Analysis and Formalization of Data Curation Process for Open Science

by

Yasuyuki Minamiyama

Dissertation

submitted to the Department of Informatics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The Graduate University for Advanced Studies, SOKENDAI

September 2022

Evaluation Committee

Dr. Hideaki Takeda (Chair)
Professor, National Institute of Informatics/ SOKENDAI

Dr. Asanobu Kitamoto
Professor, National Institute of Informatics/ SOKENDAI

Dr. Hitoshi Okada
Associate Professor, National Institute of Informatics/ SOKENDAI

Dr. Ikki Ohmukai
Associate Professor, The University of Tokyo

Dr. Masao Takaku
Associate Professor, University of Tsukuba

Abstract

With the trend of open science, there are growing expectations for research data sharing and reuse across fields. For research data reuse, data curation is essential to make the data interpretable and reusable. In some leading fields, field-specific tasks and procedures have been developed to implement systematic data curation. However, even in those leading fields, the research data reuse is often closed within the field. For interdisciplinary reuse of research data, research data must be interpretable by researchers from different fields. The problem here is the difference in data curation, which depends on the field. This variation in the data curation by field reduces the interpretability of research data. Without a method to overcome this challenge, open science will not be realized. This study will analyze and formalize the practices of data curation in each field for interpreting the data curation process. Furthermore, we will conduct practical studies for interpreting the data curation process with a formalized method.

In Chapter 2, we provide an overview of the research background and related areas. First, we review the current status of research data sharing and reuse with its social positioning. Then we discuss the significant studies from the data curation field. Finally, we show our approach to address the issues.

In Chapter 3, we analyze and formalize the knowledge of the data curation process to provide an interdisciplinary interpretation across different fields. The granularity of tasks and procedures that constitute data curation building blocks is not formalized, so it is impossible to identify common tasks and procedures across different fields. Here we propose an approach using ontology theory and techniques to interpret data curation tasks and procedures across fields. Our proposed ontology will allow data re-users to interpret the tasks and procedures performed with a formalized method.

The formalized interpretation leads to eliminating potential risks such as improperly misusing metadata. This study contributes to building a knowledge framework for a common understanding of the data curation process in different fields.

In Chapter 4, we deal with the processes in the Data Evaluation category formalized in Chapter 3 through practical implementations. To interpret research data from a field to be independently, the data and their documentation must be reviewed from an interdisciplinary perspective and revised as necessary. In contrast, many data repositories either do not have a policy for evaluating research data and encouraging improvement, or the evaluation process is done only from the perspective of a specific field. This chapter focuses on publishing mechanisms for data papers that divert the journal peer-review and clarifies the mechanism from a process perspective. Furthermore, we conduct a technical implementation for interpreting the mechanism using the reference model for data publishing. This study is a practical interpretation of the relationship and boundaries between the existing quality assurance process and the peer-review process in the Data Evaluation category.

In Chapter 5, we deal with the processes in the Appraisal category formalized in Chapter 3, focusing on the processes related to setting conditions of use. These processes have different variations due to the differences in legal restrictions and disciplinary norms by jurisdiction. Also, the conditions of use granted by data providers are more diverse. As a result, it is difficult for data re-users to interpret the results of the process accurately. This chapter investigates the actual processing status related to setting conditions of use in different fields and clarifies the correspondence of information tied to each formalized process. Furthermore, we conduct a practical implementation for stepwise interpretation of these processing results. This study reframes the processes in the Appraisal category and complements Chapter 4 practices each other.

In Chapter 6, we discuss the results of this study and prospects. Through this study, we provided our framework for understanding data curation activities as processes from an interdisciplinary perspective. This framework allows the data curation process to be interpreted in a decoupled way of the original research context. Data re-users will be able to formally assess the increased interpretability by verifying that the processes included in the framework have been properly executed. Furthermore, we have demonstrated some practical implementations through our problem-solving approach as a stepwise formalization to the level of interpretation. We believe that as we move forward with these formalized efforts, we can improve the interpretability of research data and thereby contribute to the realization of open science.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

In recent years, with the trend of open science, there have been many efforts to share and reuse research data on the Web (Kowalczyk and Shankar, 2011). The main purpose of researchers sharing research data is to improve research efficiency, increase verifiability, and generate new knowledge by research data reuse (Piwowar, 2011; Tenopir et al., 2011; OECD, 2015b). Research data reuse is an essential act for researchers to achieve open science (Fecher et al., 2015).

Research data reuse occurs when the data provider processes the research data to make it interpretable and reusable (Peer et al., 2014), and the data re-user uses the processed research data. The set of activities that make research data interpretable and reusable is called data curation (Sun and Khoo, 2017). The sequence of data curation processes includes various tasks such as cleaning, documenting, standardizing, formatting, and associating metadata with relevant research data and codes (CASRAI, 2019a). The high-quality metadata given by these tasks and mutual understanding of the tasks makes published research data interpretable.

The practice of data curation has been developed mainly in fields such as life sciences (Venkatesan et al., 2019), earth sciences (Gray et al., 2002), and social sciences (Johnson, 2008). Through its historical efforts, tasks and procedures have been developed in these fields to implement systematic data curation (Ball, 2012). With the increasing reliable and interpretable research data, the research style of reusing other' s research data is becoming the norm (Hemphill et al., 2022). However, even in those leading fields, research data reuse is often closed within the field (Yoon, 2016). For further development of open science, it is necessary to expand this norm across fields.

For interdisciplinary reuse of research data, research data must be interpretable by researchers from different fields (UNESCO, 2021). The problem here is the difference in data curation, which depends on the field. This variation in the data curation by field reduces the interpretability of research data. Without a method to overcome this challenge, open science will not be realized.

In order to interpret the variation in the data curation by field in common, it is necessary to interpret the data curation process from an interdisciplinary perspective. This study will analyze and formalize the practices of data curation in each field for interpreting the data curation process. Furthermore, we will conduct practical studies for interpreting the data curation process with a formalized method.

1.2 Objectives

This study aims to interpret the data curation process in various fields from an interdisciplinary perspective. To achieve this purpose, we set the following two objectives:

Objective 1: Analysis and formalization of knowledge representing interdisciplinary data curation process

Objective 2: Practical studies to interpret the interdisciplinary data curation process

Objective 1 is to formalize knowledge regarding the data curation process to build a framework for unifying interpretations. The process targeted for formalization will be reviewed for interdisciplinary understanding by data curators in each field.

Objective 2 is to interpret the data curation process with the framework through

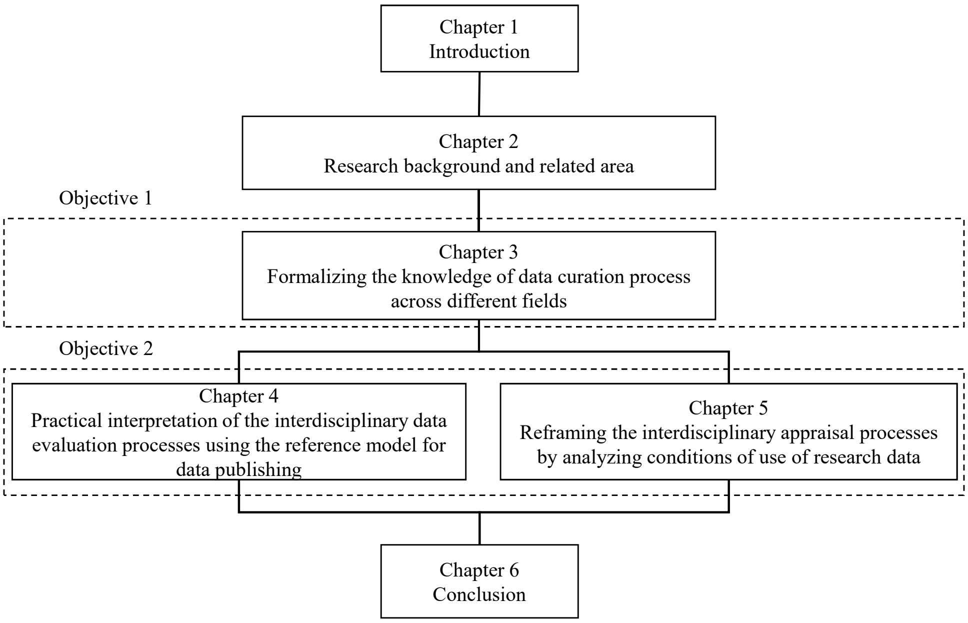

Chapter 1
Introduction
Chapter 2
Research background and related area
Objective 1
Chapter 3
Formalizing the knowledge of data curation process across different fields
Objective 2
Chapter 4
Practical interpretation of the interdisciplinary data evaluation processes using the reference model for data publishing
Chapter 5
Reframing the interdisciplinary appraisal processes by analyzing conditions of use of research data
Chapter 6
Conclusion

Figure 1.1: The structure of this thesis.

practical implementations. It also includes exploring the possibility of providing appropriate technical support.

1.3 Thesis outline

Figure 1.1 shows the structure of this thesis.

Chapter 2 Research background and related area

In Chapter 2, we provide an overview of the research background and related areas. First, we review the current status of research data sharing and reuse with its social positioning. Then we discuss the significant studies from the data curation field. Finally, we show our approach to address the issues.

Chapter 3 Formalizing the knowledge of the data curation process across different fields

In Chapter 3, we analyze and formalize the knowledge of the data curation process to provide an interdisciplinary interpretation across different fields. The granularity of tasks and procedures that constitute data curation building blocks is not formalized, so it is impossible to identify common tasks and procedures across different fields. Here we propose an approach using ontology theory and techniques to interpret data curation tasks and procedures across fields. Our proposed ontology will allow data re-users to interpret the tasks and procedures performed with a formalized method. The formalized interpretation leads to eliminating potential risks such as improperly misusing metadata. This study contributes to building a knowledge framework for a common understanding of the data curation process in different fields.

Chapter 4 Practical interpretation of the interdisciplinary data evaluation processes using the reference model for data publishing

In Chapter 4, we deal with the processes in the Data Evaluation category formalized in Chapter 3 through practical implementations. To interpret research data from a field to be independently, the data and their documentation must be reviewed from an interdisciplinary perspective and revised as necessary. In contrast, many data repositories either do not have a policy for evaluating research data and encouraging improvement, or the evaluation process is done only from the perspective of a specific field. This chapter focuses on publishing mechanisms for data papers that divert the journal peer-review and clarifies the mechanism from a process perspective. Furthermore, we conduct a technical implementation for interpreting the mechanism using the reference model for data publishing. This study is a practical interpretation of the relationship and boundaries between the existing quality assurance process and the peer-review process in the Data Evaluation category.

Chapter 5 Reframing the interdisciplinary appraisal processes by analyzing conditions of use of research data

In Chapter 5, we deal with the processes in the Appraisal category formalized in Chapter 3, focusing on the processes related to setting conditions of use. These processes have different variations due to the differences in legal restrictions and disciplinary norms by jurisdiction. Also, the conditions of use granted by data providers are more diverse. As a result, it is difficult for data re-users to interpret the results of the process accurately. This chapter investigates the actual processing status related to setting conditions of use in different fields and clarifies the correspondence of information tied to each formalized process. Furthermore, we conduct a practical implementation for stepwise interpretation of these processing results. This study reframes the processes in the Appraisal category and complements Chapter 4 practices each other.

Chapter 6 Conclusion

In Chapter 6, we discuss the results of this study and prospects.

2
Research background and related area

In Chapter 2, we provide an overview of the research background and related areas. First, we review the current status of research data sharing and reuse, which serves as the research background, and its social positioning. Then we discuss the significant studies from the data curation field. Finally, we show our approach to address the issues.

2.1 Definition

In this Section, we define “Researcher,” “Data provider,” and “Data user” in this thesis.
“Researcher” is a person who conducts research using research data. Both a data provider and a data user are included.
“Data provider” is either a person or an institution with research data to provide a data user.
“Data user” is also a person or an institution that uses research data. It is sometimes

described as a "data re-user" to emphasize the use in different fields.

2.2 Research background

2.2.1 Research data

The term research data is defined as "Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results" (CASRAI, 2019b). It includes not only observation data and experimental data, which are generally easy to imagine, but also digitized data of historical documents, data of documentary photographs and images, text data of reprints, and so on. The term research data is used diverse ways in different fields: such as intermediate data, samples, and specimens may also be described as research data.

Among the various types of research data, research data to validate research findings and results is essential to ensure reproducibility. Research data validating research findings and results are also called evidence data; the evidence data should be stored in the same manner as academic papers as it constitutes part of the research. In recent years, some academic publishers have established data availability policies and made registering evidence data in data repositories a condition for starting the peer-review (CHORUS, 2021).

Research data treated as evidence in the research process are generally produced through the life of research projects. In fields such as life sciences (Koso, 2013) and earth sciences/astronomy (Murayama and Hayashi, 2014), where data acquisition costs are high, there has been the promotion of systematic collection and data sharing using large-scale data repositories. Such project-based data are not only used as primary sources for research but also as archives of research activities.

2.2.2 Research data sharing

Research data sharing is broadly defined as "providing access for use and reuse of data" (Tenopir et al., 2011). There are two possible methods of research data sharing: private data sharing among researchers and public data sharing through journal supplements and/or repositories (Borgman, 2015). Private data sharing among researchers are reported to be more common than public data sharing through journal supplements and/or repositories (Faniel and Jacobsen, 2010; Wallis et al., 2013). But having said that, public data sharing has an advantage over private data sharing in terms of long-term access and scalability (OECD, 2017). For this reason, more expectations have been placed on public data sharing.

