Research and application of data management based on Data Management Maturity Model (DMM)

Yang Baolong

National Academy of Innovation
Strategy, CAST
No.222, Huizhongbei Road, Beijing
861087413445
yangbaolong@cast.org.cn

Wu Hong

National Academy of Innovation
Strategy, CAST
No.222, Huizhongbei Road, Beijing
861087413443
wuhong@cast.org.cn

Zhang Haodong

National Academy of Innovation
Strategy, CAST
No.222, Huizhongbei Road, Beijing
861087413445
Hdzhang06@126.com

ABSTRACT

Through the analysis and contrast of the different Data Management Maturity Model, such as DCAM, DMM, DCMM and the model of IBM, we try to make empirical research under the framework of data management maturity model. This article take a project whose main research object is about the academic career of scientists and with massive unstructured data for example, through analysis of the goal, management processes and influence factors of this project in detail, we built up an evaluation system for data management for such projects under the framework of DCMM. It is expected to have a positive significance to the evaluation of similar data management capability.

CCS Concepts

•Software and its engineering → Capability Maturity Model

Keywords

Data management; maturity model; Measurement and evaluation; Unstructured data

1. INTRODUCTION

In Jim Gray's last talk to the Computer Science and Telecommunications Board on January 11, 2007, the fourth paradigm of scientific research has been proposed [1], and nowadays data-driven scientific research has increasingly become the mainstream of scientific research. The debate between Ali Rahimi and Yann LeCun which just happened in NIPS2017 also causes the big discussion about theoretical basis of deep learning [2]. Almost at the same time, a viewpoint collision of the Computing Capability, Big Data and Algorithms, which is the most fundamental driving factor to artificial intelligence also becomes a hot topic in domestic related fields [3]. All these phenomena prove the important of data in this fast digital age, along with this, the significance of data management should be given more and more attention in every research field. To improve software quality just rely on making evaluation to each stage in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICMLC 2018, February 26–28, 2018, Macau, China

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-6353-2/18/02...\$15.00

DOI: <https://doi.org/10.1145/3195106.3195177>

the practice of software development process, which is far from the meeting the needs of data as a strategic resource. In this paper, by comparing the domestic and foreign data management maturity models, as one project which is oriented to the data collection and analysis and processing of growing-up of China modern scientists and has lasted for 10 years as an example (this project is named as Project S for short in the following), a data management maturity framework taking the above content as the research object is put forward to. The purpose of this framework is to promote the improvement of the data management capability of this project and also to explore the DMM model for engineering applications.

2. RESEARCH STATUS OF DATA MANAGEMENT MATURITY MODEL

2.1 An overview of data management maturity model

The publication of the book "The IBM Data Governance Unified Process" [4] formally proposed the concept of data governance capability maturity model in October 2010. Drawing the experience from CMMI, this book defines the domain of data governance in detail, and divides the field into Data Risk Management and Compliance, Data Value Creation, Organizational Structures & Awareness, Data Stewardship, Policy, Data Quality Management, Data Life-cycle Management, Data Security & Privacy, Data Architectures, Classification & Metadata Management and Audit Information Logging & Reporting.

The Data Management Maturity (DMM) Model was developed using the principles and structure of CMMI Institute's Capability Maturity Model Integration (CMMI)—a proven approach to performance improvement and the gold standard for software and systems development for more than 20 years [5], which is put forward by the Software Engineering Institute (SEI) of Carnegie Mellon University. The DMM Model contains six major functional domains which include Data Management Strategy, Data Quality, Data Operations, Platform and Architectures, Data Governance and Supporting Processes.

Data management Capability maturity Model (DCAM) was issued by the EDM Council which is a non-profit trade association founded by the financial industry to elevate the practice of data management as a business and operational priority [6]. The model was created based on the collective experience of data practitioners across the global financial information industry which focuses on the practical capabilities needed to ensure trust and confidence in data content as a factor of input into business processes. DCAM's data management capability maturity model is

mainly divided into eight functional domains, which are Data Management Strategy, Data Management Business Case, Data Management Program, Data Operations, Data Quality, Data

Governance, Data Architecture and Technology Architecture respectively.

