Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics

Jian Qin, Ph.D., Professor
School of Information Studies
Syracuse University

Kevin Crowston, Ph.D., Distinguished Professor
School of Information Studies
Syracuse University

Arden Kirkland, MSLIS, Adjunct
School of Information Studies
Syracuse University

Corresponding author contact:
Jian Qin, jqin@syr.edu, Phone: (315)443-5642
School of Information Studies
Syracuse University
311 Hinds Hall
Syracuse, NY 13244

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.

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). Improved research data management (RDM) was recognized almost a decade ago (Gray, 2009) as a critical area with action needed across the data lifecycle. More recently, the importance of RDM has been raised to a new level with policy mandates for data management and sharing and the fast growth of digital data, both in volume and complexity. However, low awareness of—or lack of—data management is still common among researchers (Akers & Doty, 2013).

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; Akers & Doty, 2013). 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, intellectually intensive, and costly. Personnel performing these tasks must be trained both in technology and in subject fields, and be able to effectively communicate with different stakeholders. In this sense, 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. Research projects need more concrete guidance to analyze and assess the processes of RDM.

To support assessment and improvement of RDM practices that increase its reliability, we developed a capability maturity model for RDM (CMM for RDM). The documentation of this model is presented at a Wiki site (http://rdm.ischool.syr.edu/xwiki/bin/view/Main/) and organized into six sections:

Section 0: Introduction
Section 1: Data management in general
Section 2: Data acquisition, processing and quality assurance
Section 3: Data description and representation
Section 4: Data dissemination
Section 5: Repository services and preservation

This paper summarizes the development of this model and, by using two scenarios of research data management, demonstrates the roles and methods for eScience librarians in planning and implementing RDM services.

Overview of the 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. The CMM framework 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, Weber, Chrissis, & Curtis, 1993). 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). In this context, reliability refers to an organization's ability to develop quality software on time and within budget by executing processes in a consistent and predictable fashion. By analogy, our CMM is intended to improve the consistency and predictability of RDM.

The CMM has evolved over time, but the basic structure remains the same. It includes four main elements: key practices, key process areas, maturity levels and generic processes. We introduce each in turn.

Key practices and process areas

In the original CMM, specific software development practices are clustered into 22 *specific process areas*, that are, "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). 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 described in the model are the expected (though not required) way to achieve those goals.

Maturity levels

Perhaps the most well-known aspect of the CMM is its 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." Maturity levels serve as indicators of process capability, while key process areas are where goals will be achieved (or failed). The maturity levels are defined by the organization's ability to achieve the following levels of performance for each process area:

1. Achieve specific goals: the processes are performed;
2. Institutionalize a managed process: 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 are 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: a description of the process is maintained and improvement information is collected;
4. Institutionalize a quantitatively managed process: quantitative objectives are established and subprocess performance is stabilized; and
5. Institutionalize an optimizing process: continuous process improvement is ensured and root causes of defects are identified and corrected.

The intuitive level describes an organization without defined processes. In the original CMM, an organization at this level succeeds in developing software (i.e., the specific software-related goals are achieved), but in an *ad hoc* and unrepeatable way, making it difficult or impossible to

plan or predict the results of a future development project with any confidence. This lack of predictability about future efforts is what the CMM calls process immaturity. As the organization increases in maturity, moving up the levels, processes become more defined, institutionalized and standardized. As a result, when a new project starts, it has clear processes to draw on, allowing the organization to be more assured of project results. Note that it is possible for different process areas to be at different levels of maturity (e.g., to have a well-defined process for tracking changes but no clear process for eliciting customer needs). By identifying areas in need of improvement, the CMM thus describes an evolutionary improvement path from *ad hoc*, immature processes to disciplined, and mature processes with improved software quality and organizational effectiveness (CMMI Product Team, 2006, p. p. 535).

Generic processes

In addition to the specific process areas, those related 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. In the original Capability Maturity Model, maturity levels contain key process areas that are organized by so called common features. 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, Weber, Chrissis, & Curtis, 1993, p. p. 37). These common features are shown in Table 1.

