospheric Administration (NOAA, http://www.noaa.gov/) regularly collect data about the global ecosystems and process them into data products for scientific research and learning. The research lifecycle and data management lifecycle at this level will be different from those at the individual project level where teams of scientists have specific goals to solve specific problems. The scale of data and kinds of requirements for data management will vary along the stages of the whole research lifecycle. National research centers are publicly funded agencies and have the obligation of preserving and providing access to ecosystems data they collected. Hence generating data products and providing ways to discover and obtain data is crucial for them. Another example is the type of research projects carried out at academic institutions. These research projects may be funded by federal funding agencies or private foundations and can be collaborative among institutions or within a department/college of an institution. The data collected and generated from these projects are specialized and subject to the control and regulation of different data policies and compliance, which creates a different set of issues and requirements for data management and use/reuse from those generated by the national research centers.

Regardless of the context and nature of research, research data need to be stored, organized, documented, preserved (or discarded), and made discoverable and (re)usable. The amount of work and time involved in these processes is daunting, both intellectually intensive and costly. The personnel performing these tasks must be highly trained both in technology and in subject fields and able to effectively communicate between different stakeholders. In this sense, the lifecycle of research and data management is not only a technical domain but also a domain requiring effective management and communication. To be able to manage research data at community, institution, and project levels without reinventing the wheel, it is critical to build technical, communication, personnel, and policy capabilities at project and institutional levels and gradually evolve the maturity levels.

0.2 Background of the Capability Maturity Model

This document presents suggestions for assessing and improving research data management in the form of a capability maturity model. The original Capability Maturity Model (CMM) was developed at the Software Engineering Institute (SEI) at Carnegie Mellon University to support improvements in the reliability of software development organizations, that is, in their ability to develop quality software on time and within budget. More specifically, it was “designed to help developers to select process-improvement strategies by determining their current process maturity and identifying the most critical issues to improving their software quality and process” (Paulk et al., 1993, p. 19).

The model has evolved over time, but the basic structure remains roughly the same. It includes four key concepts: key practices, key specific and generic process areas and maturity levels. The development of the CMM was based on the observation that in order to develop software, organizations must be capable of reliably carrying out a number of key software development *practices* (e.g., eliciting customer needs or tracking changes to products), that is, they must be able to perform them in a consistent and predictable fashion. In the original CMM, these practices are clustered into 22 *specific process areas*, that is, “related practices in an area that, when implemented collectively, satisfy a set of goals considered important for making improvement in that area” (CMMI Product Team, 2006, Glossary). For example, eliciting customer needs is part of requirements development; tracking changes to products, part of configuration management. Achieving the goals is mandatory for good performance; the practices given are the expected (though not required) way to achieve those goals. The process areas are further grouped into four categories: support, project management, process management and engineering.

In addition to the specific process areas, those related specifically to software engineering, the SEI CMM included a set of *generic goals* and subgoals that describe the readiness of the organization to implement any processes reliably, namely:

1. achieve specific goals (i.e., the processes are performed),

2. institutionalize a managed process (i.e., the organization has policies for planning and performing the process, a plan is established and maintained, resources are provided, responsibility is assigned, people are trained, work products are controlled, stakeholders are identified, the processes is monitored and controlled, adherence to process standards is assessed and noncompliance addressed and the process status is reviewed with higher level management);

3. institutionalize a defined process (i.e., a description of the process is maintained and improvement information is collected),

4. institutionalize a quantitatively managed process (i.e., quantitative objectives are established and subprocess performance is stabilized), and

5. institutionalize an optimizing process (i.e., continuous process improvement is ensured and root causes of defects are identified and corrected).

As with the software-specific goals, these goals are required for a fully reliable organization; for each, there is a set of practices that are the expected though not required way to accomplish these goals.

References

CMMI Product Team. (2006). CMMI for Development Version 1.2. CMU/SEI-2006-TR-008. Pittsburgh, PA, USA: Carnegie Mellon Software Engineering Institute. Retrieved from http://repository.cmu.edu/sei/387

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. (1993). Capability maturity model, Version 1.1. IEEE Software, 10(4): 18-27. Retrieved from http://www.computer.org/csdl/mags/so/1993/04/s4018-abs.html

0.3 Research Data Management Maturity Levels

Perhaps the most well-known aspect of the CMM is five levels of *process or capability maturity*, which describe the level of development of the practices in a particular organization, representing the “degree of process improvement across a predefined set of process areas” and corresponding to the generic goals listed in the previous section. The initial level describes an organization with no defined processes: in the original CMM, meaning that software is developed (i.e., the specific software related goals are achieved), but in an ad hoc and unrepeatable way, making it impossible to plan or predict the results of the next development project. As the organization increases in maturity, processes become more refined, institutionalized and standardized, achieving the higher numbered generic processes and meaning that the organization can be assured of project results. The CMM thus described an evolutionary improvement path from ad hoc, immature processes to disciplined, mature processes with improved software quality and organizational effectiveness (CMMI Product Team, 2006, p. 535).

Our goal in this document is to lay out a similar path for the improvement of research data management. RDM practices as carried out in research projects similarly range from ad hoc to well-planned and well-managed processes (D'Ignazio & Qin, 2008; Steinhart et al., 2008). The generic practices described above provide a basis for mapping these maturity levels into the context of RDM, as illustrated in Figure 1 and described below.

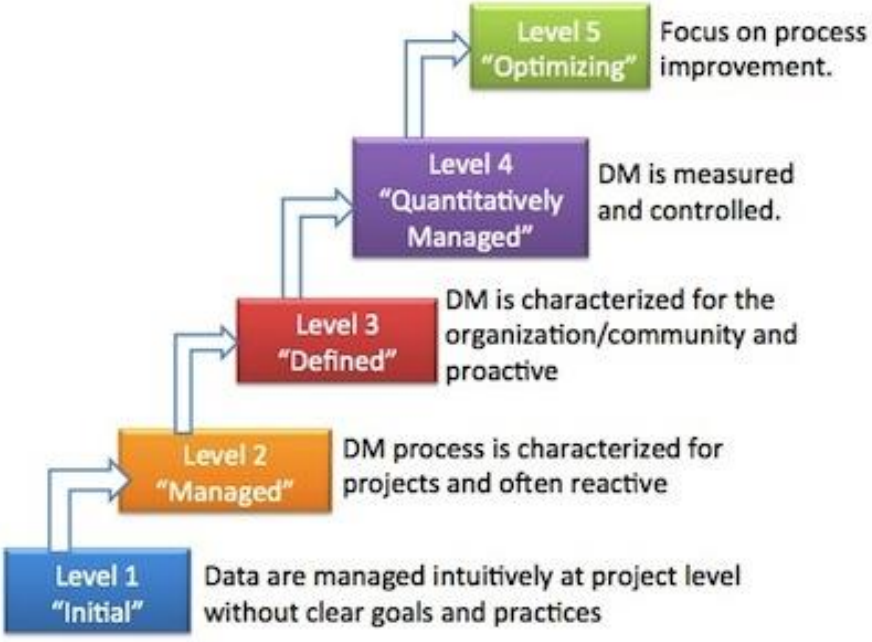
Level 5
"Optimizing"
Focus on process improvement.
Level 4
"Quantitatively Managed"
DM is measured and controlled.
Level 3
"Defined"
DM is characterized for the organization/community and proactive
Level 2
"Managed"
DM process is characterized for projects and often reactive
Level 1
"Initial"
Data are managed intuitively at project level without clear goals and practices

Figure 1. Capability maturity levels for research data management

0.3.1 Level 1: Initial

The initial level of the CMM describes an organization with no defined or stable processes. Paulk et al. describe this level thusly: "In an immature organization,… processes are generally improvised by practitioners and their managers during a project" (1993, p. 19). At this level, RDM is needs-based, ad hoc in nature and tends to be done intuitively. Rather than documented processes, the effectiveness of RDM relies on competent people and heroic efforts. The knowledge of the field and skills of the individuals involved (often graduate students working with little input) limits the effectiveness of data management. When those individuals move on or focus elsewhere, there is a danger that RDM will not be sustained; these changes in personnel will have a great impact on the outcomes (e.g., the data collection process will change depending on the person doing it), rendering the data management process unreliable.

0.3.2 Level 2: Managed

Maturity level 2 characterizes projects with processes that are managed through policies and procedures established within the project. At this level of maturity, the research group has discussed and developed a plan for RDM. For example, local data file naming conventions and directory organization structures may be documented. However, these policies and procedures are idiosyncratic to the project meaning that the RDM capability resides at the project level rather than drawing from organizational or community processes definitions. For example, in a survey of science, technology, engineering and mathematics (STEM) faculty, Qin and D'Ignazio (2010) found that respondents predominately used local sources to decide what metadata to create when representing their datasets, either through their own planning, in discussion with their lab groups or somewhat less so through the examples provided by peer researchers. Of far less impact were guidelines from research centers or discipline-based sources. Government requirements or standards also seemed to provide comparatively little help (Qin and D'Ignazio, 2010). As a result, at this level, developing a new project requires redeveloping processes, with possible risks to the effectiveness of RDM. Individual researchers will likely have to learn new processes as they move from project to project. Furthermore, aggregating or sharing data across multiple projects will be hindered by the differences in practices across projects.

0.3.3 Level 3: Defined

In the original CMM, "Defined" means that the processes are documented across the organization and then tailored and applied for particular projects. Defined processes are those with inputs, standards, work procedures, validation procedures and compliance criteria. At this level, an organization can establish new projects with confidence in stable and repeatable execution of processes, rather than the new project having to invent these from scratch. For example, projects at this level likely employ a metadata standard with best practice guidelines. Data sets/products are represented by some formal semantic structures (controlled vocabulary, ontology, or taxonomies), though these standards may be adapted to fit to the project. For example, the adoption of a metadata standard for describing datasets often involves modification and customization of standards in order to meet project needs.

In parallel to the SEI CMM, the RDM process adopted might reflect institutional initiatives in which organizational members or task forces within the institution discuss policies and plans for data management, set best practices for technology and adopt and implement data standards. For example, the Purdue Distributed Data Curation Center (D2C2, http://d2c2.lib.purdue.edu/) brings researchers together to develop optimal ways to manage data, which could lead to formally maintained descriptions of RDM practices. Level 3 organizations can also draw on research-community-based efforts to define processes. Examples include the Hubbard Brook Ecosystem Studies (http://www.hubbardbrook.org/), the Long Term Ecological Research Network (LTER, http://www.lternet.edu/) and Global Biodiversity Information Facility (GBIF, http://www.gbif.org/). Government requirements and standards in regard to research data are often targeted to higher level of data management, e.g., community level or discipline level.

0.3.4 Level 4: Quantitatively Managed

Level 4 in the original CMM means the processes have quantitative quality goals for the products and processes. The processes are instrumented and data are systematically collected and analyzed to evaluate the processes.

For the level 3 capability maturity to reach level 4, the quantitatively managed RDM processes, institutions and projects will "establish quantitative objectives for quality and process performance and use them as criteria in managing processes" (CMMI Product Team, 2006, p. 37). These quantitative objectives are determined based on the goals and user requirements of RDM. For example, if one of the goals is to minimize unnecessary repetitive data entry when researchers submitting datasets to a repository, then it might be useful to ask data submission interface users to record the number of times a same piece of data (author name, organization name, project name, etc.) is keyed in. An analysis of unnecessary repetitions in data entry may inform where in the RDM process the efficiency of data entry may be improved. The key here is to collect the statistics while action is being taken rather than after the fact. This means that a quantitatively managed maturity level has better predictability of process performance, because "the performance of processes is controlled using statistical and other quantitative techniques, and is quantitatively predictive" (CMMI Product Team, 2006, p. 38).

0.3.5 Level 5: Optimizing

Level 5, Optimizing, means that the organization is focused on improving the processes: weaknesses are identified and defects are addressed proactively. Processes introduced at these levels of maturity address generic techniques for process improvement.

While CMM has been around for two decades and applied in various contexts for improving processes and performance, it just began to draw attention from the research data management community. RDM is still a relatively new domain and much of the research has been devoted to the specific fields and practices such as metadata and data repositories. Examples of using CMM for data management processes and other goals began to emerge in the last couple of years (see note 1), with slightly different focus and interpretations. This document takes a holistic view of RDM and uses the CMM lens to examine RDM processes in the hope that we can identify the weaknesses of RDM and find ways to improve RDM processes.

References

Brooks Jr, F. P. (2010). *The design of design: Essays from a computer scientist*. Pearson Education.

CMMI Product Team. (2006). CMMI for Development Version 1.2. CMU/SEI-2006-TR-008. Pittsburgh, PA, USA: Carnegie Mellon Software Engineering Institute. Retrieved from http://repository.cmu.edu/sei/387

D'Ignazio, J., & Qin, J. (2008). Faculty data management practices: A campus-wide census of STEM departments. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1–6. doi:10.1002/meet.2008.14504503139 . Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/meet.2008.14504503139/abstract

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. (1993). Capability maturity model, Version 1.1. IEEE Software, 10(4): 18-27. Retrieved from http://www.computer.org/csdl/mags/so/1993/04/s4018-abs.html

Qin, J. & D'Ignazio, J. (2010). The central role of metadata in a science data literacy course. *Journal of Library Metadata*, 10(2), 188-204. doi:10.1080/19386389.2010.506379. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/19386389.2010.506379

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., et al. (2008). *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library* (Working Paper). Retrieved from http://hdl.handle.net/1813/10903

0.4 Structure of this Document

In the original Capability Maturity Model, maturity levels contain key process areas that are organized by common features. Maturity levels serve as indicators of process capability while key process areas are where goals will be achieved (or failed). Common features address the implementation or institutionalization of key practices. The common features are defined in the original CMM as "attributes that indicate whether the implementation and institutionalization of a key process area is effective, repeatable, and lasting" (Paulk et al., 1993, p. 37). The organization of key RDM practice areas is based on the five common features specified in the original CMM:

Table 1. Common features in the Capability Maturity Model (Paulk et al., 1993, p. 38)

Commitment to Perform	Commitment to Perform describes the actions the organization must take to ensure that the process is established and will endure. Commitment to Perform typically involves establishing organizational policies (e.g., the rules for data management) and senior management sponsorship.
Ability to Perform	Ability to Perform describes the preconditions that must exist in the project or organization to implement the process competently. Ability to Perform typically involves resources, organizational structures and responsibilities, and training.
Activities Performed	Activities Performed describes the roles and procedures necessary to implement a key process area. Activities Performed typically involve establishing plans and procedures (i.e., the specific actions that need to be performed), performing the work, tracking it, and taking corrective actions as necessary.
Measurement and Analysis	Measurement and Analysis describes the need to measure the process and analyze the measurements. Measurement and Analysis typically includes examples of the measurements that could be taken to determine the status and effectiveness of the Activities Performed.
Verifying Implementation	Verifying Implementation describes the steps to ensure that the activities are performed in compliance with the process that has been established. Verification typically encompasses reviews and audits by management and software quality assurance.

There are five chapters in this document for the key process areas in research data management: 1) data management in general; 2) data acquisition, processing and quality assurance; 3) data description and representation; 4) data dissemination; and 5) repository services and preservation. Each key process area is further divided into a number of sub-areas. The description of these sub-areas includes definition of key concepts, rationale/importance, examples, and recommended practice.

The organization of the process areas follows the structure of the common features listed in Table 1. However, we made one change from the original CMM model. In our analysis of RDM practices, we found limited evidence of quantitative measurement or validation of processes, which we suggest reflects the current state of maturity of RDM. As a result, in this document we have combined Measurement and Analysis and Verifying Implementation as one practice area.

This document is built on a wiki platform to enable registered users to make contributions. Initially, registered users can comment. Crowdsourced editing will be deployed when a governance structure such as a review committee is established. Please view the pages "How to Use this Site" and "Guide for Authors and Editors" for more information.

References

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). *Capability Maturity Model for Software, Version 1.1* (No. CMU/SEI-93-TR-024). Software Engineering Institute. Retrieved from http://resources.sei.cmu.edu/library/asset-view.cfm?assetID=11955

1. Data Management in General

Overall goal: Have a high-quality research data management process.

The overall goal of data management is to collect and maintain high quality data to support research. A mature research data management process bears a number of signposts: an organization-wide commitment to ensuring a high quality management and maintenance process as reflected in a set of practices that establish the overall data management process, effective communication to, and training of, existing and new staff for maintaining the ability to perform the research data management processes, and clearly defined processes, roles, and responsibilities that are kept updated and controlled for improvement as well as cost-benefit analysis.

1.1 Commitment to Perform

***Commitment to Perform** describes the actions the organization must take to ensure that the process is established and will endure. Commitment to Perform typically involves establishing organizational policies and senior management sponsorship.*

1.1.1 Identify stakeholders

The goal of identifying stakeholders is to establish a shared understanding of who are the data owners, contributors, managers, and users affected by data management. Stakeholders include not only those who create and manage data but also entities that are data users, funding agencies, or home institutions of contributing researchers (DataOne, 2011).

