A Capability Maturity Model for Scientific Data Management

Kevin Crowston & Jian Qin

School of Information Studies, Syracuse University

July 2010

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 CMM describes key process areas and practices necessary for effective SDM. The CMM further characterizes organizations by the level of maturity of these processes, meaning the organizational capability to reliably perform the processes. We suggest that this framework will be useful to organizations in evaluating and planning improvements to their SDM practices.

Introduction

E-Science, the application of information and communication technologies (ICT) to support scientific work, is 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). As a "transformed scientific method", e-science puts scientific data management at the center stage in the whole research cycle, which includes data capture, data curation, data analysis and data visualization {Gray, 2007}. In this paper, we propose a capability maturity model for scientific data management (SDM) practices, with the goal of supporting assessment and improvement of these practices in order to make SDM more reliable.

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 data management (DM) policies and practices, enabling data repositories for data curation, discovery and reuse. Data management in these disciplines often has significant funding support, which ensures the personnel and technology infrastructure necessary for running a DM operation.

A less often considered type of e-science is in "small science" research, that is, projects involving a single PI and a few students. These scientists also often depend upon e-science tools to generate and manage data. 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. 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.

The importance of SDM has been raised to a new level, as demonstrated by US National Science Foundation's mandate that future 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. While lack of awareness may be 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), another important factor is the lack of a theoretical 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. In other words, SDM at research project level needs an operational framework to which they can adopt and evolve the data management program. 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. In the follow section, we first describe the origins of the notional of a CMM, then make an initial proposal for the application of this idea to SDM. We conclude by describing possible uses for a CMM for SDM.

A Capability Maturity Model 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, i.e., their ability to develop quality software on time and within budget. 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 CMM includes a number of key concepts: maturity levels, key process areas, key practices and common features, as shown in Figure 1.

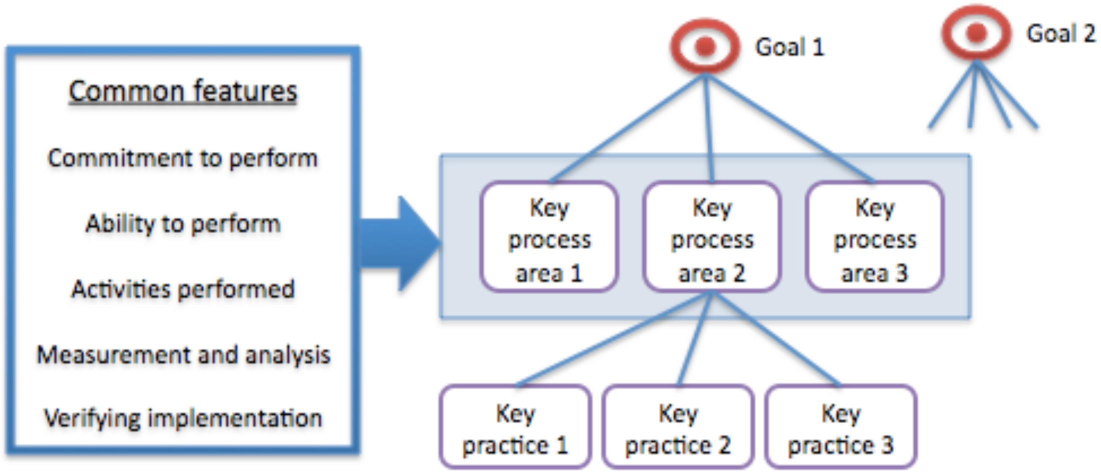


Figure 1. Operationalization of CCM through goals, key process areas, and key practices (from the SEI CMM).{CMMI Product Team, 2006}

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., requirements determination or configuration management). In the model, these practices are clustered into key 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). Reliable organizations have the capability to reliably execute these practices, that is, to perform them in a consistent and predictable fashion. In the model, the level of development of each process area is described in terms of "common features". In the following section, we develop an initial list of key practices for SDM before going on to discuss the maturity of SDM practices.

