IFLA WLIC 2017 WROCŁAW

Submitted on: 29.05.2017

Context-based Roles and Competencies of Data Curators in Supporting Data Lifecycle: Multi-Case Study in China

Zhenjia Fan
Department of Information Resources Management, Business School, Nankai University, Tianjin, P. R. China
E-mail address: fanzhenjia@nankai.edu.cn

Copyright © 2017 by Zhenjia Fan. This work is made available under the terms of the Creative Commons Attribution 4.0 International License:
http://creativecommons.org/licenses/by/4.0

Abstract:
[Purpose/Significance] This essay describes the state quo of data curation in three kinds of given contexts including enterprises, research institutes and libraries in China in data-driven era and discusses on roles of data curation in such given contexts, which would have significance both in theory and practice.
[Method/Process] This essay adopts multi-case study method, with data curation and data governance models, to analyse the roles and their competencies of data curators based on different contexts. Case collection would cover significant enterprises as Neusoft, one of the largest IT companies in China, and research institutions as China Academy of Sciences, and libraries as National Library of China and several academic and university libraries. Via the data lifecycle analysis on different cases, the critical roles such as data supervisor, data steward and data custodian in insuring data quality and efficiency of data reuse would be outlined. Based on observation and interview of participants from corresponding roles, the General Competency Framework (GCF) of different roles required would be put forward. After then, suggestions for empowering data curators would be raised according to GCF.
[Implication/Conclusion] Besides digital archiving and preservation, more emphasis should be on data reutilization in the field of data curation. In different contexts of data curation practices, roles of data curation are not equivalent to interest relators in context of data governance. Different roles of data curators would take their own parts in the process of data curation and should be specified according to data curation in given contexts.
[Originality/Value] The General Competency Framework and empowerment policy suggestions might generate significance for the fields of data curation and data governance.

Keywords: Data Curation, Data Lifecycle, Data Curator, Data Literacy

Introduction

As one of the key activities for innovative enterprises, research institutes and universities, data curation has improved the efficiency and quality of R&D data management in the era of big data. In the field of data curation practice, different agents such as university libraries, enterprises, government data centers and other institutions based on specific

business situations have accumulated a lot of valuable cases of real operations, which has strong reference value to both theory and practice of data curation. Based on this problem context, this study tries to summarize the key roles and competencies from multi-case study approach. This essay describes the state quo of data curation in three kinds of given contexts including enterprises, research institutes and libraries in China in data-driven era and discusses on roles of data curation in such given contexts, which would have significance both in theory and practice.

Literature Review

The process of data curation involves different stakeholders who take the corresponding roles. In order to grasp the theoretical basis, it is essential to complete the literature review covered topics such as data governance, data curation and data literacy.

Data Governance and Framework

Data governance can be considered as a set of activities including planning, supervision, and enforcement that governs the process and methodologies that are carried out to ensure and improve the quality of data. Otto (2011) points out data asset is the object of data governance and data management, and the common purpose is to maximize the value of the data, but data management needs to be carried out under the guidance of data governance. The Data management, as a data management practice aimed at improving the level of data reuse, need to follow a certain governance framework and guidance strategy. The data governance model proposed by the IBM Data Governance Committee regards business objectives as the most critical factor in data governance, and the three factors of organizational structure and awareness, policy, and data stakeholders are Objective of business objectives.

In library science, Gu (2016) emphasizes data management from the perspective of data life-cycle, which consists of data acquisition, data sharing, data reuse and data appreciation. The perspective has been achieved with the data management and specific business of the basic convergence. Huang & Lai(2016) combed the library as the core data stakeholders and summarized library, data centre, research institutions, government and public sectors, policy-making institutions, fund management organizations and data publishers as the main categories of data governance subjects.