Research data sharing is also perceived positively by the research community. In a recent international survey, 86.7% to 89.6% said they are willing to share data with broader research groups, although there are differences by field (Tenopir et al., 2020). Relatedly, there has been many discussion about research data sharing policies. One of the most well-known community norms for research data sharing is the FAIR principles (Wilkinson et al., 2016), which stands for Findable, Accessible, Interoperable, and Reusable. The FAIR Principles are expanding internationally, as exemplified by the EU's Horizon 2020 programme, which explicitly states that data management will be aligned with the FAIR principles (Commission, 2016). That being said, the FAIR principles are somewhat abstract and subject to diverse interpretations. Therefore, efforts to measure the FAIR-compliance degree are underway internationally, including the development of a FAIR maturity model (Bahim et al., 2020) and services to evaluate the degree of FAIR (Wilkinson et al., 2019; Clarke et al., 2019).

2.2.3 Research data reuse

Research data reuse occurs when the data provider processes the research data to make it interpretable and reusable (Peer et al., 2014), and the data re-user uses the processed research data. When research data reuse, researchers evaluate information such as relevance to their own research, trustworthiness, and reproducibility of research data (Faniel and Jacobsen, 2010). However, the information needed for evaluation is often tacit knowledge; researchers contact data providers as needed to discuss how the data will be created, cleaned, processed, analyzed, and reported (Zimmerman, 2008; Wallis et al., 2013). Prior research also suggests that some contextual information is essential to most data re-users, including information about the physical, technological, and social environment in which the data was collected (Baker and Yarmey, 2009; Chin and Lansing, 2004).

Despite the high level of consensus regarding research data sharing, there has not been much progress in research data reuse. The previous research data sharing survey introduced in Section 2.2.2 also asked about the experience of research data reuse generated by others. According to this survey, the percentage of researchers who regularly reuse research data varies widely from 13.6% to 50.4% (Tenopir et al., 2020).

There have been several empirical studies on research data reuse in each field: engineering (Howard et al., 2010), astronomy (Sands et al., 2012), cancer epidemiology (Rolland and Lee, 2013), archaeology (Faniel et al., 2013), social sciences (Faniel et al., 2016; Yoon, 2014) in addition to the examples already mentioned. Through these studies, the difficulty of documenting context information is highlighted as challenge when research data reuse (Yoon, 2016).

2.2.4 Increasing social significance for research data sharing and reuse

Traditionally, research data on the basis of research papers have been shared to the extent necessary for verification, either as figures and tables or supplementary information. Efforts to share broader research data, such as project-based data have been limited to leading fields mentioned above. However, with the expansion of international collaborative research and the trend of open science, there are growing expectations for research data sharing and reuse across fields (Kowalczyk and Shankar, 2011). Starting with the G8 Science Ministers Statement in 2013 (G8, 2013), a full-scale activities for open research data have begun. In a recent example, UNESCO defines the ideal way of open research data as “available in a timely and user-friendly, human-and machine-readable and actionable format, in accordance with principles of good data governance and stewardship, notably the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, supported by regular curation and maintenance “ (UNESCO, 2021). Even in Japan, the Cabinet Office published the report "Promoting Open Science in Japan -Opening up a new era for the advancement of science-" (Cabinet Office,

2015) in 2015. This report made the keyword "open science" widely known in Japan. Moreover, the 5th Science and Technology Basic Plan includes promoting open science as a fundamental national stance (Cabinet Office, 2016). This national policy has been continued by the recent 6th Science, Technology, and Innovation Basic Plan (Cabinet Office, 2021). The research data sharing and reuse across fields are not only advancing research in the academic community but are also beginning to have strong social significance.

2.3 Related area

2.3.1 Data curation

Research data sharing and reuse are positioned as part of research activities. Since research data are collected, organized, and stored according to the interests of the researchers or research group, research data are deeply embedded in the research context. A knowledge gap exists in interpreting research data between the research group and third parties outside the community. To expand the possibilities research data reuse, it is necessary to make the data interpretable in a detached way from the research context.

Efforts to ensure the long-term continuous access and reusing digital information have been practiced mainly in the field of data curation. Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities and policies enable data discovery and retrieval, maintain data quality and add value, and provide for re-use over time (of Illinois Urbana-Champaign, 2017).

2.3.2 Data curation activities, actions, and processes

The term data curation has two origins: as a subset of digital curation in information science and preprocessing in other scientific databases. In the former context, digital curation as a discipline was established around the 1990s and has its origins in archive/records management and library science (Higgins, 2011). Data curation activities include selection and evaluation by creators and archivists and developing digital repositories as a technical foundation (Lee and Tibbo, 2007). In the latter, data curation is an older concept describing the process of selecting, normalizing, annotating, and integrating data from journals, reports, or other databases (Buneman et al., 2006). In fields such as astronomy, genomics, ecology, and economics that use data-driven methods, data curation has long been a specialized skill (Gray et al., 2002).

Despite these differences in origin, the two have been treated as almost identical concepts (Ball, 2010). According to the early definition (Lord and Macdonald, 2003), curation is defined as follows:

Curation: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.

Thus, the latter process is incorporated into one of the activities as a "Higher levels of curation. “ In a recent study, the components for expressing data curation activities are organized into four categories: work, action, curator, and repository (Johnston et al., 2016). The term ’action’ is used as a concept that includes processes directed toward data preprocessing. It can be said that the terms ’action’ and ’process’ are interpreted in an integrated manner with different origins.

2.3.3 Theoretical model

Data curation activities are modeled in various fields; two representative theoretical models that support the field are the Reference Model for an Open Archival Information System (OAIS reference model) and the DCC curation lifecycle model. These two models provided a specialized language for the field, complementing each other.

OAIS reference model

The OAIS reference model was developed for use in facilitating a broad, discipline independent, consensus on the requirements for an archive or repository to provide long-term, preservation of digital information (Consultative Committee for Space

Data Systems, nd). This model describes the technical workflow for digital preservation and "establishes a common framework of terms and concepts" (Consultative Committee for Space Data Systems, 2012).

The OAIS reference model adopts a mechanism for storing multiple pieces of information placed around the content itself in a single package. This package information is called the Information Package, and consists of two major components: Content Information and Preservation Descriptive Information. OAIS approach is considered valid for preventing loss of meaning by identifying and preserving the information contained in both components.

Originally, the OAIS reference model was developed based on the space physics field practices. However, its usefulness was found early on. After a draft of the OAIS reference model became available, the CEDARS project was conducted on its applicability to the digital curation field (Jones, 2009). After the initial validation is complete, the OAIS reference model was widely recommended, adopted, and critiqued as a theoretical construct in the digital curation field. OAIS reference model is now an ISO standard (ISO, 2012), with version 2 being the latest as of May 2022.

DCC curation lifecycle model

The DCC curation lifecycle model was developed as a training tool to help curators understand the processes involved in successful curation and develop curation and preservation methodologies for their organizations (Higgins, 2008). This lifecycle model is unique because it provides widely used active verbs to describe the set of activities and the sequence that needs to be performed (Higgins, 2018). DCC curation lifecycle model was developed by reviewing previous lifecycle models; its functions, roles, and responsibilities appear to be strongly influenced by the OAIS reference model (Higgins, 2008; Lord and Macdonald, 2003). This conceptual model played a central role in driving data curation activities across multiple fields (Tibbo, 2012). In 2020, ver. 2 is being developed, which reviews issues such as multidisciplinary practices, the emergence of big data, and AI to address changes in social conditions after the development of ver. 1 (Choudhury et al., 2020).

2.4 Gaps and Challenges

Data curation's definition and the lifecycle are gradually expanded in response to the digital information revolution of the last two decades. Despite sharing these foundations, data curation activities in each field have developed independently. The problem here is the difference in data curation activities, especially at the level of actions or processes. For example, manual data cleaning and related tasks are often tacit knowledge and not documented in data curation records (Claerbout, 2010). Even if they were recorded, the granularity of the recorded information varies widely among the fields (Mayernik et al., 2013). Moreover, even if the granularity of recorded information is partially the same, identification is often difficult due to different representations of tasks and procedures (Borgman, 2007). This variation in the data curation actions or processes by field reduces the interpretability of research data activities in different fields. Without a method to overcome this gap, interdisciplinary reuse of research data will not be promoted.

2.5 Approach

To overcome the differences in data curation activities in different fields, it is necessary to unify interpretations regarding data curation actions or processes. This study aims to provide a framework for understanding data curation activities as processes from an interdisciplinary perspective. Data re-users will be able to formally assess the increased interpretability by verifying that the processes included in the framework have been properly executed. To build an interdisciplinary data curation process framework, this study will adopt a stepwise formalization approach to the level of interpretation.

For objective 1, we adopt a knowledge engineering approach. Methodologies for clarifying and systematically expressing certain knowledge have been studied mainly in the knowledge engineering field. The theory and practice of the knowledge engineering field are also likely to help formalize the data curation process framework.

For objective 2, we adopt a problem-solving type approach. The formalized framework needs to be elaborated based on practice. By focusing on the differences from an interdisciplinary perspective, we can analyze the data curation process in more detail. Also, a practical implementation may lead to the discovery of areas for improvement and the interpretation of the entire process. By interpreting with a formalized data curation process, it will be possible to provide appropriate technical support.

3

Formalizing the knowledge of data curation process across different fields

In Chapter 3, we analyze and formalize the knowledge of the data curation process to provide an interdisciplinary interpretation across different fields. The granularity of tasks and procedures that constitute data curation building blocks is not formalized, so it is impossible to identify common tasks and procedures across different fields. Here we propose an approach using ontology theory and techniques to interpret data curation tasks and procedures across fields. Our proposed ontology will allow data re-users to interpret the tasks and procedures performed with a formalized method. The formalized interpretation leads to eliminating potential risks such as improperly misusing metadata. This study contributes to building a knowledge framework for a common understanding of the data curation process in different fields.

3.1 Introduction

As outlined in Chapters 1 and 2, the differences in data curation tasks and procedures across different fields are a significant obstacle to realizing research data reuse. In particular, the different granularity of the process results in a significant reduction in interpretability. Without having a method for interpreting the data curation process across fields, it will be difficult to put it into practice on open science.

In order to interpret the tasks and procedures performed in different fields at the same granularity, it is necessary to identify the tasks and procedures in an interdisciplinary method as a knowledge. Methodologies for clarifying and systematically expressing certain knowledge have been studied mainly in the knowledge engineering field. Among them, ontology theory has been established and widely supported for constructing a conceptual system of knowledge (Uschold and Gruninger, 1996). Ontology theory and techniques has a possibility for interdisciplinary understanding for structural knowledge sharing of the data curation tasks and procedures.

This study investigates the practices of data curation conducted in each field. Then, to formalize the data curation tasks and procedures in different fields, we propose an approach using ontology theory and techniques that are considered helpful for structural knowledge representation. This study contributes to building a knowledge framework for a common understanding of the data curation process in different fields.

3.2 Literature Review

The data curation is commonly described with a research data lifecycle model (Kowalczyk, 2018). In a research data lifecycle model, the decisions involved in a set of data curation are divided into abstracted steps (Wallis et al., 2008). By performing data curation according to a lifecycle model, the data provider can perform each data curation task and procedure with accuracy and the data re-user can understand in detail the methodology and workflow used (Ball, 2012).

Two frameworks, knowledge creation and knowledge transfer, are presented as perspectives to better understand the data curation that takes place at each stage of the life cycle model (Humphrey, 2006). Regardless of the theoretical framework, the actual model is a mixture of both. Table 3.1 shows an example of the fields and steps involved in a representative research data lifecycle (ICPSR, 2021; Faundeen et al., 2014; Oostdijk et al., 2013; DataONE, 2013; Chao et al., 2014; Johnston et al., 2016; Higgins, 2008; Griffin et al., 2018; UK Data Archive, nda).

Table 3.1: List of data curation activities by field.

Name of Institutions / Communities	CLARIN-NL	Data Curation Network	DataONE	Digital Curation Centre	DPCVocab	EMBL Australia Bioinformatics Resource	ICPSR	UK Data Archive	U.S. Geological Survey
Fields	Humanities / Linguistics	Multiple	Earth Sciences	Multiple	Earth sciences/Life sciences	Life sciences	Social sciences	Social sciences	Earth Sciences
Steps	A: Identification and assessment B: Development of a curation plan C: Curation D: Validation E: Archiving	Ingest Appraise / Accept Curate Access Preserve	Plan Collect Assure Describe Preserve Discover Integrate Analyze	Conceptualise Create or receive Appraise / Select Dispose Ingest Preservation action Store Access, use and reuse Transform	Ingest Representation Provenance management Systems management Data storage Policies Preservation Public access provision	collecting integrating processing analyzing storing sharing publishing finding	Proposal development and data management plan Project start-up Data collection and file creation Data analysis Preparing data for sharing Depositing data	Transfer of data Assigning processing standard Data processing Documentation processing Metadata creation Additional user information Publishing data Delivering data Preserving data	Plan Acquire Process Analyze Preserve Publish/ Share

The "Steps" row contains the steps defined by each organization, starting from the top. The steps defined by each field differ in terms of granularity. It is not easy to standardize decisions at each step throughout the life cycle of research data (Borgman, 2007).

Each step in the life cycle consists of multiple tasks and procedures necessary to make decisions. The tasks and procedures included in each field are more diverse than the steps themselves, and there is no comprehensive list of tasks and procedures performed in data curation across fields. In one of the few efforts to formalize definitions of tasks and procedures across fields, the Data Curation Network has drafted a glossary of terms to be used in a survey of important data curation activities in the U.S. (Johnston et al., 2016). This glossary is based on the existing glossary provided by the Digital Curation Centre (DCC), Society of American Archivists (SAA), CASRAI, RDA Data Foundation and Terminology Group, Digital Preservation Coalition (DPC), RDC (Research Data Canada), ICPSR, and practices in U.S. university libraries. Such efforts can be evaluated as potentially helpful in capturing the data curation tasks and procedures at the level of activities and supporting knowledge sharing. However, there still some issues: Some definitions do not distinguish between an 'action' performed by data curators and a 'process' performed by appropriate tools or repositories. An ambiguous expression may lead to misinterpretation in different fields where data curation activity is performed differently.

