Functional Requirements of Data Repository for DMP Support and CoreTrustSeal Authentication

Sun-Tae Kim*

ARTICLE INFO

Article history:
Received 13 December 2019
Revised 04 December 2019
Accepted 19 December 2019

Keywords:
Data Management Plan,
CoreTrustSeal,
Data Repository,
Functional Requirement,
DMP

ABSTRACT

For research data to be shared without legal, financial and technical barriers in the Open Science era, data repositories must have the functional requirements asked by DMP and CoreTrustSeal. In order to derive functional requirements for the data repository, this study analyzed the Data Management Plan (DMP) and CoreTrustSeal, the criteria for certification of research data repositories. Deposit, Ethics, License, Discovery, Identification, Reuse, Security, Preservation, Accessibility, Availability, and (Meta) Data Quality, commonly required by DMP and CoreTrustSeal, were derived as functional requirements that should be implemented first in implementing data repositories. Confidentiality, Integrity, Reliability, Archiving, Technical Infrastructure, Documented Storage Procedure, Organizational Infrastructure, (Meta) Data Evaluation, and Policy functions were further derived from CoreTrustSeal. The functional requirements of the data repository derived from this study may be required as a key function when developing the repository. It is also believed that it could be used as a key item to introduce repository functions to researchers for depositing data.

1. Introduction

Nowadays, it is common to use the results of scientific research to make government policy decisions. The government proves that it is a reasonable policy decision using data supporting the determined policy. Researchers have also begun to use data as a central tool for research. This is the picture of the fourth-generation research paradigm, which uses research data as a key tool for research. Already, major advanced countries such as the United States, Britain, Germany, and Australia are leading the way in which research data can be stored, managed and reused systematically. The research content of researchers is published in academic papers, and research data supporting the research is being produced, preserved and shared under the Data Management Plan (DMP). Europe's DMPOnline and the US' DMPTool platform are provided as services to support researchers' DMP writing.

* Professor, Jeonbuk National University, Republic of Korea (kim.suntae@jbnu.ac.kr)
International Journal of Knowledge Content Development & Technology, 10(1): 7-20, 2020.
http://dx.doi.org/10.5865/IJKCT.2020.10.1.007

As of August 2019, 2,385 data repositories are globally registered on Re3data.org a registry that registers worldwide research data repositories. These data repositories are also trying to gain CoreTrustSeal certification to give researchers confidence that they are stable repositories.

In line with these international trends, domestic research is also active. The definitions of 'Research Data' and 'Data Management Plan' were recently included in the 'REGULATIONS ON MANAGEMENT OF NATIONAL RESEARCH AND DEVELOPMENT PROJECTS' in Korea. In addition, new items were added for a selection of R&D tasks. Universities and Government-funded Research Institutes are paying attention to how DMP is applied, as the regulation goes into effect on Sept. 1, 2019. There is also a growing interest in the data repositories needed to realize the items required by DMP. Recently, the Korea Institute of Science and Technology Information (KISTI) is building and distributing national research data platforms and DMP platforms, and research data repositories for use by institutions are being developed and distributed.

Data repositories described in DMP should ensure data collection, systematic management, and data reuse. Therefore, data repositories must be certified to gain trust from researchers so that they can firmly establish themselves in the academic information ecosystem. Therefore, this study is designed to derive functional requirements that these data repositories should have.

2. Methodology and Limits

In this study, the purpose of the study is to derive functional requirements that data repositories must have for DMP realization and CoreTrustSeal certification. This is to derive the requirements for a repository where researchers can trust and deposit their data. For DMP analysis, DMP forms registered with DMPOnline were used for analysis. DMPOnline is the platform European researchers use to register DMP. DMPTool was excluded from this study because it is only open to researchers in the Americas. In addition, the DMP platform under development by the Korea Institute of Science and Technology Information (KISTI) was excluded because there are relatively few registered DMPs. The functional requirements of the repository through the CoreTrustSeal analysis were analyzed using 'CoreTrustSeal Certification 2017-2019'. Because the CoreTrustSeal certification is established on a three-year basis, the 2017-2019 version analyzed in this study is the latest version. The limitation of this study is that extracting the functional requirements of the data repository from the text of a document described in free text is bound to rely on subjective judgments from the researcher.

