Research on Data Management Plan Formulation Specifications and Their Operational Data Curation Model *Postprint

Authors: Liu Feng, Zhang Xiaolin

Date: 2017-10-11T00:00:00+00:00

Abstract

[Objective] Propose a set of detailed constituent specifications for scientific data management plans; and construct a data curation model from an operational perspective. [Method] Investigate and statistically analyze the data management plan specifications of major international research management institutions; and supplement them based on the current needs and characteristics of scientific data management. [Result] Developed detailed constituent specifications for data management plans comprising 8 major basic elements and 39 sub-elements, and constructed a data-management-plan-driven data curation model. [Conclusion] The detailed constituent specifications for data management plans can completely and accurately standardize and guide scientific data management activities, and can also effectively control and constrain the data curation process throughout the entire research lifecycle at the operational level.

Full Text

ChinaXiv Partner Journal, Issue 266, 2016, No. 1

Research on the Specification of Data Management Plan and Its Operational Data Curation Model*

Liu Feng[1],[2],[3], Zhang Xiaolin[1]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [Objective] This study proposes a detailed structural specification for scientific data management plans and constructs an operational data curation

model based on it. [Methods] We investigated and statistically analyzed the data management plan specifications of major international research management organizations, supplementing them with current requirements and characteristics of scientific research data management. [Results] The study developed a detailed structural specification for data management plans comprising eight major basic elements and 39 sub-elements, and constructed a data curation model driven by data management plans as the core component. [Conclusions] The detailed structural specification of data management plans can comprehensively and accurately regulate and guide scientific data management activities, while also effectively controlling and constraining the data curation process throughout the entire research lifecycle at the operational level.

Keywords: Data Management Plan; Data Curation; Research Lifecycle
Classification Number: G250

This work was supported by the National "Twelfth Five-Year" Science and Technology Support Program Project "Research on Cloud Storage and Cloud Computing Technologies for Rural Information Services and Platform Construction" (Project No. 2013BAD15B02), the Chinese Academy of Sciences "Twelfth Five-Year" Informatization Project "Scientific Data Resource Integration and Sharing Engineering" (Project No. XXH12504), and the Chinese Academy of Sciences Computer Network Information Center 135 Planning Key Cultivation Direction Special Project "Research Big Data Resource Management and Service Platform and Its Key Technologies" (Project No. NIC_{{{PY}}_{{1405}}}).

Corresponding author: Liu Feng, ORCID: 0000-0002-5816-2067, E-mail: liufeng@cnic.cn.

As summarized by the UK Digital Curation Centre (DCC), effective management of scientific data offers numerous benefits. These include improved data discovery and understanding when needed, maintaining project continuity despite personnel changes, avoiding redundant data collection or processing, providing more opportunities for scientific collaboration, enhancing individual research impact, and strengthening the traceability, reusability, and verifiability of scientific data. Scientific data management plans represent the most effective means to help researchers achieve these advantages.

A scientific data management plan is a formal document that outlines how scientific data will be effectively processed during and after a research project. It is not static but is continuously enriched and refined throughout the project lifecycle to become more accurate. Through scientific data management plans, researchers can comprehensively grasp the entire process of data generation, processing, sharing, and application, conveniently track research progress, and make targeted decisions, thereby ensuring effective data management throughout the entire research process.

In recent years, the data-centric, data-driven nature of research has become increasingly prominent. To ensure research integrity, international funding organizations and research institutions have placed growing emphasis on scientific

data management and sharing policies in research projects. For example, since January 2011, the U.S. National Science Foundation (NSF) has required the submission of a "Data Management Plan (DMP)" with project proposals. The UK Research Councils (RCUK), comprising seven research councils, and the Wellcome Trust have each issued their own data management plan requirements. Organizations such as the Data Observation Network for Earth (DataONE), the UK Digital Curation Centre (DCC), and the Inter-university Consortium for Political and Social Research (ICPSR) have also proposed detailed data management plan specifications from their respective perspectives.

2 Component Analysis

Current data management plan specifications issued by various organizations exhibit both similarities and differences in their main components due to diverse requirements, posing challenges to a complete and accurate understanding of scientific data management plans. Therefore, it is necessary to comprehensively review the structure of scientific data management plans.