Explicit identification of stakeholders is important because research data management processes are increasingly complex and so involve entities with different roles, specializing in different aspects of data management. For example, data managers are responsible for data storage, management, backup, and access. Research team members need to document data collection and processing methods and parameters, validate and verify data quality, and maintain information on workflows and data flows for provenance and quality control purposes. Technology staff need to assure that the infrastructure services are in good order to support the data management activities. However, organizations may not have all of these stakeholders and responsibilities can be differently distributed.

Furthermore, the tasks and interests in data management among these different groups may or may not cross with one another. For example, Mullins (2007) reported that, after extensive interviews with scientists in biology, earth and atmospheric science, astronomy, chemistry, chemical engineering, plant science, and ecological sciences, it became clear that no single method or process would suffice the needs for data management across all disciplines. Their extensive conversations with stakeholders led them to identify the need to foster collaboration between domain scientists as well as librarians/archivists, computer scientists, and infrastructure technologists. In addition to project level stakeholders, three types of data sharing intermediaries may have a role in supporting data management at various stages of the research data life cycle: data archives (all stages), institutional repositories (end of research life cycle), and virtual organizations.

As a result, explicit identification of stakeholders is necessary to ensure that the design of the processes meets their different needs and to ensure implementation efficiency and usefulness of data management. As in Mullins (2007), identification of stakeholders may start with discussion with key informants, such as researchers or sponsored program office staff, and then use snowball sampling to identify additional stakeholders. The results of these efforts may be confirmed by a follow-up survey.

1.1.2 Develop user requirements

The goal of developing user requirements is to describe the goals the data management systems and practices achieve for various user groups, without going into details about how those goals are to be achieved. For example, researchers may require that data management ensures that data are available for future analysis, while potential reusers of data may require effective data description to enable them to find and make sense of the data.

Developing user requirements for research data management must consider a wide array of factors because differences in disciplinary or research fields and types of research significantly affect the workflows, data flows, and data management and use practices. These differences in turn will affect the user requirements for data management services and tools and will result in idiosyncrasies of the systems and services supporting the data management tasks. For example, the requirements for storing and describing a real-time stream of data are different than for survey data. In a collaborative data management situation, user requirements must take into consideration the technical standards for data formats, sampling protocols, variable names, and data discovery interfaces, among other things (Hale et al., 2003).

User requirements for research data management may be identified through analyzing data flows, workflows, leading data management problems, and researchers' data practices. These requirements can be represented at a high level in use cases, user scenarios or personas (Cornell University Library, 2007; Lage, Losoff, & Maness, 2011). A key point in this process is that user requirements mean not only clear-cut project objectives but also goals for the data management services to serve a longer term and wider scope of research data management.

1.1.3 Establish quantitative objectives for data management

The goal of establishing quantitative objectives for data management is to provide a set of measures of the data management process and quantitative targets for those measures. For example, a simple metric is the quantity of data collected and the cost of the collection process. In doing a survey, a goal might be a certain sample size (number of surveys completed) and a target set based on the research needs and the project's budget for data collection. An alternative metric is the quality of the data, with a target of a no more than a certain error rate. A goal for data privacy might be that there be no unintentional data releases. For data sharing, a goal might be that new users can gain access to the data within a certain time period.

Establishing quantitative objectives is important to provide a basis for measuring the effectiveness of the data management process and for assessing improvements to the process. Picking inappropriate measures can be counterproductive if it leads people to focus on achieving the wrong goals. For example, if a data repository used only number of datasets collected as a measure of the data archiving process, it might fail to ensure the datasets are well documented or useful, resulting in a large collection of useless data. It is likely that a portfolio of measures will need to be developed, addressing the different goals of the process.

At present, this goal seems rarely to be explicitly addressed in data management.

Establishing quantitative objectives can be done following common practices in management (e.g., key performance indicators and balanced scorecard) and in research project assessments (e.g., outcome-based assessment).

1.1.4 Develop communication policies

Developing communication policies relates to communication channels and procedures among the constituencies. This makes communication efficient and clear. Communication channels are specific to organizational contexts, and can be facilitated by communication technologies such as websites, ticketing systems, discussion forum, mailings, wikis, social media, etc.

Developing communication policies is dependent on the scale and context of data management. For example, a community level data management project needs to maintain proper channels to communicate with internal functional groups and external constituencies about the decisions, procedures, and policies about the process and products. These may be a call for comments and suggestions on a metadata schema, policy on data publication and use, or the approval process for contributed data sets. A research group may also install communication policies that will clearly specify the reporting channels for data management operations.

Whether a data management project is at a community level or research group level, the objectives and expectations should be clearly defined and communicated. This is especially important when multiple partners are involved because documenting the nature of collaborative partnership supports open communication (Hale et al., 2003). Policies for data management, use, and services are an instrument of communication. Providing them on an institution or project's websites as separate documents offers open communication with the community members and constituencies. Data service providers should maintain open and effective communication venues for the community. For example, Cornell's Research Data Management Service Group uses their website to provide communication channels for their community on different levels (https://confluence.cornell.edu/display/rdmsgweb/Home).

Rubric

Rubric for 1.1 - Commitment to Perform

Level 0 This process or practice is not being observed	No steps have been taken to establish organizational policies or senior management sponsorship for stakeholder or end user needs, quantitative objectives, or communication policies
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Stakeholder and end user needs, objectives, and communication have been considered minimally by individual team members, but nothing has been quantified or included in organizational policies or senior management sponsorship
Level 2: Managed DM process is characterized for projects and often reactive	Stakeholder and end user needs and objectives have been recorded for this project, but have not taken wider community needs or standards into account and have not resulted in organizational policies or senior management sponsorship
Level 3: Defined DM is characterized for the organization / community and proactive	The project follows approaches to stakeholder and end user needs and objectives that have been defined for the entire community or institution, as codified in organizational policies with senior management sponsorship
Level 4: Quantitatively Managed	Quantitative quality goals have been established regarding stakeholder and end user needs and objectives, and are codified in

DM is measured and controlled	organizational policies with senior management sponsorship; both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding stakeholder and end user needs and objectives are evaluated on a regular basis, as codified in organizational policies with senior management sponsorship, and necessary improvements are implemented

References

Cornell University Library. (2007). Cornell University Library personas. Retrieved from http://hdl.handle.net/1813/8302

DataONE. (2011). Recognize stakeholders in data ownership. Retrieved from https://www.dataone.org/best-practices/recognize-stakeholders-data-ownership

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder. Portal: Libraries and the Academy, 11(4): 915-937. doi:10.1353/pla.2011.0049. Retrieved from http://www.press.jhu.edu/journals/portal_libraries_and_the_academy/portal_pre_print/current/articles/11.4lage.pdf

Mullins, James. (2007). Enabling international access to scientific data sets: Creation of the Distributed Data Curation Center (D2C2). Purdue University, Purdue E-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1100&context=lib_research

1.2 Ability to Perform

***Ability to Perform** describes the preconditions that must exist in the project or organization to implement the process competently. Ability to Perform typically involves resources, organizational structures, and training.*

1.2.1 Develop and implement a budget

Effective data management incurs costs (Hale et al, 2003). Budgeting for data management helps ensure allotment of sufficient financial resources to support data management activities. Budget considerations vary with the type, scope, scale, and timeframe of the data management context. Those who collect data need adequate financial resources to manage local data during the life cycle of the project (DataOne, 2011a; Hale et al., 2003). Local data management costs might include data management personnel, database systems, servers, networks, and security for project data that is shared over a network (Hale et al., 2003).

Another type of data management cost is synthesis and integration of data, and collaboration necessary to support this synthesis (Hale et al., 2003). The creation of metadata using a standardized metadata format is a cost for data that is publically shared beyond the scope of a research project.

Organizations with missions aimed at disseminating and preserving data budget for data management beyond the timeframe of specific research projects. When data centers are underfunded, their focus becomes managing their own data rather than addressing the broader needs of those they serve.

As new data management models emerge, the budget for data management also needs to take the memberships or subscriptions of data repository services into consideration. This has become a trend that, on the one hand, disciplinary data repositories are seeking self-sustainable solutions through devising economic models that will charge institutions for services (Sheaffer, 2012). On the other hand, institutions that are initiating or have established data management services will need funding to start up the RDM services and keep them in operation once they become part of the regular tasks.

Budgeting should include not only allotment of hardware and software, but also near- and long-term RDM service payments and staff with the appropriate technical expertise. In their ethnographic study of data and work practices across three science cyberinfrastructure projects in the environmental sciences Mayernik et al. (2011) found that "human support is valuable in the development of data management plans, but is only available in institutions that specifically provide funding for it" (p. 421).

1.2.2 Staffing for data management

Staffing for data management refers to identifying the levels and types of expertise needed for achieving immediate and/or near-term data management objectives. A data management lifecycle involves different tasks at different stages that demand a combination of varying levels and types of expertise and skills. For example, the Data Preservation Alliance for the Social Sciences (DATA-PASS at http://www.data-pass.org) is a broad-based partnership of data archives for acquiring, cataloging, and preserving social sciences data. The partnership involves existing data repositories, academic institutions, and government agencies. As such the communication among partners, technical system architecture, and policies are inherently complicated. Having a capable staff will be extremely important to meet the constantly shifting data curation activities (Walters & Skinner, 2011).

Staffing needs should be reviewed carefully and each role/position's responsibilities specified clearly. This is not only important for hiring the right personnel but also important for developing a suitable training program "to ensure that the staff and managers have the knowledge and skills required to fulfill their assigned roles" (Paulk et al., 1993, p. 12).

1.2.3 Develop collaborations and partnerships

Stakeholder involvement in data management processes often takes the form of collaboration and/or partnership. When resources can be effectively shared, partnerships can reduce hardware and software costs, lead to better data and data products, and reduce many technical barriers by agreeing on core data standards and the flow of data (Hale et al., 2003). Collaboration and partnership are often a process of community building that, if managed properly, can contribute to sustaining a community of RDM practice.

Collaboration and partnership can be managed by creating agendas and schedules for collaborative activities, documenting issues, and developing recommendations for resolving relevant stakeholder issues. In addition, activities in collaboration and partnership may also include problem solving, information and experience sharing, resource/assets reuse, coordination, visits, and creation of documentations. Over time a community of RDM practice can be built, which in turn will strengthen the collaboration and partnership.

1.2.4 Train researchers and data management personnel

A key indicator for mature data management processes is that training programs are provided so researchers and staff understand data management processes well and have the capability to perform data management activities. Examples of training programs include:

- Providing online guidance and workshops for data management
- Training in data documentation best practices
- Training in the unique tools and methods used in a research field

The purpose of training programs is two-fold: for researchers, the training program is to develop the skills and knowledge of individuals so that they can adopt the best practices in managing their data; and for data managers, the training program will build the institutional capability by having capable personnel to perform infrastructural and technical services for data management.

Planning for training typically involves identification of training needs, training topics, requirements and quality standards for training materials, training tasks, roles, and responsibilities, and required resources. Schedules for training activities and their dependencies also need to be laid out in the training program. Training programs may also be offered by conference workshops, professional development events, or educational programs outside of one's institution. These venues are useful for training the trainers who will provide internal training programs and services.

1.2.5 Develop RDM tools

Research data management tools are software programs that help researchers effectively manage data during a research lifecycle. The nature of research types determines the requirements for such tools. Computational intensive research fields such as astrophysics use workflow management systems to capture metadata for provenance and output management, which is a highly automated process (Brown et al., 2006). Geodynamics data, on the contrast, often reside in spreadsheet files and sometimes are mixed with researchers' annotation text. It will be difficult to manage this type of data with completely automatic tools due to the inconsistent data recording practice (Qin, D'Ignazio, & Baldwin, 2011). Developing RDM tools in a sense is also a process of developing and establishing best practices in RDM.

Tools for RDM include off-the-shelf applications, such as data repository management systems and metadata editors created for specific standards, along with those developed in-house. Before deciding whether to adopt an off-the-shelf tool or develop one in-house, a comprehensive analysis should be conducted to understand not only the local requirements but also the need for links to community data management infrastructure and standards. This means that tools adopted or developed should consider key functions for immediate data management needs such as storage, annotation, organization, and discovery, and at the same time the "staging" functions for effective data deposition and dissemination in community, national, and international data repositories.

More often than not software tools for RDM have been developed (Michener, 2006). Adoption of such tools means adopting the mechanisms to systematically capture the integration process (DataONE, 2011b). RDM projects vary in scope and nature as the data they deal with change from discipline to discipline and from project to project. Whether tools are adopted or developed for ad hoc or long-term needs, support for researchers to use these tools should be an integral part of the tool adoption/development process (Mayernik et al., 2011).

1.2.6 Establish a data management plan

A data management plan (DMP) documents the definitions, procedures, methods, and best practices for a project or organization to maintain a consistent practice of RDM. Careful planning for data management before you begin your research and throughout the data's life cycle is essential (DataONE, 2011c) because it can increase project efficiency and optimize the reliability of the data that are collected by minimizing errors.

The most common DMPs are the kind prepared as part of a grant proposal because of the mandate from funding agencies such as the U.S.National Science Foundation (NSF), the Institute for Museum and Library Services (IMLS), or the National Endowment for the Humanities Office of Digital Humanities (NEH-ODH). Examples of this type of DMP can be found from funding agencies' websites as well as many research universities' websites, e.g., the Research Cyberinfrastructure (RCI) at UC San Diego provides a list of DMP samples for major NSF disciplinaries (http://rci.ucsd.edu/dmp/examples.html). Also, the DMP Tool website has a list of templates based on specific funder requirements (https://dmp.cdlib.org/pages/funder_requirements).

Resources for DMP development:

1. Disciplinary-based NSF DMP templates: http://dmconsult.library.virginia.edu/dmp-templates/
2. DMP Tool hosted at California Digital Library: https://dmp.cdlib.org/

Rubric

Rubric for 1.2 - Ability to Perform

Level 0 This process or practice is not being observed	No steps have been taken to provide organizational structures or plans, training, or resources such as budgets, staffing, or tools
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Structures or plans, training, and resources such as budgets, staffing, or tools have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Structures or plans, training, and resources such as budgets, staffing, or tools have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows includes structures or plans, training, and resources such as budgets, staffing, or tools that have been defined for the entire community or institution

Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding structures or plans, training, and resources such as budgets, staffing, or tools, and practices in these areas are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding structures or plans, training, and resources such as budgets, staffing, or tools are evaluated on a regular basis, and necessary improvements are implemented

References

Brown, D.A, Brady, P.R., Dietz, A., Cao, J., Johnson, B., & McNabb, J. (2006). A case study on the use of workflow technologies for scientific analysis: Gravitationalwave data analysis, in I.J. Taylor, E. Deelman, D. Gannon, and M.S. Shields(Eds.), Workflows for e-Science, chapter 5, pp. 41–61. Berlin: Springer-Verlag.

DataONE. (2011a). Define roles and assign responsibilities for data management. Retrieved from https://www.dataone.org/best-practices/define-roles-and-assign-responsibilities-data-management

DataONE. (2011b). Document the integration of multiple datasets. Retrieved from https://www.dataone.org/best-practices/document-integration-multiple-datasets

DataONE. (2011c). Plan data management early in your project. Retrieved from https://www.dataone.org/best-practices/plan-data-management-early-your-project

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

Mayernik, S. M., Batcheller, A. L., Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. In: *Proceedings of iConference 2011, February 8-11, 2011, Seattle, WA,* pp. 417-425. New York: ACM Press.

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3–7. doi:10.1016/j.ecoinf.2005.08.004. Retrieved from http://www.sciencedirect.com/science/article/pii/S157495410500004X

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). *Capability Maturity Model for Software, Version 1.1* (No. CMU/SEI-93-TR-024). Software Engineering Institute. Retrieved from http://resources.sei.cmu.edu/library/asset-view.cfm?assetID=11955

Qin, J., D'Ignazio, J., & Baldwin, S. (2011). A workflow-based knowledge management architecture for geodynamics data. A White paper submitted to NSF GEO/OCI EarchCube Charrette meeting. Retrieved from http://earthcube.ning.com/group/user-requirements/forum/topics/white-paper-a-workflow-based-knowledge-management-architecture

Sheaffer, P. (2012). Creating a sustainable business model for a digital repository: the Dryad experience. ASIS&T Research Data Access and Preservation Summit 2012, Baltimore, MD. Retrieved from http://www.slideshare.net/asist_org/creating-a-sustainable-business-model-for-a-digital-repository-the-dryad-experience-peggy-schaeffer-rdap12

Van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. (2011). Managing and Sharing Data: A Best Practice Guide for Researchers. (3rd ed.) Essex, England: University of Essex. Retrieved from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Walters, T. & Skinner, K. (2011). New roles for new times: Digital curation for preservation. Retrieved from http://www.arl.org/focus-areas/workforce/1086

1.3 Activities Performed

***Activities Performed** describes the roles and procedures necessary to implement a key process area. Activities Performed typically involve establishing plans and procedures (i.e., the specific actions that need to be performed), performing the work, tracking it, and taking corrective actions as necessary.*

In the general data management process area, the activities performed involve turning the requirements, collaborations/partnerships, plans, and procedures into written documents that state shared consensus and understanding of the goals and actionable plans within an institution or a research group. Different kinds of activities performed will reflect different levels of capability maturity in research data management.