Figure 1. One flowchart example of DMMI model

In order to regulate the domestic industry data management and application, and also promote the application ability, Chinese start the data model capability maturity evaluation work in 2014. In 2016, a national standard of Data Capability Maturity Model capability maturity DCMM was published, which includes eight domains, they are Data Strategy, Data Governance, Data Infrastructure, Data Standard, Data Security, Data Application, Data Quality, Data Life Cycle Management, respectively. DMMI model, consists of seven core processes, namely business architecture, data dictionary, technical architecture, data governance frame work, data dimension & quality reports, data cleansing and mistake proofing; refer to Figure 1.

2.2 Comparison of Data Maturity Model

From the above analyses can be seen, IBM is a pioneer in Data Maturity Model research, whose model lays particular stress on the perspective of technology and possesses finer granularity of services, for example, metadata and data auditing are included in this model. So the IBM's model owns strong operability, and is more suitable for evaluation on enterprise data management, and especially can be used for its own repeated iterative and optimization.

DMM model which advocated by CMMI Institute is based on 20 years of software maturity research accumulation and massive application cases, and its service goal is very clear. Because there is very close relationship between the data management model evaluation and the CMMI, whose evaluation has been widely applied around the world, therefore, the prospect of DMM model's application is very bright.

The data maturity model of EDM can directly give the evaluation ranking of the enterprise in the aspects of its application, which possesses distinctive features, and it is widely used in the financial industry.

As for the Data Capability Maturity Model (DCMM), which is self-developed by China. In the process of standard formulation process, we fully learned the development experience of domestic advanced industry, mainly financial industry, and also combined with the contents of DAMA and DMBOK. The national standard of DCMM is formally published in 2016 as GB/T 33136-2016. This standard covers data management from technology, organization, system and process. Take account in the internal research status, data standard, data application and data security, these three aspects are proposed as separated capability domains [7], and is also classified corresponding to the four functional domains.

3. DATA MANAGEMENT SERVICES AND APPLICATIONS BASED ON PROJECT S

3.1 Current status of Project S

Project S aims to collect information about scientists who are over 80 years old, with rich academic experience and significant contributions to the development of science and technology in New China. Since the launch of the pilot program in 2010, with the close cooperation of all parties and the support of scientists and their families, in strict accordance with the project implementation plan, we conscientiously carried out the data collection, cataloging and research work of the academic growth data of the elderly scientists. Till now, the collection of 496

scientists' academic growth data has been conducted, and nearly 300TB scientists' raw data have been collected. The project has obtained a huge amount of scientific and historical documents, which has laid a solid foundation for the in-depth study of the law of the growth of modern Chinese scientists.

The project is a huge and complex system work with no processors. Collected data includes the oral information of old scientists and their physical materials such as biographies, certificates, letters, manuscripts and writings, as well as the important photographs of participating in official activities, academic activities, foreign affairs, social activities and family life, and other video, audio tapes, video tapes, CDs, audio-visual materials. These unstructured data greatly deepens the complexity of management and application.

The scientist raw materials' database and National Museum for Modern Chinese scientist (<http://www.mmcs.org.cn>) have been launched from 2013. In early 2016, the website of this museum which is the first comprehensive and 3D-displayed online portal in China to show the academic growth of Chinese modern scientists was released to public. It provides the first comprehensive and three-dimensional displayed website to show the academic growth of Chinese scientists. By the end of November 2017, 330 scientists' information have been displayed online, the hardware with 1 PB storage capability and 300TB data migration is finished, as well as filename-based retrieval system, all these laid a good foundation to support further data analysis and mining of this project.

3.2 Data Management Maturity Model (DMM) based on Project S

The DDM model adopts phased representation methods and includes five levels: initial level, managed level, defined level,

quantitative management level and optimization level according to the maturity of the whole process capability. Project S possesses the largest database of historical data of modern Chinese Scientists and 15 classified items covering the scientists' raw data about their academic career. How to measure its maturity in data collection, management, security and service capability under the framework of DDM is our main concerning, and now we divide the data management maturity into 5 stages in Project S; refer to Table 1.