Data Curation and Roles

The concept of "curation" originated from the field of museum science and has been applied in the field of data management. It soon led to the recognition and attention of LIS disciplines. Data management objects, including observation, calculation, experiment, derived and other means to obtain data, can be expressed as text, number, image, video, audio, software, algorithms, reports, models and other forms. The Digital Curation Center (DCC) defines it as all activities that maintain, preserve, and value-add value during the lifecycle of digital data. In addition to the "preservation" of the data, the emphasis is data value-added and reuse. In this process, the collaboration between data researchers, publishers, managers and users is highlighted (Heidorn, et al, 2008).

The model of data curation and management mainly includes two categories: one is based on the concept of logical link between the conceptual model, such as Detailed Data Life-cycle Model by Wang & Shen (2014) who proposed six stages of 14 specific steps of the and Wu (2016)’s Data Management Quality Control Model. The other is business model based on the data management practice, such as DCC data regulation model (Heidorn, 2008)

and so on. At present, the research results of data curation have been discussed from the concept, theoretical introduction to the development of empirical research paradigm, focusing on the analysis of specific cases or combining existing models to try to analyse real problems.

Data Literacy

Similar to the concept of information literacy, data literacy is becoming one critical concept in LIS, which is related to data management through all the lifecycle of data management. According to Koltay (2015), Hagen-MacIntosh(2016) and Carlson & Johnston (2016), data literacy can be summarized as following: data literacy is embedded in R&D data flow and related to data life-cycle; R&D management and data utilization are main perspectives in data literacy; practical skills such as data analysis, description and tools would be main focus in data literacy.

Research Design

In view of the limitations of the existing achievements, this study explores the analysis approach of "context-business-role" to summarize the roles from real facts.

The basic situation of this study is as following: (1) what is the context of data curation? (2) what are stakeholders of data curation? (3) what are the key roles in the data curation process? Based on the above questions, this study adopts multi-case study method to summarize the roles and competencies.

Case study can be used to describe and analyze the case in depth, and through management behavior and results can help to tap the implied impact factors in management cases. In view of the lack of effective localization of data governance and data management and related results, this study attempts to summarize the theoretical framework from the practice of data curation in different context.

This article selected Neusoft Corporation (hereinafter referred to as "Neusoft"), Chinese Academy of Sciences (hereinafter referred to as "CAS") and Nankai University (hereinafter referred to as "NKU") as the main case objectives. Participatory observation, in-depth interview and document investigation has been adopted in data collection.

According to the requirements of "Triangle" (Yin, 2013), this study is based on data collection from multi-time, multi-channel and multi-source, and constructs the complete evidence chain. In addition, open interviews and participatory observations were used in fieldwork to facilitate a holistic understanding of the case. The above two aspects can guarantee the validity of the case data. In addition, this study preserves the case database and contact information, which helps to support the same research by other researchers. The reliability of this study can be ensured at the case study level.

Discussion

Based on the case study, the data governance framework needs to establish a clear organizational structure and division of responsibilities in the context of data curation. It is concluded that the R&D data curation related to 9 specific subjects, including the competent authorities, assessment experts, auditors are outside the main body, scientific research manager, research manager, R&D personnel, financial staff are internal subjects, project leader, the person in charge of the collaboration unit belongs to the contact subjects. According to the actual authority in the data curation, the three main subjects of data supervisor, data custodian and data steward are the core roles. Other roles also include data creators and data users, which can be converted to each other. The basic roles of the relevant roles see Figure 1.