3.3 Research questions

This study aims to interpret the data curation activities in an interdisciplinary perspective and formalize the data curation activities in different fields. To achieve this objective, we set the following research questions (RQ):

RQ1: What are the commonalities and differences in the data curation tasks and procedures being performed in different fields?

RQ2: How to formalize the data curation activity’ s structure?

3.4 Preliminary analysis and survey

To clarify the commonalities and differences in the data curation tasks and procedures being performed in different fields, we conducted a preliminary analysis and survey in the following steps.

3.4.1 Analysis of existing data curation vocabulary

To clarify the commonalities and differences in data curation tasks and procedures, we need a working framework that can be used as a yardstick for different fields. As observed in Section 3.2, the Data Curation Network defines 47 vocabularies for the most important data curation activities derived from multiple lexical analyses. These vocabularies have been used in various fields of investigation (Hudson-Vitale et al., 2017) and are highly comprehensive; we have chosen to use the Data Curation Network vocabulary as our working framework for these reasons. Firstly, we analyzed the vocabularies using the IPO (Input - Process - Output) model to interpret from a process perspective. Table 3.2 shows a list of the 47 vocabulary sets subjected to analysis and the control structure expressed at the definition level.

Table 3.2: Results of Input – Process - Output analysis of data curation activity vocabularies.

No	Activity	Input information	Process	Output information	Control structures
1	Authentication	Data depositor identity information	Authenticate the identity of data depositors	Data depositor's identity authentication results	Sequential
2	Chain of custody	Data files	Generate data file provenance information	Data file provenance information	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
3	Deposit agreement	Deposit agreement application information	Verify that deposited agreement file is fit for data repository's policies and conditions	Verification results of deposited agreement file	Sequential
4	Documentation	Information describing any necessary information to use and understand the data	Generate all information describing any necessary information to use and understand the data	Data document file	Sequential
5	File Validation	Data files	Generate and verify checksums for data files	Checksum verification result of the data files	Sequential
			Verify the data file format	File Format Verification Results	
6	Metadata	Information about a data set that is structured for purposes of search and retrieval	Generate necessary information about a data set that is structured for purposes of search and retrieval	Metadata file for purposes of search and retrieval	Sequential
7	Rights management	Data document file	Verify that retention and copyright rights inherent in data files are consistent with policies and conditions for access and reuse	Verification results on data file ownership and copyright	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
8	Risk management	Data files/Data document file	Verify that external constraints contained in data files are consistent with policies and conditions	Verification results of external constraints contained in the data files	Sequential
9	Selection	Verification results of deposit agreement/file format/data file ownership and copyright/external constraints contained in the data files	Verify that the results of the various verifications conform to the collection policy of the repository	Results of acceptance/rejection decision	Sequential
10	Arrangement and description	Data files	Re-organize data files according to standards and policies set by the repository	Data files (re-organized)	Sequential
11	Code review	Computer code	Verify the computer code	Verification results of the computer code	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
12	Contextualize	Data document file/Metadata file for purposes of search and retrieval	Generate link information related to data files	Link information related to data files	Sequential
13	Conversion (Analog)	Analog data	Convert information into machine-readable format	Data files (converted into machine-readable format)	Sequential
14	Curation log	Execution results of the data curation process and executor information	Record changes made to the data and executor information during the data curation process	Information that records the execution results of the data curation process and executor information	Sequential
15	Data cleaning	Data files	Detect and fix (or remove) defects and errors in data files	Data files (cleaned)	Sequential
16	Deidentification	Data files	"Redact or remove personally identifiable or protected information (e.g., sensitive geographic locations) contained in data files"	Data files (de-identificated)	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
17	File format transformations	Data files	"Transform files into open, non-proprietary file formats"	Data files (transformatted)	Sequential
18	Transcoding	Data files	Encode audio/video files in ways that optimize reuse and long-term preservation actions	Data files (encoded)	Sequential
19	File inventory or manifest	Data files	"Verify the number of data files, file types (extensions), and file sizes periodically"	Verification results of data files	Sequential
20	File renaming	Data files	Rename data files	Data files (renamed)	Sequential
21	Indexing	Data document file/Metadata file for purposes of search and retrieval	Crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability	Metadata files that conform to the repository's standard format	Sequential
22	Interoperability	Data files	Format the data using a disciplinary standard	Data files (formatted)	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
23	Peer-review	Data files/Data document file/Computer code	Validation of data files/data document file/computer code according to discipline-specific criteria by peers	Validation results of data files/data document file/computer code by peers	Sequential
24	Persistent Identifier	Data files/Metadata files that conform to the repository's standard format	Generate persistent identifier for data files	Persistent identifier for data files	Sequential
			Set up redirection when necessary	Redirect URL for data files	
25	Quality assurance	Data files/Data document file/Computer code	Validate data files/data document file/computer code according to the standards set by the repository	Validation results of data files/data document file/computer code	Sequential
26	Restructure	Data files	Organize and/or reformate poorly structured data files	Data files (restructured)	Sequential
27	Software registry	Data document file/Metadata file for purposes of search and retrieval	Maintain copies of modern and obsolete versions of software (and any relevant code libraries)	Copies of modern and obsolete versions of software (and any relevant code libraries)	Occasional

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
28	Contact information	Data document file/Metadata file for purposes of search and retrieval	Generate contact information for the data depositor and/or contact person	Contact information for the data depositor and/or contact person	Occasional
			Update contact information for the data depositor and/or contact person	Latest contact information for the data depositor and/or contact person	
29	Data citation	Metadata files that conform to the repository's standard format	Display of a recommended bibliographic citation	Recommended bibliographic citation text	Sequential
30	Data visualization	Data files/Data document file	Generate visualized data	Visualized data	Sequential
31	Discovery Services	Information on applying for connection to the discovery services/Metadata files that conform to the repository's standard format	Connect external discovery services	Discovery Service connection results	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
32	File download	Identifying information of authorized third parties/Metadata files that conform to the repository's standard format	Generate access URLs to data files by authorized third parties	Access URLs to data files by authorized third parties	Sequential
33	Full-text indexing	Data files	Generate text inherent in data file in search-engine-optimized formats	Full text information of the data files	Sequential
34	Metadata brokerage	Information on harvesting requests for metadata search and discovery services/Metadata files that conform to the repository's standard format	Set harvesting requests for metadata search and discovery services	Results of harvesting settings for metadata search and discovery services	Sequential

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
35	Restricted access	Access permission information/Access URLs to data files by authorized third parties	Set access permissions for data files based on access permission information	Access URLs to data files by authorized third parties restricted by access authority information	Sequential
36	Embargo	Embargo period information/Access URLs to data files by authorized third parties	Set an appropriate embargo period	Access URLs to data files with the embargo period set	Sequential
37	Terms of use	Metadata files that conform to the repository's standard format	Display information about the requirements or conditions for use provided to the end user of the data files	Information on the requirements or conditions for use of data files	Sequential
38	Use analytics	Data files/Data document file/Metadata files that conform to the repository's standard format	"Generate information on the frequency of data views, requests, and downloads"	Various usage information about data files	Occasional

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
			Generate reuse metrics information such as data citations and impact measures for the data over time		
39	Cease data curation	Information on data file storage and disposal plans	Plan for any contingencies that will ultimately terminate access to the data	Data Storage and Disposal Policy	Occasional
40	Migration	Data files	Transform obsolete file formats to new formats	Data files (migrated)	Occasional
41	Emulation	Copies of current versions of software (and any relevant code libraries)	Store and/or provide software to use the data files available in legacy systems	Software for emulation	Occasional
42	Secure storage	Data files	Back up data files on a regular basis	Backup data files	Occasional
43	File audit	Data files	Verify the digital integrity of data files	Verification results of digital integrity of data files	Occasional

Continued on next page

Continued from previous page

No	Activity	Input information	Process	Output information	Control structures
44	Repository certification	A set of information about repository certification	Verify the technical and administrative capabilities of the repository by a trusted third-party accreditation body	Trusted third-party review results for repositories	Occasional
45	Succession planning	Information about the repository's long-term management plan	Develop a succession plan for the repository	Succession plan for the repository	Occasional
46	Technology monitoring and Refresh	Technical information about repository	Validate the performance of the repository against the latest technical requirements	Verification results of technical information	Occasional
47	Versioning	Data files	Generate version information for data files	Version information for data files	Occasional

End of table

Next, we classified the control structure of the vocabulary into two categories based on the pairs of input and output information extracted from each vocabulary: Sequential processing, in which the output information of activity becomes the input information of a different activity (35 vocabularies); and occasional processing, in which activities are carried out independently from the time series (12 vocabularies). This classification is consistent with existing models (Higgins, 2008) so we judged it to be appropriate as a working framework. However, the following three points should be noted:

1) Lack of vocabulary corresponding to the output information Some of the “generation” activities corresponding to the output information are not defined. For example, several activities have “data files” as input information, such as "Chain of Custody" or "File Validation," but the vocabulary for activities that output data files is not defined.

2) Lack of a vocabulary with different hierarchies There are parallel and sequential processes that require multiple input information for some output information. However, some activities that aggregate multiple input information are not exist. For example, activities that have data files as input information ("Arrangement and Description," "Conversion," "Data Cleaning," "Data Visualization," "De-identification," "File Format Transformation," "File Renaming," and "Interoperability") are a series of activities that aggregate these activities to create an individual processed data file. However, “File Download" targets the processed data file that aggregates a series of these activities.

3) No staffing/software information is included Each vocabulary does not include staffing information, so it is difficult to know the roles required to perform these activities. Also, some vocabularies are assumed to process by a repository software, which may make a difference depending on the software implemented.

3.4.2 Field survey

We conducted a field survey for several organizations conducting data curation activities in Japan. The purpose of the survey was to evaluate the actual status of data curation in each field using the working framework created in 3.4.1. As a supplement, we also intended to verify the validity of the working framework. First, we conducted interviews with the data curators at each organization. Table 3.3 shows an overview of the surveyed repositories.

Table 3.3: List of surveyed repositories.

Organization name	Repository name	Name abbreviation	Repository type	Field	Repository Description
The Center for Global Environmental Research, Earth System Division, National Institute for Environmental Studies	Global Environmental Database	GED	Institutional	Global environmental issues	The Center for Global Environmental Research (CGER) at the National Institute for Environmental Studies (NIES) has created a Global Environmental Database (GED), which comprises data and research results collected and compiled from natural and social sciences. The GED serves as a fundamental database related to global environmental problems with an emphasis on global warming and climate change.
Center for Statistics and Information, Rikkyo University	Rikkyo University Data Archive	RUDA	Institutional	Social sciences	Rikkyo University Data Archive "RUDA" aims to collect, organize, and store social survey data which are valuable public assets, and make them widely available for research purposes such as academic secondary analysis and educational use in classes.

Continued on next page

Continued from previous page

Organization name	Repository name	Name abbreviation	Repository type	Field	Repository Description
Japan Agency for Marine-Earth Science and Technology	Data and Sample Research System for Whole Cruise Information	DARWIN	Institutional	Marine-earth science	On the "Data and Sample Research System for Whole Cruise Information (DARWIN)" the Japan Agency for Marine-Earth Sciences (JAMSTEC) disseminates information for data, rock samples, and sediment core samples obtained by its research vessels and submersibles, and links to related databases.
Japan Science and Technology Agency National Bioscience Database Center	Life Science Database Archive	NBDC archive	Institutional	Life science	The Life Science Database Archive maintains and stores the datasets generated by life scientists in Japan in a long-term and stable state as national public goods. The Archive makes it easier for many people to search datasets by metadata (description of datasets) in a unified format, and to access and download the datasets with clear terms of use (see here for detailed descriptions).

Continued on next page

Continued from previous page

Organization name	Repository name	Name abbreviation	Repository type	Field	Repository Description
National Museum of Japanese History	Knowledgebase of Historical Resources in Institutes	khirin	Institutional	Japanese history	"khirin (https://khirin-ld.rekihaku.ac.jp)" is the information infrastructure system that has been developed by the National Museum of Japanese History. "khirin" is an attempt to provide access to historical materials held by universities and museums on their networks as well as to offer data in a stable and sustainable manner in collaboration with the Japan Search.
National Institute for Materials Science	Materials Data Repository	MDR	Institutional	Materials science	MDR : Materials Data Repository is a data repository that hosts materials research data and publications. Discover various data and publications using metadata tailored for materials. MDR is operated by the National Institute for Materials Science (NIMS), Japan.

Continued on next page

Continued from previous page

Organization name	Repository name	Name abbreviation	Repository type	Field	Repository Description
National Museum of Ethnology	Digital Picture Library for Area Studies	DiPLAS	Project	Ethnology	The purpose of this project is to support the representatives of Grant-in-Aid for Scientific Research projects conducting research in various regions of the world (including Japan), and to contribute to the research advancement by promoting the digitization and creating photographic materials database.
The Research Organization of Information and Systems, National Institute of Polar Research; Tohoku University; Nagoya University; Kyoto University; Kyushu University	Inter-university Upper atmosphere Global Observation NETwork	IUGONET	Project	Upper atmospheric physics	"We have three action plan in the second term (FY2015-) as follows: To provide the infrastructure and opportunity of the upper atmospheric research for users, in particular, in emerging countries. To provide our products and nowhow for other fields and nurture human resources who can develop future database and utilize it. To promote the use of various data in a wide range of fields and support the advanced integration science.

End of table

In selecting interviewees, we collected as various fields of practice as possible. On this basis, we limited our interviewees who can provide the following verification method: some form of documentation and/or the data curator's review. As a result, we conducted these interviews with people committing these repositories; four institutional repositories, i.e., Global Environmental Database (GED), Data and Sample Research System for Whole Cruise Information (DARWIN), Knowledgebase of Historical Resources in Institutes (khirin), and Materials Data Repository (MDR)) and two project-based repositories, i.e., Digital Picture Library for Area Studies (DiPLAS) and Inter-university Upper atmosphere Global Observation NETwork (IUGONET) from August to November 2020. We conducted additional interviews with those committing two institutional repositories, i.e., Rikkyo University Data Archive (RUDA) and Life Science Database Archive (NBDC archive) in August 2021. Each repository adopts various data curation models based on the nature and characteristics of the research data in each field. By comparing the models through an abstracted process, it is possible to extract commonalities and differences in structure. Each interview survey took about 1.5 to 2 hours. We used a topic guide to share the specific phase of data curation activities with the interviewee. In the topic guide, we set nine questions referred to the previous study categories (Johnston, 2017). The interview results were assigned to our working framework under the authors' responsibility and checked by each interviewer. The topic guide template used for the interviews is shown in Appendix 1.