1.3.1 Manage RDM Requirements

Two aspects of RDM requirements are crucial for RDM. The user aspect of RDM requirements focuses on the functionalities that an RDM system or platform can offer for researchers to perform their data management tasks throughout the research lifecycle, so that they can save time while achieving RDM goals. The technical aspect of RDM requirements refers to the technologies and organizational support that make these functionalities possible. RDM requirements may change over time as new projects and new data emerge. Documenting RDM requirements and keeping them updated will establish a common understanding between researchers and RDM processes. This agreement with researchers is the basis for planning and managing the RDM processes.

Developing RDM requirements can be done through a wide variety of channels (as described in 1.1.2 Develop user requirements), but managing RDM requirements goes further than requirements gathering. The goal is to establish a baseline for use by research data management processes and keep RDM plans, outcomes, and activities consistent with the RDM requirements from users and systems.

Requirements management encompasses four core activities:

- ***Elicitation***: requirements are obtained from stakeholders and other sources and refined in great detail.

- ***Documentation***: the elicited requirements are documented by using natural language or conceptual models.
- ***Validation and negotiation***: documented requirements are validated against predefined criteria and negotiated with stakeholders.
- ***Management***: validated requirements are properly structured and prepared so that they can be used by different roles, to maintain consistency after changes, and to ensure their implementation (Pohl & Rupp, 2011).

1.3.2 Manage Collaborations and Partnerships

Collaborations and partnerships in RDM may take place at all organizational levels and among any number of community members. Large-scale collaborations and partnerships include examples such as DataONE (https://www.dataone.org/) and the Laser Interferometer Gravitational-Wave Observatory (LIGO, http://www.ligo.caltech.edu/). There are also regional, disciplinary-based collaborations (e.g., the Hubbard Brook Ecosystem Study, http://hubbardbrook.org/) and many within-institutional-unit collaborations for research data management (e.g., Cornell University's Research Data Management Service Group, https://confluence.cornell.edu/display/rdmsgweb/Home). The goals of collaboration and partnership management are to keep the collaborators and partners aware of the shared purpose, gain consensus on problem solving, engage them in the process, and ensure sharing between the parties involved.

Maintaining communication policies (described in 1.1.4 Develop communication policies) is crucial in managing collaborations and partnerships. Regular meetings should be held and other communication methods used for awareness, sharing, motivating, and engaging purposes. Whether collaboration scale is large or small, decisions reached and notes taken during meetings or through asynchronous channels should be carefully documented and shared among collaborators and partners.

1.3.3 Create Actionable RDM Plans

Discussion of a data management plan as part of the activities performed refers to one that is operational, created when a new research project starts or when an institution takes a data management initiative. In the case that a project is funded by a grant from NSF or another funding agency, the DMP submitted with the proposal will need to be expanded with operational specifics for the project staff to follow and execute. The operational DMP for a new research project should specify essential management tasks that may not have been included in the proposal-stage DMP, including data storage structures, backup schedules, naming conventions for data files and folders, and procedures for data processing and transformation, in addition to the high-level descriptions in a proposal-stage DMP.

1.3.4 Develop Workflows and Procedures

A workflow is defined as a "set of tasks involved in a procedure along with their interdependencies and their inputs and outputs" (Ailamaki, Ioannidis, & Livny, 1998, p. 1). Data management workflows consist of tasks to be performed and procedures that ensure the consistent performance of the tasks. For example, the objective of a file naming convention is to establish patterns of file names for searching and identifying data input and managing data output. A workflow for data input and output will involve defining naming conventions, assigning names to output data, depositing them to appropriate file locations, and creating appropriate annotations. These tasks should follow standard procedures so that data output is managed with consistency, upon which scientific experiments or computational runs will depend, to obtain the input data.

In developing workflows for data management, staff need to define each key process area clearly, as these will then be used to identify tasks to be performed and procedures to ensure consistency in performing the tasks.

Rubric

	Rubric for 1.3 - Activities Performed
Level 0 This process or practice is not being observed	No steps have been taken for managing the workflow during the research process, such as managing functional requirements, managing collaboration, creating actionable plans, or developing procedures
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Workflow management during the research process, such as managing functional requirements, managing collaboration, creating actionable plans, or developing procedures, has been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Workflow management during the research process, such as managing functional requirements, managing collaboration, creating actionable plans, or developing procedures, has been recorded for this project, but has not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to workflow during the research process, such as managing functional requirements, managing collaboration, creating actionable plans, or developing procedures, that have been defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding workflow during the research process, such as managing functional requirements, managing collaboration, creating actionable plans, or developing procedures, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding workflow during the research process, such as managing functional requirements, managing collaboration, creating actionable plans, or developing procedures, are evaluated on a regular basis, and necessary improvements are implemented

References

Ailamaki, A., Ioannidis, Y.E., & Livny, M. (1998). Scientific workflow management by database management. In: Proceedings of the Tenth International Conference on Scientific and Statistical Database Management, Capri, Italy, July 1-3, 1998. Retrieved from http://www.cs.cmu.edu/~natassa/aapubs/conference/scientific-workflow-management.pdf

Pohl, K. & Rupp, C. (2011). (Requirements Engineering Fundamentals: Study Guide for the Certified Engineering Exam. Sebastopol, CA: O'Reilly Media.

1.4 Process Assessment

***Process Assessment** includes Measurement and Analysis and Verifying Implementation. Measurement and Analysis describes the need to measure the process and analyze the measurements, and typically includes examples of the measurements that could be taken to determine the status and effectiveness of the Activities Performed. Verifying Implementation describes the steps to ensure that the activities are performed in compliance with the process that has been established, and typically encompasses reviews and audits by management and quality assurance.*

Process assessment involves establishing measures and control of the effectiveness and quality of data management so that the RDM processes are continuously improved. This key process area is based on the activities performed that are well defined as the result of level-3 maturity in the RDM capabilities. The fact that a research project or organization (group, institution, or community) is capable of conducting process assessment signifies a level-4 capability maturity, i.e., the managed level. It is important to point out that a higher level of capability maturity must have achieved the previous level of maturity because the previous level of maturity is the foundation for achieving the next level of capability maturity.

The first step in process assessment is to set quantitative quality goals for both RDM outcomes and processes. Effectiveness and quality are measured for important RDM process activities. Identifying these measures is an intensive process and better conducted across all projects as part of an organizational measurement program. In other words, effectiveness and quality measures tend to be project-neutral and should be able to be applied to all projects in process assessment for RDM.

The second step in process assessment focuses on continuous process improvement. The effectiveness and quality measures established through the first step will be used to identify weaknesses and strengthen the process proactively, with the goal of preventing the occurrence of defects. Data on the effectiveness of the RDM process is used to perform cost benefit analyses of RDM.

There is very little available in the literature to generalize the characteristics of level 4 and level 5 of capability maturity in RDM. The measurement and quality management for RDM is therefore defined in terms of analogy to the original CMM (Paulk et al., 1993).

1.4.1 Measurement and Analysis

The goal of RDM varies because the nature and characteristics of research types and data differ from discipline to discipline. Data flows and stages in field observations and lab experiments will be different from those in computer simulations or computational intensive types of research, for example. The involvement of researchers and data professionals in data flows and stages is also different, e.g, data collection during a field visit will be usually conducted by researchers while datasets ready for curation are handled by data mangers or librarians. The measurements for process assessment should maintain a focus on effectiveness and quality while recognizing these differences and complexities. The following therefore is targeted to establishing the measurements regardless who (researchers, data staff, or librarians) perform it:

- *The amount of effort that went into the process*, e.g., how many redundant runs were performed to complete the processing.
- *Time spent on a task*, e.g., how long it took to verify/check data, code data, or transform data.
- *Presence (or absence) of process data collection*: when data about process effectiveness is collected on the spot, it is easier to do than after the fact. It is tedious to do it afterwards and the data can easily become inaccurate.
- *Data points produced*: e.g., number of survey responses generated, number of data frames segmented.

Measurements can be constructed from the perspective of input, output, and throughput, or from the perspective of workflows. The amount of effort, for example, can be considered as an input measurement, while data points produced would be an output measurement. Effectiveness is getting things right. Process measurements can help to identify problems, especially the causes of the problems. If you observe the missing data is high, then it makes sense to look for what caused the missing data.

1.4.2 Verifying Implementation

According to the original CMM, "Verifying Implementation describes the steps to ensure that the activities are performed in compliance with the process that has been established. Verification typically encompasses reviews and audits by management and software quality assurance" (Paulk et al., 1993, p. 38). Verifying implementation in the context of RDM focuses on reviews and audits of the key processes areas against the established policies and procedures (which are mainly reflected in the commitment to perform, ability to perform, and activities performed). The goal is to identify whether there is any weakness in the process and how it can be strengthened.

Rubric

Rubric for 1.4 - Process Assessment

Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for measurement, analysis, or verification of the research process in general
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Measurement, analysis, or verification of the research process in general have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Measurement, analysis, or verification of the research process in general have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to measurement, analysis, or verification of the research process in general that have been defined for the entire community or institution

Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established including measurement, analysis, and verification of the research process in general, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding measurement, analysis, or verification of the research process in general are evaluated on a regular basis, and necessary improvements are implemented

References

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). *Capability Maturity Model for Software, Version 1.1* (No. CMU/SEI-93-TR-024). Software Engineering Institute. Retrieved fromhttp://resources.sei.cmu.edu/library/asset-view.cfm?assetID=11955

2. Data acquisition, processing and quality assurance

Overall goal: Reliably capture research data in a way that facilitates use, preservation and reuse.

The first stage in the data lifecycle is to collect the data along with data documentation. Data collection is the process of capturing observations of the world—physical, biological, behavioural or social—in a form that can be used for analysis. Observations are of some property or properties (e.g., presence or absence, mass, behaviour, structure, attitude) of one or more units of observation (e.g., an organism, artifact, sample, group or organization). Data documentation means the description created by the researcher of how the data were collected (e.g., conditions, parameters, techniques, etc.), the initial processing of the data, and of the data themselves (e.g., formats, units, etc.). An important subgoal of this stage is to ensure the quality of the data and the data documentation as they are captured and processed.

Given a phenomenon of interest, it may be possible to record the properties of all of the relevant observational units (e.g., the single case being studied in depth or all of the organisms in an experiment). However, as the scale and number of units in the study increases, it may not be feasible to record more than a fraction of the units, requiring some process for sampling, i.e., for choosing which units to measure. Temporally, data collection may be one-off, i.e., at a single point in time, or repeated at more or less regular intervals, with greater or finer temporal spacing. Finally, data collection might be made simultaneously of multiple properties of each unit of observation, or of only a few.

Observations can be recorded as verbal or textual reports, yielding qualitative data. Qualitative observations might be left free-form or coded into a fixed set of categories, e.g., the species of an observed organism or one particular behavior or structural characteristic from a set, with more or less formal rules for translating the observation into the categories. Often data from observations are recorded as quantitative measurements. Measurement is the process of converting the observed properties to numbers, that is, symbols representing points along a scale. While conceptually a measure might take on any value, in practice there are only a finite number of possible symbols available to represent the value. Measurements can be made on scales with different properties, from an ordinal scale that simply distinguishes ordered values (e.g., the life stage of an organism that could be represented as A, B, C and so on) to a ratio scale that imposes ordering, equal spacing and a zero point (e.g., a count, length or intensity).

Adopting a realist perspective, a measurement can be thought of as the true value plus some amount of error. Error can arise from many different sources. Some error is inherent in the measurement process itself, e.g., quantization error due to the spacing of points on the measurement scale. Such error is lower for a more precise measurement, i.e., one with a finer gradation of points on the scale. Error can also be introduced by the specific measurement process, e.g., the instruments used may have some inherent inaccuracy, or from accidents in the measurement. Finally, if observations are aggregated, e.g., to create estimates of an average value in a population, then there will be statistical uncertainty in the estimate due to sampling.

2.1 Commitment to Perform

Commitment to Perform *describes the actions the organization must take to ensure that the process is established and will endure. Commitment to Perform typically involves establishing organizational policies and senior management sponsorship.*

2.1.1 Develop data quality control policies

The goal of developing data quality control policies is to establish a shared understanding of the goals, rules and responsibilities for data quality assurance (Hook et al., 2010). The policies should provide a clear definition of what quality data means in the context of the research given the data to be collected.

Developing data quality policies is important to ensure that different actors in the data collection process have common understandings of the goals and rules for ensuring data quality and that there are clear responsibilities for these actions.

Quality might refer to the level or nature of error in the measurements, e.g., whether the error is randomly distributed (noise) or systematic (bias) and the expected magnitudes of the error. Data quality policies should also address the coverage of the data, e.g., how wide a geographic, temporal or conceptual range is covered, how fine the geographic or temporal sampling and how representative the sample. Policies should reflect the desired tradeoffs between these characteristics. For example, it may be that one project determines that it is more valuable to have a broader geographic scope of data collection, trading off the need to sample within that region, while another elects to emphasize repeated measurement at regular time intervals, trading off geographic scope, while a third emphasizes the precision and accuracy of measurements, trading off the volume of data collected.

2.1.2 Develop data documentation policies

The goal of developing data documentation policies is to establish a shared understanding of the goals, rules and responsibilities for creating data documentation. The policies should provide a clear definition of what data documentation needs to be collected along with the data, what that documentation should include, and who is responsible for collecting the documentation (DataONE, 2011).

Developing data documentation policies is important to ensure that different actors in the data collection process have common understandings of the goals and rules for collecting data documentation and that there are clear responsibilities for these actions.

For example, when collecting field observations, data documentation might include such details as the observation protocol followed. Lab data documentation might similarly describe the equipment used as well as the protocols followed. Human subjects data documentation should include details about required institutional review board protections, such as informed consent requirements.

For more discussion about data documentation, please see 3.1.1 Develop metadata policies.

Rubric

Rubric for 2.1 - Commitment to Perform

Level 0 This process or practice is not being observed	No steps have been taken to establish organizational policies or senior management sponsorship for data quality or documentation
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Data quality and documentation have been considered minimally by individual team members, but nothing has been codified or included in organizational policies or senior management sponsorship
Level 2: Managed DM process is characterized for projects and often reactive	Data quality and documentation have been addressed for this project, but have not taken wider community needs or standards into account and have not resulted in organizational policies or senior management sponsorship
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to data quality and documentation that have been defined for the entire community or institution, as codified in organizational policies with senior management sponsorship
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding data quality and documentation, and are codified in organizational policies with senior management sponsorship; both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding data quality and documentation are evaluated on a regular basis, as codified in organizational policies with senior management sponsorship, and necessary improvements are implemented

References

DataONE. (2011). Develop a quality assurance and quality control plan. Retrieved from https://www.dataone.org/best-practices/develop-quality-assurance-and-quality-control-plan

Hook, L. A., Vannan, S. K. S., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from http://daac.ornl.gov/PI/BestPractices-2010.pdf

2.2 Ability to Perform

Ability to Perform *describes the preconditions that must exist in the project or organization to implement the process competently. Ability to Perform typically involves resources, organizational structures, and training.*

2.2.1 Develop data file formats

Typically collected data for a research study form a data set that includes a set of data files, where each data file includes a set of data items representing the observed data as well as data about how those data were collected. The project should define and document the formats of the files that will store collected data, both at the level of whole files and for the specific data items within a file (Hook et al., 2010).

It is important to develop data file formats carefully to ensure that data are stored consistently both within and across files (Hook et al., 2010). Data need to be represented in consistent formats to facilitate integration with data in other data files and data sets (Hale et al., 2010, and DataONE, 2011a). Documentation of data file formats is necessary to ensure that data creators store data correctly and data users interpret data correctly.

At the whole file level, electronic data files should be stored in non-proprietary formats, e.g., a simple text format such as tab- or comma-separated values (CSV) (DataONE, 2011j) or a more complex format such as NetCDF (Network Common Data Form) or Hierarchical Data Format (HDF). More complex formats offer additional features, such as error correcting codes to detect and recovery from errors in the underlying data store. Use of software such as spreadsheets (e.g., Excel) that save data in proprietary formats limit how data can be used and increase the risk of the data becoming unreadable due to file corruption or changes in the software (DataONE, 2011h). Data that are stored in a proprietary format should include documentation of the specific software and versions used to create it (Hook et al., 2010). The format of multimedia files such as sound, images or video should similarly be documented.