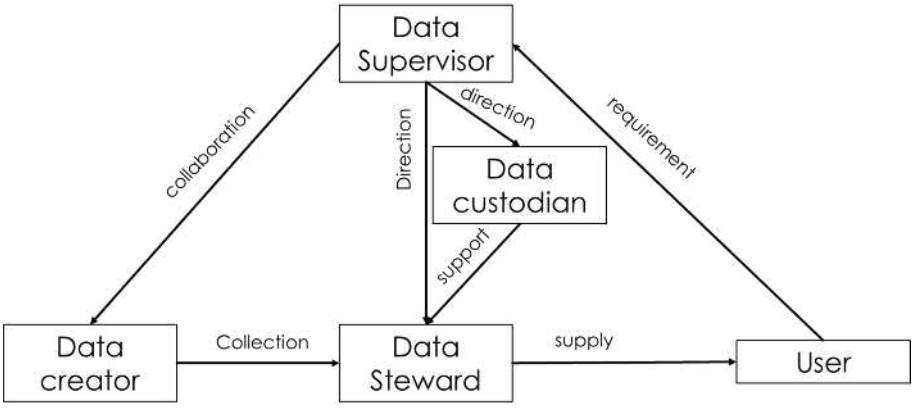

Fig 1 General Model for Roles of Data Curation

Depended on qualitative research of multi-cases, both structural and agency factors will affect data curation, and data curation can lead to achievements. That's to say, we can benefit from data curation in practice. And achievements can feedback to context of data curation. If the utility can bring more freedom for stakeholders, we can call it as development. So data curation can be regarded as a bridge among different stakeholders.

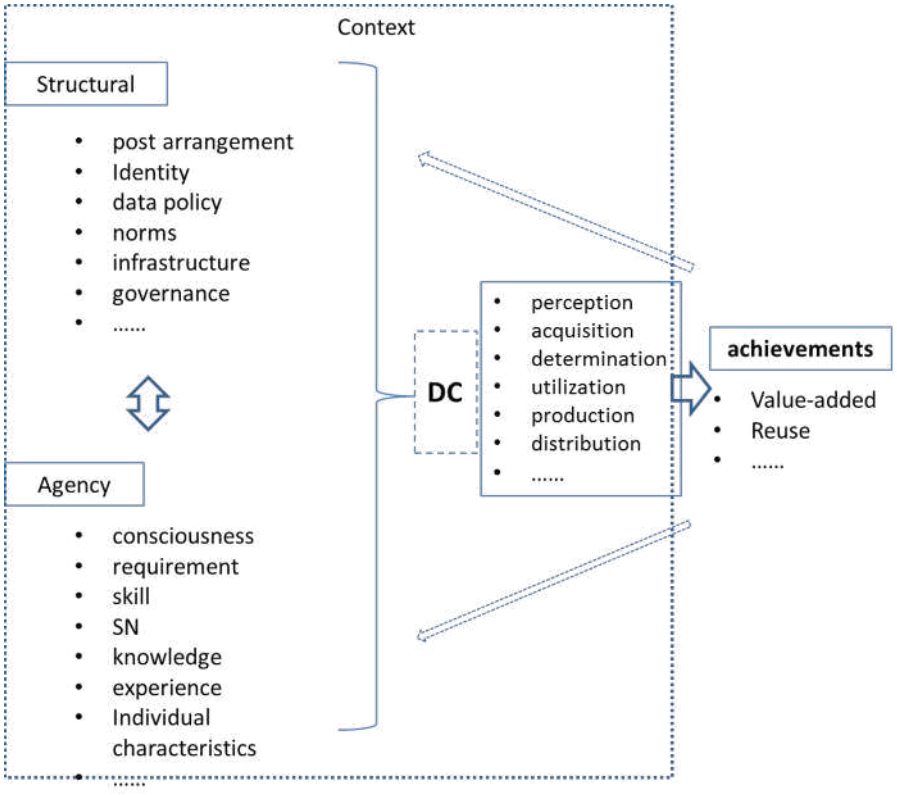

Fig 2 General Competency Framework

According to coding analysis of interview for data curators, the core literacy in data curation can be concluded as following:

- Data planning: be familiar with data types, volumes, formats, metadata, standard, reservation and access, etc.
- Data creation and collection: be familiar with data sources, tools, skills and evaluation criteria.
- Data processing: be familiar with the description, cleaning and cataloguing for

data.
- Data analysis: be familiar with metadata and visual tools to service the decision-making.
- Data preservation: be familiar with long-term preservation, data security and problem solving.
- Data sharing and reuse: be familiar with data sharing platform, policy and regulations.
- Common competency: including communication skills, background knowledge, collaboration abilities, etc.