To this end, we conducted a detailed statistical analysis of data management plans from the aforementioned major research management organizations, as shown in Table 1 . Through review and statistical analysis of data generation, organization, storage, rights protection, sharing and reuse, archival preservation, and resource support, we ultimately identified eight major categories for data management plans: data production context, data organization specifications and strategies, data storage and security management, data ethics and intellectual property, data sharing and service practices, data reuse management, data archiving and long-term preservation, and data resource support plans.

Among these, data production context primarily includes relevant context for data collection or production (tools, environment, experimental methods, etc.) and potential future user needs. Data organization specifications and strategies determine how data will be organized into files, including data and metadata storage formats and standards, and what types of data products will be generated, considering which relational databases or other data organization strategies are most suitable. Data storage and security management clarifies data storage locations, methods, and management responsibility chains; backup and version control methods and tools; and related security and confidentiality measures. Data ethics and intellectual property focuses on data resource ownership, data copyright and permission management information, and explicitly states data sharing rights requirements, including data use licenses, confidentiality, and other ethical considerations. Data sharing and service practices describe plans for sharing data and data access and publication policies, determining the conditions, scope, and timing of data publication, as well as how to cite data. Data reuse management describes strategies for how data is typically accessed or shared, such as specific terms of use for disclaimers and conditions for data application in other uses or products. Data archiving and long-term preservation identifies datasets with long-term value and describes archival preservation

methods to ensure the long-term value of key datasets. Data resource support plans outline the resource information needed to implement individual data management plans, describing requirements for additional funding, software, hardware, technical expert support and training, and how to obtain these resources.

3 Detailed Component Analysis

The eight basic elements of data management plans only provide a generalized outline of the basic structure. However, to comprehensively and thoroughly understand data management plans, we must decompose these elements in detail. Therefore, based on the content of the basic elements, we systematically refined each element to form a series of sub-elements. Simultaneously, considering current requirements and characteristics of scientific research data management, we incorporated sub-elements such as data retention periods, embargo periods, priority rights declarations, and data citation. We also categorized elements as mandatory or recommended based on their essential level in data management plans. The refined elements of data management plans are presented in Table 2 .

4 Data Curation Model Design

As demonstrated by the detailed element decomposition in Table 2, data management plans cover all aspects of the entire lifecycle of data curation, from data production and sharing services to data archiving and preservation. They provide crucial specifications and guidance for data management in scientific research. From a current application perspective, data management plans remain at the level of normative documentation. To further manifest their guiding and controlling role, we constructed an operational, data management plan-driven framework model for the entire research lifecycle, as shown in Figure 1 [Figure 1: see original paper].

The framework model consists of three main components. First, the Data Curation Engine (DCE) serves as the core control module for the entire lifecycle of research data, responsible for constraining and controlling the complete lifecycle management process from data production to data sharing and archival preservation. The DCE can be further subdivided into sub-modules for user and permission management, data management plan management, and workflow management. Second, the Data Management Plan (DMP) constitutes the core input to the data curation engine, with the detailed specifications of the data management plan forming the essential elements of the engine. Third, Data Production Management (DPM), Data Sharing Management (DSM), and Data Archiving Management (DAM) are abstracted as important functional operation modules for the lifecycle flow of research data, accepting drive and control from the data curation engine. Each module can be further divided into sub-modules: DPM includes data collection and processing, data organiza-

tion, and data storage; DSM includes data rights control and data sharing and publication; DAM includes data archival backup and long-term preservation.

The framework model demonstrates that the data management plan is the core driver, with its specifications forming the essential elements of the entire model. However, to truly realize the core constraining and controlling role of data management plans, we must filter out quantifiable control elements from the detailed specifications. Table 3 presents the filtered quantifiable control elements of data management plans.

Scientific data management plans are documents that effectively standardize the entire data management process. Currently, major data regulatory organizations have proposed specification requirements based on their own needs, but no unified standard has been formed, posing challenges for unified and standardized management of scientific data in research projects. Based on the specifications of major data management organizations and statistical analysis, this paper proposes a detailed structural specification for scientific data management plans, aiming to comprehensively and thoroughly summarize the main elements and specification requirements of general scientific data management plans.