3. Literature Review

Various studies have been conducted both at home and abroad to derive functional requirements for data repositories. Kim (2018) analyzed the characteristics of research data, analyzed the repository platforms, and analyzed the community requirements and mapped and aggregated each of the functional requirements to derive functional requirements for the research data repository. 75 detailed functional requirements were derived for 13 categories. Repository Platforms for Research Data Interest Group

of the Research Data Alliance (2016) derived repository requirements based on 11 case studies for digital repository platforms. Functional requirements were derived based on a total of 13 categories, including Metadata. The University of Leeds in the U.K. undertook the RoaDMaP project to implement the school's official data store. A workgroup was created as part of the process, in order to derive functional requirements for the repository, the workgroup analyzed the requirements of stakeholders such as Funder, Depositor, End user (re-user), Data repository manager, Legal / ethics, Business manager, and IT/Systems. Quisi-Peralta et al. (2018) proposed a data repository that is automatically generated using mobile for people with disabilities and users with language and communication difficulties.

In addition to studies on functional requirements of repositories, studies on data repository description levels and subject distribution were conducted as follows. Cho and Park (2019) conducted a cluster analysis of 305 data repositories in the humanities and social sectors registered at re3data.org. Kim and Choi (2017) investigated the quality of metadata that describes the repository itself in a data repository registered at re3data.org. Meanwhile, as a prior study related to repository certification, Corrado (2019) studied trust issues in digital repositories in the context of repositories with digital preservation functions.

Looking at the above previous studies, it is judged that there exists a study case of functional requirements of the data repository. However, it was found that the researchers did not have enough research to derive the functional requirements needed to support the data repository for the contents described in the DMP. Data repositories also require authentication to become data repositories that can be trusted by researchers. Therefore, studies are also needed to derive functional requirements for repositories that meet repository certification criteria.

4. Theoretical background

4.1 DMP

Advanced countries' efforts to manage and reuse research data have long been underway. In particular, institutions that support research funds have provided policy support. Research funding agencies in major developed countries include both research records and data as eligible content for research data policies. In addition, the policy includes time limits and data management plan (hereinafter DMP), data access and sharing, long-term preservation, and monitoring items to present an environment for access to research data and research records. The support functions of these institutions include providing guidelines, data repository, data management, and service costs. Most research funding agencies provide publishing repositories. Amid such changes in the research environment, there is a growing perception that data, one of the products of research for which public funds have been injected, is also a national asset. In order to systematically collect, manage, preserve, service and reuse research data, which is a national asset, DMP is becoming common. The DMP is a document that the researchers conducting the research must submit to receive research funding and describes the types of data and metadata produced during the research, the approach and sharing

plan to the data, and the policies and archiving plan for reuse. DMPs are generally short, high-level descriptive plans that prescribe the data to be generated by a research project, how that data will be stored (securely, as required), who will have access, what documentation and metadata will be created with the data, and preservation intentions if the data are to be preserved long-term(Burnette, Williams, & Imker, 2016). Many libraries, including Edinburgh University, Purdue University, and the University of California, offer DMP consulting and writing support services. Among the DMP content are problems that researchers cannot solve on their own. Most of these problems can be solved with the help of libraries in institutions to which researchers belong.

Meanwhile, according to the "Regulations on the Management of National R&D Projects" in Korea, data management plans refer to plans for production, preservation, management and joint utilization of research data. The regulations also describe the definition of research data. According to the definition, research data is factual data produced through various experiments, observations, investigations, and analyses conducted in the course of R&D tasks, which is essential for verification of research results. According to the applicable regulations, the fidelity of research data production, preservation and management under the data management plan and the possibility of joint use can be considered when the head of the central administration evaluates R&D tasks deemed necessary for DMP application.