Next, we read and referred each organization' s data curation process manuals and related documents for the rationale for the activities to ensure consistency with the interview results. We mapped the specific description of the activities and the data curators' information onto the working framework for those activities for which we were able to identify a description of the rationale for the activities. The description of the rationale for the activities is shown in Appendix 2.

We can learn from the interview results concerning commonalities and differences among data curation activities in different organizations and fields. Firstly, we observed how much data curation activities can be interpreted in the same way. Figure 3.1 shows the implementation rates of data curation processes in eight repositories. We calculated the implementation rates using the following procedure: First, we classified each data curation process mapped to the working framework into three levels: 1. Implemented, 2. Partially implemented, and 3. Not implemented. The processes classified as "2.

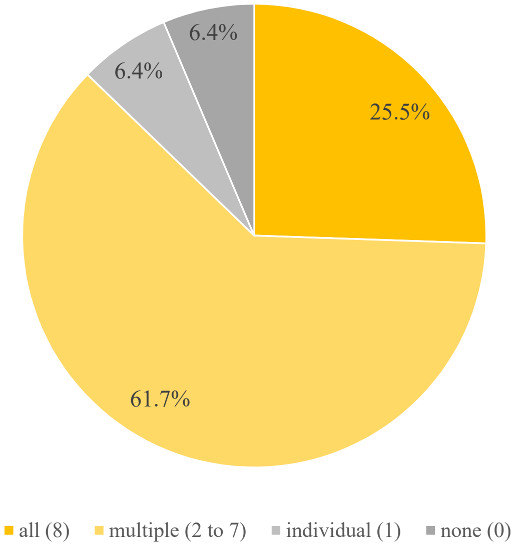


Figure 3.1: Implementation rates of data curation processes in eight repositories.

Partially implemented" were mainly found when the vocabulary included multiple processes such as "generating and verifying checksums of data files" and "verifying file formats," as in "File Validation." Next, we aggregated the implementation number of organization by each activity. We counted "2. Partially implemented" as one organization. Finally, we classified the implementation number by four categories (all / multiple / individual / none) from the perspective of interpretability. As a result, we found that about 87.2% of the processes are interpretable across multiple fields. Among them, about a quarter of the processes were found to be fully interpretable across all fields.

Secondly, we observed the variety of staffing. Table 3.4 shows an overview related to the staffing of each repository.

Table 3.4: List of roles and number of appearances in eight repositories.

Repository name (abbreviated)	Roles	Number of appearances
khirin	Researcher	4

Continued on next page

Continued from previous page

Repository name (abbreviated)	Roles	Number of appearances
	Related committee	2
	Center for Integrated Studies of Cultural and Research Resources	27
	Photographer	2
	System administrator	1
	Department of Rekihaku museum	6
	Department of internal database	10
	External organization	1
DiPLAS	Researcher	2
	Technical staff	10
	System administrator	15
	Data provider	1
	Project staff	8
	Digitization support staff	1
	Operation support staff	1
	Graduate students	1
	Review board	1
Materials Data Repository	Researcher	6
	Data system group	14
	Data service team	13
	System administration division	1
DARWIN	Researcher	9
	Data Management group	42
	Technician	9
	Navigation planning department	2
GED	Data provider	14
	Data curator	29
	Technical support staff	1

Continued on next page

Continued from previous page

Repository name (abbreviated)	Roles	Number of appearances
	Web application developer	1
RUDA	RUDA manager	33
	Research assistant	10
	Researcher	5
	System administrator	1
	Related committee	2
IUGONET	IUGONET manager	23
	Researcher	16
NBDC archive	Contact information staff	9
	Researcher	14
	Data curator	17
	System operator	6
	Repository manager	1

End of table

The roles defined by each repository are different, and there is no noticeable trend in the number of appearances. Each repository’ s data curation activities are carried out in different ways. For example, there are three staffing patterns in the “Data Cleaning” process: the data holders themselves, the data curator(s), and the 2 or 3 parties working together. It should be noted that some of these processes are covered by support systems or tools.

3.4.3 Section summary

First, we conducted a vocabulary analysis of data curation activities as a preliminary study to interpret commonalities and differences in data curation activities. As a result, we obtained a working framework to compare the data curation activities. Next, we conducted a field survey using the above working framework and found that about 87.2% of the processes are interpretable across multiple fields. In contrast, there were no similarities in staffing by fields. To accurately compare data curation activities in different fields, the working framework should be improved by including the staffing as well as the three issues pointed out in 3.4.1.

3.5 Development of Data Curation Process Ontology

3.5.1 Formalizing commonalities and differences in data curation activity' s structure

This section explores methods for formalizing commonalities and differences in data curation activities. Based on the preliminary analysis and the survey results, there needs an expressive method of the structure including the relationships among Input-Output objects, processes, hierarchical relationship among activities, and staffing in order to accurately represent the data curation activity' s structure in different fields. Since these relationships are complicated, it is not impossible to represent the relationships in a simple tabular form. Some model is needed to describe these relationships adequately.

In this study, we adopt ontology as a model representation. We have developed an ontology that collects and structures knowledge to represent the data curation activity' s structure. Ontology is one of the methods for constructing conceptual systems used in the knowledge engineering field. The ontology theory provides a framework for knowledge sharing by clearly defining concepts and describing the logical relationships between concepts. Developing an ontology makes it possible to manage processes in which people and information systems are mixed.

3.5.2 Development process

In order to develop an appropriate ontology, it is recommended to follow some ontology developing procedure. Developing an ontology is not an easy task since explicating and formalizing the conceptual system behind the target system need a very complex and abstract thinking and reasoning. To ease the task, several procedures to develop an ontology are proposed. As for the ontology development procedure, we followed the seven steps proposed by Noy McGuiness (Noy and McGuinness, 2001). In the actual work, we made several iterations between Step 4 and Step 6 to maintain consistency with the hierarchical relationship. The OWL description of the ontology is shown in Appendix 3.

Step 1: Determine the domain and scope of the ontology

In this step, we determine the domain and scope of the ontology to design an ontology. The decisions to be made include those for the domain to be covered by the ontology, the intended use of the ontology, the problem that the ontology to be developed can solve, and the maintainer of the ontology.

In our ontology, we aim to represent commonalities and differences in data curation activity' s structure. The domain to be covered by this ontology is that of data curation. Providing the structured data curation activity in a machine-readable format can support the knowledge-sharing process between humans and information systems in a scalable manner. It would be desirable to maintain the ontology through the collaboration of data curators in each field and ontologists who deal with knowledge sharing in information systems.

Step 2: Consider reusing existing ontologies

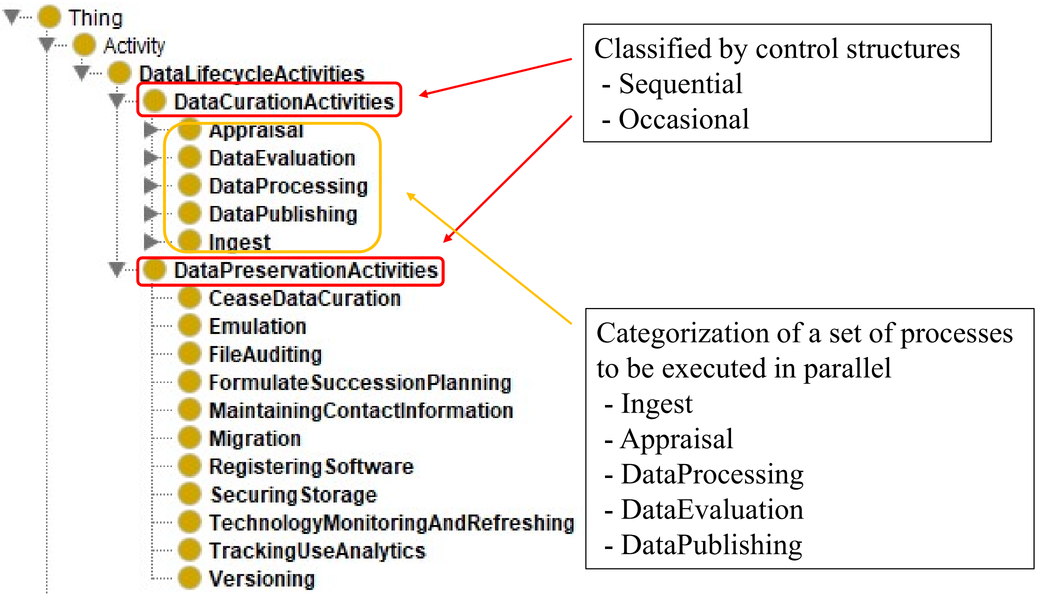
In this step, we consider the reusing existing ontologies. The description of this ontology is based on the vocabulary of the PROV ontology endorsed by W3C (Lebo et al., 2013). The PROV ontology is an ontology that provides a set of classes, properties, and restrictions that can be used to represent and exchange provenance information generated by different systems and different contexts. Since this study aims to interpret data curation activity' s structure in different fields, including different systems, we determined the PROV ontology to be compatible for representing the data curation activity' s structure.

As a basic structure of the PROV ontology, the information is represented by three classes and their relationships: Activity, Entity, and Agent. In the case of data curation activities, the data curation process can be represented as the "Activity" class, the input information and output information as the "Entity" class, and the staffing as the "Agent" class.

We mainly used the relationships defined in the PROV ontology to describe the relationships among Activities, Entities, and Agents. For some of the relationships, we used the FOAF ontology (http://xmlns.com/foaf/spec/) as a compliment.

Step 3: Enumerate important terms in the ontology

In this step, we enumerate important terms in describing the data curation activity’ s structure. First, we added the Data Curation Network vocabulary used in our preliminary analysis. Also, we added four additional vocabularies to organize the input-output information pairs. The vocabularies we added are "SubmitData," "ActualDataProcessing," "MetadataProcessing," and "CreatingLandingPage.” We extracted input information, output information, and the staffing role from the vocabulary in an abstract form to express the relationship between the data curation activity’ s structure. The criteria for the extraction are described in detail in Step 4.



Classified by control structures
- Sequential
- Occasional

Categorization of a set of processes to be executed in parallel
- Ingest
- Appraisal
- DataProcessing
- DataEvaluation
- DataPublishing

Figure 3.2: Data Curation Process Ontology structure.

Step 4: Define the classes and the class hierarchy

In this step, we define the classes and hierarchical relations of the ontology. Figure 3.2 shows the overall picture of this ontology’ s classes and hierarchical relations. Before determining the logical hierarchical relationship between the classes, we performed a categorical division of the activities; as shown in the preliminary analysis in Section 3.4.1, the extracted processes are a mixture of sequential and occasional processes. To separate the two types of activities with different control structures, we divided the classes into ‘Data Curation Activities’ for the sequential processes and ‘Data Preservation Activities’ for the occasional processes.

Next, we examined the logical structure of the ‘Data Curation Activities’ . Figure 3.3 shows the list of classes associated with each category. We set the following five categories under ‘Data Curation Activities’ : "Ingest," "Appraisal," DataProcessing," "DataEvaluation," and "DataPublishing.” We have already known that some sets of the data curation activity are performed in parallel from the preliminary analysis in Section 3.4.1. When handling this ontology, categorizing the process sets to be performed parallel helps interpretation. We set 22 processes under the five categories. In addition, four of the 22 processes have subclasses.

Step 5: Define the properties of classes-slots

In this step, we define the properties of the class-slots. Table 3.5 shows the list of properties used in this ontology.

This study adopted eight properties from the PROV ontology and one from the FOAF ontology. In describing the relationships in this ontology, we kept the description to the minimum necessary. In particular, the relationship between Activity and Entity is limited to "used" and "generated." In the reality of the data curation activity’s structure, the relationship between Activity and Entity is far more diverse. For example, "CodeReview (Activity)" has the relationship of reviewing "sourceCode (Entity)."

But having said that, describing the elaborate relationship intends to complicate the properties’ semantics. Since the complexity of semantics may affect the data curation activity’ s structure in different fields, we adopted the above policy as the first step in this ontology.

