A Capability Maturity Model for Scientific Data Management: Evidence from the Literature

Kevin Crowston
Syracuse University School of Information Studies
crowston@syr.edu

Jian Qin
Syracuse University School of Information Studies
jqin@syr.edu

ABSTRACT

In this paper, we propose a capability maturity model (CMM) for scientific data management (SDM) practices, with the goal of supporting assessment and improvement of these practices. The model describes key process areas and practices necessary for effective SDM. Appropriate SDM practices were identified by content analysis of papers about SDM and include both those specific SDM practices and generic process management practices. The CMM further characterizes organizations by the level of maturity of these processes, meaning the organizational capability to reliably perform SDM processes. The model may be useful to organizations in evaluating and planning improvements to their SDM practices.

Keywords

Scientific data management, capability maturity model.

INTRODUCTION

Science is increasingly data-intensive, highly collaborative and highly computational at a large scale. The tools, content and social attitudes for supporting multidisciplinary collaborative science require "new methods for gathering and representing data, for improved computational support and for growth of the online community" (Murray-Rust, 2008). Given these developments, scientific data management (SDM) is now at center stage in the research cycle, leading to a data lifecycle: data capture, curation, analysis and visualization (Gray, 2007), sharing and preservation, discovery and reuse. In this paper, we propose a capability maturity model (CMM) for SDM, with the goal of supporting assessment and improvement of SDM practices to increase the reliability of SDM.

Currently, SDM practices vary greatly depending on the scale, discipline, funding and type of projects. "Big science" research—such as astrophysics, geosciences, climate science and system biology—generally has established well-defined SDM policies and practices, with supporting data repositories for data curation, discovery and reuse. SDM in these disciplines often has significant funding support for the necessary personnel and technology infrastructure. By contrast, in "small science" research, that is, projects involving a single PI and a few students, SDM is typically less well developed. However, even in these fields, such practices are still needed: 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 science projects.

The importance of SDM 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, SDM is an interdisciplinary field: communities of practice involve scientists, information technology professionals, librarians and graduate students, each bringing their domain-specific culture and practices to bear on SDM. 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 SDM. The goal of this paper is to present the first steps towards development of such a model, in the form of a Capability Maturity Model (CMM) for SDM.

A CMM FOR SCIENTIFIC DATA MANAGEMENT

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, 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 model, 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, 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).

Finally, the CMM described five levels of process or capability maturity for organizations as a whole, representing the "degree of process improvement across a predefined set of process areas" and corresponding to the generic goals listed above. The initial level describes an organization with no defined processes: software is developed, 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, as implemented by the higher numbered generic processes. 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 paper is to lay out a similar path for the improvement of scientific data management.

IDENTIFYING KEY PRACTICES AND PROCESS AREAS FOR SCIENTIFIC DATA MANAGEMENT

While the organizational maturity levels are most well-known aspect of the SEI CMM, its heart is the description of the key practices clustered in a set of process areas. To create a CMM for SDM, we therefore sought to identify and cluster key SDM practices. As SDM represents an emerging interdisciplinary research field, the processes and practices areas are still being explored and understood. We therefore undertook a content analysis of SDM practices as described in the literature to develop this part of the model.

Research Method

Data collection for this paper involved selecting a set of published articles that describe SDM practices, either as the main topic or as part of the description of a particular study or research group. Articles were found in journals and conference proceedings devoted to data curation, preservation and management. The selection of articles followed a "purposeful sampling" method. The goal of purposeful sampling is to yield "insights and in-depth understanding rather than empirical generalizations" (Patton, 2002, p. 230). Since our study goal was to identify as many data practice areas as possible in order to understand and gain insights into science data management processes, we felt that it would be more effective to select an article set and examine the content until the practice areas reach saturation. The main limitation of this sampling approach is that it does not provide a basis for estimating the relative frequencies of the practices, though given our data source, such inference would be limited to the frequency of mention even in the best case.

We selected approximately 20 articles and reports in the areas of data curation, data management, and data science. The selected articles were imported into the content analysis software NVivo. The authors then read through the articles separately, coding paragraphs in the articles that described some kind of SDM practice. After this individual coding process, the two authors discussed each of the described practices found in the articles and collected together different descriptions that seemed to refer to the same practice. The practices were then grouped based on the goal they achieved into the key process areas. An initial set of process areas was created based on the data lifecycle and the SEI CMM model, but new areas were added based on the practices found, and some areas that seemed overlapping were collapsed.