These DMPs contain the requirements of the funders. In particular, issues related to data storage, retention, and reuse are related to data repositories. Thus, the functions of the data repository can be extracted from the details described in the DMP. In this study, DMP items registered in DMPOnline were analyzed.

4.2 Research Data and Data Repository

ROAR has classification codes for the repository type such as Database, Demonstration, Institutional, Journal, Learning, Multi, Opendata, Other, Researchdata, Subject, Theses, and Webobservatory. This classifies the repository depending on the subject area repository, journal repository, institutional repository, learning material repository, research data repository, dissertation repository and web laboratory repository (Kim & Lee, 2014). In order to meet the DMP needs, the existing operating repositories should naturally be expanded to data repositories. Data repositories are designed to manage and serve all data produced in the course of a study, compared to repositories that manage and serve research records (papers, patents, reports, trend data, etc.) based on the research findings. Data repositories contain contents that were managed by existing repositories, including observational and observational data, experimental data, simulation data, extraction and compilation data, and so on. Observational and observational data are mainly generated in real time and are difficult to reproduce. Experimental data are mainly produced in laboratory equipment and are reproducible but can be expensive. Simulation data are produced as experimental models, and input data, such as models and metadata, have more important characteristics than result data. Extraction and compilation data are reproducible but can be expensive. Therefore, studies of data management and retention policies that apply the characteristics of various data managed by data repositories are also actively underway.

4.3 DSA & CoreTrustSeal

Researchers need reliable data repositories to implement DMP. CoreTrustSeal is a typical authentication tool for reliable data repositories. The ICSU World Data System (ICSU-WDS) and Data Seal of Approval (DSA) have created a CoreTrustSeal organization that authenticates the data repository. Later, in September 2017, Data Seal of Approval was integrated into CoreTrustSeal. Expenses for data repository authentication are required from 2018. The CoreTrustSeal certification basis is based on the Seal of Approval. Seal of Approval was developed by DANS (Data Archiving and Networked Services) to ensure that archived data will be found, understood and available in the future. The first version of Data Seal of Approval was announced at the conference in 2008. This is a rebuttal of the long-standing need for certification standards for repositories. The purpose of the DSA is to protect data without new standards, regulations, and costs, to ensure data quality and reliable data management. For this reason, data producers, digital data producers, data archive agencies, and data consumers have become interested in DSA's quality guidelines. The benefits of stakeholder groups through DSA include: First, data producers can be assured that data and related materials are preserved and reusable in a reliable manner. Second, it is possible to convince research funding agencies that data are being preserved in a reusable state and that their input results are not lost. Third, it enables data consumers to judge data repositories. Fourth, it helps data repositories efficiently archive data and distribute it. Therefore, it can be considered reasonable to use CoreTrustSeal, which is based on DSA, as an authentication tool for data repositories that researchers can trust. In this study, the contents of CoreTrustSeal were analyzed to derive the functional requirements required for the data repository.

5. Results and Discussion

5.1 DMP analysis

Table 1 is a study of DMP items registered with DMPOnline. As of 2019, DMP items from NSF, IMLS and USGS organizations registered with DMPOnline were investigated and prepared. Table 1 is a collection and analysis of preparation guidelines that can be referenced when describing DMP items. The Class column in Table 1 indicates the group name of DMP elements in DMPOnline. NSF-BIO includes DMP items specific to specific research areas, such as Pre-Cruise planning items. In the case of IMLS-SA, technical information is required to include additional software and system dependencies, documentation, and examples for programming language, platforms, software, and interoperability. USGS requires a more specific description of how data is collected, depending on the number of cases. For example, it is asked to separate the collection of data that already exists from the production of new data. It also requires a description of whether the characteristics of the data are static or dynamic and requires the presentation of data processing procedures and scientific workflows. It also requires the presentation of procedures for data quality assurance. Based on the analysis of the DMP form with the above characteristics, the functional requirements required