Key Process Areas for Scientific Data Management

A key contribution of the SEI CMM was to describe the key practices needed for software development, clustered in a set of process areas that identify related goals, objectives and practices. The CMM for SDM will similarly identify key areas of skills and expertise necessary for accomplishing the SDM goals, thus providing an analytical tool for improving the effectiveness of SDM. A full description of the CMM would include a set of practices and key process areas necessary for SDM performance. However, SDM represents an emerging interdisciplinary research field, and its processes and practices are still being explored and understood. In this section, we present some preliminary suggestions for practices and process areas.

Science data management resolves around the life cycle of science data, which includes data collection, processing, organization, curation, distribution and use. Possible key process areas for SDM include:

- *User requirements development*: systematic study of users—data contributors/producers, users, and managers—about their needs and requirements for data management functions.
- *Data management planning*: decision making on policy issues such as data retention, preservation, access, sharing and publishing; funding issues, such as economic models and cost administration; and technical design issues such as system architecture and software and hardware infrastructures.
- *Workflow management*: procedures and quality control mechanisms at different stages of data management, from raw data to final data products and metadata.
- *Documentation management*: technical documentation (documents and diagrams related data and metadata formats and standards, rules and codes, etc.), system documentation (system architecture, database and XML schemas, etc.), user documentation (help, user guide and/or best practices) and policy documentation (see *Data management policy development* below).
- *Semantic metadata development*: metadata standard adoption and/or application profile development, knowledge organization systems development, metadata records generation and harvesting, interoperability, and quality control.
- *Technology solution management*: design and implementation of technical, operation, and system architecture, survey and selection of enabling technologies for all the stages of science data life cycle and appraisal and migration of technologies.
- *Data management policy development*: development of legal (copyright, intellectual property, privacy, confidentiality), social (ethics and access), technical (retention, preservation, metadata) and use (discovery, publishing, citation, distribution) policies.

In addition to these SDM specific processes, projects also need defined processes to address more generic management issues, including:

- *Integrated project management*: project deliverables that reflect all the project requirements and components along with horizontal and lateral coordination, the final outcome and due dates that meet the planned costs, schedule, and quality, as well as user satisfaction, collective contribution and new working relationships with professional and technical project staff, and effective blending of cost, schedule, and quality considerations in the project cycle (Barkley, 2006).
- *Processes and quality assurance*: processes to verify the execution of process.
- *Organizational training*: training on policy, technology, metadata and best practices. Staff training has different goals and objectives than user training as well as different focus on content, though the areas may be the same.
- *Best practices and guidelines development*: documentation on decisions made on each stage of the data life cycle; best practices on key areas such as data retention, preservation and metadata creation.
- *Evaluation and analysis*: assessment of effectiveness, efficiency and impact.

For each of these areas, key practices need to be defined to flesh out the model.

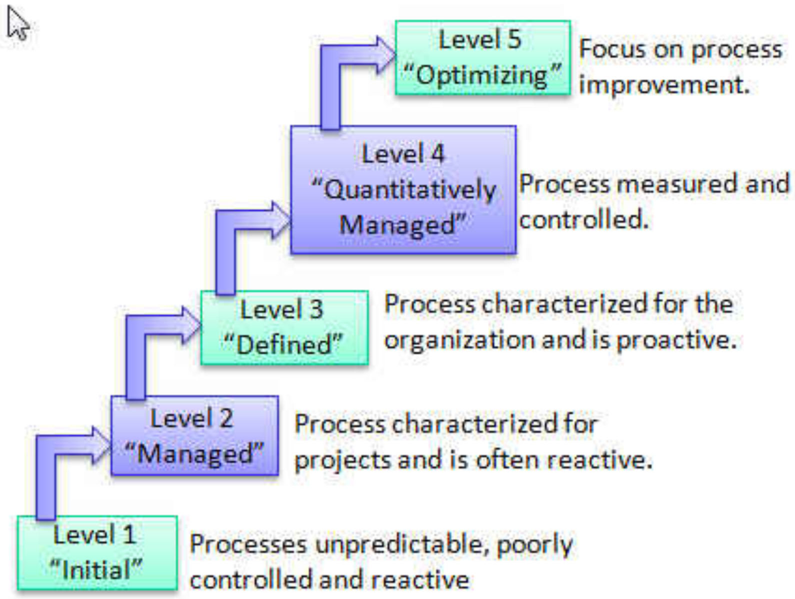
Finally, the SEI CMM included five generic features that are describe the readiness of the organization to implement effective practices that can be applied to SDM, namely:

