IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A CAPABILITY MATURITY MODEL FOR SCIENTIFIC DATA MANAGEMENT: A LITERATURE REVIEW PAPER

(1) Dipak Kumar Bisoi (M.Tech.CSE) (Student)
Gandhi Engineering College (GEC, Bhubaneswar) Odisha

(2) Dr. Prakash Kumar Pathak (Professor of CSE)

Gandhi Engineering College (GEC, Bhubaneswar) Odisha

ABSTRACT

In this research paper, I propose a capability maturity model (CMM) for Scientific Data Practices (SDP) with the core aim of helping hands to assessment and improvement to such practices for building up of effective SDM by content analysis of papers about SDM practices and the reliable management policies. CMM works further by characterizing organizations by level of maturity of these processes. This model will be helpful for data management improvement evaluating SDM practices.

Keywords – Big Data, Maturity model, Business Analytics, Business value, Scientific Data Management.

INTRODUCTION

Data is omni present, generated and transferred from many heterogeneous origins and consistently come in varigate structures and formats. This existence has resulted in the emergent of " BIG DATA". Almost every organization seems to focus or manipulate or execute data for competitive advantage as well as solving accurate enterprise problems.

The explosion of data in terms of 3 V's i.e , volume, variety and velocity – is categorizing new models and reshaping industries. The business heads cannot stroll their way forward in the large big data's age. The transformation of business is a tough job and this complexity begins with understanding the opportunities unique to your business and also focusing on ability to maturity models by strategically pursue Big Data programs aligned to specific goals of, or related to business.

BIG DATA MATURITY MODEL

Maturity/ Analytics come from experience with the age cofactor similarly Maturity Models helps or is keen to help organizations grasp trends of data and data to achieve success. It mainly focuses and considers data and data sets of an organization to predict and analyze the capacity of data, decisions, data events too.

As data's quantity will undoubtedly increase day by day which will largely have impact on large organizations as there data would be extremely large and will be corely effective may be in major loss so the real understanding to leaders or business should initially be known to factors related to it hence the problem can be resolved and Big data can also have an impact widely on operational processes of an organization. SDM is at centre stage research cycle, leading to a data lifecycle. A scientific data management is nothing but a software that pre acts as a management of software wholely as a system (DMS) which includes capturing, cataloguing and achieving data either generated by laboratory instruments or applications generated. For handling SDM if we take capability maturity model (CMM) then CMM is keenly in need to be understood, it is a methodology for developing and refining an organization's software development process.

A CMM FOR SDM

The original CMM was developed at software Engineering Institute at Carnegie Mellon University to support improvements in reliable software development organizations, that too in time and budget criteria. The core reason to design was to help developers to select improved strategies by examining maturity and even identify critical issues to improve their quality of software.

Through this model has been changing in a quick ratio but the basic structure is roughly the same. Key concepts which has key practices, key specific and generic process areas and maturity models. For developing successful development of CMM the keen observation in terms of software, organizations must be capable of carrying out with more and more number of key software development practices. In this model, these key software development practices are into 22 specific process areas as clusters i.e, when implemented collectively satisfy the goal which is to be considered vital for making improvement in that specific area. These process areas are grouped together in 4 categories – support, project management, process management and engineering.

THE PROCESS IMPLEMENTED RELIABLY, NAMELY:-

- Gain Success of specific goals (processes are performed)
- Establishing a managed process (i.e, the organization has policies for planning and performing the process)
- Establish a defined process (process maintenance and improvement is collected)
- Establish quantitative managed process (Quantitative processes are made and performance is stabilized)
- Establish an optimizing process (Improvement of process continuously and root causes to problems are identified and pre solved)

Finally, the CMM describes its levels of processes and capability maturity for whole organizations symbolizing the margin of improved process and set of its data.

As for a start, initially there is an organization with no defined process, software being developed but to extempore making it impossible to plan or predict. As per the increase in maturity level processes become more established, furnished and standardized, as implications. Thus, CMM described the way to make processes mature enough (from immature) in terms of software quality and organizational effectiveness. My goal in this paper is to lay out the similar path for scientific data management.

Identifying Key Practices –

For creating CMM for SDM, we initially need to identify and cluster key SDM practices. As SDM represents yet to come emerging interdisciplinary research field, the processes and practices areas are explored and observed properly understandable looking forward for content analysis of such practices in form of literature to develop this part of the model.

RESEARCH METHOD –

Data collected for this paper involved selecting a set of published articles either corely or a part of research where such practices are considered for the other topics for result. Articles were observed and identified through journals or conference proceedings devoted to data curation, preservation and management. The aim of sampling the data for this research paper yields to insight understanding rather focusing on emphirical or other form of generalizations. Preferably, for this paper we chose to select an article set and examine content until practices reach saturation point. The main limitation or further future assessed for this approach that it does not provide the basis for relative frequencies of such practices, such inference would be limited to the frequency of mention even in best case.