Findings and Implications

Data supervisors and data stewards often have background of data creators and data users. The role of data creators, data managers, and data users are often relatively easy to distinguish, but data stewards and data custodians typically behave as research manager positions or librarians, and it is necessary to further subdivide these two roles in post settings.

Combining with the status of data curation at universities, certain problems of the practitioners on data literacy are pointed out: lack of policy, unclear responsibilities, mismatch in quality and demand of staff's data literacy, lack of professional education and others. Aiming at the problems mentioned above, this study put forwards corresponding countermeasures and suggestions. First, clarify the job responsibilities and professional quality, and speed up standards setting and policy construction. Secondly, establish data literacy training system, and strengthen data consciousness and data ethics education. Thirdly, we should promote innovation on data curation and realize the mutual promotion of theory and practice. Finally, cooperate with professional departments in university to establish relevant courses and to cultivate professionals on data curation.

Based on the real business situation, this study combines the stakeholders involved in the data governance framework to sort out the role of the subject of scientific research data management and management, and completes the role recognition of the subject of data curation. The main conclusions are as following:

- Stakeholders involved in the data governance framework are not fully related to the subject of data curation;
- The framework of data curation needs to establish a clear organizational structure and division of responsibilities, and clarify the synergistic relationship between different stakeholders;
- Data curation as a specific business data management, core roles, including data supervisors, data stewards and data custodians should be subdivided in post settings.

Because of the inherent limitations of the case study, this study explores the roles of the subject of data management in the framework of data governance, and needs to be tested for more case studies. In addition, based on the data management framework, different data management role should have the data literacy framework, as well as data continuity, master data maturity, data management performance measurement and other aspects of how to achieve different management role coordination research issues, pending More in-depth study.

Originality/value

The General Competency Framework and empowerment policy suggestions might generate significance for the fields of data curation and data governance.

Acknowledgments

This study is supported by *the Fundamental Research Funds for the Central Universities* (Project: Innovation Driven Enterprise Data Governance) in Nankai University.

References

CARLSON, J, JOHNSTON, L.(2015). Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers. Purdue Information Literacy Handbooks. West Lafayette, Indiana: Purdue University Press.

GU L. (2016). Data Governance: Data Governance: Opportunity for the Library. Journal of Library Science in China, 42(9): 40-56. (In Chinese)

HAGEN-MCINTOSH, J. (2016). Information and Data Literacy: The Role of the Library. Oakville, ON, Canada Waretown, NJ, USA: Apple Academic Press.

HEIDORN P B, TIBBO H R, CHOUDHURY G S, et al.(2008). Identifying best practices and skills for workforce development in data curation[J]. Proceedings of the American Society for Information Science & Technology, 44(1):1-3.

HUANG R, LAI T. (2016). Analysis of the Library's Participation in Scientific Data Management From the Perspective of Stakeholders. Library and Information Service, 60(3): 21-25, 89. (In Chinese)

KOLTAY, T. (2015). Data Literacy: In Search of a Name and Identity. Journal of Documentation, 71(2): 401–415.

OTTO B. (2011). Data Governance [J]. Business & Information Systems Engineering, 3(4): 241-244.

WANG F, SHEN J. (2014). Advances in Data Curation Abroad: Research and Practice. Journal of Library Science in China, 40(4): 116-128. (In Chinese)

WU J, CHEN Y, HU M. (2016). Research on the Process Reference Model of Scientific Data Quality Management in e-Science Environment. Journal of the China Society for Scientific and Technical Information, 35(3): 237-245. (In Chinese)

YIN R K. (2013). Case Study Research: Design and Methods[M].5th ed. Thousand Oaks, CA: Sage.