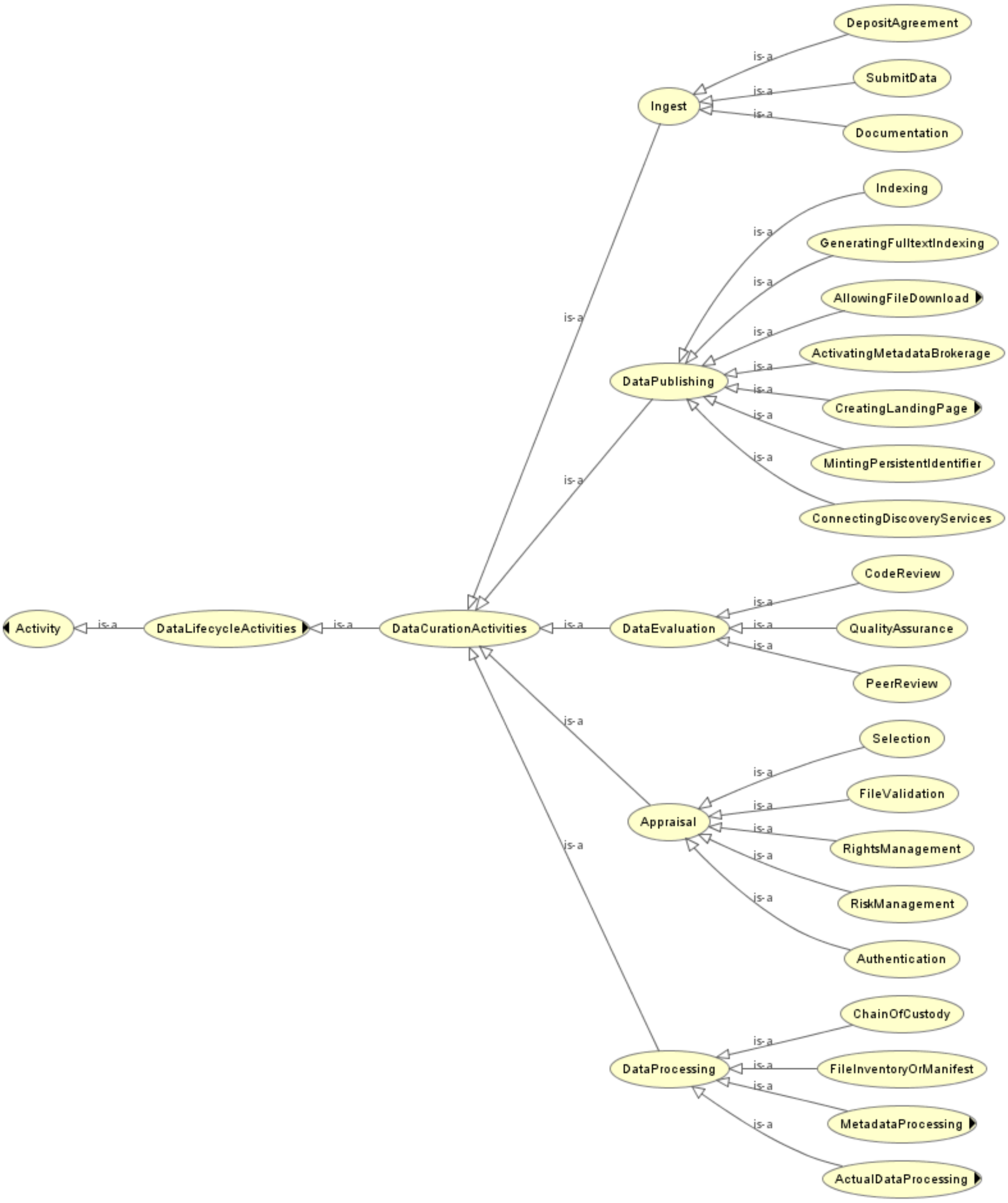


Figure 3.3: List of classes by category for sequential data curation processes.

Table 3.5: List of properties used in Data Curation Process Ontology.

prefix	property
prov	used
	generated
	wasAssociatedWith
	influenced
	wasInformedBy
	wasAttributedTo
	wasEndedBy
	wasStartedBy
foaf	primary topic

Step 6: Define the facets of the slots

In this step, we define the value type, allowed values, the number of the values (cardinality), and other features of the values as the facets that can be set for each slot. Since facets’ values can vary depending on the type of research data handled, it is necessary to accumulate data based on actual output information. Here, we have set tentative values for constraint types that align with the actual situation obtained from preliminary survey results in Section 3.4.1.

Step 7: Create instances

In this step, we create an instance corresponding to the class of this ontology. Since this ontology abstracts the commonalities and differences of data curation activity’s structure, it does not deal with the description of instances, which are individual phenomena. The description of the actual data curation activity’s structure is treated in Section 3.6.

3.5.3 Basic structure of Data Curation Process Ontology

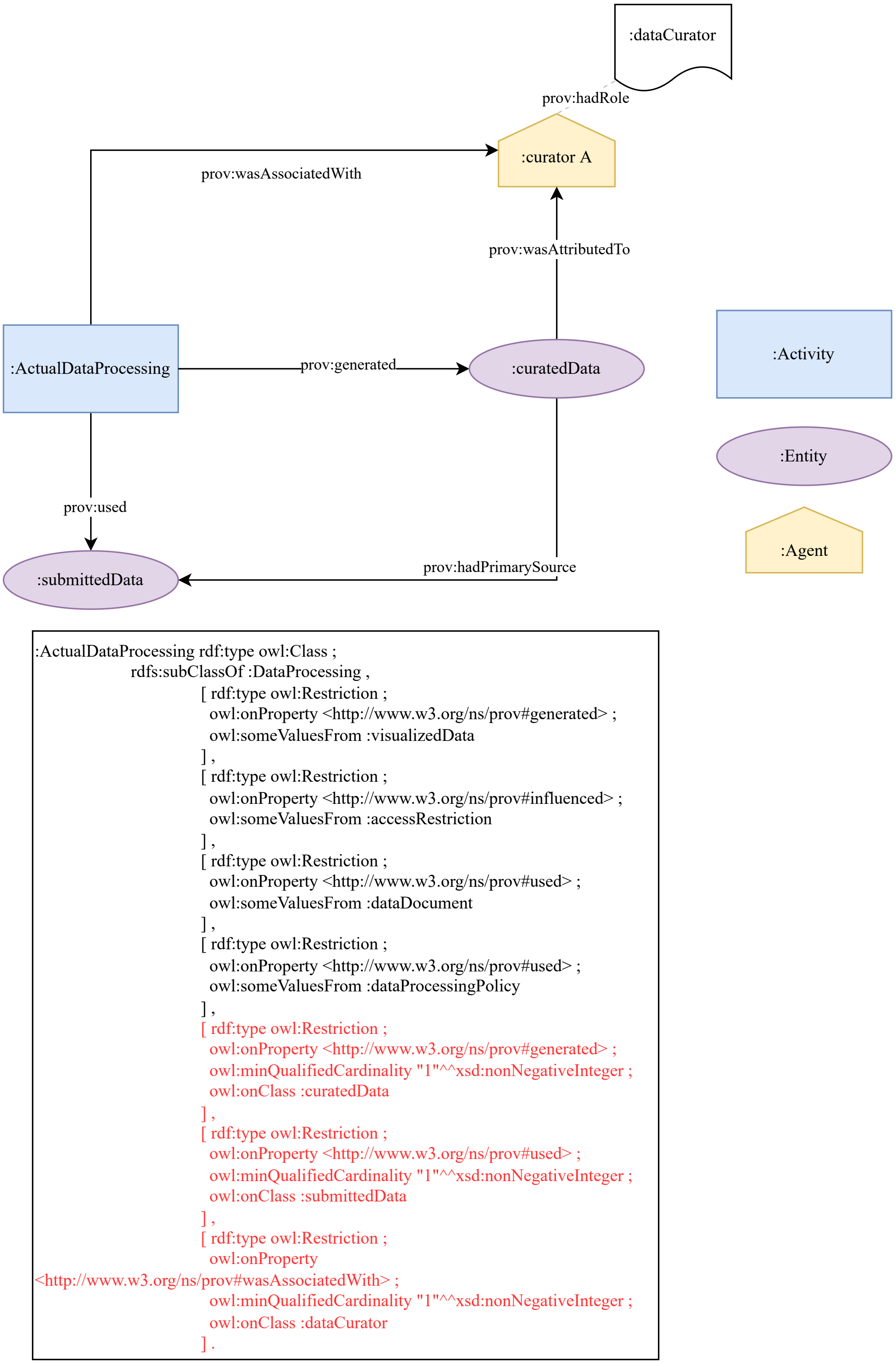
In this section, we explain the basic structure of the ontology created in Section 3.5.2. Figure 3.4 shows the basic structure of the configured class and an example of its output in OWL format.

In this figure, "ActualDataProcessing (Activity)" is represented as generating "curated data (Entity)" by using "submittedData (Entity)." Also, "curator A (Agent)" is represented as a performer of "ActualDataProcessing." The corresponding OWL description is shown in red.

In organizing the hierarchical relationship between activities, we defined the structure so that the relationship with Entity can be expressed only by "used" and "generated" as described Figure 3.4 “ActualDataProcessing" is a vocabulary introduced to maintain this structure, with subclasses "ArrangementAndDescription," "Conversion," "DataCleaning," "DataVisualization," "Deidentification," "FileFormatTransformation," "FileRenaming," "Interoperability," and "Restructure". Similarly, "MetadataProcessing" has "Contextualization" and "MetadataGeneration” , "CreatingLandingPage" has "DisplayingDataCitation" and "SettingTermsOfUse" as subclasses. This ontology has 184 classes and 313 subclasses generated as of version 1.

3.6 Case study

This section discusses a case study using the Data Curation Process Ontology. We used this ontology to describe the actual data curation activity’ s structures. Figure 3.5 through Figure 3.12 show the data curation activity’s structure covered in Section 3.4.



```
:ActualDataProcessing rdf:type owl:Class ;
                rdfs:subClassOf :DataProcessing ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty <http://www.w3.org/ns/prov#generated> ;
                              owl:someValuesFrom :visualizedData
                            ] ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty <http://www.w3.org/ns/prov#influenced> ;
                              owl:someValuesFrom :accessRestriction
                            ] ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty <http://www.w3.org/ns/prov#used> ;
                              owl:someValuesFrom :dataDocument
                            ] ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty <http://www.w3.org/ns/prov#used> ;
                              owl:someValuesFrom :dataProcessingPolicy
                            ] ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty <http://www.w3.org/ns/prov#generated> ;
                              owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                              owl:onClass :curatedData
                            ] ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty <http://www.w3.org/ns/prov#used> ;
                              owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                              owl:onClass :submittedData
                            ] ,
                            [ rdf:type owl:Restriction ;
                              owl:onProperty
<http://www.w3.org/ns/prov#wasAssociatedWith> ;
                              owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                              owl:onClass :dataCurator
                            ] .
```

Figure 3.4: Example of the basic structure of "ActualDataProcessing" class and an output example of "ActualDataProcessing" class in OWL format.

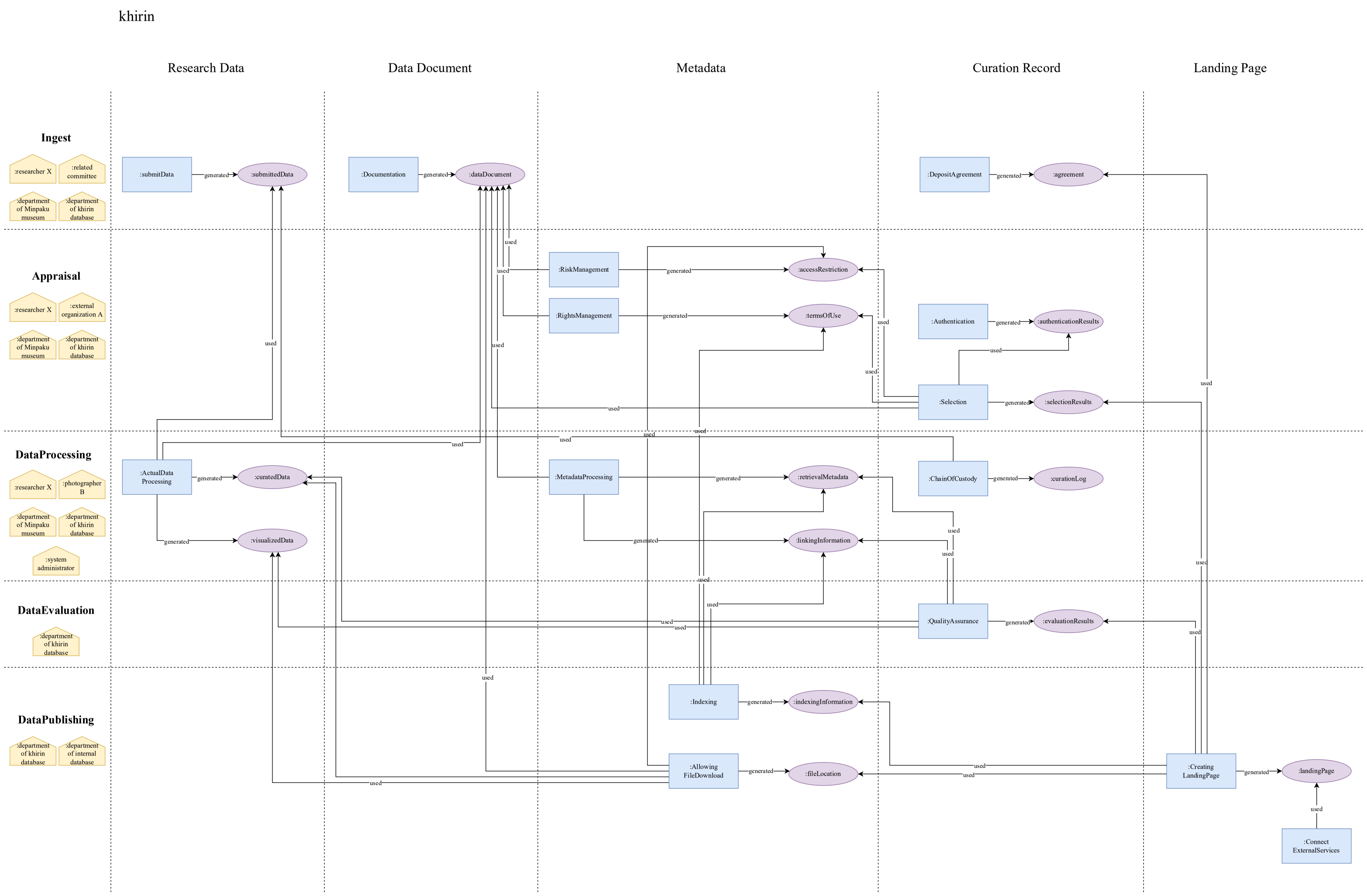


Figure 3.5: khirin data curation activity's structure.