It is also important to document the layout of data within each file. Observational data files are generally structured like spreadsheets, with rows and columns and a value at the intersection of each row and column, each row representing an observation and each column, data about the observation (e.g., time or location) or a type of data collected.

The format of the file should be such that only rows are added for additional observations, not columns (Borer et al., 2009). Each row should have one column or set of columns that uniquely identify the observation (a key field) (Borer et al., 2009).

Each column of a data file should represent a single type of data (DataONE, 2011h). Storing multiple values in a single cell complicates data analysis (Borer et al., 2009). Each column should have a header that describes the variable in that column (Borer et al., 2009). Data and annotations of data should be stored in separate columns (Hook et al., 2010). A separate column should also be used for data qualifiers, descriptions and flags (DataONE, 2011i).

Format for representing collected data items should be clearly defined. The data type and precision (i.e., how many digits) should be selected to be appropriate for the data in each column (DataONE, 2011g). It is important to establish these formats to ensure that stored data are consistently recorded and can be unambiguously interpreted, and to reduce the complexity of processing data.

A consistent set of data types should be used across a data set (DataONE, 2011e). Date and time formats in particular should be consistent across the data set (DataONE, 2011b). If the date or time associated with an observation is not completely known (e.g., only date but not time for certain observations), then separate columns should be used to separate the parts that are known (DataONE, 2011b). If data are collected at diverse locations, it may be necessary to capture the timezone of times (Hook et al., 2010). Location information in a data set should all use the same coordinate system and representation (Hook et al., 2010). Categorical values should be represented by a consistent set of terms or codes (DataONE, 2011k). These should not be specific to a particular column or data file but should be consistent across the data set. Missing values should be represented in a consistent way across a data set (DataONE, 2011f).

The format of observations stored in a single file should be consistent. Ideally, each observation would correspond to one row in the file. An optimal data format has data in each column rather than being sparse, with many blank cells (DataONE, 2011d). Mixing different kinds of data (e.g., from different types of observations) in a single file complicates further processing or integration of the data. If many observations of different types of measurements are collected, each measurement should be recorded in a separate file (Hook et al., 2010).

2.2.2 Develop data quality control procedures

Projects should develop and document procedures for controlling the quality of data collected (DataONE, 2011c). Procedures can address control of quality in both data collection and capture.

Having documented procedures is important to ensure that data quality tasks are performed consistently and correctly.

The specific tasks required are highly dependent on the type of data and the observations. For example, a simple procedure is to establish reasonable ranges for data items and to double check recorded values that fall outside these ranges. If a batch of data are entered (e.g., from a hand-written data collection form), a simple check is that the number of items entered match the number recorded in the original document. Slightly more complicated is the technique of "casting out nines": repeatedly adding up all of the digits entered and comparing the sum to the sum of the digits in the original document. For some kinds of data, it may be possible to audit a sample of data to ensure that they were collected and recorded correctly and to estimate the proportion of erroneous data in the unaudited dataset.

Procedures should be reviewed periodically to ensure that they are up to date, complete and effective (DataONE, 2011c).

Rubric

	Rubric for 2.2 - Ability to Perform
Level 0 This process or practice is not being observed	No steps have been taken to provide for resources, structure, or training with regards to file formats or quality control procedures
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Resources, structure, and training with regards to file formats or quality control procedures have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Resources, structure, and training with regards to file formats or quality control procedures have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the	The project provides resources, structure, and training with regards to file formats or quality control procedures as defined for the entire community or institution

organization/community and proactive	
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established for resources, structure, and training with regards to file formats or quality control procedures, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding resources, structure, and training, with regards to file formats or quality control procedures, are evaluated on a regular basis, and necessary improvements are implemented

References

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America*, *90*(2), 205–214. http://dx.doi.org/10.1890/0012-9623-90.2.205

DataONE. (2011a). Consider the compatibility of the data you are integrating. Retrieved from https://www.dataone.org/best-practices/consider-compatibility-data-you-are-integrating

DataONE. (2011b). Describe formats for date and time. Retrieved from https://www.dataone.org/best-practices/describe-formats-date-and-time

DataONE. (2011c). Develop a quality assurance and quality control plan. Retrieved from https://www.dataone.org/best-practices/develop-quality-assurance-and-quality-control-plan

DataONE. (2011d). Document your data organization strategy. Retrieved from https://www.dataone.org/best-practices/document-your-data-organization-strategy

DataONE. (2011e). Ensure basic quality control. Retrieved from https://www.dataone.org/best-practices/ensure-basic-quality-control

DataONE. (2011f). Identify missing values and define missing value codes. Retrieved from https://www.dataone.org/best-practices/identify-missing-values-and-define-missing-value-codes

DataONE. (2011g). Maintain consistent data typing. Retrieved from https://www.dataone.org/best-practices/maintain-consistent-data-typing

DataONE. (2011h). Preserve information: keep your raw data raw. Retrieved from https://www.dataone.org/best-practices/preserve-information-keep-your-raw-data-raw

DataONE. (2011i). Separate data values from annotations. Retrieved from https://www.dataone.org/best-practices/separate-data-values-annotations

DataONE. (2011j). Use appropriate field delimiters. Retrieved from https://www.dataone.org/best-practices/use-appropriate-field-delimiters

DataONE. (2011k). Use consistent codes. Retrieved from https://www.dataone.org/best-practices/use-consistent-codes

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

Hook, L. A., Vannan, S. K. S., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from http://daac.ornl.gov/PI/BestPractices-2010.pdf

2.3 Activities Performed

***Activities Performed** describes the roles and procedures necessary to implement a key process area. Activities Performed typically involve establishing plans and procedures (i.e., the specific actions that need to be performed), performing the work, tracking it, and taking corrective actions as necessary.*

2.3.1 Capture / Acquire data and data documentation

Capturing how data are collected or digitized, what they mean, and how the data are structured is at the center of data documentation. Maintaining good data documentation is crucial for data reuse (UK Data Archive, 2014). Data documentation is also vital when the data are used by researchers who are unfamiliar with the data and/or were not involved in data collection.

Procedures need to be established for data and data documentation, both for what should be collected and documented and how it should be collected and documented. Once procedures are established, they should be followed to standardize the data collection process. Recording of data should be done as soon as possible after data are collected to minimize the opportunities to introduce error (Columbia Center for New Media Teaching and Learning, n.d.). Each unique measurement should be recorded only once to minimize data collection effort and to avoid possible transcription errors (Borer et al, 2009).

Data should not be recorded with higher precision than was actually collected (DataONE, 2011c). Measurement uncertainty should be recorded if known (DataONE, 2011a). If actual measurements can not be obtained and an estimated value is recorded, a note identifying the estimate and estimation technique should also be recorded (DataONE, 2011b).

A note should be made if the date and time recorded with a record represents the date of data collection or date of data recording if those two are not the same.

If data are collected from human subjects (e.g., via interviews or a survey), then the necessary informed consent documents should be collected at the same time.

Rubric

	Rubric for 2.3 - Activities Performed
Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for the workflow of collecting and documenting data
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	The workflow for collecting and documenting data has been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	The workflow for collecting and documenting data has been addressed for this project, but has not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to the workflow of collecting and documenting data that have been defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding the workflow of collecting and documenting data, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding the workflow of collecting and documenting data are evaluated on a regular basis, and necessary improvements are implemented

References

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America*, *90*(2), 205–214. http://dx.doi.org/10.1890/0012-9623-90.2.205

Columbia Center for New Media Teaching and Learning. (n.d.). Responsible conduct of research: Data acquisition and management: Foundation text. Retrieved from http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/index.html#3_B

DataONE. (2011a). Describe measurement techniques. Retrieved from https://www.dataone.org/best-practices/describe-measurement-techniques

DataONE. (2011b). Identify values that are estimated. Retrieved from https://www.dataone.org/best-practices/identify-values-are-estimated

DataONE. (2011c). Store data with appropriate precision. Retrieved from https://www.dataone.org/best-practices/store-data-appropriate-precision

UK Data Archive. (2014). Create and manage data: Documenting your data. Retrieved from http://www.data-archive.ac.uk/create-manage/document

2.4 Process Assessment

***Process Assessment** includes Measurement and Analysis and Verifying Implementation. Measurement and Analysis describes the need to measure the process and analyze the measurements, and typically includes examples of the measurements that could be taken to determine the status and effectiveness of the Activities Performed. Verifying Implementation describes the steps to ensure that the activities are performed in compliance with the process that has been established, and typically encompasses reviews and audits by management and quality assurance.*

2.4.1 Measurement and Analysis

Measurement and analysis of data acquisition and processing provides specific practices and procedures that guide this process area. It should keep in mind that the goal of measurement and analysis is to provide "general guidance about measuring, analyzing, and recording information that can be used in establishing measures for monitoring actual performance of the process" (CMMI Product Team, 2006). Projects should develop and implement metrics for the data acquisition, processing and quality assurance processes. Example metrics include the quantity of data being collected or the observed error rate at different points in the process. A small sample of data might be intensively quality checked to provide an estimate of the level of undetected errors in the data collected.

2.4.2 Assure data quality

Data quality should be assessed as data are collected, and the data quality process is documented. Checking for data quality as the data are collected ensures that only valid data are recorded and that erroneous values are either recollected or at least eliminated from further analysis.

At a minimum, data items must be consistent with the data type of the column.

Data should be inspected after data collection to check for validity (e.g., plotting for visual examination). Times and dates should be checked to be sure they are valid (DataONE, 2011b). Locations coordinates can be mapped and checked to ensure that they are valid (DataONE, 2011b). Values recorded by instruments should be inspected to check that they are within a sensible range for the property being measured and for the instrument (e.g., within the detection limits of the equipment) (DataONE, 2011b).

Data can be transcribed by two or more people and the values compared to ensure accuracy (DataONE, 2011a). Newly collected data can be compared to data from other data sets with similar data. Comparison to historic ranges can help identify anomalous values that require further examination. However, outliers should not be removed without careful consideration that they do not represent a true measurement.

Supervisors should review and sign off on data to signify completeness and accuracy (Columbia Center for New Media Teaching and Learning, n.d.).

Codes should be recorded in the data file to represent the quality of data at the time quality is assessed (DataONE, 2011b). Problematic data should be flagged to indicate known issues (DataONE, 2011c). Any ancillary data used to assess data quality should be described and stored (DataONE, 2011b).

2.4.3 Check data integration from other sources

If data from other sources are used, the quality of those other sources should be reviewed (Hale et al., 2003). In addition, the license or permissions for those data should be reviewed to ensure that the use is allowed. Finally, the source of the data should be recorded to ensure that the data can be cited as appropriate.

Rubric

	Rubric for 2.4 - Process Assessment
Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for measurement, analysis, or verification of data collection and documentation
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Measurement, analysis, and verification of data collection and documentation have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Measurement, analysis, and verification of data collection and documentation have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to measurement, analysis, and verification of data collection and documentation that have been defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding measurement, analysis, and verification of data collection and documentation, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding measurement, analysis, and verification of data collection and documentation are evaluated on a regular basis, and necessary improvements are implemented

References

CMMI Product Team. (2006). *CMMI for Development, Version 1.2* (No. CMU/SEI-2006-TR-008). Pittsburgh, PA, USA: Carnegie Mellon Software Engineering Institute. Retrieved from http://repository.cmu.edu/sei/387

Columbia Center for New Media Teaching and Learning. (n.d.). Responsible conduct of research: Data acquisition and management: Foundation text. Retrieved from http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/index.html#3_B

DataONE. (2011a). Double-check the data you enter. Retrieved from https://www.dataone.org/best-practices/double-check-data-you-enter

DataONE. (2011b). Ensure basic quality control. Retrieved from https://www.dataone.org/best-practices/ensure-basic-quality-control

DataONE. (2011c). Mark data with quality control flags. Retrieved from https://www.dataone.org/best-practices/mark-data-quality-control-flags

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

3. Data description and representation

Overall goal: Describe and represent data to facilitate future discovery and use.

Data description and representation is a process of capturing information that enables users to find, understand, and use/reuse data. In a broad sense even an email exchange between colleagues explaining how data can and cannot be used is a type of informal metadata (Edwards et al, 2011). The focus of this section of the CMM for RDM is on metadata process areas that involve adopting metadata standards, generating metadata descriptions for data, and best practices.

Metadata can be applied to different levels of interrelated research data outputs, from those that are more granular to those that are more global, such as:

- a variable, parameter, or column heading field in a database
- a file
- a study

During the active phase of a research project researchers might be most attuned to documentation and management of data at granular levels (i.e. variables and files). However, the metadata in a data archive needs to have contextual information about the study as a whole that is not common knowledge to those beyond the project in which the data were produced.

Metadata has different functions that can carry differing requirements. It is generally true that there is less immediate need for metadata the closer one is to the context of data creation. A researcher who just took a measurement has the units of measurement in her head, and researchers on collaborative projects have informal opportunities for communicating about data. When data gets farther from the context of creation, documentation of contextual details becomes increasingly important. There is a sense in which documentation of contextual information has a life cycle of its own, which roughly correspond with different functions metadata serves:

- active management of data during a project,
- preservation and discovery once data have been shared in an archive,
- reuse of data or replication of analysis performed in a study, and
- assessment of the impact of research outputs.

Different stakeholders might value different metadata functions. For example, researchers are typically concerned with active management of data during a project, and librarians tend to value preservation and discovery once data have been shared in an archive. Consequently, different stakeholders may have deeply different conceptions of metadata requirements. A life cycle approach to data management, which takes the function of metadata throughout its life cycle into account, can be helpful in attending to differences in perspective.

Fortunately, one metadata element can often serve multiple functions (Riley, 2014), and documentation of data at different levels of granularity can reap benefits at other levels. Practices that can improve project level data management (e.g. variable documentation) can also increase opportunities for discovery when the study data is archived (e.g. ICPSR is a data archive that offers a variable search capability). Similarly, practices that improve discovery for secondary users also facilitate self-discovery for data creators who may not remember project details at a later date.

References

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. doi:10.1177/0306312711413314. Retrieved from http://pne.people.si.umich.edu/PDF/EdwardsEtAl2011ScienceFriction.pdf

Riley, Jenn. (2014). Metadata services. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals*. West Lafayette, Indiana: Purdue University Press.

3.1 Commitment to Perform

***Commitment to Perform** describes the actions the organization must take to ensure that the process is established and will endure. Commitment to Perform typically involves establishing organizational policies and senior management sponsorship.*

For data description and representation, the commitment to perform includes committing to documenting project activities to facilitate replication, generating standard-compliant metadata specifications and schemas, and using controlled vocabularies to facilitate discovery.

3.1.1 Develop metadata policies

Metadata policies support the creation of metadata that fits the data and conforms to the standards and best practices of the relevant research community (Riley, 2014). An example of a national level metadata policy is the National Science Foundation's suggestion that data management plans include "standards to be used for data and metadata format and content."

It is clear that not every stage of a research lifecycle, hence the data lifecycle as well, requires comprehensive metadata descriptions. Metadata policies should provide guidelines on when to create metadata descriptions and what types of metadata are mandated or optional. The content of these guidelines may vary widely depending on the scope of a research project and the nature of data. For example, at the project level, the metadata policy would focus more on workflows and procedures, while at the institutional level, the policies can become more general and function as guidelines for what should be done rather than how it should be done.

There are also differences between data documentation and metadata descriptions. Raw data files and intermediary data files, for example, may not have formal metadata descriptions but documentation should be provided for data creation/collection processes, errors or issues identified, etc. so that users can have sufficient information to decide whether the data is suitable for their research. Metadata is considered as a "subset of data documentation, which provide standardized, structured information explaining the purpose, origin, time references, geographic location, creating author, access conditions, and terms of use of a data collection" (Corti et al., 2014, p. 38).

Most research data is not currently described with metadata that meets an authoritative standard. Tenopir et al. (2011) found that 78 percent of researchers either do not use metadata schema at all, or use an ad hoc, homegrown metadata format to describe their data. The limitation of not describing a study's data using an authoritative standard is that opportunities for discovery and reuse are diminished.

Commitment to metadata can occur on the part of institutions that support research, and in a more grassroots way by researchers themselves. However, there is a relationship between institutional commitment to metadata and default researcher metadata practices (Mayernik et al., 2011). When there is a permanent or semi-permanent institutional commitment to metadata "researchers themselves may or may not have experience and expertise in creating and working with formal metadata, but will likely have experts… to provide help and support in making data available to wider audiences. This human support is valuable in the development of data plans, but is only available in institutions that specifically provide funding for it" (Mayernik et al., 2011, p.421).

Rubric

	Rubric for 3.1 - Commitment to Perform
Level 0 This process or practice is not being observed	No steps have been taken to establish organizational policies or senior management sponsorship regarding metadata development
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Metadata development has been considered minimally by individual team members, but nothing has been quantified or included in organizational policies or senior management sponsorship
Level 2: Managed DM process is characterized for projects and often reactive	Metadata development policies have been recorded for this project, but have not taken wider community needs or standards into account and have not resulted in organizational policies or senior management sponsorship
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to metadata development that have been defined for the entire community or institution, as codified in organizational policies with senior management sponsorship
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding metadata development, and are codified in organizational policies with senior management sponsorship; data are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding metadata development are evaluated on a regular basis, as codified in organizational policies with senior management sponsorship, and necessary improvements are implemented

References

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). Managing and Sharing Research Data: A Guide to Good Practice. Los Angeles, CA: SAGE.