[Insert Table 1 here]

In other words, for each process area (e.g., *eliciting customer needs*), in addition to having practices for achieving the goal (e.g., *eliciting requirements*), a high-performing organization is expected to also have processes that establish its commitment to perform those practices (e.g., *establishing policies*), that establish its ability to perform the practices (e.g., *providing funding or training*), that measure performance (e.g., *counting how many requirements are elicited* or *assessing the quality of requirements*), and that verify implementation (e.g., *verifying that the practices were followed* or *verifying that requirements were correctly elicited*). The notion is that simply performing the activities in one case is not sufficient to ensure that it will be possible to perform them again reliably on a future project; and that without data about the performance of the activities, it is not possible to plan improvements.

Towards a CMM for RDM

While the SEI CMM was to help organizations lay out a path for improving software development, our goal is to lay out a 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), with an increasingly high demand for RDM services across disciplines (Barsky, 2017-07). The CMM has been around for two decades and the concept has been applied in various contexts for improving processes and performance. The RDM community has had application examples such as the Community Capability Model Framework (CCMF) (Lyon, Ball, Duke, & Day, 2012) and each has a slightly different focus and interpretation. In the following subsections we will discuss how we apply the various elements of the CMM model to the practices of RDM.

Key practices and process areas for RDM

The CMM for RDM (the Model thereafter) we developed includes a description of the key practices for RDM. This Model is available on the wiki site at https://goo.gl/nUMd7X. The Model consists of six sections, with section 0 introducing background and rationale. Sections 1-5 describe five key RDM process areas: 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 follows a structure of goal, key concepts, rationale/importance, examples, and recommended practice. As with the software-development goals in the SEI CMM, the sub-areas are the goals that must be achieved by a fully reliable research data management organization. For each goal, the SEI CMM defines a set of practices that are viewed as best practices, but not necessarily the only or required way to accomplish the goals. For example, in CMM for RDM, a goal for *data management in general* is to *have trained personnel to carry out the data management*; example practices to achieve this goal include *providing workshops* or *online training*. A goal for *data acquisition* is to *develop data quality control procedures*; an example practice is to *determine reasonable ranges for data items and to plan to check data against these limits as they are collected*.

RDM process maturity

Capability maturity levels in the context of RDM are illustrated in Figure 1. There is one important change in moving the concept of maturity levels from the SEI CMM to the context of RDM. The SEI CMM focused on defining processes within a single software development organization. In contrast, RDM might be institutionalized at the level of a research community or discipline, not an organization, as discussed below.

[Insert Figure 1 here]

Level 1: Intuitive

This level in the SEI CMM was labeled as "intuitive," which Paulk et al. describe as: "In an immature organization, …processes are generally improvised by practitioners and their managers during a project" (Paulk, Weber, Chrissis, & Curtis, 1993, p. p. 19). In an institution with such immature RDM activities, 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 personnel and bold 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.

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, either the research group or

project managers have discussed and developed a plan for RDM (e.g., in the data management plan included in a research proposal, or in the planning for an initiative RDM project in a library). For example, local data file naming conventions and directory organization structures may be documented.

However, at this level of maturity, 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. This level of maturity characterizes RDM in many settings. For example, in a survey of science, technology, engineering and mathematics faculty, Qin and D'Ignazio (2010) found that respondents predominantly used local sources for decisions about what metadata to create when representing their datasets, either through their own planning, or in discussion with their lab groups. Guidelines from research centers or discipline-based sources were of far less impact. Government requirements or standards also seemed to provide comparatively little help (Qin & D'Ignazio, 2010). A similar phenomenon was also reported in Whitmire, Boock, & Sutton (2015).

As a result, at this level, developing a new project requires rediscovering practices or redeveloping processes, with possible risks to the effectiveness of RDM. Individual researchers will also likely have to learn new practices as they move from project to project. For RDM service librarians, disciplinary idiosyncrasies may also require them to learn new practices as RDM initiatives are deployed more broadly within the institution. 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 processes are documented across the organization and then tailored and applied for particular projects. Defined processes are those with inputs, standards, work practices, validation procedures, and compliance criteria. As an example, projects at level 3 likely employ a widely-recognized metadata standard and apply best practice guidelines for its use.