FINDINGS: SDM PRACTICES AND PROCESS AREAS

From our analysis we identified a large number of key practices for SDM. A number of these seemed to be specific practices for SDM, which we clustered into four process areas based on the high-level goal the practice helped achieve. These practices are shown below in Table 1. For each process area, we give the high level goal and list and briefly describe the practices we clustered under this heading.

Key Process Area	Practice	Example text
1: Data acquisition, processing and quality assurance Goal: Reliably capture and describe scientific data in a way that facilitates preservation and reuse	1.1 Capture/acquire data	"Geospatial data is received by the project either as a data download or as a set of files delivered on optical or magnetic media." (Morris & Tuttle, 2008)
	1.2 Process and prepare data for storage, analysis and distribution	"The Cornell Science Data Center (CSDC) will provide initial processing, analysis and archiving of the science data." (Grayzeck & Acton, 2002)
	1.3 Assure data quality (e.g., validate and audit data)	"…planned quality assurance and back-up procedures for data." (Van den Eynden et al., 2010)
2: Data description and representation Goal: Create quality metadata for data discovery, preservation, and provenance functions.	2.1 Develop and apply metadata specifications and schemas	"… adopted a non-geospatial metadata standard called the Ecological Metadata Language (EML) as a metadata specification for coordinating data access via a community catalogue." (Karasti & Baker, 2008)
	2.2 Contextualize, describe and document data	"…preserve the contextual metadata that establishes when and by whom the data was created." (Data Working Group, 2008)
	2.3 Document data, software, sensors and mission	"The software and associated documentation for the Level 2 data are archived both at the ISOC and at the SPDF and NSSDC." (Schwadron, 2007)
	2.4 Create descriptive and semantic metadata for datasets	"…create preliminary metadata for research data sets… complete a more detailed metadata record using a form-based editor…" (Data Working Group, 2008)
	2.5 Design mechanisms to link datasets with publications	"Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences." (http://datadryad.org/)
	2.6 Ensure interoperability with data and metadata standards	"…create preliminary metadata for research data sets," (Data Working Group, 2008)
	2.7 Ensure compliance to standards	"…carries out validation of data content, adequacy of documentation, and adherence to archiving standards…" (Schwadron, 2007)
3: Data dissemination Goal: Design and implement interfaces for users to obtain and interact with data	3.1 Identify and manage data products	"Level 3 data products consist of H-ENA All-sky flux maps…" (Schwadron, 2007)
	3.2 Encourage sharing	"…the above projects are aimed to develop methods and tools for marine data integration and sharing…" (JCOMM/IODE, 2010)
	3.3 Distribute data	"The data archive at the ISOC is maintained through the IBEX database. In addition, there is a Distributed Archive that is designed, generated, validated, packaged and distributed by the ISOC." (Schwadron, 2007)
	3.4 Provide access (e.g., by creating and piloting service models)	"…we need to create and pilot curation service models...Georgia Tech is piloting data curation services via the Fedora-based Islandora application..." (Walters, 2009)

4: Repository services/ preservation Goal: Preserve collected data for long-term use	4.1 Store, backup and secure data (e.g., by backing up databases, preserving datasets and enforcing security of data systems)	"High-security Firewalled ISOC computer which contains the base telemetry and archived science products; Normal security firewalled RAID system and computer for web interface and software processing contained outside the high-security firewall...." (Schwadron, 2007)
	4.2 Manage schedules for archive generation, validation and delivery	"The delivery schedule of five separate delivery dates for different portions of the mission will facilitate validation..." (Grayzeck & Acton, 2002)
	4.3 Curate data	"Neuroscience may be a leading example of a scientific domain that will curate its data in a diffused fashion…" (Walters, 2009)
	4.4 Perform data migration	"The library also tested the demand for and feasibility of a file format and media migration service…" (Data Working Group, 2008).
	4.5 Build digital preservation network	"...operates a distributed digital preservation network...and focuses largely on humanities and social science primary resources in digital form including datasets..." (Walters, 2009)
	4.6 Validate data archives	"… validation of the compliance of the archive… will be overseen by the PDS, in coordination with the Science Team." (Grayzeck & Acton, 2002)
	4.7 Package and deliver data archives	"The final data delivery will incorporate the entire archive, including the earlier data deliveries." (Grayzeck & Acton, 2002)