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). *How Institutional Factors Influence the Creation of Scientific Metadata.* Paper presented at the iConference '11, Seattle.

Riley, Jenn. (2014). Metadata services. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals.* West Lafayette, Indiana: Purdue University Press.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6), e21101. doi:10.1371/journal.pone.0021101. Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101

3.2 Ability to Perform

***Ability to Perform** describes the preconditions that must exist in the project or organization to implement the process competently. Ability to Perform typically involves resources, organizational structures, and training.*

The ability to perform in the data description and representation process area refers to the readiness of metadata artifacts and tools as well as the readiness of staff and procedures that are essential for performing data description and representation.

3.2.1 Develop or adopt metadata specifications and schemas

A large number of metadata standards are available for adoption. Whether to develop new metadata specifications or adopt an existing standard requires a good knowledge of the standards relevant to the description needs. Metadata policies (See Section 3.1) provide guidelines for decision making about what data should be described by agreed-upon metadata standards or schemas, and when. Metadata specifications define how data should be described with the goal of helping future users find, identify, select, obtain, and appropriately understand and use information from a dataset. Metadata specifications are usually a collection of elements, controlled vocabularies, encoding schemas, and best practice guidelines.

Regardless of whether the work involves developing new specifications or adopting existing standards, careful analyses of data types and status at different stages of the research lifecycle must be performed to understand description and user requirements. For example, active data files that may change by the minute will be fine with just rudimentary metadata embedded in the file (descriptive file names, creator's name, time stamps, and other technical metadata), while a dataset as the final data product from a project will need comprehensive metadata to describe the research context and key metadata values.

In practice, metadata standards are rarely followed exactly as they are. Modifications will most likely be necessary when adopting a metadata standard(s). The resulting metadata specifications from modifying one or more metadata standards are called metadata "Application Profiles" (AP). Zeng & Qin provide a detailed discussion of different approaches to designing metadata application profiles (2014). Many projects and communities have created numerous metadata application profiles and many of these APs can be located through metadata directories or registries, e.g., the Digital Curation Centre in the UK (http://www.dcc.ac.uk/) hosts a metadata directory for science disciplines at http://www.dcc.ac.uk/resources/metadata-standards. Sometimes informal, "homegrown" metadata practices are used, which is better than using no metadata schema at all. Whenever possible, use a previously created schema that complies with an authoritative community standard. Use of these services can help prevent "reinventing the wheel" when designing metadata specifications and schemas.

In addition to easing retrieval, the use of standards makes documentation more consistent in general. The use of a schema will greatly improve the interoperability of the information collected.

3.2.2 Select and acquire tools

Tools for producing metadata should be selected and evaluated for feasibility. Metadata standards often come with tools. Some standards have multiple tools. An example of a type of tool is the workflow management system astrophysicists use that automates capture of metadata. Automated tools typically cannot capture all of the necessary metadata. A best practice is to make use of tools currently in use in a research community for generating metadata (Riley, 2014).

3.2.3 Develop strategies for generating metadata based on community practices

Metadata descriptions may be created for a collection of data, the study that generated the collection of data, or individual data sets and files. For computationally-intensive research fields such as astrophysics, much of the required metadata may be captured automatically for data files and datasets, but in field and experimental research fields such as ecology and geodynamics, a large amount of human intervention has to go into the metadata creation process. A best practice for generating metadata is to leverage existing documentation practices within a community of researchers (Riley, 2014).

One strategy for generating metadata to facilitate discovery and long-term preservation is to rely on researchers to perform this activity themselves. Thus far this approach has had limited success (Tenopir, 2011), and has inhibited the deposit of data in repositories with useful metadata (Riley, 2014). This is often a default approach for generating metadata due to limited resources.

There are efforts to automate the generation of metadata via software tools, though this capability is not fully realized for most research communities. An example of an ability to perform issue is ensuring flexible data services for virtual datasets (DataONE, 2011).

A best practice in many contexts is to conceptualize metadata creation as a shared responsibility, that is facilitated by librarian support (Riley, 2014). For example, the ICPSR data repository asks researchers to provide descriptive study information, but also devotes significant staff resources to enhancing researcher metadata to make it more fully interoperable with DDI (Data Documentation Initiative) metadata (a social science metadata standard), and transforming data into multiple data formats (for three common statistical software platforms) to make it widely accessible.

Researcher interest in documentation of data is greatest when it assists with everyday project data management (Jahnke & Asher, 2012). A best practice is to integrate metadata creation into researcher workflows during the active phase of research projects, leveraging researcher interest in project data management (Jahnke & Asher, 2012).

3.2.4 Arrange staffing for creating metadata

Roles in creating metadata vary with the scale and nature of the research context. Large, heavily funded projects often have internal infrastructure with dedicated data management personnel; smaller projects are more likely to benefit from support from data supports services offered by an academic library (Ray, 2014).

Often there are two levels of metadata that are of concern for research data: annotation on the spot that researchers do in the context of everyday data management, and high-level bibliographic metadata afforded by librarian expertise. When metadata is conceptualized as a shared responsibility, project researchers themselves might produce on the spot metadata, and need training in best practices; a librarian might then later produce bibliographic metadata to facilitate discovery.

To support documentation of everyday data management it can be helpful for researchers to commit to putting aside time at the end of each work session, and at project milestones, to document project activities (Long, 2009).

3.2.5 Provide training for researchers and librarians

When metadata creation is conceptualized as a shared responsibility, training can be helpful for both researchers and librarians (Riley, 2014). Training for researchers can be in the form of general information appropriate for a broad range of researchers delivered at key points in the research life cycle. For example, DMPTool (https://dmp.cdlib.org/) offers guidelines for generating metadata at https://dmptool.org/dm_guidance as part of data management planning; with regard to discipline specific training on data management practices, Colorado Clinical and Translational Sciences Institute (CCTSI) offers education in data management best practices (http://cctsi.ucdenver.edu/CommunityEngagement/Resources/DataSharingGuidelines/Pages/DataManagement.aspx) for translational biomedical research via a website with videos (http://cctsi.ucdenver.edu/RIIC/Pages/DataManagement.aspx).

A promising approach to researcher data management education is the TIER protocol developed by Ball and Medeiros at Haverford College (http://www.haverford.edu/TIER/). This approach to researcher education is to experientially teach data management practices that produce replicable analysis through the structure of deliverables required for student research projects. The rationale is that if budding researchers learn data management when they learn research methods, sound documentation practices are not perceived as a hardship.

When metadata support is offered as a service delivered by subject liaison librarians, training for librarians can come via online resources. Examples include the Digital Curation Centre's curation resources (http://www.dcc.ac.uk/resources) and training materials (http://www.dcc.ac.uk/training), and Purdue University's Data Profile Toolkit (http://datacurationprofiles.org/). Librarians can also pursue more in-depth professional development, or formal education such as the five library schools in the United States that offer data curation programs (Riley, 2014).

3.2.6 Assess community data and metadata practices

The provision of metadata services requires understanding of existing research community metadata practices, in addition to metadata structures associated with libraries (Ray, 2014). Purdue University’s data curation profiles, which are generated via interviews, are one such approach for librarians to increase their knowledge of existing practices. Another approach is to use small pilot studies early on in development of data curation services (Westra, 2014).

Rubric

Rubric for 3.2 - Ability to Perform

Level 0 This process or practice is not being observed	No steps have been taken to provide organizational structures or plans, training, or resources such as staffing and tools for metadata development
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Structures or plans, training, and resources such as staffing and tools for metadata development have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Structures or plans, training, and resources such as staffing and tools for metadata development have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows includes structures or plans, training, and resources such as staffing and tools for metadata development that have been defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding structures or plans, training, and resources such as staffing and tools for metadata development, and practices in these areas are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding structures or plans, training, and resources such as staffing and tools for metadata development are evaluated on a regular basis, and necessary improvements are implemented

References

DataONE. (2011). Ensure flexible data services for virtual datasets. Retrieved from https://www.dataone.org/best-practices/ensure-flexible-data-services-virtual-datasets

Jahnke, L., Asher, A., & Keralis, S. D. (2012). The problem of data. Council on Library and Information Resources (CLIR) Report, pub. #154. ISBN 978-1-932326-42-0 Retrieved from http://digitalcommons.bucknell.edu/fac_pubs/52/

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, Texas: Stata Press Books.

Ray, J. M. (2014). Introduction to research data management. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences)*. West Lafayette, Indiana: Purdue University Press.

Riley, Jenn. (2014). Metadata services. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101.

doi:10.1371/journal.pone.0021101. Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101

Westra, Brian (2014). Developing Data Management Services for Researchers at the University of Oregon. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press.

Zeng, M. L. & Qin, J. (2014). Metadata. Chicago, IL: ALA Neal Schuman.

3.3 Activities Performed

Activities Performed *describes the roles and procedures necessary to implement a key process area. Activities Performed typically involve establishing plans and procedures (i.e., the specific actions that need to be performed), performing the work, tracking it, and taking corrective actions as necessary.*

3.3.1 Generate metadata according to agreed upon procedures

Follow agreed upon procedures for generating metadata for variables, files, and studies to ensure the ability of future users to find, identify, select, and obtain data. There is not a single set of metadata that applies in all situations, but consider which elements are important for lower levels of granularity and higher-level description of the dataset as a whole.

3.3.1.1 Document variables

Document individual data items such as variables (columns in structured tabular data), with names, labels and descriptions. Examples of elements of variable documentation are data type; units of measurement; formats for date, time, and geography; method of measurement, coverage (e.g. geographic, temporal), and codes and classification schemes (e.g. codes for missing data, or flags for quality issues or qualifying values). ICPSR offers extensive guidelines for variable documentation based on the DDI standard for quantitative social science data. DataOne (2011) offers guidelines based on best practices in the natural and physical sciences.

Document variables in the data file, and in a separate file. Long (2009) offers guidelines for naming and describing variables and values (p. 143-194). For structured, tabular data, a well-documented data dictionary provides a concise guide to understanding and using the data. An example of a data dictionary is available from the Colorado Clinical and Translational Sciences Institute: http://cctsi.ucdenver.edu/RIIC/Documents/Data-Management-Figure-3.pdf.

For qualitative data, offering structured contextual information in a separate data list provides users with a guide to the data. The UK Data Archive has examples and templates for data lists: http://www.data-archive.ac.uk/create-manage/document/data-level?index=2

Use a controlled (standardized) vocabulary. Sometimes there is a sufficiently high degree of standardization in a research community to make it possible to report data in standardized ways (time, taxonomy, for example). This promotes interoperability of metadata, which is desirable when possible. When this degree of standardization does not exist, documentation of the language used on a study is next best.

3.3.1.2 Document files

Describe the contents of data files. It may be helpful to create a separate document describing how files are structured and technical information on the files (e.g. the version of the software).

File formats that are stable, and interoperable with other systems, are desirable.

Long (2009) offers extensive recommendations on file management best practices (p. 18-30, 125-141). Long also offers templates for planning a directory structure and for creating a data registry here: http://www.indiana.edu/~jslsoc/web_workflow/wf_chapters.htm.

3.3.1.3 Document the study

Describe the research project. Common elements in study level documentation are author (principal investigator, researchers); funding; rationale for the project; data sources used; context of data collection; data collection methods; information on confidentiality; access and use conditions, transformation of data, and its structure and format. Examples of guidelines for study level documentation are available at the UK Data Archive at http://www.data-archive.ac.uk/create-manage/document/study-level and ICPSR (based on the Data Documentation Initiative (DDI) metadata schema) at http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html.

When the dataset or collection is a complex object that consists of multiple files, describe their organization in an index, table of contents, or a readme file.

- ICPSR suggests a table of contents: http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html
- DataOne offers guidelines: https://www.dataone.org/best-practices/describe-overall-organization-your-dataset
- In the TIER (Teaching Integrity in Empirical Research) protocol the guide to the dataset as a whole is conceptualized as a readme file: http://www.haverford.edu/TIER/protocol/#readme

Provide a mechanism for identity control that uniquely identifies the data in a machine readable way. One system for providing identity control is via the International DOI Foundation (IDF)'s Digital Object Identifier system, (DOI).

Provide a citation. There is not complete consensus on the elements that make up a complete data citation. However, Brase et al. (2014) say the Digital Curation Centre's 11 elements of a data citation are well-supported by literature: http://www.dcc.ac.uk/resources/how-guides/cite-datasets#x1-5000. DataOne offers citation guidelines here: https://www.dataone.org/best-practices/provide-citation-and-document-provenance-your-dataset.

Provide documentation of analysis when information for replication is desired (Long, 2009). Documentation of analysis is not necessarily required to support discovery and secondary use of a dataset, as secondary use may explore a completely different research question than the original analysis. Replication repositories or journal data sharing policies may require documentation of analysis. For example, Nature Publishing Group's data policy is here: http://www.nature.com/authors/policies/availability.html.

Rubric

	Rubric for 3.3 - Activities Performed
Level 0 This process or practice is not being observed	No steps have been taken for managing the workflow of metadata creation during the research process
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Workflow management for metadata creation during the research process has been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Workflow management for metadata creation during the research process has been recorded for this project, but has not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to workflow for metadata creation during the research process as defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding workflow for metadata creation during the research process, and both metadata and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding workflow for metadata creation during the research process are evaluated on a regular basis, and necessary improvements are implemented

References

Brase, J., Socha, Y., Callaghan, S., Borgman, C.L., Uhlir, P.F., Carroll, B. (2014). Data citation: Principles and practice. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press.

DataONE. (2011). Best Practices. Retrieved from https://www.dataone.org/best-practices

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, Tex.: Stata Press.

3.4 Process Assessment

***Process Assessment** includes Measurement and Analysis and Verifying Implementation. Measurement and Analysis describes the need to measure the process and analyze the measurements, and typically includes examples of the measurements that could be taken to determine the status and effectiveness of the Activities Performed. Verifying Implementation* *describes the steps to ensure that the activities are performed in compliance with the process that has been established, and typically encompasses reviews and audits by management and quality assurance.*

As discussed in Chapter 1 - Data Management in General, process assessment involves identifying needed measurements and analysis and using the measurements for verification. For the data description and representation process area, the measurement of performance is related to the quality of metadata and ability of metadata schemas to communicate with other standards and systems.

3.4.1 Measuring and verifying implementation

Measurement in the data description and representation process includes two aspects: one is the performance of metadata generation/creation and the other is the quality of metadata as the product of this process. Quantitative measures for assessing the performance typically include the time taken to complete describing a dataset or documenting the study context and data, workflow steps from start to finish in metadata description, time spent in finding relevant sources in order to enter accurate metadata in the record, and unnecessary repetitions in data entry. The data for these measures should be collected in action to ensure the reliability of data because such very specific data values tend to become forgotten and affect the accuracy of measurement.

The quality of metadata can be measured by the criteria below:

- Completeness: the portion of elements in a description record that actually contain values (non-empty elements).
- Correctness in content, format, input, browser interpretation, and mapping.
- Consistency in data recording, source links, identification and identifiers, description of sources, metadata representation, and data syntax.
- Duplication rate in integrated collections. (Zeng & Qin, 2014)

Performance assessment in this process area is closely tied to the quality of metadata. A problematic workflow in metadata creation may hinder the discovery of potential issues and miss the opportunity to correct the process sooner to prevent the problem from becoming worse. Data for the quality of metadata descriptions should be regularly collected and procedures established to ensure the capturing of data that will later be used to assess both the process performance and quality of metadata.

Data collected against the measurements for performance and quality will be used to verify the implementation of the policies, schemas, and operations. The verifying process can be formal as described in the original CMMI document (Paulk et al., 1993). The Australian National Data Services (ANDS, 2011) and the DMVitals project at the University of Virginia Library (Sallans & Lake, 2014) are examples of two initiatives in the data management community exploring strategies for supporting verification of implementation.

Verification also includes making sure that the metadata schema(s) developed conform to standards and internal verification by building documentation verification steps into one's daily practice and into the project workflow at key milestones (Long, 2009). One strategy Long uses for ensuring internal compliance with agreed upon documentation standards is designating a project team member to be responsible for checking verification.

Rubric

	Rubric for 3.4 - Process Assessment
Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for measurement, analysis, or verification to ensure quality and compliance with metadata standards
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Measurement, analysis, or verification to ensure quality and compliance with metadata standards have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Measurement, analysis, or verification to ensure quality and compliance with metadata standards have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to measurement, analysis, or verification to ensure quality and compliance with metadata standards as defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established including measurement, analysis, and verification to ensure quality and compliance with metadata standards, and both metadata and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding measurement, analysis, or verification to ensure quality and compliance with metadata standards are evaluated on a regular basis, and necessary improvements are implemented

References

Australian National Data Service (2011). Research data management framework: Capability maturity guide http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.pdf

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, Texas: Stata Press Books.