A key point about level-3 processes is that they are institutionalized at a level beyond a single project. For example, many university libraries have established an organization unit and implemented RDM service programs, e.g., the Research Data Management Service Group at Cornell University Library and the Research Data Services + Sciences at the University of Virginia Library. The fact that these RDM services and programs operate under defined mission, procedures, best practices, and policies symbolizes the institutionalization of RDM, and hence can be considered a defined process -- a level 3 capability maturity. As a result, PIs can develop new projects with confidence in stable and repeatable execution of RDM processes, rather than the new project having to invent practices and processes from scratch.

RDM processes might also be created through cooperation across institutions to develop research community-wide best practices for technology and adopt and implement 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. Institutional infrastructure, such as a data repository, could be the basis for organizational-standard practices, e.g., for data storage or backup. When RDM is performed at a community or discipline level, it is more likely to serve as part of the

infrastructural service for that community. Examples include the Interuniversity Consortium for Political and Social Research (ICPSR, http://www.icpsr.umich.edu/icpsrweb/) and Dryad (http://datadryad.org/), which are not only a data repository for the disciplinary community but also have well defined metadata schemas and tools, best practices, and policies necessary for building and managing the data collections. Well-defined RDM processes at an Institutional level, typically RDM services offered by academic libraries, should also define relations between community-level and institutional-level RDM.

Level 4: Quantitatively Managed

Level 4 in the original CMM means the processes have quantitative quality goals for the products and processes. At this level, processes are instrumented and performance data are systematically collected and analyzed to evaluate the processes. For the level 3 capability maturity to reach level 4, institutions and projects must "establish quantitative objectives for quality and process performance and use them as criteria in managing processes" (CMMI Product Team, 2006, p. p. 37).

In the context of RDM, 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 submit 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. 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. p.38).

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 specific techniques for process improvement. In other words, not only are the data collected, but there is also systematic attention to using the data to suggest process improvements. To continue the above example, an analysis of unnecessary repetitions in data entry may inform where in the RDM process the efficiency of data entry may be improved.

Generic practices

Finally, our organization of the process areas follows the structure of the SEI CMM common features listed in Table 1. However, we made one change from the original CMM model. In our analysis of RDM practices during the development of CMM for RDM (or the Model), we found limited evidence of quantitative measurement or validation of processes, which we suggest reflects the current state of maturity of RDM (i.e., few if any level 4 or 5 organizations). As a result, in the Model we have combined the areas of *Measurement and Analysis* and *Verifying Implementation* as one practice area, labelled *Process Assessment*. In other words, we organized data management practices within each practice area (i.e., within each section of the Model) into four groups: commitment to perform, ability to perform, tasks performed, and process

assessment. For example, a commitment to perform a goal for *data management in general* is to *identify stakeholders for the data management*. An ability to perform a task for *data acquisition* is to *develop data file formats that will be used*. A task performed for *data description* is to *create metadata for the data collected*. And a process assessment task in the *repository services and preservation* area is to *validate backups*. Again, for each of these goals, the Model describes possible practices that might be adopted to achieve the goals.

Assessment Rubrics

The intent of the Model is to help organizations assess their current level of performance of RDM and to identify opportunities for improvement. To help organizations assess RDM processes, we developed a rubric for each of the generic practice areas in each section. In keeping with the maturity levels defined in the Model, these rubrics provide a description of an organization at each level of maturity for that area. As an example, the rubric for *activities performed* for *data dissemination* is shown below (Table 2).

[Insert Table 2 here]

By applying these rubrics, a project or organization can identify areas where their RDM practices are strong or weak, and thus prioritize actions for process improvement. We also provide a full rubric (Kirkland, Qin, & Crowston, 2014) to help data managers see the "big picture" of the Model as a whole. In this table, the rows represent specific practices and the columns represent the levels of maturity. A brief statement in each cell provides a description of what that practice might look like at that maturity level. Using this format, data managers who are already familiar with the Model sections can easily move through the list of practices and circle or highlight the level of maturity that applies to that practice for the project or institution being evaluated.

It is important to note that in its current form the rubric functions as a qualitative rather than quantitative measure, as different practices hold different weights, a factor that is further differentiated depending on the project. Use of the rubric helps to demonstrate the ways in which this is an aspirational model, helping research data managers to visualize and implement the higher levels of maturity to which they aspire. Another version of the rubric [citation blinded for review] provides blank space for research data managers to write notes about the level of each practice, to help guide future improvements. This rubric can make it clear which areas are at the lowest levels of maturity and should be prioritized for remediation.