Table 1. Specific process areas for scientific data management

The key process areas and practices in each of the key process areas can be mapped to the data lifecycle stages, hence provide a framework for creating a checklist for data management in the planning stage and later for evaluating and assessing the performance and impact of data management. Although these key processes are developed according to data management workflows (thus bearing a sequential logic), it does not prevent them from being applied for particular purposes and projects in managing, preserving, and providing access to data. For example, the well-known digital curation lifecycle (DCC, 2011) defines digital data broadly to include digital surrogates of physical artifacts as well as born digital objects. The lifecycle includes steps such as conceptualize, create, access and use, appraise and select, dispose, ingest, preservation action, and reappraise. Many of these practices can be mapped with the key process areas in Table 1.

Generic practices and process areas

In addition to the specific practices discussed above, we also found a number of practices that seemed to best fit the SEI CMM Generic Goals, as they were oriented around managing or supporting the process of SDM rather than managing data directly. These are shown in Table 2, on the following page. Note that we numbered the generic processes to match the SEI CMM model, so there are gaps in the numbering in cases where we did not observe matching practices in our data collection.

SCIENTIFIC DATA MANAGEMENT MATURITY LEVELS

Perhaps the most well-known aspect of the CMM is the maturity levels, which describe the level of development of the practices in a particular organization. SDM practices as carried out in scientific projects similarly range from *ad hoc* to well-planned and well-managed processes (D'Ignzaio & Qin, 2008; Steinhart et al., 2008). The generic practices described above provide a basis for mapping these maturity levels into the context of SDM, as illustrated in Figure 1 and described below.

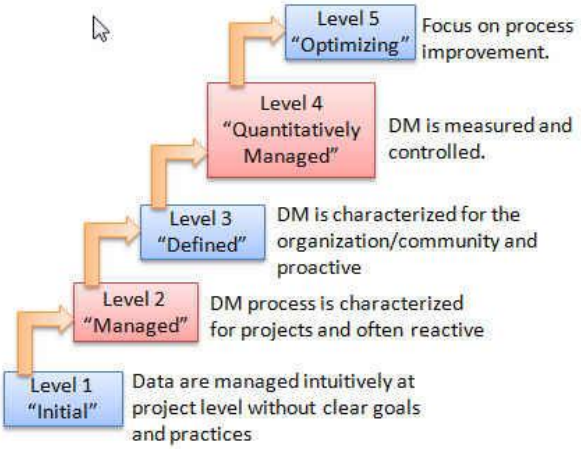

Figure 1. Capability maturity levels for SDM.

Generic process area 2: Institutionalize a managed process Goal: The data management process is institutionalized as a managed process.		
Generic Goal	**Generic Practices**	**Example text**
G2.1 The organization establishes policies for planning and performing the process	G2.1.1 Develop data release policies G2.1.2 Develop sharing policies G2.1.3 Develop policies for data rights and rules for data use G2.1.4 Develop data curation policies	"…developing local infrastructure and related policies in several areas…" (Data Working Group, 2008) "Policies are needed to govern submissions, selection, usage, and levels of service, at a minimum." (Witt, 2009)
G2.2 A data management plan is established and maintained	G2.2.1 Assess faculty data practices G2.2.2 Analyze data flows G2.2.3 Identify leading data management problems G2.2.4 Develop user requirements G2.2.5 Identify staffing needs	"Georgia Tech has devised and implemented its assessment techniques and has been interviewing groups of researchers." (Walters, 2009) "Identify new skills that will be needed … to support activities in the area of data curation and cyberinfrastructure, as well as the positions where those skills would be needed and applied." (Data Working Group, 2008)
G2.3 Resources are provided	G2.3.1 Develop business models for preservation G2.3.2 Appraise and evaluate enabling technologies (e.g., storage technology) G2.3.3 Develop SDM tools (e.g., tools to help researchers organize data, initial technology platforms or Web interface to the science repository) G2.3.4 Manage enabling technologies for access and conformance to standards	"…implementation of tools enabling members of multiple communities to supervise the creation (manual, semi-automatic, and automatic) of metadata, as well as the analysis, use, and preservation of data." (Borgman et al., 2006) "A 4-day Summer Institute on Data Curation (for practicing librarians) is also planned for 2008, and will address the topics of digital preservation, technical aspects of data repository systems, appraisal and selection of digital data, and resource requirements for a data curation program." (Data Working Group, 2008)
G2.4 Responsibility is assigned	G2.4.1 Identify roles and responsibilities of personnel	"Teams, who may be widely distributed, have to agree upon what data will be collected, by what methods, and who has the rights and responsibilities to analyze, publish, and release those data." (Borgman, 2010) "Managing the life cycle of scientific data presents many challenges. These include deciding responsibilities…" (Lynch, 2008)
G2.5 People are trained	G2.5.1 Train researchers and data management personnel G2.5.2 Provide online guidance and workshops for data management	"…organizing DaWG forums open to all interested CUL staff (or Cornellians, for that matter), a journal club, or organization of workshops or training sessions for staff." (Data Working Group, 2008)
G2.6 Work products are controlled	G2.6.1 Changes to data are controlled G2.6.2 Data provenance metadata is captured, including documentation of changes	"Annotation of the data to record its provenance and content takes place mostly by including the information within the data…" (Martinez-Uribe, 2008)
G2.7 Stakeholders are identified	G2.7.1 Develop collaboration and partnership with research communities and learned societies	"Seek out and cultivate partnerships with other organizations" (Data Working Group, 2008)
G2.8 The processes is monitored and controlled	G2.8.1 Assess SDM effectiveness and impact	"Because the documentation and organization of scientific data sets can be time-consuming and expensive, it is important to evaluate the effectiveness of existing standards in meeting their objectives." (Zimmerman, 2003)