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). *Capability Maturity Model for Software, Version 1.1* (No. CMU/SEI-93-TR-024). Software Engineering Institute. Retrieved from http://resources.sei.cmu.edu/library/asset-view.cfm?assetID=11955

Sallans, A. & Lake, S. (2014). Data management assessment and planning tools. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals.* West Lafayette, Indiana: Purdue University Press.

Zeng, M. L. & Qin, J. (2014). Metadata. Chicago, IL: ALA Neal Schuman.

4. Data Dissemination

Overall goal: Establish the policy and technical infrastructures for users to share, discover, obtain, and interact with data.

Data generated and produced from research or large-scale data collection projects may have tremendous value for future knowledge creation. But to realize their value for research and society at large, such data must be shared through various channels. Such sharing is complex, as research data come in varying forms, may be owned by public or private entities, and may involve human subjects that require privacy and confidentiality protection. Before any data can be shared, questions must be answered about what is to be shared, who may access the data, whether any restrictions apply, and how data may be disseminated.

Dissemination of research data as one of the key process areas must have an institution's commitment to perform data dissemination in order to sustain the process. This commitment is mainly embodied by a set of policies to ensure that data dissemination is considered from the beginning of a research project. Ability to perform includes the tools (technologies) and services that will enable the institution members to carry out the data dissemination process. Activities performed delineate the practices that the institution must put in place to allow data dissemination to be performed in a consistent way so that no wheels will be reinvented. Process assessment identifies the measurements / metrics that will be used to assess how effective the key process (in this case, data dissemination) is performed and where improvement might be needed to enhance the process effectiveness.

4.1 Commitment to Perform

Commitment to Perform *describes the actions the organization must take to ensure that the process is established and will endure. Commitment to Perform typically involves establishing organizational policies and senior management sponsorship.*

Data dissemination involves two aspects: one is data submission to a repository and the other is dissemination to communities. Data submission ensures that there are data to disseminate while the dissemination part publicizes the data, distributes, and delivers them to the users who requested the data.

An important signpost for an institution's commitment to disseminating data is a technical and policy infrastructure that

1. makes data submission easy to do
2. incentivizes and normalizes the practice of data submission by widening data dissemination

The commitment to perform includes identifying what should be submitted and disseminated, through which channels, how communities should be made aware of the data availability, and how the impact should be evaluated. In addressing these issues, a group of data policies are established to ensure the institutional commitment to repository services and data dissemination.

4.1.1 Develop data sharing policies

Data sharing policies are concerned with rules and guidelines on how data should be archived, disseminated, accessed, and used. They may be developed by a research center, an institution, or a data repository and generally conform to a funding agency's policy mandates for data sharing and dissemination. Policies for data sharing vary in scope and type depending on the type of organization for which such a policy is aimed. For example, a data submission policy may specify the requirements that a standard data submission form must be used; all data must have metadata meeting the standards adopted by the repository (Black Rock Forest Consortium, 2007).

In general, policies for data sharing should cover:

- What to be shared: this item usually involves data classification based on legal and/or contractual restrictions, public or internal domains, and so on.
- Compliance: whether submitting data to a data repository is a requirement or option for the members of the organization and when such submission should be completed. This lays out the expectations for sharing data (Hale et al., 2003). For example, "Datasets will be uploaded to the data catalog for availability within PISCO within one year of collection" (from the member node description for The Partnership for Interdisciplinary Studies of Coastal Oceans, DataONE, 2013, p. 2).
- Standards: tools for capturing metadata during data submission should be based on community and/or disciplinary metadata standards for ensuring metadata quality and interoperability.
- Constraints: whether there are any legal or contractual bindings for the data to be shared and how such legal or contractual procedures should be followed. These constraints define data access capabilities needed by a community of users (DataONE, 2011a) and the likely final destination and likely mode of dissemination of the data (Hook et al, 2010).

Sharing is good for the research enterprise as a whole (Columbia Center for New Media Teaching and Learning, n.d.), and having data sharing policies ensures the institutional commitment to making it happen and to reducing the level of effort required to prepare data for sharing. (Hook et al., 2010).

4.1.2 Develop policies for data rights and rules for data use

Policies for public data and restricted data often have different sets of conditions and rules for access and use. For publicly accessible datasets, the access and use policy typically specifies acceptable use, redistribution, citation, acknowledgement, disclaimer, and terms of agreement. DataOne suggests that usage rights statements should include what are appropriate data uses, how to contact the data creators, and how to acknowledge the data source. (DataONE, 2011c).

Acceptable use: defines the scope of use, e.g., commercial or non-commercial; derivations or other forms of products based on the dataset. The policy of acceptable use lays down the basis for more specific requirements and conditions in data use or reuse. The Protein Data Bank (PDB)'s usage policy represents that of a large open data repository, which includes conditions regarding how it is available (open to all users), conditions for redistribution, and recognition of intellectual property (PDB, 2014).

Redistribution: specifies whether the data sets can be redistributed and if so what rules should be followed. Many publicly available data sets allow for redistribution but only in their original format.

Citations: citations to data sets not only credit the original data creator or principle investigator, but are also a great way to broaden the impact and raise the visibility of the data set. Policies in this area should provide example citations.

Acknowledgement: this policy specifies that data users should acknowledge any institutional support or specific funding awards referenced. The Hubbard Brook Ecosystem Study (HBES), for example, provides the acknowledgement example in its data use policy:

"Acknowledgment example: Data on [topic] were provided by [name of PI] on [date]. These data were gathered as part of the Hubbard Brook Ecosystem Study (HBES). The HBES is a collaborative effort at the Hubbard Brook Experimental Forest, which is operated and maintained by the USDA Forest Service, Northern Research Station, Newtown Square, PA. Significant funding for collection of these data was provided by [agency]-[grant number], [agency]-[grant number], etc." (HBES, 2014)

Terms of agreement: this section clearly states the rights of data owners and the responsibilities of data users.

4.1.3 Develop data confidentiality policies

Data confidentiality refers to the rules and conditions that limit the release of data for access and the access permissions and rights to data and information. Release of early data before publication can jeopardize the ability of an investigator to be the first to publish a research finding. Data that can lead to patents also cannot be shared prematurely. Data confidentiality policies help scientists balance the free exchange of some sensitive scientific data and the risk that might come with such free exchange (Columbia Center for New Media Teaching and Learning, n.d.).

Before disseminating the data, it should be determined whether the data has any confidentiality concerns (DataONE, 2011b) and if so, such concerns should be documented to determine overall sensitivity. Confidentiality policies should be developed to protect the data and establish procedures and mechanisms based on sensitivity of the data (DataONE, 2011b). The policy should also specify who should have access based on ethical, intellectual-property, and research-based considerations (Columbia Center for New Media Teaching and Learning, n.d.).

Rubric

Rubric for 4.1 - Commitment to Perform

Level 0 This process or practice is not being observed	No steps have been taken to establish organizational policies or senior management sponsorship regarding data sharing or confidentiality
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Data sharing or confidentiality has been considered minimally by individual team members, but nothing has been quantified or included in organizational policies or senior management sponsorship

Level 2: Managed DM process is characterized for projects and often reactive	Policies for data sharing or confidentiality have been recorded for this project, but have not taken wider community needs or standards into account and have not resulted in organizational policies or senior management sponsorship
Level 3: Defined DM is characterized for the organization/ community and proactive	The project follows approaches to data sharing or confidentiality that have been defined for the entire community or institution, as codified in organizational policies with senior management sponsorship
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding data sharing or confidentiality, and are codified in organizational policies with senior management sponsorship; practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding data sharing or confidentiality are evaluated on a regular basis, as codified in organizational policies with senior management sponsorship, and necessary improvements are implemented

References

Black Rock Forest Consortium. (2007). Data submission protocol. Retrieved from http://www.blackrockforest.org/docs/scientist-resources/DataResources/DataSubmission.html

Columbia Center for New Media Teaching and Learning. (n.d.). Responsible conduct of research: Data acquisition and management: Foundation text. Retrieved from http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/index.html#3_B

DataONE. (2011a). Ensure flexible data services for virtual datasets. Retrieved from https://www.dataone.org/best-practices/ensure-flexible-data-services-virtual-datasets

DataONE. (2011b). Identify data sensitivity. Retrieved from https://www.dataone.org/best-practices/identify-data-sensitivity

DataONE. (2011c). Sharing data: legal and policy considerations. Retrieved from https://www.dataone.org/best-practices/sharing-data-legal-and-policy-considerations

DataONE. (2013). Member node description: PISCO. Retrieved from http://www.dataone.org/sites/all/documents/DataONEMNDescription_PISCO.pdf

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

Hook, L. A., Vannan, S. K. S., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from http://daac.ornl.gov/PI/BestPractices-2010.pdf

Hubbard Brook Ecosystem Study. (2014). Data use policy. Retrieved from http://www.hubbardbrook.org/data/dataset.php?id=4

Protein Data Bank. (2014). Policies and references [of Protein Data Bank]. Retrieved from http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html

4.2 Ability to Perform

Ability to Perform *describes the preconditions that must exist in the project or organization to implement the process competently. Ability to Perform typically involves resources, organizational structures, and training.*

For data dissemination services, Ability to Perform includes enabling technologies, procedures, and business models that will sustain the dissemination services.

4.2.1 Manage enabling technologies for access and conformance to standards

Enabling technologies for data dissemination are not standalone, instead, they are part of the larger system that make data submission, management, discovery, and archiving possible. For the dissemination tasks in particular, the enabling technologies include those that are critical in performing dissemination functions: data discovery, consultation (with principle investigators and/or data producers), selection, and obtaining.

Data discovery systems in different disciplines may have customized search fields and options, or special filters to perform targeted data discovery and selection. Federated search is a common approach to solve the problem of data silos. These approaches and techniques for data discovery should conform to standards for cross-system discovery and interoperability.

In data dissemination there is a need for middleware applications for translating among major databases, collaborative computing tools to improve communication, and software tools for developing metadata. Advanced data centers can help smaller centers develop standards, design databases, archive their data, and construct metadata (Hale et al., 2003).

Data portals offer great potential for creating and promoting partnerships (Hale et al., 2003). Developing data portals for data dissemination should be carefully planned to ensure the sustainability.

Rubric

Rubric for 4.2 - Ability to Perform

Level 0 This process or practice is not being observed	No steps have been taken to provide organizational structures or plans, training, or resources for enabling technologies for data sharing or confidentiality
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Structures or plans, training, and resources for enabling technologies for data sharing or confidentialityt have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Structures or plans, training, and resources for enabling technologies for data sharing or confidentiality have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project includes structures or plans, training, and resources for enabling technologies for data sharing or confidentiality as defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding structures or plans, training, and resources for enabling technologies for data sharing or confidentiality, and practices in these areas are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding structures or plans, training, and resources for enabling technologies for data sharing or confidentiality are evaluated on a regular basis, and necessary improvements are implemented

References

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

4.3 Activities Performed

Activities Performed *describes the roles and procedures necessary to implement a key process area. Activities Performed typically involve establishing plans and procedures (i.e., the specific actions that need to be performed), performing the work, tracking it, and taking corrective actions as necessary.*

Policies regarding data dissemination institutionalize data dissemination and show commitment, but enabling technologies add the actual ability to perform this process.

4.3.1 Identify and manage data products

Along a research lifecycle data come in various forms and with different levels of processing. They can be categorized based on the nature of research as observational, experimental, derived (or compiled), or simulation (DataONE, 2011e). The nature of research determines what types of data will be produced and what format these data will take (DataONE, 2011c). Before these data become sharable, they must be processed, "packaged," and registered in a repository or catalog of data products. According to the level of processing, data products can range from raw data, calibrated data, or derived/calculated data to visualized and interactable data. While data sharing policies define the classification of data to be shared, this process requires a list of criteria and procedures to identify individual datasets that can be deemed as data products for sharing and any restrictions of access and usage associated with each of them.

The identification and management of data products relies heavily on the metadata descriptions (a key process area described in Chapter 3) and tools. As data products vary in their content and complexity, e.g. both a large collection of datasets and documentation files or only a single data file may be viewed as a data product, it is essential to have clear guidelines for how data products may be grouped, packaged, or aggregated. It is also necessary that data packages be represented (Jones et al., 2001). The dissemination service interfaces should be based upon Open Standards (DataONE, 2011d).

4.3.2 Encourage sharing

Shared data can improve research by providing greater spatial, temporal, and disciplinary coverage than individual organizations can offer. Data submitted to a data repository are integrated and provide a way for organizations to build repositories of cohesive, high-quality data (Hale et al., 2003). However, data sharing policies following the institution's commitment to perform data dissemination do not always function as an incentive to motivate researchers to share data. A variety of venues should be used to convey the benefits of sharing data and the protection of data confidentiality and intellectual property rights to raise the awareness among researchers. Incentives such as impact and usage metrics embedded in the dissemination service system should be implemented as a reward mechanism to encourage sharing. Create shared need for data among partners to encourage better data stewardship (Hale et al., 2003)

4.3.3 Enable data discovery

Data discovery is a key function of all data repository systems. The discovery services should take into consideration the needs of both domain experts and non-expert users. For data products that might be useful for interdisciplinary research, it is even more important for the discovery service to facilitate and support discovery functions through enabling search and browsing. In other words, make your outputs perceivable (DataONE, 2011b).

Discovery services should also allow the addition of community tagging, annotation, and comments (DataONE, 2011f). For example, researchers can share and publish data using web-based datacasting tools and services (DataONE, 2011a).

4.3.4 Distribute data

Multiple channels can be established for data distribution to allow the widest possible coverage and timely dissemination. These channels include:

- *Linking data to publications*: Dryad Digital Repository (http://datadryad.org/) and Astrophysics Data Systems (ADS) (http://adsabs.harvard.edu/index.html) are two examples of this type of services. Linking services enables bi-directional discovery, i.e., finding and obtaining data through publications or vice versa.
- *Registering the data repository in a data union catalog*: Examples includes DataBib (http://databib.org/) and the Registry of Research Data Repositories (re3data, http://www.re3data.org/). The DataONE project has built a system for searching across multiple member data repositories. Joining a union catalog or data registry allows for federated and other broader searches, which affords the data to be distributed to much wider communities.
- *Distribute information on data products through Web services*: Open Standards for Web services include RSS/Atom and Web Services Definition Language (DataONE, 2011d). Users may subscribe these services to receive timely updates on data product information.

4.3.5 Ensure data citation

Data citation embodies two notions: to credit the data creator and to enable data reuse, verification, and impact tracking (DataCite, 2014). To enable consistent practice of data citation, guidelines should be provided regarding what information should be included (content) and how the information should be presented in a data citation (style). The Socioeconomic Data and Applications Center (SEDAC) provides examples of guidelines for citing the data from this center. This guideline specifies the required information for a data citation as:

- Primary responsibility party
- Year of publication, issue, release
- Edition/Version
- Type of resource, format
- Statement of responsibility for dynamically generated data and maps
- Publisher and place of publication
- Distributor
- Availability and access
- Retrieval statement
- Unpublished data (SEDAC, 2014)

Adopting a data citation standard such as DataCite can be another way to ensure consistent data citation practice.

Rubric

	Rubric for 4.3 - Activities Performed
Level 0 This process or practice is not being observed	No steps have been taken for managing the workflow of data dissemination, including sharing, discovery, and citation
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Workflow management for data dissemination, including sharing, discovery, and citation, has been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Workflow management for data dissemination, including sharing, discovery, and citation, has been recorded for this project, but has not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to workflow for data dissemination, including sharing, discovery, and citation, as defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding workflow for data dissemination, including sharing, discovery, and citation, and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding workflow for data dissemination, including sharing, discovery, and citation, are evaluated on a regular basis, and necessary improvements are implemented

References

DataCite. (2014). Why cite data? Retrieved from https://www.datacite.org/whycitedata

DataONE. (2011a). Advertise your data using datacasting tools. Retrieved from https://www.dataone.org/best-practices/advertise-your-data-using-datacasting-tools

DataOne. (2011b). Check data and other outputs for print and web accessibility. Retrieved from https://www.dataone.org/best-practices/check-data-and-other-outputs-print-and-web-accessibility

DataONE. (2011c). Define expected data outcomes and types. Retrieved from https://www.dataone.org/best-practices/define-expected-data-outcomes-and-types

DataONE. (2011d). Ensure flexible data services for virtual datasets. Retrieved from https://www.dataone.org/best-practices/ensure-flexible-data-services-virtual-datasets

DataONE. (2011e). Identify data with long-term value. Retrieved from https://www.dataone.org/best-practices/identify-data-long-term-value

DataONE. (2011f). Provide capabilities for tagging and annotation of your data by the community. https://www.dataone.org/best-practices/provide-capabilities-tagging-and-annotation-your-data-community

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing*, *5*(5), 59–68. doi:10.1109/4236.957896. Retrieved fromhttp://www.computer.org/csdl/mags/ic/2001/05/w5059-abs.html

SEDAC. (2014). Citing our data. Retrieved from http://sedac.ciesin.columbia.edu/citations

4.4 Process Assessment

***Process Assessment** includes Measurement and Analysis and Verifying Implementation. Measurement and Analysis describes the need to measure the process and analyze the measurements, and typically includes examples of the measurements that could be taken to determine the status and effectiveness of the Activities Performed. Verifying Implementation describes the steps to ensure that the activities are performed in compliance with the process that has been established, and typically encompasses reviews and audits by management and quality assurance.*

Process assessment for data dissemination follows the general guidelines as stated in chapter 1. It should be pointed out that the assessment of the dissemination process area can be easily confused with the outcome assessment such as impact and usage of data. Assessment of the data dissemination process aims at establishing appropriate quantitative measurements so that through consistent data gathering on these measurements, the RDM personnel can assess the process systematically on a regular basis for continuous improvement.