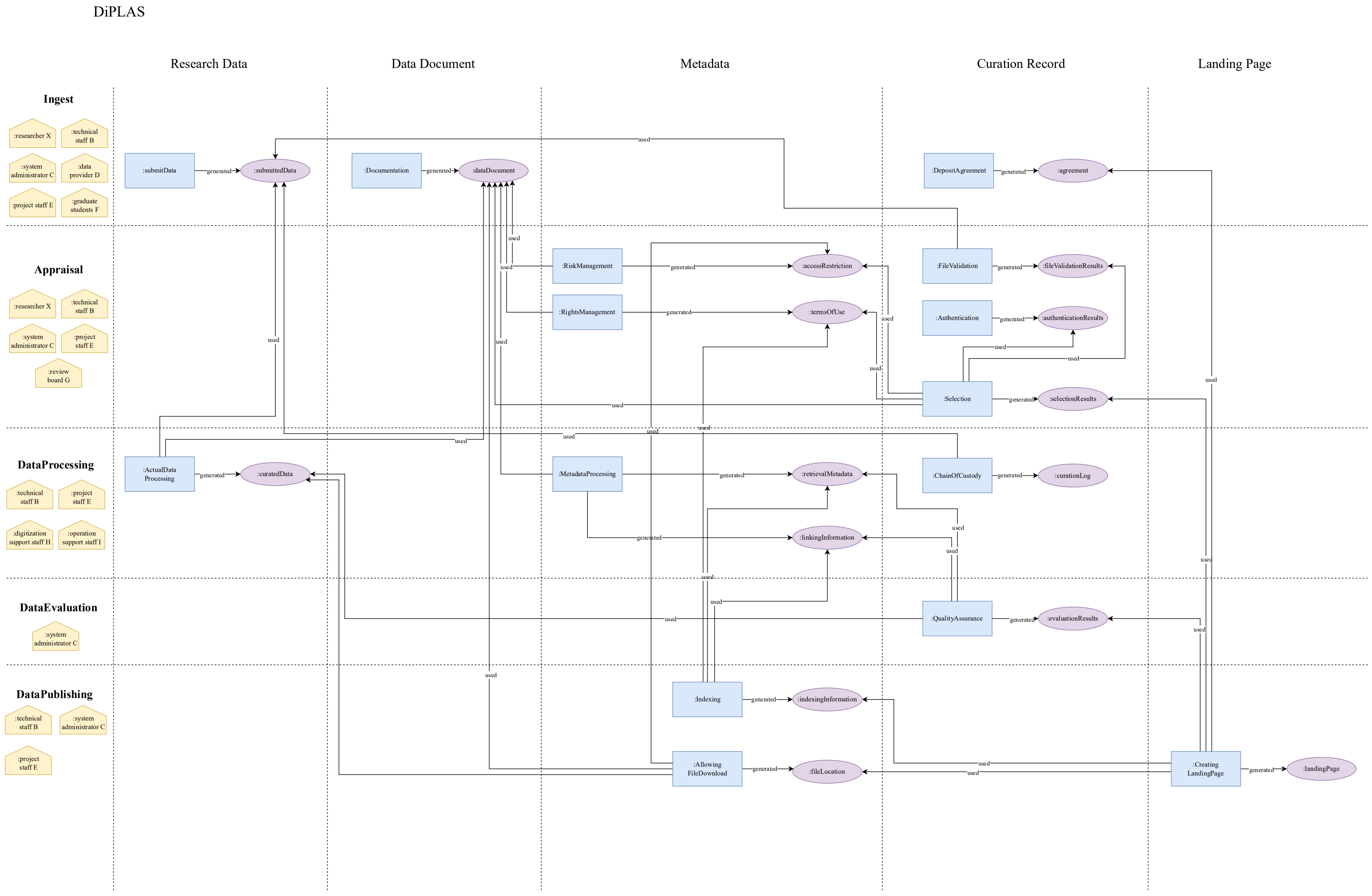


Figure 3.6: DiPLAS data curation activity's structure.

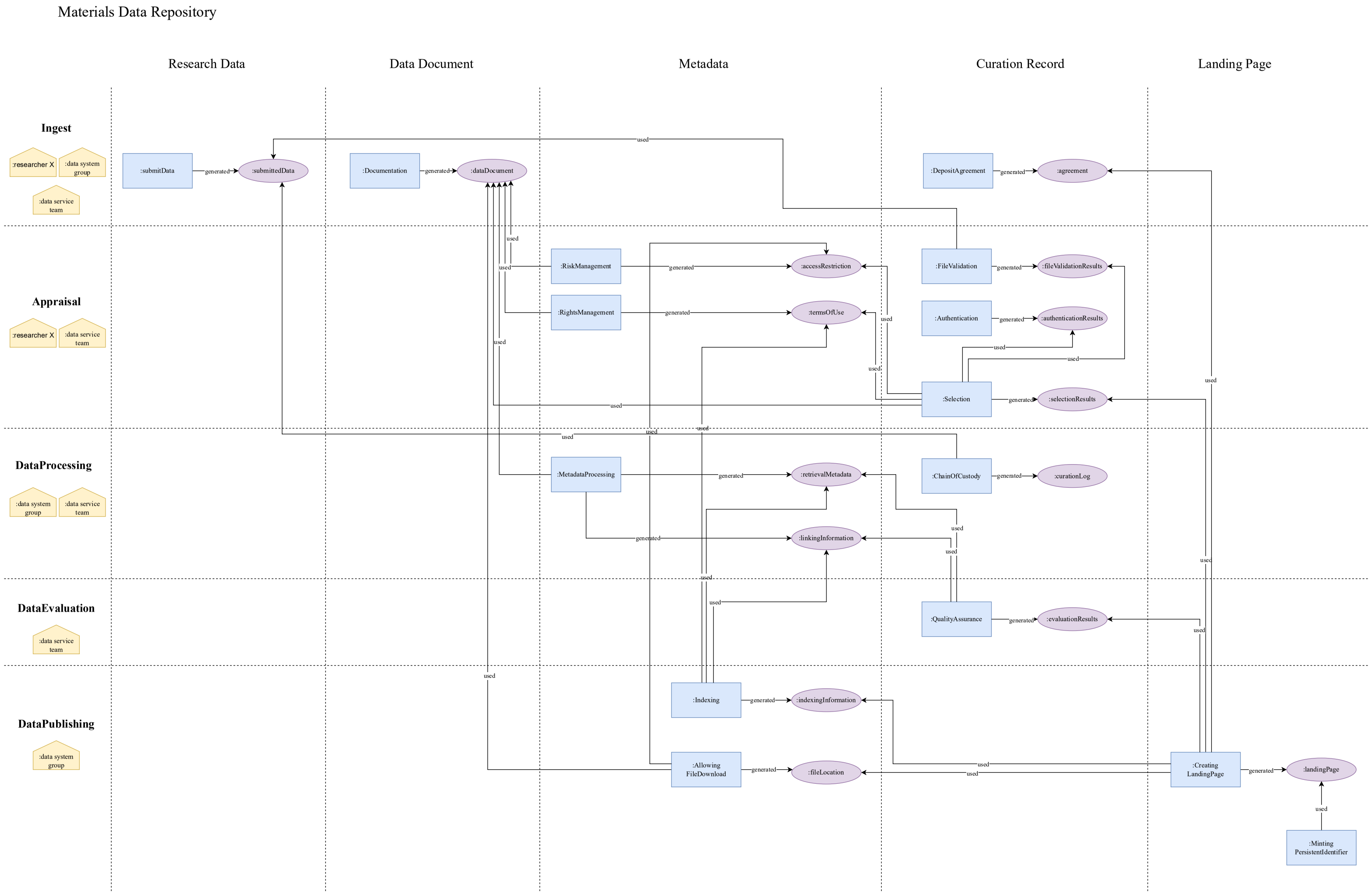


Figure 3.7: MDR data curation activity's structure.

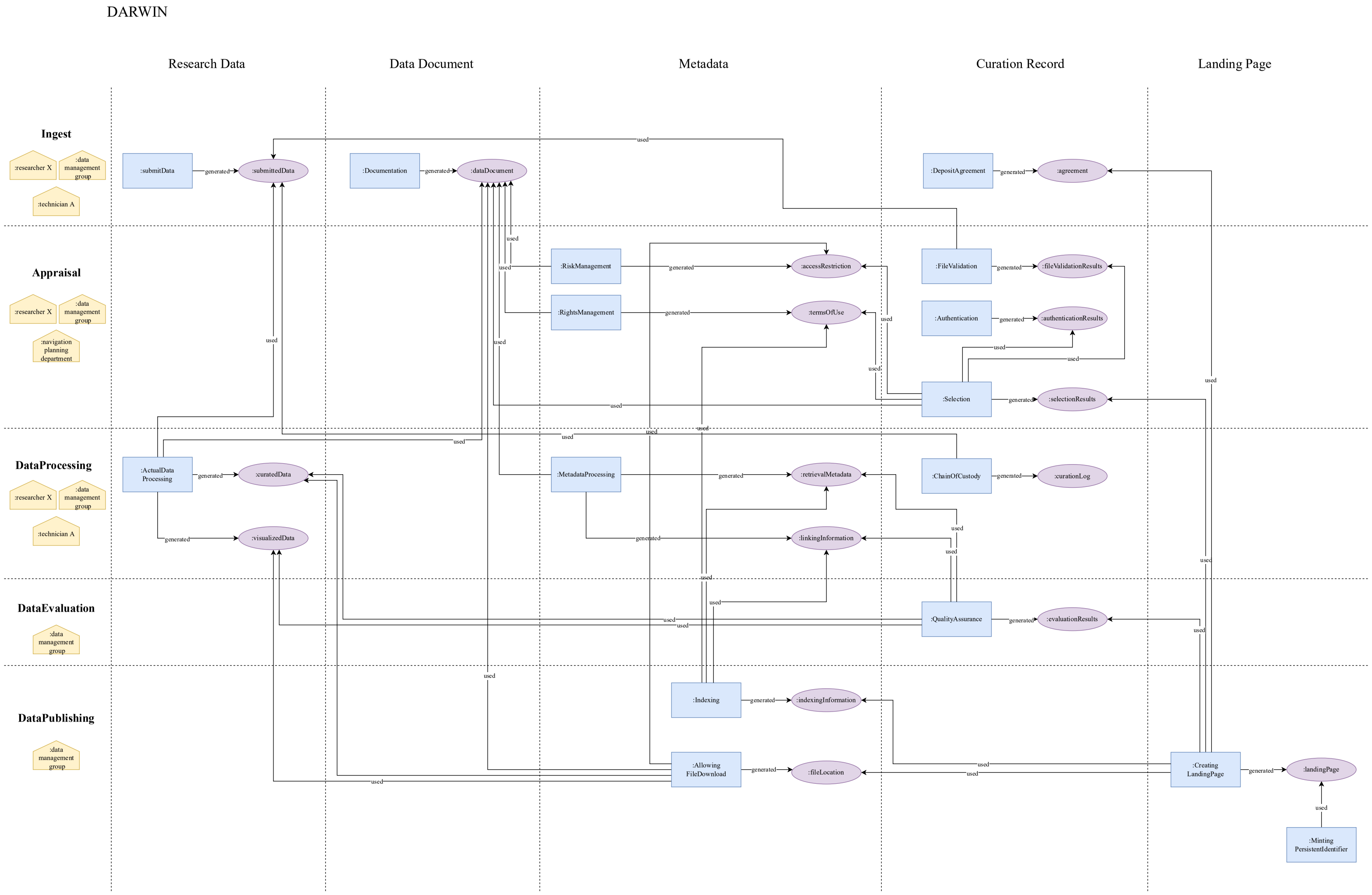


Figure 3.8: DARWIN data curation activity’s structure.

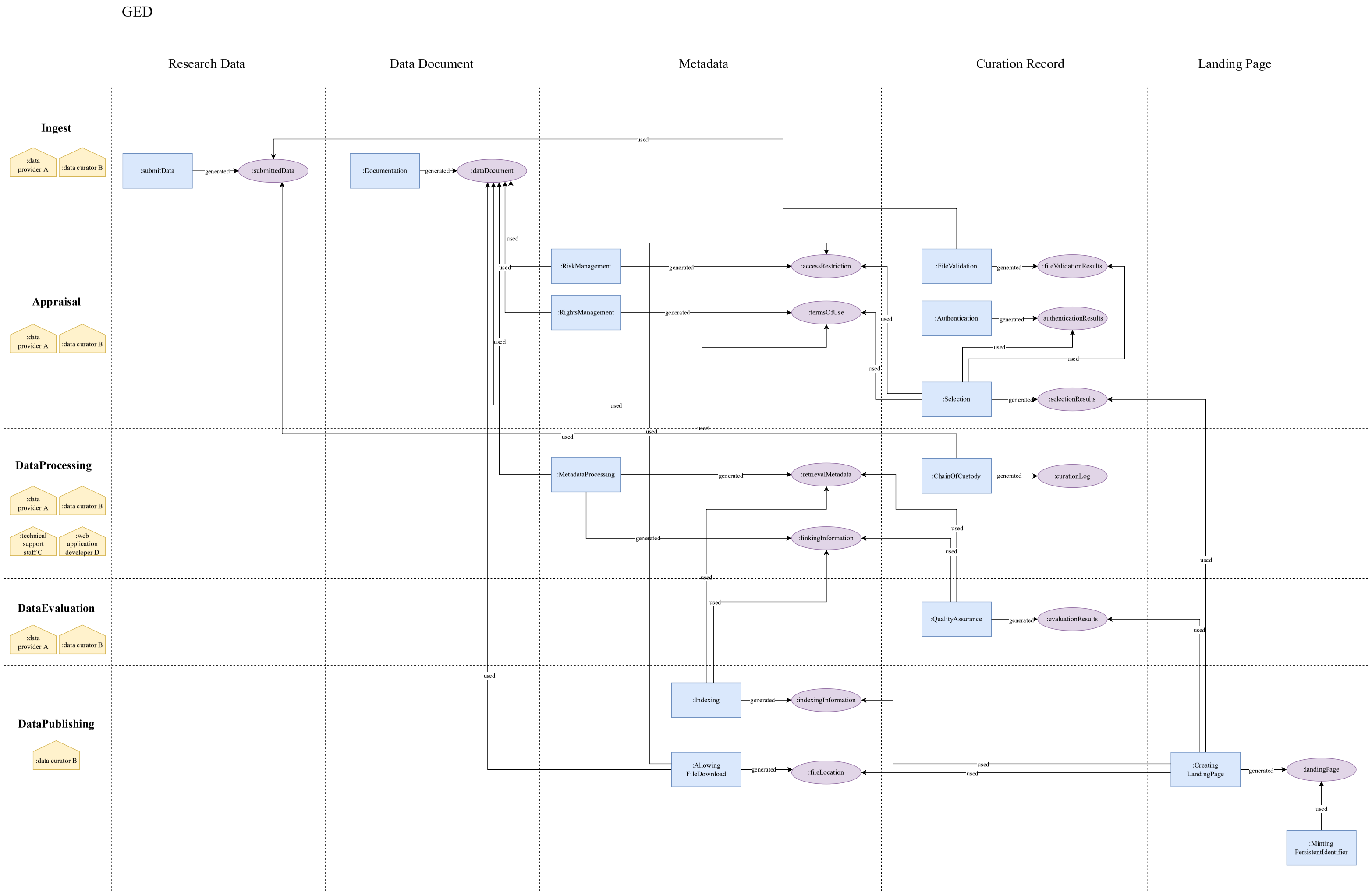


Figure 3.9: GED data curation activity's structure.