G2.9 Adherence to process standards is assessed and noncompliance addressed	G2.9.1 Enforce policy	"The retention of selected and appraised data raises other issues. These include maintenance of data quality, enforcement of data security, and, migration of data sets to current, available and maintainable, hardware and software systems." (Anderson, 2004) "Research funders should create, implement, and enforce data management, sharing, and preservation policies." (Data Working Group, 2008)

Table 2. Generic process areas for supporting scientific data management.

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, SDM is needs-based, *ad hoc* in nature and tends to be done intuitively. Rather than documented processes, the effectiveness of SDM relies on competent people and heroics. 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 the SDM 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).

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 SDM. 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 SDM capability resides at the project level rather than drawing from organizational or community processes definitions. For example, in a recent survey of science, technology, engineering and mathematics (STEM) faculty, Qin and D'Ignazio (in press) 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 SDM. 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.

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. 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 SDM 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, d2c2.lib.purdue.edu) brings researchers together to develop optimal ways to manage data, which could lead to formally maintained descriptions of SDM practices. Level 3 organizations can also draw on research-community-based efforts to define processes. Examples include the Hubbard Brook Ecosystem Studies (www.hubbardbrook.org), LTER (www.lternet.edu) and Global Biodiversity Information Facility (www.gbif.org). Government requirements and standards in regard to scientific data are often targeted to higher level of data management, e.g., community level or discipline level.

Level 4: Quantitatively Managed and Level 5: Optimizing

Level 4 in the original CMM means the processes have quantitative quality goals for the products and processes. The processes are instrumented and data is systematically collected and analyzed to evaluate the processes. 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. As noted earlier, we found no examples of these processes in our analysis, which we suggest reflects the current state of maturity of SDM.

DISCUSSION

The set of key practices and process areas described above are clearly preliminary. Future research will elaborate and extend these descriptions. But even in their current state, they lead to some interesting comparisons and show the potential benefits of applying this model. First, comparing the specific practices documented above to the data lifecycle, we note that our current set is focused on data creation. Perhaps as a result of the sources we selected, we currently do not describe any practices related to discovery or reuse of data. Future work on this model should address such practices.

Second, comparison of the generic practices we found to the SEI CMM model reveals a few interesting differences. First, we found generic practices that correspond only to Goal 2, Institutionalize a managed process, and not the higher-level goals. What should be noted here is that each level in CMM is built on top of the previous level, i.e., level 3 is built on top of level 2. It is impossible for a project's SDM activities to be ranked as level 3 if the level 2 key process areas are not performed in addition to the level 3 processes. This absence may be a function of our data collection, but we believe it also reflects the low level of development of SDM practice—at present, the processes are usually not well defined beyond the project level and rarely quantitatively managed or optimized, at least not in the sources we analyzed.