Scenarios of Application

Research data management may happen at individual researcher, project group, institutional, and community levels. The varying scales imply different objectives and tasks to accomplish in managing the data. Below are two case scenarios that demonstrate the application of the CMM for RDM: one at a project group level and the other at a community level.

Scenario 1: Assessment of project-level data management

We first show how the Model and rubric could be used to assess the maturity of RDM processes in a single project. The case scenario is a research project that investigates research collaboration network structures and dynamics via the metadata from a very large data

repository. The research project group consists of two faculty members, a Ph.D. student, and three master graduate students.

The metadata collected from the repository needed to be parsed and processed into the formats suitable for data mining. This process included several steps. First, the data were downloaded, extracted, parsed, and loaded into a relational database, all of which were done by writing computer programming code. Second, the data were checked for anomalies and errors and verified by triangulating with descriptive statistics and inspecting any cases that seemed to be unusual. Finally, analysis strategies and program codes were tested and modified until they were ready for running on the whole dataset. Different kinds of data were produced at each step: pre-processed data (i.e., raw data from the repository), processed data, output data, and compiled data, among which the processed and compiled data will be shared through either the project website or a data repository when they are ready.

Data management at the project-group level was needed to address at least two types of accountability: provenance of data for reproducible research and reliability of data to ensure the validity of research findings and conclusion. To achieve these purposes, the principal investigators (PIs) developed a specific, actionable plan for managing the data generated from the project.

As the project is in its mid-term, the timing was good to assess the data management using the Model. With the help of the rubric, the activities that took place for data documentation and management were identified as being mostly at level 2 (Table 3). In other words, while this project had documented practice for data management, these did not draw on organizational- or community-level definitions. This distinction determines the scope of data management to be concentrated on managing the data products produced from this project for quality control, reproducibility, and long-term access.

[Insert Table 3 here]

As the assessment result in Table 3 shows, the project group members were able to identify the strengths and weaknesses of the data management process. Although the Model itself did not offer specific strategies to improve the process and strengthen the weak areas, the fact that these weaknesses were made aware to the PIs is helpful for them to design solutions to improve the process.

This case scenario also offers some insights into eScience librarianship in two perspectives. The first is the presence of RDM infrastructure. It seemed that the research group in this case was not seeking or getting help with their RDM needs from the institution or the library. The PIs might have been satisfied with the state of their data management practices and the institution did not have established infrastructure or channels to communicate the need for improvement between the organizational units (or the library) and the research projects/groups. While technology is an important part of the RDM infrastructure, institutionalization of RDM has less to do with technology than with organizational culture, vision of administration, awareness, and human factors.

Another perspective for eScience librarianship is the service mode. Although many academic libraries have an established organizational unit to provide RDM services, proactive services are still a weak area. A common type of RDM services offered by academic libraries is consulting for faculty RDM needs. Without being aware of what and when RDM support is needed, such consulting service would be no different from the traditional reference service, a mode of waiting

for patrons to ask for help. How can eScience librarians be made aware of ongoing research projects and offer proactive services? This need for outreach perhaps suggests a weakness in RDM services that can be improved.

Scenario 2: Assessment of community-level data management

The second scenario examines data management in the context of a large project involving researchers from multiple institutions. Research data management at the community level bears a different mission from that at a project level. Logically, data are produced by research projects, so managing data at this level means much of the time is spent on managing active, constantly-changing data. On the other hand, project-level data management is the primary user of RDM infrastructures such as collaboration tools, data storage and sharing tools, and workflow management tools. At community level, the mission for managing data would be to provide infrastructural services for data curation, aggregation, discovery, sharing, and reuse. In this sense, the community-level data management acts as a service provider while project-level data management is the primary user of such services.