I selected around 8 articles and reports in the considered area and later it was into content analysis software NVivo. The authors read the analysis as per there area of expertise and described few kind of practice. While a few discussed and described practices found in the article and further collected together that seemed to refer same practice hence it aimed to get key process areas as result. In this way CMM model was made and followed new areas were added, accordingly some area seems overlapping were collapsed.

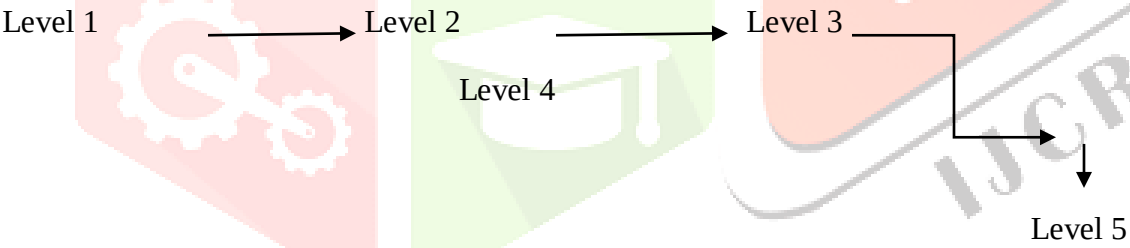
FINDINGS : SDM PRACTICES AND PROCESS AREAS:

From the analysis, it is more precisely noticed that there are large number of key practices for SDM.

Key Process Areas	Practice
➢ Data Acquisition and Data Description	1. Capture 2. Process and Prepare 3. Assure Quality 4. Develop 5. Describe 6. Create, Design and Ensure
➢ Data Dissemination and Preservation (Design and Preserve)	1. Identify 2. Encourage sharing 3. Distribute 4. Access 5. Back up 6. Preserve Package and deliver data

Scientific Data Management Maturity Model Levels

Maturity level in SDM describes the level of development of the practices in specific organization. SDM practices carried out in scientific projects from pin point to well managed processes.



To describe this figure,

Level 1 – Initial Stage

In this stage, data are managed automotive without any practices.

Level 2 – Managed Stage

Characterizations of DM processes are done.

Level 3 – Defined Stage

DM characterised for specific organization.

Level 4 – Quantitatively Managed Stage

Control and measured DM.

Level 5 – Optimizing Stage

Keen Observation on improvement of process.

Process Area 2 : Establishing managed process (Data Management work as a managed process)

Goal	Practices
1. Organization establishes policies for planning	Develop data, release data share, policies and rules for use of data, data curation policies
2. Data Management plan is made and maintened	Access data flow, user requirements staffing needs.
3. Resources provided as per need	Business model presentation, appraise, develop nd manage SDM, tools and technologies.
4. Responsibility Assigned	Identification amd assigning of roles
5. People trained	Train researchers, online guidance
6. Work Products controlled	Changes to data controlled
7. Stakeholders are identified	Develop collaboration and partnership with communities
8. Process is monitored	Access and enforce

DISCUSSION

Set of practices and processes described above are straight preparatory. Future research may elaborate these descriptions. In this present state may lead to some nice comparisions and show benefits of reflecting this model. Perhaps as a result of sources selected, I haven't described any reuse of data, in future such practices can be the topic to work upon.

By writing, found that these practices revealed few interesting differences. Firstly, we need to have few practices that established till level 2 of managed process and for further levels. Each level is made on the top of the previous level therefore it makes impossible for any level to work individually without depending on previous. But for present analysis of this paper, the processes are usually not defined beyond the project level and rarely optimized or managed at least not for the sources analyzed in this paper.

Secondly, found a number of practices related to development or management of technologies. It seems that tools are still being in developed phase. We may further find a discussion for the achieving of data for a long term use.

CONCLUSION

This model described in this research paper is still in preliminary state, but to a positive note the possibility of seeing some possible implications is the achievement on our level. Firstly, the catalogue of processes should help organisations ensuring covering all aspects of data management. The detailed version of goals, objectives and practices will provide a guide for managing and implementation practices. Secondly, data management for organizations is corely covered in this paper. Thirdly, CMM supports practices and process areas that work to support a higher level of organizations capability and core management.

Finally, I hope that as per software development, careful description of different levels of maturity may serve for organizations to improve their level of maturity enabling better level of SDM.

REFERENCE

- 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/6fs4559sBorgman.
- 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/do cs/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.