Second, we found a number of practices related to development or management of SDM technologies. We grouped these above with practices for providing resources, since tools are a kind of resource, but it seems that the tools and technologies of SDM are still being developed, making these practices more salient in this setting. We also found discussion of the need to develop a business model for long-term archiving of data. Many, if not all, of these projects discussed were initiated with support from the NSF or other funding agencies. Identifying future support is critical for these efforts to last after the funding period ends. In contrast, the SEI CMM was developed for organizations that control their own resources and can make their own funding decisions.

Third, one of the SEI CMM generic practices was not found in our review: reviewing process status with higher management. It may be that this practice is simply not that relevant in setting of SDM, given the generally high levels of autonomy enjoyed by most scientific researchers.

CMM APPLICATIONS SCENARIOS

The model introduced above can be used in different ways. First, a project can be assessed for its current level of maturity. Although SDM is still in its early development stage, it is not too early to study how to evaluate and assess SDM activities and practices. By mapping the key process areas with maturity levels, we established a framework of criteria that can be applied to analyze and assess SDM activities. We use the DataStaR (Steinhart, 2010) project as an example. Table 3 shows some of the DataStaR process activities mapped to the key process areas. This analysis indicates how the maturity of the project's process can be assessed by comparison to the framework.

Key process areas of CMM	DataStaR process activities
G2.2.1 Assess faculty data practices	Met with research group to understand their SDM needs
2.1 Develop metadata specifications and schemas G2.1 Develop policies for SDM	Developed policies, technology architecture and metadata application profile
G2.3.2 Apprise and evaluate enabling technologies G2.3.4 Manage enabling technologies for access and conformance to standards	Evaluated and customized technologies related to SDM; ensured conformance to standards
G2.5.1 Train researchers and data management personnel 4.5 Build digital preservation network	Provided guidelines for data authors; linked data sets to external repositories
2.1 Develop metadata specifications and schemas 2.6 Ensure interoperability with metadata standards	Specified metadata element set; ensured interoperability and metadata quality
G2.9.1 Enforce policy	Provided a central location for policy and guideline documents
G2.8.1 Assess SDM effectiveness and impact	Reflected on the project outcomes and challenges in published paper

Table 3. Key process areas examples (Steinhart, 2010).

As a related product of such assessment, the model can help institutions identify weaknesses in SDM processes as targets for improvement. The generic processes listed in Table 2—offer some guidance: is the organization committed to the process and capable of performing the activities? Were the activities performed effectively and was the project implemented as planned and on schedule? For example, the generic questions might be asked:

1. Is the project committed to documenting the decisions, designs, rules and best practices related to policy, technical, system and user areas?
2. Are the project personnel capable of performing the activities?
3. Are sufficient funds, resources, equipment and tools available?
4. What activities were actually performed to document decisions, designs, rules and best practices?
5. What processes are in place to measure the effectiveness of the process?
6. Was the process managed properly?
7. Are efforts in place to improve the process?

CONCLUSION

The model presented in this paper is still in a preliminary state, but it is already possible to see some possible implications. First, the catalog of processes areas should help projects and organizations ensure that they are covering all aspects of data management. The description of goals, objectives and practices will provide a guide for implementing and managing data management practices. Second, the model will provide a way to assess project and organizational data management plans. For example, the data management plan in an NSF proposal might be assessed for its coverage of the process areas and the level of maturity described. Third, the SEI CMM model includes practices and process areas that support a higher level of organizational capability, namely quantitatively managing and optimizing SDM processes. Finally, we hope that as has happened in software development, careful description of the different levels of maturity may serve as an impetus for organizations to improve their level of maturity, thus enabling better SDM.

REFERENCES

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal*, 3:191-202. http://www.jstage.jst.go.jp/article/dsj/3/0/3_191/_article

Barkley, B. T. (2006). *Integrated Project Management*. New York, NY, USA: McGraw-Hill.

Borgman, C.L., Wallis, J.C., & Enyedy, N. (2006). *Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries*. Papers, Center for Embedded Network Sensing, UC Los Angeles. http://escholarship.org/uc/item/6fs4559s

Borgman, C.L. (2010). Research data: Who will share what, with whom, when, and why? *China-North America Library Conference*, Beijing. http://works.bepress.com/borgman/238

Carlson, S. (2006). Lost in a sea of science data. *The Chronicle of Higher Education*, 52: A35.

CMMI Product Team. (2006). *CMMI for Development Version 1.2*. CMU/SEI-2006-TR-008. Pittsburgh, PA, USA: Carnegie Mellon Software Engineering Institute.