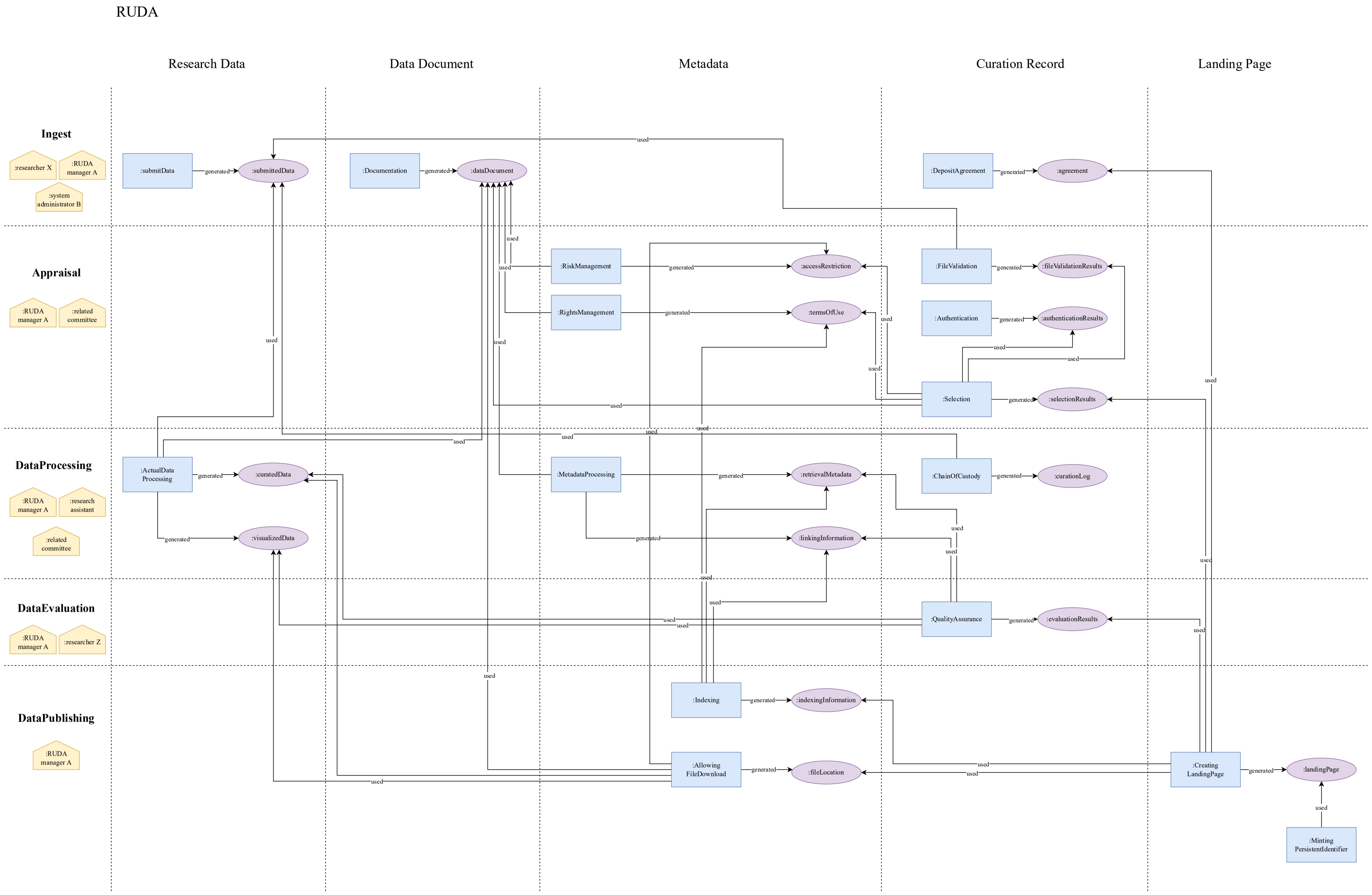


Figure 3.10: RUDA data curation activity’s structure.

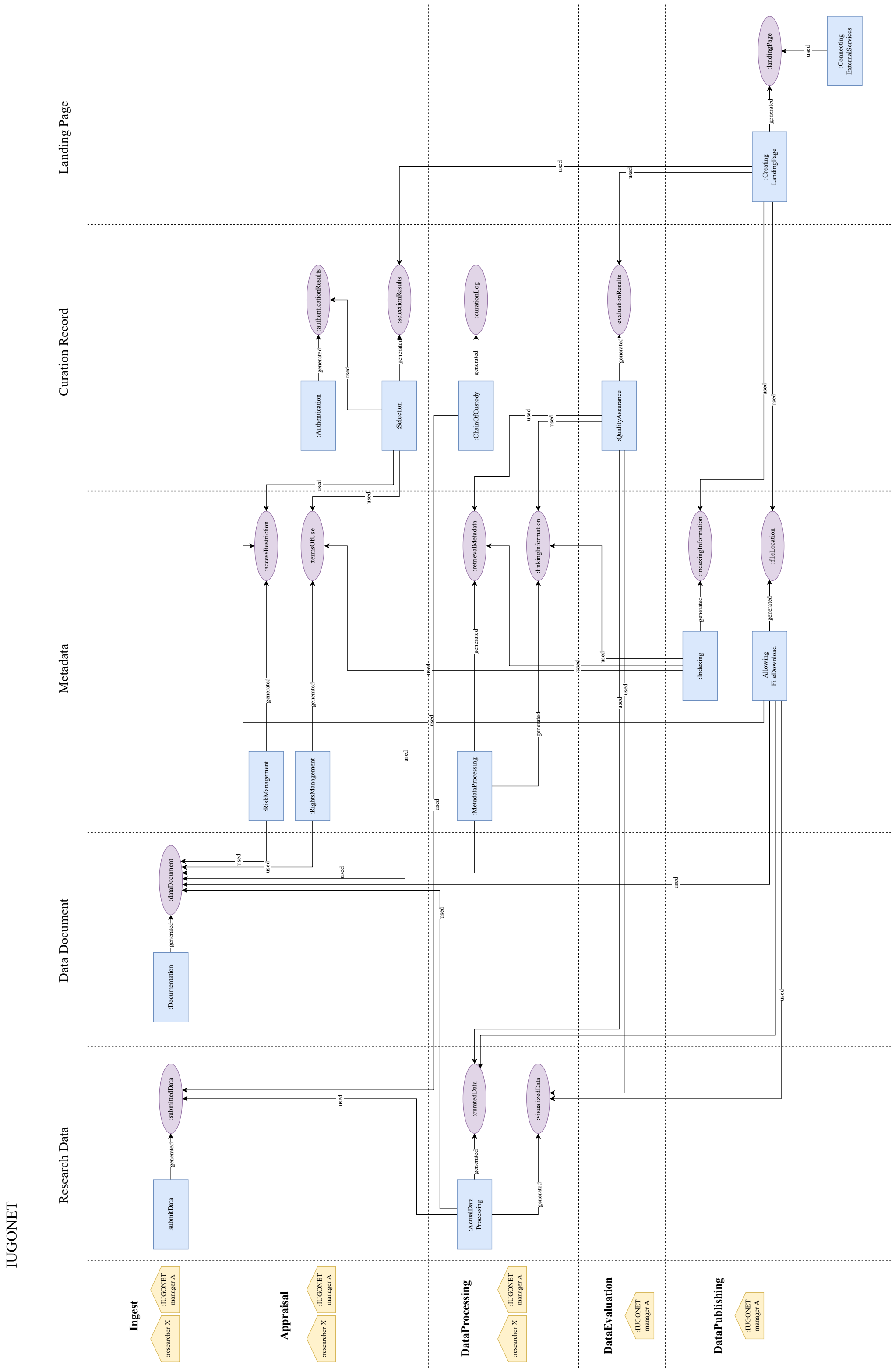


Figure 3.11: IUGONET data curation activity's structure.

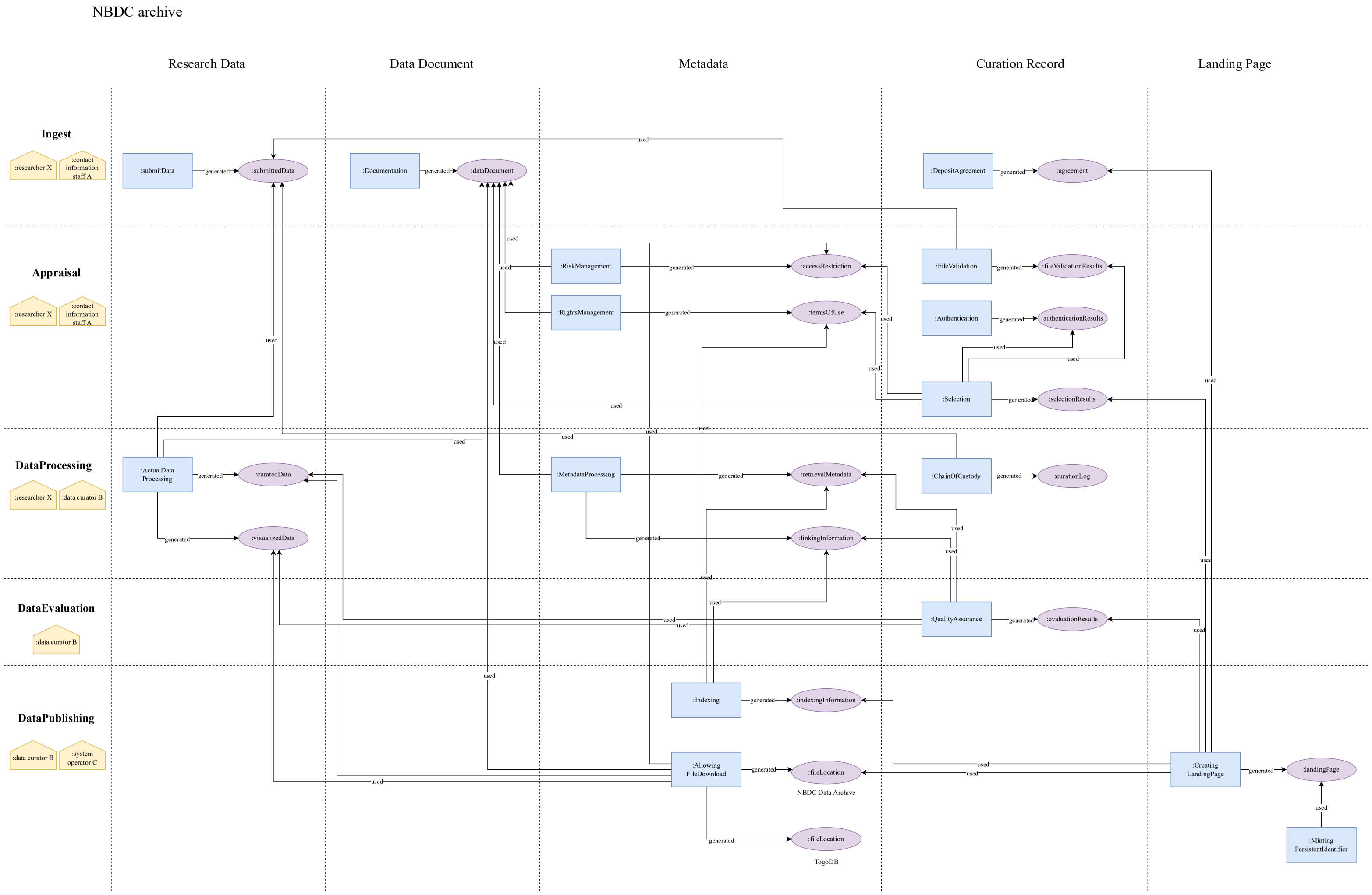


Figure 3.12: NBDC archive data curation activity's structure.

The essential information of the figure is the graph structure consisting of activity, entity, and agent. For ease of understanding, we arranged the graph with the following rules. The columns in each Figure show five typical Entities: "Research Data," "Data Document," "Metadata," "Curation Record," and "Landing Page.” The rows in each Figure show the categories of "Ingest," "Appraisal," "Data Processing," "Data Evaluation," and "Data Publishing.” A corresponding data curation process and the generated entity are placed at the intersection of the categories and the entities. The generated entity is connected to another data curation process in which the entity is used by a "used" line. Also, we described agent information on the horizontal axis in each Figure. Agent should be associated with each activity in PROV ontology. Since there are many agent-activity linkages, we described agent information in the simplified form; the agent linked to the activity is described at the left-most column on the same row.