for data repository are as follows. First, the repository must obtain the DMP identification associated with the deposited data. This is because both when the DMP was created and when the data was created and when the data was deposited in the repository are different and the managing body is different. Second, before data is available in the repository, the repository should have procedures that clarify the responsibilities of the transfer of data. On the other hand, 'A brief description of staff/organizational roles and responsibilities for implementing a data management plan' was excluded because it is a separate issue from the repository function. Third, the repository should clearly state the type of data it accepts and allow researchers to set up the data sharing target and sharing conditions. Fourth, the repository should have policies on the management, storage, and preservation of data deposited by researchers. Fifth, the repository should allow researchers to describe the minimum quality standards for data and should have a process to check if the data conforms to the specifications set by a particular domain. Sixth, the repository should allow researchers to set up settings such as legal and ethical restrictions regarding access to data. Seventh, the repository should ensure that data can be published timely and quickly. It should also allow researchers to set an Embargo Period.

5.2 CoreTrustSeal Analysis

CoreTrustSeal is a community-based, nonprofit organization that promotes a sustainable and reliable data infrastructure. At the same time, it is also the certification granted to a trusted data repository. This study focused on certification criteria that are valid from 2017 to 2019. Certification criteria are largely divided into three categories: organizational structure, digital object management, and technology. Organizational structure categories consist of the responsibilities and scope of repositories, licenses, the permanence of access, confidentiality and ethics, organizational infrastructure, and expert guides. The digital object management category consists of data integrity and reliability, evaluation, documented storage procedures, preservation plans, data quality, workflows, data discovery, and identification, and data reuse. The technology category consists of technical infrastructure and security. The following is a description of the functional requirements of the data repository derived from the analysis of the CoreTrustSeal. Table 2 shows functional requirements derived from the analysis of DMP and CoreTrustSeal. It shows the stakeholder groups associated with implementing the requirements derived at the same time. The functional requirements in Table 2 are representative representations of key keywords derived from the analysis of DMP and CoreTrustSeal.

5.2.1 Deposit, Confidentiality, and Ethics

The deposit function requested in CoreTrustSeal is a function used by researchers. It is also the starting point at which the data is secured in the repository. The steps for the repository administrator to obtain data through batch operations or to secure resources with automated systems should also be viewed as a deposit function. Of course, DMP also requires research to describe where to deposit data generated or collected in the research process. It is essential to maintain trust in researchers who agree to deposit data that includes personal and sensitive information in the repository. Therefore,

Table 1. DMP items required by funders

Class	items	NSF OCE	NSF-AGS	NSF-AST	NSF-BIO	NSF-CHE	NSF-CISE	NSF-DMR	NSF-EAR	NSF-EHR	NSF-ENG	NSF-GEN	NSF-PHY	NSF-SBE	IMLS-RD	IMLS-DC	IMLS-SA	USGS
Project Info.	Project title	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Primary research organisation	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Funding organisation	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
DMP Basic Info.	Plan name	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	ID	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Grant number	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Principal Investigator/Researcher	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Principal Investigator/Researcher ID	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Plan data contact	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Description	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Data Collection	Data Type	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O		O
	Data Format	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O		O
	Data Volume	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O		O
Documentation and Metadata	Documentation				O			O	O			O	O		O		O	O
	Metadata	O			O			O	O			O	O		O			O
Ethics and Legal Compliance	consent for data preservation and sharing						O								O	O		
	protect the identity of participants						O								O	O		
	sensitive data						O								O	O		
	copyright														O	O		
	Intellectual Property Rights (IPR)														O	O		
Storage and Backup	Storage and Backup				O		O								O			O
	Access and Security	O			O		O	O	O			O	O		O			O
Selection and Preservation	Selection	O	O	O	O	O		O	O	O	O	O	O	O	O			O
	Preservation	O	O	O	O	O		O	O	O	O	O	O	O	O	O		O
Data Sharing	Data Sharing Method	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
	Restrictions on data sharing required		O	O	O	O	O	O	O	O	O	O	O	O				O
Responsibilities and Resources	Responsibilities	O			O		O				O			O	O		O	O
	Resources						O			O				O				O