Based on the design of the detailed specifications for data management plans, we argue that data management plans should not merely serve as simple textual documentation. Their role should be more profoundly reflected in their guidance and control over the standardization and normalization of scientific data management. Therefore, this paper proposes a scientific data curation model driven by data management plans from an operational perspective. By quantitatively setting specific elements of data management plans, this model can construct a detailed control-driven data management system for the entire research lifecycle.

In China, data management plans are gradually gaining attention from major projects and data management organizations, while research on lifecycle-based data curation continues to deepen both domestically and internationally. We hope this study can provide meaningful references for future scientific data management practices.

References

[1] Strasser C, Cook R, Michener W, et al. Primer on Data Management: What You Always Wanted to Know [EB/OL]. [2015-08-18]. https://www.dataone.org/sites/all/documents/DataONE_{{{BP}}*{{Primer}}}{020212}.pdf.

[2] Guidelines on Data Management in Horizon 2020 [EB/OL]. [2015-08-18]. http://ec.europa.eu/research/participants/data/ref/h2020/grants_{manual}/hi/oa_{pilot}/h2020-hi-oa-data-mgt_{en}.pdf.

[3] NSF: Proposal Preparation Instructions of DMP [EB/OL]. [2015-08-18]. http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp.

[4] Research Councils UK [EB/OL]. [2015-08-18]. http://www.rcuk.ac.uk/.

[5] Wellcome Trust: Guidance for Researchers: Developing a Data Management and Sharing Plan [EB/OL]. [2015-08-18]. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm.

[6] ICPSR: Elements of a Data Management Plan [EB/OL]. [2015-08-18]. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/elements.html.

[7] BBSRC: Data Sharing Policy [EB/OL]. [2015-08-18]. http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf.

[8] Cancer Research UK: Data Sharing Guidelines [EB/OL]. [2013-08-18]. http://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy/data-sharing-guidelines.

[9] MRC: Guidance on Data Management Plans [EB/OL]. [2015-08-18]. http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/data-management-plans/.

[10] AHRC: Technical Plan [EB/OL]. [2015-08-18]. http://www.ahrc.ac.uk/funding/research/researchfundinggu

[11] ESRC: Research Data Policy [EB/OL]. [2015-08-18]. http://www.esrc.ac.uk/*{images}/Research*{{{Data}}} 4595.pdf.

[12] STFC: Guidelines on DMPs [EB/OL]. [2015-08-18]. http://www.stfc.ac.uk/funding/research-grants/data-management-plan/.

[13] STFC: Scientific Data Policy [EB/OL]. [2015-08-18]. http://www.stfc.ac.uk/Resources/pdf/STFC_{{{Scie

[14] Jones S. How to Develop a Data Management and Sharing Plan [EB/OL]. [2015-08-18]. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How%20to%20

Author Contributions

Zhang Xiaolin: Conceived the concept of a curation engine centered on the project lifecycle, participated in content analysis and organizational discussions, and approved the final manuscript.

Liu Feng: Designed the model, collected and analyzed data, and wrote the manuscript.

Received Date: 2015-09-07

Research on the Specification of Data Management Plan and Its Operational Model

Liu Feng[1],[2],[3], Zhang Xiaolin[1]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [Objective] Propose a set of detailed structure specifications of scientific data management plan and in accordance with a data curation model constructed from the operational perspective. [Methods] This paper carries on the research and the statistics on the scientific data management plan specification of the main research and management agencies in the world, and makes supplement combining with the requirement and characteristic of current scientific research data management. [Results] This paper forms the detailed structure specification of data management plan with 8 major basic elements and 39 sub-elements and constructs a data curation model taking data management plan as the core driver. [Conclusions] The detailed structure specification of data management plan may regulate and guide the activities of scientific data management completely and accurately, it can also be effectively controlled and restricted the data curation process of the whole life cycle of scientific research at the operational level.

Keywords: Data management plan; Data curation; Research lifecycle

Note: Figure translations are in progress. See original paper for figures.