D'Ignazio, J. A. & J. Qin. (2008). Faculty data management practices: A campus-wide census of STEM departments. In: *Proceedings of the American Society for Information Science and Technology*, October 24-29, 2008, Columbus, Ohio. (Poster)

Data Working Group. Cornell University Library. (2008). *Digital research data curation: Overview of issues, current activities, and opportunities for the Cornell University Library*. http://ecommons.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf

DCC. (2011). *What is digital curation?* http://www.dcc.ac.uk/digital-curation/what-digital-curation

Field, D., Sansone, S.A., Collis, A., Booth, T., Dukes, P., Gregurick, S.K., Kennedy, K., Kolar, P., Kolker, E., Maxon, M., Millard, S., Mugabushaka, A.M., Perrin, N., Remacle, J.E., Remington, K., Rocca-Serra, P., Taylor, C.F., Thorley, M., Tiwari B., & Wilbanks. J. (2009). 'Omics data sharing. *Science* 326:234-236.

Godfrey, S. (2008). *What is CMMI?* NASA presentation. tttp://software.gsfc.nasa.gov/docs/What%20is%20CMMI.ppt

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. Edmond, WA: Microsoft Research.

Grayzeck, E. & Acton, C. (2002). *Deep Impact project Data Management Plan*. Document #: D-21386. Pasadena, CA: Jet Propulsion Laboratory, California Institute of Technology. http://pdssbn.astro.umd.edu/missions/deepimpact/deep_impact_pdmp.pdf

JCOMM/IODE. (2010). Expert Team on Data Management Practices (ETDMP), Second Session, 6-7 April 2010, *Reports of Meetings of Experts and Equivalent Bodies, UNESCO 2010 (English)*, UNESCO, 42 pp. http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=5512

Karasti, H. & Baker, K. (2008). Digital data practices and the long term ecological research program growing global. *The International Journal of Digital Curation*, 2(3): 42-58. http://www.ijdc.net/index.php/ijdc/article/viewFile/86/57

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

Lynch, C. (2008). How do your data grow? *Nature*, 455 (4 September): 28-29.

Martinez-Uribe, L. (2008). *Findings of the scoping study interviews and the research data management workshop*. http://www.ict.ox.ac.uk/odit/projects/digitalrepository/docs/ScopingStudyInterviews-Workshop%20Findings.pdf

Morris, S.P. & Tuttle, J. (2008). *Curation and preservation of complex data: The North Carolina geospatial data archiving project*. http://www.digitalpreservation.gov/partners/ncgdap/high/curation_complex_data_report.pdf

Murray-Rust, P. (2008). Chemistry for everyone. *Nature*, 451, 648-651.

Patton, M. Q. (2002). *Qualitative Evaluation and Research Methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. (1993). *Capability maturity model, Version 1.1*. IEEE Software, 10(4): 18-27.

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.

Schwadron, N. (2007). *IBEX Project Data Management Plan*. SwRI Project 11343. San Antonio, TX: Southwest Research Institute. http://nssdc.gsfc.nasa.gov/archive/pdmp/IBEX_PDMP_200707.pdf

Steinhart, G., Saylor, J., et al. (2008). *Digital Research Data Curation: Overview of Issues, Current Activities and Opportunities for the Cornell University Library. Report of the Cornell University Library Data Working Group*. http://hdl.handle.net/1813/10903

Steinhart, G. (2010). DataStaR: A data staging repository to support the sharing and publication of research data. *International Association of Scientific and Technological University Libraries, 31st Annual Conference*. West Lafayette, Indiana: Purdue Libraries. http://docs.lib.purdue.edu/iatul2010/conf/day2/8.

Walters, T. O. (2009). Data curation program development in U.S. universities: The Georgia Institute of Technology example. *The International Journal of Digital Curation*, 3(4): 83-92. http://www.ijdc.net/index.php/ijdc/article/viewFile/136/153

Witt, M. (2009). Institutional Repositories and Research Data Curation in a Distributed Environment. *Library Trends*, 57(2). http://docs.lib.purdue.edu/lib research/104

Van den Eynden, V., Bishop, L., Horton, L., & Corti, L. (2010). *Data management practices in the social sciences*. http://www.data-archive.ac.uk/media/203597/datamanagement_socialsciences.pdf

Zimmerman, A. S. (2003). *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists*. Doctoral Thesis. University of Michigan. http://deepblue.lib.umich.edu/handle/2027.42/39373