- The initial level. This phase of project data management capabilities is at the base stage. The feature of data management is that data management is mainly completed in collection team spontaneously, without a unified data management process. This level is typical of the large amounts of isolated data island, passive data services, huge maintenance workload and weak security.
- The managed level. The data management capability is in the development and growth stage. According to the requirements of the management strategy, a management process was established and a project office was formed to clarify the responsibilities between the data management team and data research team. The rules and standards of data collection, cleaning and using are established. The public-oriented services is kicked off. Data backup and monitoring strategy are drawn up, data security has been greatly improved.
- The steady level. The data management capability is in a stable stage. A series of standardized management processes have been developed around data collection, management, and use. An accurate and consistent data will be provided to meet the needing of research. The hardware facilities is upgraded, such as VPN tunnel will be built.

Table 1. Data Management Maturity Model of Project S

Level	Collection capability	Management capability	Security capability	Service capability
The initial level	Data collection is mainly completed in collection team spontaneously, without a unified data management process about pictures, audio, video and other unstructured data.	Provide simple data storage management, MS Excel indexing.	Mobile hard disk to collect data, weak security capability	Passive service, to solve the problem case by case.
The managed level	Develop data collecting flow charts, independent project management office, data management team and data research team	Preliminary rules of data collecting, cleaning and using	Centralized unified data management, data security has been protected by developing a data backup and monitoring strategy,	Provide offline data borrowing. Clarity the searching scope and authority for manager, researcher and the public.
The steady level	A series of standardized management processes have been developed around data collecting, management and using	Provide different levels of data management tools at different levels. data quality is under control	Upgrade hardware facilities regularly, set up VPN tunnels. The data initiative safety has been promoted.	Meeting the needs for data research with online device.
The	Provide data relative offline software to	Establish a unified metadata model and	Develop Classification	Develop a general service interface for web portal.

quantitative management level	improve the collected data quality	standard of data acquisition, cleaning and using	protection system, provide quantitative measure of policy for data storage, data accessing and data services.	
The optimization level	Provide online feedback mechanism through the platform, and adjust inappropriate data acquisition rules in time.	Establish full life cycle data quality management.	Establish the Disaster Recovery Data Center to enhance the data security furtherly.	Provide full-text search capability for massive unstructured data, and provide online real-time service for date research.

- The quantitative management level. A comprehensive optimization for data management processes is conducted. A unified metadata model and standardization norms of data acquisition, cleaning and using is established. A general-purposed service interface for upper web application is developed. Data storage, data accessing, and data services' security will be measured in quantitative.
- The optimization level. A real-time optimization of data management process will be provided, along with the online feedback mechanism will be adjusted on time in order to deal with the inappropriate data acquisition rules. Life cycle data quality management criteria will be built. A full text searching capability for massive unstructured data will be provided.

4. CONCLUSION

How to build up the evaluation system under the data management maturity model, and especially applied in the complicated project is a hotspot in this fast data age. We give a rough model of data management which is mainly oriented to massive data of the academic career of scientists in China. From the four dimensions of collecting capability, management capability, security capability and service capability, we divided

the data management maturity of this project into five level. In the future, we will develop quantitative indicators referred to each level and apply into the actual process of this project to improve the data quality, and also hope to explore a complete data management method to other similar projects.

5. REFERENCES

- [1] Tony Hey, Stewart Tansley, and Kristin Tolle (2009), The Fourth Paradigm – Data-Intensive Scientific Discovery.
- [2] <https://www.facebook.com/yann.lecun/posts/10154938130592143>
- [3] <http://blog.csdn.net/cf2suds8x8f0v/article/details/78815264>
- [4] Sunil Soares (2010), The IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices.
- [5] <http://cmminstitute.com/resources/data-management-maturity-model-v10>
- [6] <http://www.edmcouncil.org/>
- [7] LI B, BIN J Z. Data management capability maturity model[J]. *Big Data Research*, 2017, 3(4): 29-36.