4.4.1 Measurement and Analysis

Assessment of the data dissemination process should stay focused on measurements that can tell how effectively and efficiently the process was performed. Example measurements include the time taken from data submission to release with full metadata description, number of venues used for dissemination, and the increase/decrease in data access that may be attributed to data dissemination efforts.

Collecting data on the dissemination process is not always straightforward. For example, once a dataset is ready for dissemination, metadata has to be created and reviewed, rights terms and access permissions defined, and venues for dissemination organized. Some of these steps may take longer to complete (e.g., the rights terms may involve legal consultation) while others may be in the form of notes rather than quantitative data. Having tools and procedures for collecting the data will not only make the data collection efficient and consistent but also enable the process assessment to occur routinely, rather than on an ad hoc basis.

4.4.2 Verifying Implementation

A higher level of capability maturity (level 4 or 5) requires that the implementation of policies and procedures be verified to ensure that the process is adequately executed with a reasonable degree of quality. For example, questions may be asked during the verification review:

- Are data being shared?
- Is the data archive accessible?
- Are confidential data secure?

Verifying implementation of policies, ability to perform, and activities performed provides the opportunity for RDM personnel to identify problems early and hopefully correct the problems early enough, before they become worse.

Rubric

Rubric for 4.4 - Process Assessment

Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for measurement, analysis, or verification to ensure accessibility and security of data
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Measurement, analysis, or verification to ensure accessibility and security of data have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Measurement, analysis, or verification to ensure accessibility and security of data have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to measurement, analysis, or verification to ensure accessibility and security of data, as defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established including measurement, analysis, and verification to ensure accessibility and security of data, and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding measurement, analysis, or verification to ensure accessibility and security of data are

evaluated on a regular basis, and necessary improvements are implemented

5. Repository Services and Preservation

Overall goal: Keep research data accessible, even as hardware, software, and storage media change.

An important function of the research data lifecycle is data preservation, drawing on a combination of technological and institutional infrastructures to ensure that data are maintained in the state expected by users. Aspects of preservation to consider include availability, consistency, privacy, integrity, and audit.

- Availability means that users are able to access the data as needed.
- Consistency means that the system behaves in the ways expected by the users.
- Privacy means that only authorized users can view data.
- Integrity means that only authorized users can change data and that data can only be changed in specified ways.
- Audit means that access and changes to the data are recorded as needed to ensure the provenance of the data.

Data preservation is a consideration across the life of a research project, though the nature and expected level of performance will evolve. For example, considering privacy, while data are being actively collected and analyzed, they might be stored locally and available only to members of the research team, while later in the project, curated datasets might be made available to the public through project, institutional or disciplinary data repositories. To ensure availability, data should be regularly backed up, more frequently if data are still being collected and analyzed. Long-term storage of data adds additional concerns about preservation of data across the inevitable changes in the underlying technologies and hosting institutions.

5.1 Commitment to Perform

***Commitment to Perform** describes the actions the organization must take to ensure that the process is established and will endure. Commitment to Perform typically involves establishing organizational policies and senior management sponsorship.*

5.1.1 Develop data preservation policies

Projects should develop data preservation policies that specify required level of access to data and needed controls on viewing and changing data. The goal of developing data preservation policies is to guide development of systems that operate as expected by users.

Development of data preservation policies is necessary to ensure that data are preserved in a cost-effective way consistent with user expectations, while maintaining desired controls on accessing and changing data.

Data preservation policies should be based on an analysis of the risks to which the data are exposed and the expectations of users. For example, a common risk facing all data systems is a loss of data due to failure of or damage to hardware, so such events should be expected and planned for. On the other hand, while commercial data may have a financial value that makes them attractive to criminals, research data might not pose such risks. Risks can be classified by likelihood of occurrence and expected impact. Likely high impact risks (e.g., a disk drive failing and destroying stored data) should be prevented (e.g., by using redundant storage so a single disk failure has no impact). Unlikely high impact risks (e.g., the building burning down) should be planned for (e.g., by keeping off site backups). Likely low impact risks (e.g., a user error in editing a data item) should be controlled (e.g., by keeping an audit trail). Unlikely low impact risks might just be ignored. Risks should be considered broadly, including technical risks (e.g., hardware or software errors), human risks (e.g., operator errors) and institutional risks (e.g., a data repository ceasing operation). Based on the risk analysis, data preservation policies should state what data are being preserved and against what risks. Identifying the likelihood and impact of risks will help ensure that resources are directed to the most important risks and that risks are not overlooked.

User expectations regarding data should be considered. For example, for a small project, it may be acceptable to lose access to data for a few days while replacing a failed server, while for others such a failure might be unacceptable, justifying the cost to maintain redundant hardware. Again, identifying user needs will help ensure that resources are spent appropriately.

Finally, data preservation policies should state who is responsible for the preservation of the data and identify acceptable and unacceptable behaviors. For example, considering data access, policies should state who can access data; considering data integrity, who can change data and under what circumstances.

5.1.2 Develop data backup policies

To backup data means to make a copy of the data that can be used in case the primary data store is damaged or lost. The goal of developing data backup policies is to provide guidance to data curators about how data should be backed up and to identify roles and responsibilities of personnel for creating, maintaining and using backups (DataONE, 2011a).

It is important to define backup policies to ensure that data are being backed up appropriately, that backups are properly protected and that responsibilities are clearly delineated.

The backup policy should describe what data need to backed up and how frequently, where backups are kept and for how long, and who can access them (DataONE, 2011b). The policy may also dictate the hardware and software to be used. If backups are not automatic, the policy should state who performs the backups. The policy should also state how and how often backups are validated and what metrics are used to evaluate backups.

5.1.3 Develop data curation policies

Projects create a variety of kinds of data, as well as data documentation and analysis scripts or tools. Data curation policies state what data should be preserved long-term and what data can be discarded. The goal of developing data curation policies is to provide guidance for data curators and users on deciding what data should be preserved.

Development of curation policies is necessary because data may have long-term value that should be preserved, but keeping all data is neither practical nor economically feasible (DataONE, 2011c). Only datasets that have significant long-term value and that cannot be recreated or that are costly to reproduce should be preserved.

In developing curation policies, consider the tradeoff between the cost of preservation due to the dataset size or repository policies against the potential value of the data to the user community (Hook et al., 2010). Funding agencies or institutions may also have requirements and policies governing contribution to repositories (DataONE, 2011c).

DataOne suggests that "raw data are usually worth preserving" (DataONE, 2011d). Data that have undergone a quality control check may be costly to recreate and so should be preserved. On the other hand, intermediate products in an analysis might be voluminous and easy to recreate and so not worth preserving. Source code is generally small and so likely worth preserving.

Rubric

Rubric for 5.1 - Commitment to Perform

Level 0 This process or practice is not being observed	No steps have been taken to establish organizational policies or senior management sponsorship for data preservation, curation, or backups
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Data preservation, curation, and backups have been considered minimally by individual team members, but nothing has been codified or included in organizational policies or senior management sponsorship
Level 2: Managed DM process is characterized for projects and often reactive	Data preservation, curation, and backups have been addressed for this project, but have not taken wider community needs or standards into account and have not resulted in organizational policies or senior management sponsorship
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to data preservation, curation, and backups that have been defined for the entire community or institution, as codified in organizational policies with senior management sponsorship
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding data preservation, curation, and backups, and are codified in organizational policies with senior management sponsorship; both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding data preservation, curation, and backups are evaluated on a regular basis, as codified in organizational policies with senior management sponsorship, and necessary improvements are implemented

References

DataONE. (2011a). Create and document a data backup policy. Retrieved from https://www.dataone.org/best-practices/create-and-document-data-backup-policy

DataONE. (2011b). Ensure integrity and accessibility when making backups of data. Retrieved from https://www.dataone.org/best-practices/ensure-integrity-and-accessibility-when-making-backups-data

DataONE. (2011c). Identify data with long-term value. Retrieved from https://www.dataone.org/best-practices/identify-data-long-term-value

DataONE. (2011d). Decide what data to preserve. Retrieved from https://www.dataone.org/best-practices/decide-what-data-preserve

Hook, L. A., Vannan, S. K. S., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from http://daac.ornl.gov/PI/BestPractices-2010.pdf

5.2 Ability to Perform

Ability to Perform *describes the preconditions that must exist in the project or organization to implement the process competently. Ability to Perform typically involves resources, organizational structures, and training.*

For data repository and presentation services, Ability to Perform includes enabling technologies, procedures, and business models that will sustain the services.

5.2.1 Appraise and select enabling technologies

Projects need to select the hardware and software technology platforms on which they will store their data. The selection process should be started early in the project to allow time to collect and evaluate information on available options, such as system documentation or experiences from other users. Larger projects may want to pilot several alternatives before making a choice. Relevant system features include functionality, in particular, support for multimedia data (DataONE, 2011f), fit to project needs (e.g., capabilities compared to the expected volume of data and number of users), ease of use, and support. Relevant hardware features include capacity, reliability and expected lifetime (e.g., for hard drives) (DataONE, 2011d).

Projects may develop their own data archives in addition to working stores for data being actively used. Rather than archiving data themselves, projects may decide to deposit data in an existing repository. Again, the process of selecting a repository should start early to provide enough time to identify and evaluate alternatives. As well, repositories may have particular requirements that will shape the project's data management plan (DataONE, 2011e). A further possibility for data preservation is joining a digital preservation network, that is, collaborating with other institutions or projects to cooperatively archive data (e.g., the Digital Preservation Network, http://dpn.org/, or Chronopolis, http://chronopolis.sdsc.edu).

5.2.2 Develop business models for preservation

Preserving data has costs that will extend long past the end of the projects that generate the data. It is therefore critical to develop business models for funding the ongoing preservation of data to ensure the long-term preservation of archived data.

Current data repositories are either funded by grants or self-supported. Funding agencies such as NSF and NIH have awarded a good number of grants to support the initiation of major data repositories (DataOne, Dataverse, GenBank, to name a few) and the long-term preservation for some of these data repositories. Business models used in the self-supported category include a wide variety of options: individual and institutional memberships, subscriptions, pay-per-submission, and voucher plans (Dryad, 2014). Generally, large reference collections of data (note 1), e.g., Genbank (http://www.ncbi.nlm.nih.gov/genbank/), the Knowledge Network for Biocomplexity (KNB) (https://knb.ecoinformatics.org/), and BioProject (http://www.ncbi.nlm.nih.gov/bioproject), are mostly supported by continued funding from the government, while resource collections of data (note 2), that are usually created by a disciplinary community for a refined scope, tend to have initial funding from the government but are increasingly required to become self-supported. The Dryad data repository so far has had a successful record in the self-supporting category.

It is the self-supported model that makes it ever more important to plan early and know what options there are to choose from. In the case of using self-supported data repositories, institutions or projects that decided to use the services can compare the cost between building an in-house repository and subscribing to data repository services. Costs to be covered include maintenance and operation of the hardware and institution infrastructure and necessary migration to new data formats and platforms.

5.2.3 Develop backup procedures and training

Projects should develop clear backup procedures. Documented procedures are necessary to ensure that data are backed up according to policy and that procedures to recover from problems are established and widely known (DataONE, 2011c). Procedures should identify all data that are to be backed up. They should set a clear schedule for making backups that is tailored to the data collection process (DataONE, 2011a). Streaming data should be backed up at regularly scheduled points in the collection process (DataONE, 2011a).

Procedures should identify who is responsible for creating the backups, including alternatives in case one person is unavailable (DataONE, 2011b). Backups may be automated, in which case someone should be responsible for regularly checking that they are being made. There may be different backup procedures for different data sets (DataONE, 2011c). Multiple versions of backups should be kept, e.g., to be able to recover from file damage that is not detected immediately.

The procedures should ensure that data backups are subject to the same protections as the original data (e.g., that confidential data are protected).

Finally, the procedures to recover from a backup copy should be described (DataONE, 2011a), both for individual files as well as for recovery from catastrophic failures. Responsibility for recovery should be assigned. Further, in the event of a failure, the recovery procedure must ensure that the backups will not be damaged by the same problem.

Personnel involved with backups should be trained in the relevant policies and procedures, including policies and procedures for data security.

Rubric

Rubric for 5.2 - Ability to Perform

Level 0 This process or practice is not being observed	No steps have been taken to provide for resources, structure, or training with regards to enabling technlogies or business models for data preservation
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Resources, structure, and training with regards to enabling technlogies or business models for data preservation have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Resources, structure, and training with regards to enabling technlogies or business models for data preservation have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project provides resources, structure, and training with regards to enabling technlogies or business models for data preservation, as defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established for resources, structure, and training with regards to enabling technlogies or business models for data preservation, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding resources, structure, and training, with regards to enabling technlogies or business models for data preservation are evaluated on a regular basis, and necessary improvements are implemented

Notes

1. Reference collections are authored by (and serve) large segments of the science and engineering community and conform to robust, well-established and comprehensive standards, which often lead to a universal standard. Budgets are large and are
often derived from diverse sources with a view to indefinite support. Retrieved
from http://www.nsf.gov/pubs/2007/nsf0728/nsf0728_4.pdf, p.23.

2. Resource collections are authored by a community of investigators, often within a domain of science or engineering,
and are often developed with community level standards. Budgets are often intermediate in

size. Lifetime is between the mid- and long-term. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728_4.pdf, p.22.

References

DataONE. (2011a). Backup your data. Retrieved from https://www.dataone.org/best-practices/backup-your-data

DataONE. (2011b). Create and document a data backup policy. Retrieved from https://www.dataone.org/best-practices/create-and-document-data-backup-policy

DataONE. (2011c). Ensure integrity and accessibility when making backups of data. Retrieved from https://www.dataone.org/best-practices/ensure-integrity-and-accessibility-when-making-backups-data

DataONE. (2011d). Ensure the reliability of your storage media. Retrieved from https://www.dataone.org/best-practices/ensure-reliability-your-storage-media

DataONE. (2011e). Identify suitable repositories for the data. Retrieved from https://www.dataone.org/best-practices/identify-suitable-repositories-data

DataONE. (2011f). Plan for effective multimedia management. Retrieved from https://www.dataone.org/best-practices/plan-effective-multimedia-management

Dryad. (2014). Pricing plans and data publishing charges. Retrieved from http://datadryad.org/pages/pricing

5.3 Activities Performed

Activities Performed *describes the roles and procedures necessary to implement a key process area. Activities Performed typically involve establishing plans and procedures (i.e., the specific actions that need to be performed), performing the work, tracking it, and taking corrective actions as necessary.*

5.3.1 Store data

A key function in data management is storing the data both for current use and for long-term archiving. Earlier sections discussed logical formats for data storage; in this section, we focus on physical storage. All storage devices, locations and access accounts should be documented and accessible to team members (DataONE, 2011a). Data should be stored in non-proprietary hardware formats (Borer et al., 2009) so that they can be read even if the original hardware is not available (e.g., many hardware RAID devices use proprietary disk formats, so a failed RAID controller must be replaced with the same model). Media should be handled and stored carefully (DataONE, 2011d). Data discs should be routinely inspected and replaced as needed (DataONE, 2011d). Storing data solely on local hard drives or servers is not recommended: keeping multiple copies of the data files in separate locations is safer (DataONE, 2011e).

5.3.2 Provide data security

Confidential data has to be stored in such a way as to restrict access to authorized personnel (Columbia Center for New Media Teaching and Learning, n.d.). Data should be secured in accordance with developed data access polices. Possible access controls include physical security on the hardware and allowing only properly authenticated users access to the data. User might have to sign license agreements governing how data are used and protected. Highly confidential data might be accessed only from particular locations, rather than being distributed to users.

