A Capability Maturity Model for Scientific Data Management

Kevin Crowston
School of Information Studies
Syracuse University, Syracuse, NY 13244
crowston@syr.edu

Jian Qin
School of Information Studies
Syracuse University, Syracuse, NY 13244
jqin@syr.edu

ABSTRACT

In this poster, we propose a capability-maturity model (CMM) for scientific data management that includes a set of process areas required for data management, grouped at three levels of organizational capability maturity. The goal is to provide a framework for comparing and improving project and organizational data management practices.

Keywords

Science data management, capability maturity model.

INTRODUCTION

Current scientific data management (SDM) practices vary greatly depending on the scale, discipline, funding, and type of projects. Although the importance of SDM has been raised to a new level, as demonstrated by NSF's mandate that proposals include a data management plan, low awareness or total lack of data management procedures is still common among research projects. While these problems 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), a more profound factor is the lack of a theoretical model upon which practices, policies and performance and impact assessment can be based.

To help address this gap, we propose a capability-maturity model (CMM) for scientific data management. The original CMM presented a framework for managing software development processes, including a set of process areas required for software development grouped at five levels of organizational capability maturity. As an organization increases in capability maturity, its processes become more refined, institutionalized and standardized, establishing a basis for process management, appraisal and improvement. The US Environmental Protection Agency (EPA) reportedly also uses CMM as a framework for assessing their SDM strategies (Petterson, 2008).

As SDM represents an emerging interdisciplinary research field, its processes and practices are still being explored and understood. A CMM in this context provides an analytical tool for classifying SDM processes and organizations, which is critical not only for improving the effectiveness of SDM and evaluating the impact and return on investment in SDM, but also for identifying key areas of skills and expertise necessary for accomplishing the SDM goals.

SCIENTIFIC DATA MANAGEMENT MATURITY LEVELS

Applying the maturity level concept in CCM, we define the SDM processes at the first three levels defined by the CMM (levels 4 & 5 address process improvement rather than data management processes per se and so are omitted in this initial discussion).

Level 1: Initial

Capability Maturity Level 1 describes an organization with no stable processes. As described in Paulk et al., "In an immature software organization, software processes are generally improvised by practitioners and their management during the course of a project" (Paulk et al, 1993, p. 19). SDM at this level is needs-based, *ad hoc* in nature and handled within a project team. Data may be managed, but the outcome depends solely on the efforts of the individuals involved. The level of knowledge of the field and skills of these task performers (often graduate students with little guidance from the team leader or other members) limits the quality of the outcome (Qin & D'Ignazio, 2008). Furthermore, there is little guarantee that the processes will be repeated reliably, since the capability resides in the individuals, who may be unavailable to contribute when needed.

Level 2: Managed

Capability Maturity Level 2 characterizes organizations with repeatable processes that are managed through established policies and procedures. As a result, the organization can reliably predict and execute data management projects with some confidence in the outcome. But the capability resides at the project level, meaning that each project establishes these procedures and policies from scratch. In the SDM context, a managed process signals that the research group has articulated plans, policies and procedures for SDM. For example, local data file naming conventions and directory organization structures may be documented and procedures set up to ensure that data is stored correctly, with clear descriptions of responsibility and measures of performance.

Level 3: Defined

In the original CMM, Capability Maturity Level 3 means that processes are documented across the organization and customized for particular projects. As a result, execution of the processes is stable and repeatable across projects. SDM at this level reflects institutional initiatives and efforts. Organizational members or task forces within the institution discuss policies and plans for data management and set best practices for technology and for standards adoption and implementation. For example, adopting a metadata standard for describing datasets involves modification of standards in order to meet institutional needs, which means the representation and organization of data not only at the physical file level but also at the dataset or product level.

KEY PROCESS AREAS

A full description of the CMM would include a set of key process areas necessary for SDM performance at each of the levels described above. The key process areas identify goals, objectives, and practices associated with the appropriate maturity level. In this poster, we present some preliminary suggestions for these areas.

SDM resolves around the life cycle of science data, which includes data collection, processing, organization, preservation, distribution and use. As noted above, level 1 relies on competent people and heroics rather than documented processes. Process areas and activities at this level of maturity are rarely seen in data curation or data management research literature but mostly exist in anecdotes.

At level 2, key process areas include process for user needs assessment, data management planning, technology management, workflow management, metadata management, documentation management and performance assessment. Although the key process areas at level 2 still tend to be reactive, the characterization has shifted from *ad hoc* reaction to managed processes. A critical difference between levels 1 and 2 is that SDM no longer remains within a research team, but rather, data professionals will be involved. Table 1 shows an example of key process areas at level 2.

Table 1. Key process areas examples (Steinhart, 2010)

Key process areas of CMM	DataStaR process activities
User needs assessment	Meet with research group to understand their data management (DM) needs
Data management planning	Develop policies, technology architecture, metadata application profile
Technology management	Evaluate and customize technologies related to DM; conform to standards
Workflow management	Provide guidelines for data authors; link data sets to external repositories
Metadata management	Specify metadata element set; ensure interoperability and metadata quality
Documentation management	Provide a central location for policy and guideline documents
Performance assessment	Reflect on the project outcomes and challenges in published paper

Level 3 would add processes for integrated project management, semantic metadata development, process and quality assurance, organizational training, best practices and guidelines development and evaluation and analysis. In the DataStaR case, the project is evolving toward the level 3, as described in Steinhart (2010). It should be noted that the higher level adds to processes for managing additional data processes for implementing, assessing and improving those processes.

DISCUSSION

The model presented in this poster 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

Key Perspectives. (2010), “Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study”, Retrieved June 17, 2010, from http://www.dcc .ac.uk/scarp

Paulk, M. C., Curtis, B., Chrissis, M. B., and Weber, C. V. (1993). *Capability maturity model, version 1.1*. TR CMU/SEI-93-TR-024, ESC-TR-93-177, Carnegie-Mellon University, Software Engineering Institute.

Petterson, L. (2008). Implementing ORF’s scientific data management (SDM) strategy. Presentation at *ORMA/IMTS EPA Conference on Managing Environmental Quality Systems April 23, 2008*. Retrieved August 20, 2010 from http://www.epa.gov/quality/qs-2008 /sdm.pdf

Qin, J. & D’Ignazio, J. (in press). The central role of metadata in a science data literacy course. *Journal of Library Metadata*.

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*. Retrieved July 15, 2010, from http://docs.lib.purdue.edu/iatul2010 /conf/day2/8