Abbreviation representation: OCE: Division of Ocean Sciences, AGS: Atmospheric and Geospace Sciences, AST: Astronomical Sciences, BIO: Biological Sciences, CHE: Chemistry Division, CISE: Computer and Information Science and Engineering, DMR: Division of Materials Research, EAR: Earth Sciences, EHR: Education and Human Resources, ENG: Engineering, GEN: Generic, PHY: Physics, SBE: 'Social, Behavioral, Economic Sciences', IMLS-RD: IMLS Research Data, IMLS-DC: IMLS Digital Content, IMLS-SA: IMLS New Software Tools or Applications, USGS: U.S. Geological Survey

Table 2. Stakeholders' Role for Data repository

Functional Requirements	Funder	Administrator	Depositor	User	DMP	CoreTrustSeal
Deposit		O	O		O	O
Confidentiality		O				O
Ethics		O	O	O	O	O
License	O	O	O		O	O
Integrity		O				O
Reliability		O	O			O
Discovery				O	O	O
Identification				O	O	O
Reuse		O		O	O	O
security		O			O	O
Archiving		O				O
Preservation		O			O	O
Technical Infrastructure		O				O
Documented Storage Procedure		O				O
Accessibility		O	O		O	O
Availability		O			O	O
Organizational Infrastructure	O	O				O
(Meta) Data Quality		O	O		O	O
(Meta) Data Evaluation	O	O	O			O
Policy	O	O				O

the repository requires Confidentiality. Ethics is related to the information contained in the research data. The DMP recommends that the data do not include personal information that participated in the survey, information that can identify the exact location of endangered species, and information on the location of archaeological sites. CoreTrustSeal requires repositories to comply with applicable discipline-specific norms and to confirm that data collection or creation has been carried out in accordance with legal and ethical standards. In addition, special procedures for managing data at risk of disclosure should be secured by repositories, and data at risk of disclosure should be stored appropriately, limiting access and deploying under appropriate conditions. In addition, it requires the repository to provide researchers with guidance on responsible use. Therefore, Ethics is a required function in terms of data repository managers, depositors, and data re-users. Table 2 shows that Deposit and Confidentiality, and Ethics requirements are relevant to managers. It also shows that Deposit and Ethics are related to the depositor. On the other hand, from the data reuse perspective, it shows that ethics is relevant. The functional requirements of Deposit and Ethics appear to be required at the same time in DMP and CoreTrustSeal, and Confidentiality shows that only CoreTrustSeal requires it.

5.2.2 License and Integrity

License is about accessing and using data. It also relates to generally accepted regulations for proper use of data. Therefore, research funding agencies may require researchers to comply with the relevant conventions, and the data depositor must comply with them. CoreTrustSeal requires the repository to present license agreements in use and usage conditions related to distribution and sensitive data protection. DMP also requires data-related licenses to be clearly presented. License, meanwhile, is deeply related to the aforementioned Confidentiality and Ethics. For Integrity, this means that no unintended changes or damage have been made since metadata and data were complete. Data repositories should ensure the integrity of digital objects. To that end, CoreTrustSeal requires documentation and versioning strategies to be prepared. When integrity is guaranteed, it can give credibility to research data contributors or researchers who reuse data. Data integrity and reliability are deeply related to the data lifecycle within the repository. Therefore, it is necessary to develop strategies that can ensure integrity and reliability in the evaluation of data collection procedures, procedures for document archiving, and methods for data preservation methods. In particular, Integrity is a function that must be guaranteed in the data repository, and Reliability is judged to be a function that must be guaranteed from the data generation stage to the data management stage.

5.2.3 Discovery, Identification, and Reuse

For Discovery, Identification, and Reuse, this is a common requirement for DMP and CoreTrustSeal. The DMP requires researchers to publish data for public access. It also requires that data be accessible using permanent identifiers without being affected by where it is loaded. CoreTrustSeal also requires data citation services as repository functions. Reuse can be guaranteed by providing data users with context information about the data. CoreTrustSeal requires the repository to provide metadata to users to enable Reuse. Therefore, when a researcher places data into the repository, it should present the metadata items that must be entered by the researcher. The metadata required to be entered by researchers should support the format used by the community. Next, ensuring reuse is deeply related to data licensing and migration. DMP also requires licensing issues to be described in relation to data reuse. It also requires a description of which metadata standard to be used. Therefore, Reuse is a functional requirement for users who must comply with the license when using the data again and for repository managers responsible for data acquisition, preservation and service.