As an example of a community-level data management, we discuss the case of the Laser Interferometer Gravitational-Wave Observatory (LIGO, https://www.ligo.caltech.edu/). LIGO research is both data intensive and computationally intensive. Programming codes or algorithms are used throughout the whole process: from data generation, calibration, processing to analyses and result reports. Within the community, projects are relatively transparent: research artifacts and documentations on which data sources were used, what parameters specified, and what software was used are openly accessible on internal websites. LIGO has a detailed data management plan (Anderson & Williams, 2016) that governs how the vast amount of data are ingested, stored, represented by metadata, and preserved, as well as operations of the underpinning technological infrastructure. It also contains policies for public access and use of the LIGO data. The LIGO data management plan represents a good case of institutionalization of RDM by establishing policies and guidelines for managing the data produced from the gravitational wave research lifecycle.

While institutionalization of RDM is a critical step in community-level RDM, it is only the first step. Effective RDM within the community relies on infrastructural services and best practices of RDM to materialize this data management plan. Meanwhile, RDM at project level is not only an ongoing process but also the underpinning for community-level RDM to prove the value of institutionalization of RDM.

While LIGO is not the only community-level RDM case, it raises at least two interesting points for RDM services for academic libraries. First, there is a need to define the relation between community-level and institutional-level RDM. The mutuality between the two is that both are infrastructure service providers, whether in technology or policy, but institutional-level RDM also plays a role of intermediary between researchers and community-level RDM. As many academic libraries have already been doing, the intermediary role includes training faculty and students, providing tools, and consulting on RDM lifecycle issues.

Second, there is an interdependent relationship between community-level and project-level RDM. Community-level RDM relies on the contribution of projects for quality data, code, and metadata for long-term access and preservation, while project-level RDM needs an effective infrastructure service to save time, increase the accuracy and effectiveness of data contribution

and documentation, reduce unnecessarily repeated or redundant work, and avoid reinventing the wheel. RDM professionals in academic libraries should be knowledgeable of the relationships between community-level and institutional-level RDM to ensure the performance of RDM services.

Next steps

The CMM for RDM described in this paper is in many ways still a work in progress and the product of a small group. We hope to open up its further development to the larger community; Our goal in writing this paper is to invite readers to join in the project. To enable interested users to contribute, the Model is built on a wiki platform.

Future work is needed in several areas. First, the set of RDM practices can be extended, both in number and in depth of description. Description of RDM practices in the current Model was based on an extensive literature review, but there are undoubtedly additional practices that could be included and the practice descriptions can be extended or additional resources added.

Second, a key concept in the notion of levels of process maturity is the degree of institutionalization of the practices. In other words, the Model embodies the notion that good RDM is not an innovation in a particular project, but rather an expected and normal way that research is done in its field or at its institution. Therefore, work is needed to identify practices that are institutionalized in this way in particular disciplines or in particular organizations. As well, practices can be identified that could be shared at that level, and work done to establish those practices as disciplinary norms.

Some likely sources for such practices are academic libraries and large research centers in different disciplines. For example, libraries in many universities have established organizational units in various names for RDM services. This is a positive sign of institutionalization of RDM. The newly released *Data Curation Network: A Network of Expertise Model for Curating Research Data in Digital Repositories* (Johnston et al., 2017) brings the institutionalization of RDM to a new height. How these academic libraries achieved the institutionalization of RDM and how such institutionalization impacted the RDM services and processes would be worthwhile case studies of the CMM for RDM. Another area of case studies would be large research centers. 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. These or other centers like them would be useful case studies for applying the Model. 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. 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. Similarly, the intertwined relation between project- and community-level RDM and between institutional-level and community-level RDM, in other projects like the LIGO project, would make a good case study for applying the CMM for RDM model to study how community-level RDM supported the project-level and institutional-level RDM, and how project-level RDM and institutional-level RDM prompted the evolution of community-RDM.

Finally, a key use of the Model is to help projects and organizations assess their current level of RDM process maturity as a guide to where improvement efforts would be most beneficial. Future work should empirically assess the utility of this guidance and use these experiences to improve the rubrics.

Acknowledgement: This project is supported by the Inter-university Consortium for Political and Social Research / Sloan Foundation Challenge Grant.

Works Cited

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation, 8*(2), 5-26.

Anderson, S., & Williams, R. (2016, August 1). *LIGO Data Management Plan, August 2016.* Retrieved March 11, 2017, from LIGO Document Control Center: https://dcc.ligo.org/M1000066/public

Barsky, E. (2017-07). *Three UBC Research Data Management (RDM) Surveys: Science and Engineering, Humanities and Social Sciences, and Health Sciences: Summary Report.* University of British Columbia, UBC Library & UBC Advanced Research Computing. Vancouver: UBC Library.