1. commitment to perform: the project has policies regarding the process and management commitment to perform the process,
2. ability to perform: the organization has the capability to perform the processes, e.g., funding, appropriate tools or trained individuals,
3. activities performed: the process is actually performed in practice,
4. measurement and analysis: the execution of the process is measured and performance analyzed, and
5. verifying implementation: quality assurance processes are in place.

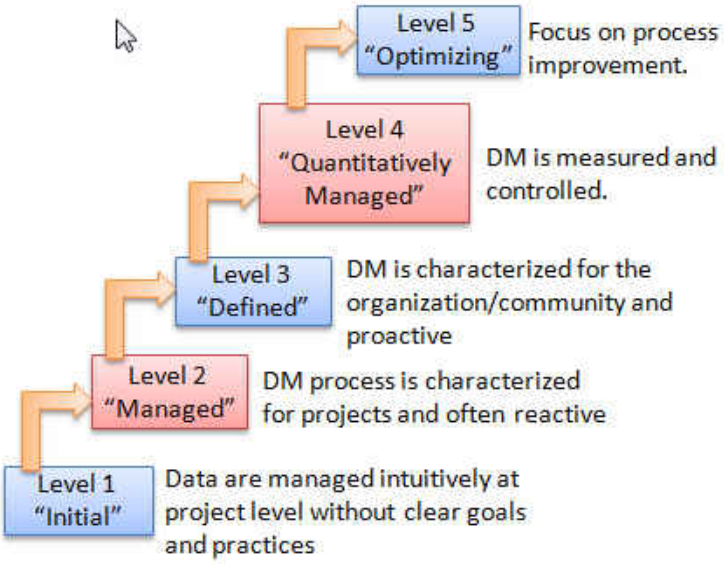
Scientific Data Management Maturity Levels

In addition to the process areas and practices and their maturity, the CMM described five levels of process maturity for software development organizations as a whole (Figure 2 (a)), representing the "degree of process improvement across a predefined set of process areas". 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 predict the results of the next development project. As the organization increases in maturity, processes become more refined, institutionalized and standardized. 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).

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). In the following section, we apply the maturity level concept to describe the difference among scientific projects in the maturity of their application of SDM (Figure 2 (b)).


Level 5
"Optimizing"
Focus on process improvement.
Level 4
"Quantitatively Managed"
Process measured and controlled.
Level 3
"Defined"
Process characterized for the organization and is proactive.
Level 2
"Managed"
Process characterized for projects and is often reactive.
Level 1
"Initial"
Processes unpredictable, poorly controlled and reactive

(a) Original CMM (Godfrey, 2008)


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

(b) CMM for SDM

Figure 2. The Capability Maturity Model characterization of process maturity level

Level 1: Initial

The initial level 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); success relies on competent people and heroics rather than documented processes. At this level, SDM is needs-based, *ad hoc* in nature and tends to be done intuitively. SDM processes are local to the project team, hindering sharing or aggregation of data. The success of DM depends entirely on the efforts and abilities of individuals involved. The knowledge of the field and skills of these task performers (often graduate students working with little input) limits the effectiveness of data management; changes in personnel will have a great impact on the outcomes.