5.3.3 Control changes to data files

The original data set should be preserved in its original state (Borer et al., 2009; DataONE, 2011f; Hook et al., 2010). Unaltered images should be preserved at the highest resolution possible. (DataONE, 2011e).

Changes to data files should be controlled, that is, appropriate tools, such as version control tools, should be used to keep track of the history of changes to the data files (Hook et al., 2010). Changes should be made only by users authorized by the developed data access policies. The nature of and reasons for the changes should be recorded. In particular, users should be aware of, and document, any changes in the coding scheme (Hook et al., 2010). A further danger of using applications such as spreadsheets to store data is that these programs are designed to facilitate making changes to the data, while for scientific data, changes should be controlled.

It may be appropriate to provide multiple versions of data products with defined identifiers for unambiguous reference, reflecting the state of the data at different points in time (DataONE, 2011g).

5.3.4 Backup data

Data, processing codes, and documentation should be regularly backed up (Hook et al., 2010) according to the defined procedures to ensure that there are at least two (and preferably more) copies of all important data. Backup devices should be selected for and regularly checked for reliability. Backups should be regularly tested for completeness and correctness to ensure that backup copies have the same content as the original data file (DataONE, 2011c). Backups might include periodic full backups (i.e., all files) as well as more frequent incremental backups (i.e., backing up only data that have changed since the last backup). The backups should also be checked to ensure that they are secure and and that only those who need access to backups have proper access (DataONE, 2011c). Contact information should be available for the persons responsible for the backed up data (DataONE, 2011c).

A copy of the backup should be kept at a trusted off-site location (DataONE, 2011b). As well, keeping backup copies of data off-line will help ensure that they will are not affected by any system problems or software errors that damage the primary copy (Borer et al., 2009). Copies of physical data stores such as lab notebooks and samples should also be regularly stored off-site for safe keeping (Columbia Center for New Media Teaching and Learning, n.d.).

5.3.5 Curate data

Data should be selected for long-term storage according to the developed curation policies and copied to the appropriate repositories. Data that are not selected for long-term storage should be disposed of on a determined schedule. The disposition of datasets should be recorded.

5.3.6 Perform data migrations

In a long-running project, it may be necessary to migrate data to newer hardware or software formats. Such migrations should be carefully planned so they are not disruptive to the research process. When new hardware is installed, it is prudent to keep the old hardware with its copy of the data until the new device “settles in” and is deemed reliable (DataONE, 2011d).

When new versions of software are released, it is prudent to continue using the version of the software that was originally used to create a data file to view and manipulate the file contents (DataONE, 2011f). If it is necessary to use a newer version of a software package to open files created with an older version of the application, first save a copy of the original file in case there are problems with the migration. Implementation of new versions of software should be coordinated across a research group to avoid compatibility problems.

Rubric

	Rubric for 5.3 - Activities Performed
Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for the workflow of data preservation, including storage, security, version control, and migration
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	The workflow of data preservation, including storage, security, version control, and migration, has been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	The workflow of data preservation, including storage, security, version control, and migration, has been addressed for this project, but has not taken wider community needs or standards into account and has not been codified
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to the workflow of data preservation, including storage, security, version control, and migration, that have been defined for the entire community or institution
Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding the workflow of data preservation, including storage, security, version control, and migration, and both data and practices are systematically measured for quality

Level 5: Optimizing Focus on process improvement	Processes regarding the workflow of data preservation, including storage, security, version control, and migration, are evaluated on a regular basis, and necessary improvements are implemented

References

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America*, *90*(2), 205–214. http://dx.doi.org/10.1890/0012-9623-90.2.205

Columbia Center for New Media Teaching and Learning. (n.d.). Responsible conduct of research: Data acquisition and management: Foundation text. Retrieved from http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/index.html#3_B

DataONE. (2011a). Advertise your data using datacasting tools. https://www.dataone.org/best-practices/advertise-your-data-using-datacasting-tools

DataOne. (2011b). Backup your data. https://www.dataone.org/best-practices/backup-your-data

DataONE. (2011c). Ensure integrity and accessibility when making backups of data. https://www.dataone.org/best-practices/ensure-integrity-and-accessibility-when-making-backups-data

DataONE. (2011d). Ensure the reliability of your storage media. https://www.dataone.org/best-practices/ensure-reliability-your-storage-media

DataONE. (2011e). Plan for effective multimedia management. https://www.dataone.org/best-practices/plan-effective-multimedia-management

DataONE. (2011f). Preserve information: keep your raw data raw. https://www.dataone.org/best-practices/preserve-information-keep-your-raw-data-raw

DataONE. (2011g). Provide version information for use and discovery. https://www.dataone.org/best-practices/provide-version-information-use-and-discovery-0

Hook, L. A., Vannan, S. K. S., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from http://daac.ornl.gov/PI/BestPractices-2010.pdf

5.4 Process Assessment

Process Assessment *includes Measurement and Analysis and Verifying Implementation. Measurement and Analysis describes the need to measure the process and analyze the measurements, and typically includes examples of the measurements that could be taken to determine the status and effectiveness of the Activities Performed. Verifying Implementation* *describes the steps to ensure that the activities are performed in compliance with the process that has been established, and typically encompasses reviews and audits by management and quality assurance.*

5.4.1 Measurement and Analysis

Projects should develop and implement metrics for the data storage and preservation process. Example metrics include the amount of data being stored vs. the available storage space, hardware failure rates, how long data backups take to complete, or how long it takes to recover from a backup.

5.4.2 Validate data storage

Projects should routinely check the integrity of data stored on hard drives, discs or tapes (DataONE, 2011c). Such checks are particularly important if data are being collected automatically over time. For example, a checksum might be stored for each file and periodically checked to ensure that the files haven't changed. The readability of files might be checked as part of the regular backup procedure.

5.4.3 Validate backups

Data backups should be regularly checked to be sure that the backups are being made and that the backup copies are identical to the original data (DataONE, 2011a), e.g., by periodically retrieving the backup file, opening it on a separate system, and comparing it to the original file (DataONE, 2011b). Drills should be run periodically to validate the procedures for recovering data and systems from the backups.

Rubric

Rubric for 5.4 - Process Assessment

Level 0 This process or practice is not being observed	No steps have been taken to establish procedures for measurement, analysis, or verification of data storage or backups
Level 1: Initial Data are managed intuitively at project level without clear goals and practices	Measurement, analysis, and verification of data storage and backups have been considered minimally by individual team members, but not codified
Level 2: Managed DM process is characterized for projects and often reactive	Measurement, analysis, and verification of data storage and backups have been recorded for this project, but have not taken wider community needs or standards into account
Level 3: Defined DM is characterized for the organization/community and proactive	The project follows approaches to measurement, analysis, and verification of data storage and backups that have been defined for the entire community or institution

Level 4: Quantitatively Managed DM is measured and controlled	Quantitative quality goals have been established regarding measurement, analysis, and verification of data storage and backups, and both data and practices are systematically measured for quality
Level 5: Optimizing Focus on process improvement	Processes regarding measurement, analysis, and verification of data storage and backups are evaluated on a regular basis, and necessary improvements are implemented

References

DataOne. (2011a). Backup your data. https://www.dataone.org/best-practices/backup-your-data

DataONE. (2011b). Ensure integrity and accessibility when making backups of data. https://www.dataone.org/best-practices/ensure-integrity-and-accessibility-when-making-backups-data

DataONE. (2011c). Ensure the reliability of your storage media. https://www.dataone.org/best-practices/ensure-reliability-your-storage-media

Bibliography

Ailamaki, A., Ioannidis, Y.E., & Livny, M. (1998). Scientific workflow management by database management. In: Proceedings of the Tenth International Conference on Scientific and Statistical Database Management, Capri, Italy, July 1-3, 1998. Retrieved from http://www.cs.cmu.edu/~natassa/aapubs/conference/scientific-workflow-management.pdf

Australian National Data Service (2011). Research data management framework: Capability maturity guide. Retrieved from http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.pdf

Black Rock Forest Consortium. (2007). Data submission protocol. Retrieved from http://www.blackrockforest.org/docs/scientist-resources/DataResources/DataSubmission.html

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America*, *90*(2), 205–214. http://dx.doi.org/10.1890/0012-9623-90.2.205

Brase, J., Socha, Y., Callaghan, S., Borgman, C.L., Uhlir, P.F., Carroll, B. (2014). Data citation: Principles and practice. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press.

Brooks Jr, F. P. (2010). *The design of design: Essays from a computer scientist*. Pearson Education.

Brown, D.A, Brady, P.R., Dietz, A., Cao, J., Johnson, B., & McNabb, J. (2006). A case study on the use of workflow technologies for scientific analysis: Gravitational wave data analysis, in I.J. Taylor, E. Deelman, D. Gannon, and M.S. Shields(Eds.), Workflows for e-Science, chapter 5, pp. 41–61. Berlin: Springer-Verlag.

Carlson, S. (2006). Lost in a sea of science data. *The Chronicle of Higher Education*, *52*(42). Retrieved from http://chronicle.com/weekly/v52/i42/42a03501.htm

CMMI Product Team. (2006). *CMMI for Development, Version 1.2* (No. CMU/SEI-2006-TR-008). Pittsburgh, PA, USA: Carnegie Mellon Software Engineering Institute. Retrieved from http://repository.cmu.edu/sei/387

Columbia Center for New Media Teaching and Learning. (n.d.). Responsible conduct of research: Data acquisition and management: Foundation text. Retrieved from http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/index.html#3_B

Cornell University Library. (2007). Cornell University Library personas. Retrieved from http://hdl.handle.net/1813/8302

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). Managing and Sharing Research Data: A Guide to Good Practice. Los Angeles, CA: SAGE.

DataONE. (2011). Best Practices. Retrieved from https://www.dataone.org/best-practices

DataONE. (2013). Member node description: PISCO. Retrieved from http://www.dataone.org/sites/all/documents/DataONEMNDescription_PISCO.pdf

D'Ignazio, J., & Qin, J. (2008). Faculty data management practices: A campus-wide census of STEM departments. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1–6. doi:10.1002/meet.2008.14504503139 . Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/meet.2008.14504503139/abstract

Dryad. (2014). Pricing plans and data publishing charges. Retrieved from http://datadryad.org/pages/pricing

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. doi:10.1177/0306312711413314. Retrieved from http://pne.people.si.umich.edu/PDF/EdwardsEtAl2011ScienceFriction.pdf

Faniel, I. M., & Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *International Journal of Digital Curation*, *6*(1), 58–69. doi:10.2218/ijdc.v6i1.172. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/163/231

Gray, J. (2007). Jim Gray on eScience: A transformed scientific method. In K. M. Tolle, D. Tansley, & A. J. G. Hey (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. 5–12). Edmond, WA: Microsoft Research. Retrieved from http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf

Hale, S. S., Miglarese, A. H., Bradley, M. P., Belton, T. J., Cooper, L. D., Frame, M. T., et al. (2003). Managing Troubled Data: Coastal Data Partnerships Smooth Data Integration. *Environmental Monitoring and Assessment*, *81*(1-3), 133–148. doi:10.1023/A:1021372923589. Retrieved from http://link.springer.com/article/10.1023%2FA%3A1021372923589

Hook, L. A., Vannan, S. K. S., Beaty, T. W., Cook, R. B., & Wilson, B. E. (2010). Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center. Retrieved from http://daac.ornl.gov/PI/BestPractices-2010.pdf

Hubbard Brook Ecosystem Study. (2014). Data use policy. Retrieved from http://www.hubbardbrook.org/data/dataset.php?id=4

Jahnke, L., Asher, A., & Keralis, S. D. (2012). The problem of data. Council on Library and Information Resources (CLIR) Report, pub. #154. ISBN 978-1-932326-42-0 Retrieved from http://digitalcommons.bucknell.edu/fac_pubs/52/

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing*, *5*(5), 59–68. doi:10.1109/4236.957896. Retrieved from http://www.computer.org/csdl/mags/ic/2001/05/w5059-abs.html

Karasti, H., & Baker, K. S. (2008). Digital data practices and the long term ecological research program growing global. *International Journal of Digital Curation, 3*(2), 42–58. doi:10.2218/ijdc.v3i2.57. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/86

Key Perspectives. (2010). *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability*. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/projects/scarp

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder. Portal: Libraries and the Academy, 11(4): 915-937. doi:10.1353/pla.2011.0049. Retrieved from http://www.press.jhu.edu/journals/portal_libraries_and_the_academy/portal_pre_print/current/articles/11.4lage.pdf

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, Texas: Stata Press Books.

Lyon, L., Ball, A., Duke, M., & Day, M. (2012). Community capability model framework. White Paper. UKOLN, University of Bath & Microsoft Research. Retrieved from http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf

Mayernik, M. S. (2010). Metadata tensions: A case study of library principles vs. everyday scientific data practices. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–2. doi:10.1002/meet.14504701337. Retrieved from http://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/337_Final_Submission.pdf

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). New York, NY, USA: ACM. doi:10.1145/1940761.1940818. Retrieved from http://doi.acm.org/10.1145/1940761.1940818

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3–7. doi:10.1016/j.ecoinf.2005.08.004. Retrieved from http://www.sciencedirect.com/science/article/pii/S157495410500004X

Mullins, J. (2007). Enabling international access to scientific data sets: Creation of the Distributed Data Curation Center (D2C2). Purdue University, Purdue E-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1100&context=lib_research

Murray-Rust, P. (2008). Chemistry for everyone. *Nature*, 451, 648-651. Retrieved from http://www.nature.com/nature/journal/v451/n7179/full/451648a.html

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993a). Capability maturity model, Version 1.1. IEEE Software, 10(4): 18-27. doi:10.1109/52.219617. Retrieved from http://www.computer.org/csdl/mags/so/1993/04/s4018-abs.html

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993b). *Capability Maturity Model for Software, Version 1.1* (No. CMU/SEI-93-TR-024). Software Engineering Institute. Retrieved from http://resources.sei.cmu.edu/library/asset-view.cfm?assetID=11955

Pohl, K. & Rupp, C. (2011). Requirements Engineering Fundamentals: Study Guide for the Certified Engineering Exam. Sebastopol, CA: O'Reilly Media.

Protein Data Bank. (2014). Policies and references [of Protein Data Bank]. Retrieved from http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html

Qin, J., & D'ignazio, J. (2010). The Central Role of Metadata in a Science Data Literacy Course. *Journal of Library Metadata*, *10*(2-3), 188–204. doi:10.1080/19386389.2010.506379. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/19386389.2010.506379

Qin, J., D'Ignazio, J., & Baldwin, S. (2011). A workflow-based knowledge management architecture for geodynamics data. A White paper submitted to NSF GEO/OCI EarchCube Charrette meeting. Retrieved from http://earthcube.ning.com/group/user-requirements/forum/topics/white-paper-a-workflow-based-knowledge-management-architecture

Ray, J. M. (2014). Introduction to research data management. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences)*. West Lafayette, Indiana:Purdue University Press.

Riley, Jenn. (2014). Metadata services. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press.

Sallans, A. & Lake, S. (2014). Data management assessment and planning tools. In J. Ray (Ed.),*Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press. Retrieved from http://books.google.com/books?id=qZStAQAAQBAJ&pg=PA87&dq=dmvitals&source=gbs_toc_r&cad=3#v=onepage&q=dmvitals&f=false

Sheaffer, P. (2012). Creating a sustainable business model for a digital repository: the Dryad experience. ASIS&T Research Data Access and Preservation Summit 2012, Baltimore, MD. Retrieved from http://www.slideshare.net/asist_org/creating-a-sustainable-business-model-for-a-digital-repository-the-dryad-experience-peggy-schaeffer-rdap12

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., et al. (2008). *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library* (Working Paper). Retrieved from http://hdl.handle.net/1813/10903

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. doi:10.1371/journal.pone.0021101. Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101

UK Data Archive. (2014). Create and manage data: Documenting your data. Retrieved from http://www.data-archive.ac.uk/create-manage/document

Van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. (2011). Managing and Sharing Data: A Best Practice Guide for Researchers. (3rd ed.) Essex, England: University of Essex. Retrieved from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Walters, T. O. (2009). Data curation program development in U.S. universities: The Georgia Institute of Technology example. *International Journal of Digital Curation, 4*(3), 83–92. doi:10.2218/ijdc.v4i3.116. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/136

Walters, T. & Skinner, K. (2011). New roles for new times: Digital curation for preservation. Retrieved from http://www.arl.org/focus-areas/workforce/1086

Westra, B. (2014). Developing Data Management Services for Researchers at the University of Oregon. In J. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals (Charleston Insights in Library, Information, and Archival Sciences).* West Lafayette, Indiana: Purdue University Press.

Willis, C., Greenberg, J., & White, H. (2012). Analysis and Synthesis of Metadata Goals for Scientific Data. *Journal of American Society for Information Science and Technology*, 1505–1520. Retrieved from http://scholarship.law.duke.edu/faculty_scholarship/2713

Zeng, M. L. & Qin, J. (2014). Metadata. Chicago, IL: ALA Neal Schuman.