CMMI Product Team. (2006, August). *CMMI for Development, Version 1.2: Improving Processes for Better Products.* Retrieved March 11, 2017, from Software Engineering Institute: 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 Association for Information Science and Technology. 45*, pp. 1-6. Hoboken, NJ: Wiley.

Gray, J. (2009). 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.

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2017). Data Curation Network: A network of expertise model for curating research data in digital repositories. http://hdl.handle.net/11299/188654

Key Perspectives. (2010, January 18). *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability.* Retrieved March 11, 2017, from Digital Curation Centre: http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf

Kirkland, A., Qin, J., & Crowston, K. (2014). *A Rubric for the CMM4RDM.* Retrieved March 11, 2017, from Capability Maturity Model for Research Data Management: http://rdm.ischool.syr.edu/xwiki/bin/view/Blog/A+Rubric+for+the+CMM4RDM

Lyon, L., Ball, A., Duke, M., & Day, M. (2012). *Community Capability Model for data-intensive research.* University of Bath, UKOLN. Bath, England: University of Bath.

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

Paulk, M. C., Weber, C. V., Chrissis, M., & Curtis, B. (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*(2-3), 188-204.

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., . . . Westbrooks, E. L. (2008, June 20). *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library.* Retrieved March 11, 2017, from Cornell's Digital Repository: https://ecommons.cornell.edu/handle/1813/10903

Whitmire, A. L., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: implications for data services development from a faculty survey. *Program*, *49*(4), 382-407.

Table 1. Common features in the Capability Maturity Model (Paulk, Weber, Chrissis, & Curtis, 1993, p. 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. Typical Commitment to Perform activities involve establishing organizational policies (e.g., the rules for data management) and senior management sponsorship.
Ability to Perform	Ability to Perform describes activities that ensure 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.

Table 2. Portion of the rubric (4.3 - Activities Performed) with corresponding level of maturity

Level of Maturity	**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: Intuitive 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.

Table 3. Assessment of project-level data management based on Maturity Level 2 rubrics with items in the first two areas

Maturity level 2 (Managed: DM process is characterized for projects and often reactive) and assessment criteria from the Model	**Item**	**Assessment results of project data management**
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.	1.1 - Commitment to Perform	*Stakeholders*: project team members, funder, institution *End user*: funding agency, policy makers and researchers, graduate students *Commitment*: Made documentation of data and workflows as a policy and communicated to the team members about enforcing the policy.
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.	1.2 - Ability to Perform	*Tool designation*: 1) Used Evernote to document strategies for coding, workflows for data processing and analysis, and any comments and questions; 2) All programming codes (R, Python, etc.) were properly annotated. *Shared space*: all team members had edit access to project shared space in Evernote, Dropbox folders, Google Drive, and servers designated to the project.
Workflow management during the research process, such as managing functional	1.3 - Activities Performed	Workflows: Master graduate research assistants report worked closely with the Ph.D. research assistant to define and

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.		assign specific tasks for weekly milestones. The results were reported at the weekly project meeting for discussion and steering by the PIs.
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.	1.4 - Process Assessment	Not yet developed or performed.
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.	2.1 - Commitment to Perform	Data quality were checked and cross-checked by using various methods, and the approaches, methods, and procedures were documented using Evernote. No community standards were applied in documenting the data since such documentations were mostly notes or annotations.
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.	2.2 - Ability to Perform	Training: Team members were trained to perform documentation tasks and applying trial-error method on small sample sets before deploying the code to whole dataset. Files: code files, data files to be fed to the code, and output files were named descriptively and kept together with metadata indicating the dependencies between the files.
The workflow for collecting and documenting data has been addressed for this project, but has not taken wider community needs or standards into account and has not been codified.	2.3 - Activities Performed	Although the workflow for creating documentation was created, the workflow for collecting documentations was not yet set up.
Measurement, analysis, and verification of data collection and documentation have been recorded for this project, but	2.4 - Process Assessment	(Almost) All data collection and documentation have been recorded, but not yet taken community standards into account.

have not taken wider community needs or standards into account.		

Figure 1. Capability maturity levels for research data management