5.2.4 Security, Archiving, and Preservation

Security doesn't just mean data security. It also includes security for users who use the repository and the services it provides. CoreTrustSeal requires the repository to have procedures for rapid recovery when the repository is shut down. The description required by DMP is limited to where to deposit and preserve research data. However, it is clear that it is an important requirement for researchers to consider in selecting repositories to which their data will be deposited. Archiving

means collecting data as a key function of the repository and loading it in storage, mostly for preservation purposes. CoreTrustSeal recommends that you deploy an archive according to a documented workflow. Therefore, qualitative and quantitative checks on data should be conducted through clear communication with data depositors. It also requires documentation of the decision-making process for data transformation. Meanwhile, the DMP does not require researchers to describe archive-related activities. However, it requires the description of a physical archive to which the data is deposited. Eventually, researchers who search for archives to which data is deposited will find a repository that performs the archives according to documented procedures. Preservation is an activity needed to ensure continuous access to data. CoreTrustSeal requires that repositories have a continuous plan to ensure data preservation. A mid/long-term plan is required to ensure continuous availability and accessibility of data, including the retention period guaranteed for the data. This is a required function with documented archive activities for preservation. DMP also requires researchers to describe data preservation plans. However, the content related to preservation is difficult for individual researchers to describe. Therefore, it is necessary to present to researchers a preservation plan prepared by the library or data center. These functions are required as library consulting functions related to DMP.

5.2.5 Technical Infrastructure and Documented Storage Procedure

CoreTrustSeal requires that repositories work on a reliable software base. It calls for hardware and software configuration to allow permanent access and reuse of data. DMP does not require researchers to describe technical infrastructure. However, whether a researcher is a technically reliable repository or not is considered an important factor when selecting a repository to which data is deposited. Meanwhile, CoreTrustSeal requires a documented storage procedure from the repository. It requires that procedures be documented from the moment researchers deposit data to the moment it is reused. For example, data deposit procedures, policies related to the preservation of data storage devices, strategies for maintaining backup and multiple copies, and data recovery procedures are required. The DMP does not require researchers to describe documented storage procedures.

5.2.6 Accessibility and Availability, and Organizational Infrastructure

Accessibility and availability mean that the data loaded into the data repository is accessible and available for both now and in the future. CoreTrustSeal requires mid-term (3−5 years) and long-term (>5 years) continuity plans from repositories to ensure continuous access and retention of data. For example, a plan is required to prepare for situations such as the suspension of funding for a repository. DMP is also required to describe how it will provide access to research data. In order to provide access to research data, a data depositor's consent and a data access interface in the repository are required. Availability means the state of reusable data. Therefore, it is deeply related to the data preservation process in the repository. Organizational Infrastructure is about the physical resources and personnel that the repository needs to perform its tasks effectively. CoreTrustSeal requires employees with funds and expertise to perform its repository role. This

is essential to ensure long-term stability of data and sustainability of services. For these reasons, the governance structure and decision-making process of the repository are also expanded and required. Due to the nature of the demand, it is not a function required by the DMP. This is where research funding agencies should think about how to support material and human resources to foster repositories.