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, the DM capability resides at the project level rather than drawing from organizational or community processes definitions. 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, in press). 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, sharing data across multiple projects may 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. As a result, an organization can establish new projects with confidence in stable and repeatable execution of processes. SDM at this level of maturity similarly draws on process definitions from beyond individual projects. 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). However, these standards are 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/efforts, in which organizational members/task forces within the institution discuss policies and plans for data management, set best practices for technology and adopt and implement data standards. Outputs at this level might include institutional repositories, e.g., Cornell's DataStar project (http://datastar.mannlib.cornell.edu/). Level 3 can also draw on research-community-based efforts to define processes. Examples include the Hubbard Brook Ecosystem Studies (http://www.hubbardbrook.org/), LTER (http://www.lternet.edu/), and Global Biodiversity Information Facility (http://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. In the remainder of this paper, we will focus our attentions on the lower 3 levels of the CMM for which SDM-specific processes can be identified.

CMM Applications Scenarios

Once fleshed out, the model introduced above could be used in different ways. First, a project can be assessed for its current level of maturity. By mapping the key process areas with maturity levels, we established a framework of criteria that can be applied to analyze and assess DM activities. Table 1 maps each key DM process area mentioned earlier in this paper into a maturity level (including only the first 3 levels).

We suggest that currently for many science projects, DM is only at level 1, meaning that data is managed through efforts of individuals, but DM is not institutionalized. When those individuals move on, or focus elsewhere, there is a danger that the DM will not be sustained. Level 2 describes a project with policies and procedures for data management that ensure that the project can reliably manage its data. However, these policies and procedures are idiosyncratic to the project. Level 3 means that the data management processes are documented across the organization or field and so are repeatable across projects.

Table 1. Key process areas coresponding to the maturity levels in SDM

Maturity level	Characterization	Key Process areas
3 Defined	Process standardization	User* requirements development Data management policy development Integrated project management Semantic metadata development Technology solution management Process and quality assurance Organizational training Best practices and guidelines development Evaluation and analysis
2 Managed	Basic data management	User* needs assessment Data management planning Enabling technology management Workflow management Metadata management Documentation management Performance assessment
1 Initial	Ad hoc	Competent people and heroics

Note: * implies data contributors/producers, users, and managers.

Table 2. Key process areas examples (Steinhart, 2010).

Key process areas of CMM	DataStaR process activities
User needs assessment	Met with research group to understand their data management (DM) needs
Data management planning	Developed policies, technology architecture, and metadata application profile
Technology management	Evaluated and customized technologies related to DM; ensured conformance to standards
Workflow management	Provided guidelines for data authors; linked data sets to external repositories
Metadata management	Specified metadata element set; ensured interoperability and metadata quality
Documentation management	Provided a central location for policy and guideline documents
Performance assessment	Reflected on the project outcomes and challenges in published paper

To illustrate the application of the CMM for SDM, we use the Cornell DataStaR project as an example. As described in Steinhart (2010), “DataStaR is a platform, as well as a set of services provided by librarians, intended to support research data sharing and publication” (Steinhart, 2010, p. 4). Some major process activities performed during the DataStaR project are listed in Tables 2, which have been grouped into key process areas for maturity level 2. The fact that DataStaR has coordinated with other digital repository projects and developed some semantic organization for datasets indicates that the project is evolving toward the level 3 of maturity. It should be noted that the higher levels of the CMM adds to processes for managing data, additional processes for implementing, assessing and improving those processes.

As a related product of such assessment, the model can help institutions identify weaknesses in SDM processes for improvements. The common features displayed in Figure 1—commitment to perform, ability to perform, activities performed, meansurement and analysis, and verifying implementation—offer some guidance: is the organization committed to the process, capable of performing the acitivies? How effectively were the acitivies performed and was the project implemented as planned and on schedule? For example, in the process area of documentation management, the common feature-based questions might be asked:

- Is the project committed to documenting the decisions, designs, rules and best practices related to policy, technical, system and user areas?
- Are the project personnel capable of performing the documentation activities?
- Are sufficient funds, resources and equipment available?
- What activities were actually performed to document decisions, designs, rules and best practices
- What processes are in place to measure the effectiveness of documentation?
- Was the documentation managed properly?
- 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.

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

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

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)

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.

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

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

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. (in press). The central role of metadata in a science data literacy course. *Journal of Library Metadata*.

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.