5.2.7 (Meta) Data Quality, (Meta) Data Evaluation, and Policy

CoreTrustSeal requires cooperation with data depositors so that data users can evaluate data and metadata quality. Especially for data repositories used by multidisciplinary communities, this is a required function because the quality cannot be assessed by data alone. This is also a function required by DMP. The DMP requires that metadata standards be presented that describe data. It also requires a description of how to maintain the quality levels required by the community to ensure data quality. Meanwhile, data and metadata evaluation is a requirement that relates to research funding agencies, repository managers, and data contributors. The DMP does not require researchers to describe matters related to the assessment. CoreTrustSeal, on the other hand, requires data and metadata to be allowed according to defined criteria or standards. This is to ensure that data is understandable to the data reuse user. Therefore, quality control check means are required to ensure completeness and understandability of data submitted with specific requirements. In addition, the repository is required to secure a process to verify whether metadata is provided or not. Policies are deeply related to the functions of research funding agencies and repositories. Research funding agencies may require researchers who receive funding to comply with policies on research data retention and publication. In order to comply with the policy, researchers will select a repository that conforms to the policy. Therefore, data repositories should have policies related to data retention and publishing and should provide consulting to researchers on data preservation and publishing required by DMP. CoreTrustSeal requires more specific policies as follows: First, it calls for policies for data collection and preservation, second, policies for data backup and copy, and third, policies for developing repository collections.

Data repository functions derived based on DMP and CoreTrustSeal are as shown in Table 2. A total of 20 functions were derived, including the Deposit function. All of the functions required by DMP are included in those derived from CoreTrustSeal. Therefore, it can be judged that the core functions that should be provided in the repository first. Deposit, Ethics, License, Discovery, Identification, Reuse, Security, Preservation, Accessibility, Availability, and (Meta) Data Quality functions extracted from DMP are judged to be functions that should be implemented primarily in the implementation of data repositories.

Table 3 compares the functional requirements of the data repository derived from this study with the findings of Kim (2018). As the table shows, all functions derived from DMP and CoreTrsutSeal include all functions derived from previous studies of Kim (2018). In addition, 13 requirements were derived from the previous study, but this study was further detailed and expanded to 20 requirements.

Table 3. Comparison of Functional Requirements for Implementing Data Repository from Previous Studies

Class (RDA IG 2016*** & Kim 2018)	Functional Requirements for DR**	CFR4DR*
Submission Ingestion and Management	Deposit	O
Data and product quality	(Meta) Data Quality	O
Metadata		
Authentication and authorization	License	O
	Security	O
Policy support	Policy	
	Ethics	O
Location	Technical Infrastructure	
preservation and sustainability	Preservation	O
	Archiving	O
	Documented Storage Procedure	
	Integrity	
	Availability	O
Publication	Discovery	O
Data access	Accessibility	O
Identifiers	Identification	O
User interface	Reuse	O
Data organization	(Meta) Data Evaluation	
	Reliability	
Integration and Interworking	Organizational Infrastructure	

* CFR4DR: Core Functional Requirements for Data Repository
** DR: Data Repository
*** RDA IG 2016: Repository Platforms for Research Data Interest Group of the Research Data Alliance (2016)

6. Conclusion

This study derived the functional requirements of a data repository for DMP implementation and CoreTrustSeal certification. Deposit, Ethics, License, Discovery, Identification, Re-use, Security, Archiving, Preservation, Accessibility, Availability, and (Meta) Data Quality have been derived as functional requirements for data repositories to support DMP implementation. Deposit, Confidentiality, Ethics, License, Integrity, Reliability, Discovery, Identification, Reuse, Security, Archiving, Preservation, Technical Infrastructure, and Documented Storage Process, Accessibility, Availability, Organizational Infrastructure, (Meta) Data Quality, (Meta) Data Evaluation, Policy as functional requirements required by CoreTrustSeal was derived. The CoreTrustSeal certification requirements were analyzed to include both DMP requirements. Therefore, the functions of Deposit, Ethics, License, Discovery, Identification, Reuse, Security, Preservation, Accessibility, Availability, and (Meta) Data Quality extracted from DMP were proposed as features that should be implemented

first in implementing data repositories. Meanwhile, the CoreTrustSeal 2020-2022 certification basis is under development. Therefore, further analysis of this is needed as a future study.

With the recent revision of 'REGULATIONS ON MANAGEMENT OF NATIONAL RESEARCH AND DEVELOPMENT PROJECTS', interest in a data repository is increasing in South Korea. Libraries should provide consulting services regarding researchers' DMP writing. Prior to this, however, data repository implementation and policy development must precede. The functional requirements of the data repository derived from this study may be required as a key function in the development of the data repository and may be used as a key item to introduce repository functions to researchers to deposit data. It's the age of data science, where data is the central tool for research. There is also an active open science movement to share and reuse research findings. Such environmental changes require changing and expanding the roles of libraries and librarians. In addition to traditional library services, they are demanding the role of data libraries and data librarians. Therefore, it is hoped that the functional requirements of the data repository presented in this study will be used to construct the data library.

References

Burnette, M. H., Williams, S. C., & Imker, H. J. (2016). From Plan to Action: Successful Data Management Plan Implementation in a Multidisciplinary Project. *Journal of eScience Librarianship, 5*(1), 6. https://doi.org/10.7191/jeslib.2016.1101

Cho, J., & Park, J. D. (2019). A Study on Analysis of Research Data Repository in Humanities and Social Sciences. *Journal of the Korean Biblia Society for library and Information Science, 30*(2), 69-87. https://doi.org/10.14699/kbiblia.2019.30.2.069

Core Certified Repositories. (n.d.). https://www.coretrustseal.org/why-certification/certified-repositories/

Core Trustworthy Data Repositories Extended Guidance. (n.d.). https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf

Corrado, E. M. (2019). Repositories, Trust, and the CoreTrustSeal. *Technical Services Quarterly, 36*(1), 61-72. https://doi.org/10.1080/07317131.2018.1532055

Kim, J., Kim, S., & Choi, S. (2019). The Functional Requirements of Core Elements for Research Data Management and Service. *Journal of the Korean Society for Library and Information Science, 53*(3), 317-344. https://doi.org/10.4275/KSLIS.2019.53.3.317

Kim, S. (2018). Functional requirements for research data repositories. *International Journal of Knowledge Content Development & Technology, 8*(1), 25-36. https://doi.org/10.5865/IJKCT.2018.8.1.025

Kim, S., & Choi, M. S. (2017). Registry Metadata Quality Assessment by the Example of re3data. org Schema. *International Journal of Knowledge Content Development & Technology, 7*(2), 41-51. https://doi.org/10.5865/IJKCT.2017.7.2.041

Kim, S., & Lee, W. (2014). Global data repository status and analysis: based on Korea, China

and Japan. *Library Hi Tech, 32*(4), 706-722. https://doi.org/10.1108/LHT-06-2014-0064

Quisi-Peralta, D., Robles-Bykbaev, V., Saquicela-Galarza, V., Bernal-Merchán, E., Suquilanda-Cuesta, P., & Lopez-Nores, M. (2018, February). Data repository of mobile applications for people with disabilities in the area of communication and language using data mining techniques. In *2018 International Conference on Electronics, Communications and Computers (CONIELECOMP)* (pp. 225-231). IEEE. https://doi.org/10.1109/CONIELECOMP.2018.8327203

Repository Platforms for Research Data Interest Group of the Research Data Alliance. (2016). Matrix of use cases and functional requirements for research data repository platforms. https://doi.org/10.15497/rda00033

RoaDMaP of Leeds University. (2013). Functional Requirements for a Data Repository. https://leedsunilibrary.files.wordpress.com/2013/06/repositoryfunctionalrequirementsv1-1_web_1_.xlsx

[About the author]

Suntae Kim is an assistant professor in the Library and Information Science Department of Jeonbuk National University. He received his PhD Degree in Library and Information Science from the Jeonbuk National University. Before his current appointment, He worked for Korea Institute of Science and Technology Information (KISTI) as Principle Researcher and he worked as a computer program developer at Linksoft which developed the NDSL and NOS. His research interests include research data management, research data platform, research data sharing, semantic web, metadata. Dr Suntae Kim can be contacted at: kim.suntae@jbnu.ac.kr.

Acknowledgement

This research was supported by “Research Base Construction Fund Support Program” funded by Jeonbuk National University in 2019