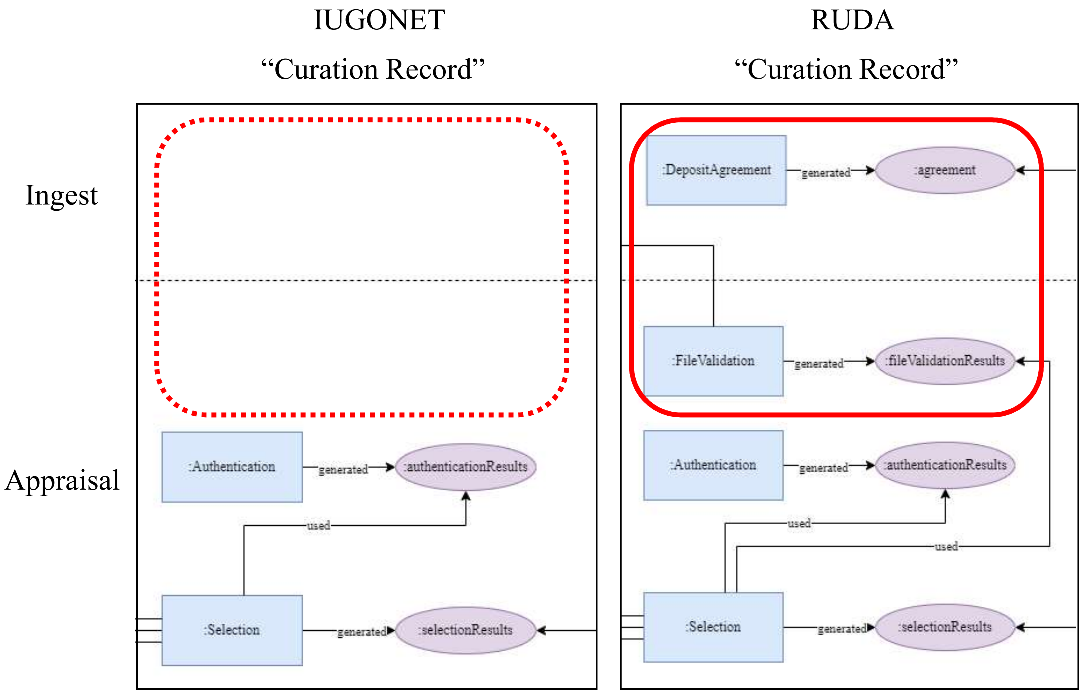


Figure 3.13: Comparison of data curation activity’ s structures in different fields.

As such, this format can describe the data curation activity’s structures in multiple fields in a single model. First of all, this format enables users from different fields to trace the provenance information of each entity. More consistent with the intent of this chapter, it also contributes to comparing the commonalities and differences between the data curation activity’ s structure in different fields by describing them by this ontology. Figure 3.13 shows an example of the “Curation Record (Entity)” comparison between IUGONET data curation activity’ s structure in Figure 3.11 (left) and RUDA in Figure 3.10 (right). The comparison shows that there are no "DepositAgreement" in the Ingest category and no "FileValidation" in the Appraisal category in the IUGONET data curation processes. The reason why these process has not been implemented in IUGONET is that the IUGONET is a metadata distribution service that relies on the data provider for data access. There is no need to verify the data or obtain permission for publication. Therefore, the “Authentication” process is positioned as more important duty of the data curator in terms of comparison with other fields.

3.7 Application: Specification of ontology-based data curation process support functions

In this section, we discuss the specification of a data curation process support function using this ontology. Table 3.6 shows the mapping to the functions possessed by the repository software WEKO3 (https://rcos.nii.ac.jp/en/service/weko3/), which is widely used in Japan.

Table 3.6: Functional mapping with WEKO3.

Category	Data Curation Processes	Function name (WEKO3)	Shareable process flag	Remarks
Ingest	SubmitData Documentation Deposit-Agreement	Item registration (No function) (No function)	* * *	
Appraisal	Authentication FileValidation	Log-in Item registration	* *	Except for file format validation

Continued on next page

Continued from previous page

Category	Data Curation Processes	Function name (WEKO3)	Shareable process flag	Remarks
	RightsManagement	(No function)	*	
	RiskManagement	(No function)	*	
	Selection	(Partly) Workflow	*	Except for selection criteria support
DataProcessing	ActualDataProcessing	(for journal article) Cover page creation	*	Required processes vary by field
	ChainOfCustody	Workflow	*	
	FileInventoryOrManifest	(No function)	*	
	MetadataProcessing	Item registration/Item linking	*	
DataEvaluation	CodeReview	(No function)		
	PeerReview	(No function)		
	QualityAssurance	Item approval	*	
DataPublishing	ActivatingMetadataBrokerage	OAI-PMH harvesting / ResourceSync	*	
	CreatingLandingPage	LandingPage displaying	*	
	GeneratingFulltextIndexing	Full-text indexing		
	Indexing	Index creation	*	
	AllowingFileDownload	Download URL creation	*	
	ConnectingDiscoveryServices	OAI-PMH harvesting / ResourceSync / Google Scholar metadata / schema.org	*	

Continued on next page

Continued from previous page

Category	Data Curation Processes	Function name (WEKO3)	Shareable process flag	Remarks
	MintingPersistentIdentifier	DOI registration / CNRI handle	*	

End of table

Table 3.6 shows the WEKO3 functions corresponding to the Data Curation Process Ontology and the flags indicating the interpretable processes by multiple organizations analyzed in Section 3.4.2 as the shareable processes. WEKO3 is a data publishing platform for researchers to publish research data and related materials. WEKO3 supports basic data registration routes such as "SubmitData" and "FileValidation," and also supports a wide range of metadata registration, editing, and publishing functions such as "MetadataProcessing," "ChainOfCustody," "QualityAssurance," and "DataPublishing." Whereas, WEKO3 does not support some shareable processes related to data itself in each field such as "Documentation," "RightsManagement," "RiskManagement," "Selection," and "ActualDataProcessing." It is recommended to provide the support functions for these processes to support data curation in different fields by WEKO3 comprehensively.

Thus the formalized processes enable analysis system functions to identify the shareable data curation processes in different fields. Moreover, the ontology supports integrating with external tools to handle missing processes. Data curators often use multiple data curation systems, then the ontology can help them to allocate support functions among the systems.

3.8 Results and remaining issues in chapter 3

As the first step to interpret the tasks and procedures performed in different fields at the same granularity, we investigated the practices of data curation conducted in each field. As a result, we found that about 87.2% of the processes are interpretable across multiple fields. Also, we realized that there needs a suitable model to describe the structure such as the relationships among Input-Output objects, processes, and staffing to accurately represent the data curation activity' s structure in different fields.

Based on the preliminary analysis and survey results, we developed the Data Curation Process Ontology to formalize the data curation activity’ s structure in different fields. To verify the usefulness and validity of this ontology, we described and compared the several actual data curation activity’ s structures. Users can visualize the data curation activity’ s structure and the provenance information of the research data by using the format presented in Section 3.6. It is the important contribution of this study to compare the activity’ s structure of eight diverse repositories in a single model. Also, we showed that the ontology can use the specification of data curation process support functions by systems. This ontology can also help to allocate data curation process among multiple systems and improve the data curation process during data integration in different fields. With the formalized method enabled by the ontology, we can expect to expand the data curation process in various fields from an interdisciplinary perspective. Thus this study contributed to building a knowledge framework for a common understanding of the data curation process in different fields.

A possible direction for utilizing ontology is to evaluate the quality of research data. The ontology provides a well-formalized structure for the data curation activities and may function as a framework for the process management of research data. To achieve process management of research data, it is necessary to accumulate case study practices based on a formalized process definition at an application level. Technical implementation in terms of formalized processes will help validate and solidify the framework.

4

Practical interpretation of the interdisciplinary data evaluation processes using the reference model for data publishing

In Chapter 4, we deal with the processes in the Data Evaluation category formalized in Chapter 3 through practical implementations. To interpret research data from a field to be independently, the data and their documentation must be reviewed from an interdisciplinary perspective and revised as necessary. In contrast, many data repositories either do not have a policy for evaluating research data and encouraging improvement, or the evaluation process is done only from the perspective of a specific field. This chapter focuses on publishing mechanisms for data papers that divert the journal peer-review and clarifies the mechanism from a process perspective. Furthermore, we conduct a technical implementation for interpreting the mechanism using the reference model for data publishing. This study is a practical interpretation of the relationship and boundaries between the existing quality assurance process and the peer-review process in the Data Evaluation category.

4.1 Introduction

In Chapter 4, we deal with the processes in the Data Evaluation category formalized in Chapter 3 through practical implementations. The Data Evaluation category includes Code review, Quality Assurance, and Peer-review processes. Generally, these processes review the quality of the research data and associated documentation and take the necessary actions to make the subject independently interpretable (Peer et al., 2014; Consultative Committee for Space Data Systems, 2012). The target objects are data files, data documents, metadata, codes, etc., and may include data papers (Lawrence et al., 2011). The data evaluation is essential in maintaining control over scholarly communication.

The data evaluation processes have been put into practice in some leading fields. These experiences have shown several gaps and challenges in operationalizing the evaluation process. First, many data repositories either do not have a policy for evaluating research data and encouraging improvement, or the evaluation process is done only from the perspective of a specific field (Peer et al., 2014). Dataset quality seems to be a mixed bag due to the diversity of the data repository' s policy (Ruggles, 2017). Also, describing metadata is time-consuming and often not recognized as a research contribution (Edwards et al., 2011). As a result, it lacks positive motivation for researchers. This situation is one of the primary reasons for the lack of progress in depositing research data even in the leading fields.

Mechanisms that include incentives for researchers regarding data curation are needed to overcome challenges in the data evaluation process. In recent years, peer-review of data using the media type of data papers has come into the spotlight (Hrynaszkiewicz and Shintani, 2014; Assante et al., 2016). Traditionally, peer-review has been the standard mechanism across all fields as a framework for accrediting and evaluating research papers (Spier, 2002). Peer-review of data papers could be a more effective mechanism for evaluating research data.

This chapter focuses on publishing mechanisms for data papers that divert the journal peer-review and clarifies the mechanism from a process perspective. Furthermore, we conduct a technical implementation for interpreting the mechanism using the reference model for data publishing. This study is a practical interpretation of the relationship and boundaries between the existing quality assurance process and the peer-review process in the Data Evaluation category.

4.2 Literature review

4.2.1 Peer-review of data

Peer-review of data is not yet a clearly defined concept; its meaning varies depending on the peer-reviewed scenarios (Mayernik et al., 2015). There are three main scenarios in which data may be subject to peer-review: 1) when data is published in a data repository, 2) when data is published as an appendix to a paper, or 3) when data is published as a data paper (Lawrence et al., 2011; Mayernik et al., 2015). These scenarios can be positioned as a framework for data publishing (Candela et al., 2015; Pampel et al., 2012), with variations depending on the division of roles among researchers, repository managers, and publishers involved in publishing data (Peer et al., 2014; Lawrence et al., 2011).

1) When data is published in a data repository, the data curator reviews the data in terms of file interpretability and accessibility, and long-term preservation (Peer et al., 2014). This process includes dataset dimension checks, validity checks, confidentiality checks, metadata checks and enhancement, and format transformation checks (UK Data Archive, ndb). As already mentioned, many data repositories either do not have a policy for evaluating research data and encouraging improvement, or the evaluation process is done only from the perspective of a specific field (Peer et al., 2014).

2) When data is published as an appendix to a paper, the peer-reviewer reviews to substantiate the statements contained in the paper (Grootveld and van Egmond, 2012). The items checked to depend on the journal's peer-review policy. In most cases, data documents and metadata are not subject to peer-review, making it difficult to interpret the data independently.

3) When data is published as a data paper, the peer-reviewer reviews from the

perspectives of both 1) and 2). The data paper describes mainly ancillary information, including where the research data is stored. Unlike ordinary research papers, a data paper rarely describes some findings or interpretations derived from research data (Candela et al., 2015). The peer-reviewers will review the data documents described in the data paper and confirm that the data have been processed as described. In this case, the challenge is establishing a system to realize the peer-review of data.

4.2.2 Data paper in data publishing

The following section will then provide an overview of the mechanisms that make the peer-review process of data papers available. Reflecting the recent open science trends, many data papers are published in an open access form (NISTEP, 2015). Dataset tied to the data paper can be easily utilized by a wide range of research fields now. Having said that, the concept of data publication itself is not new. There have been many attempts to publish research data in reports or chronologies (Beagrie, 2008). While the open access initiative has given new meaning to the data publication (Klump et al., 2006), data journals that mainly deal with data papers have emerged as a unique media format (Candela et al., 2015). Data journals handle research data using the same business model as a typical research paper (NISTEP, 2015). This means that the publication functions of registration, certification, awareness, archiving, and rewarding are applied to research data (Van de Sompel et al., 2004). According to this perspective, the following benefits can be expected.

a) Enriching description regarding research data

The authors can describe more detailed information in the data paper compared with existing research papers. Table 4.1 shows the general elements of the data paper's description (Candela et al., 2015). Note that the precision and emphasis of the required descriptions vary from journal to journal. These descriptions include information that is rarely described in general research papers. For example, "Format," "Microattribution," "Provenance," and "Reuse" are likely to be excluded from general research papers.

Increased information could simply improve reproducibility and ensure research transparency. Also, project-isolated and not yet cited research data can be stored and traced step-by-step on a common platform via a data paper. It is expected to preserve and reuse datasets as research contribution, 80% of which are said to be lost in 20 years (Vines et al., 2014).

Table 4.1: Data paper description style. Source: Candela, L et al., 2015

No	Element	Description
1	Availability	Name of the data provider, DOI/URI, etc.
2	Competing interests	Individual/Organizational relationship related to the dataset
3	Coverage	Spatial/Temporal
4	Format	File format, Encode, Language, etc.
5	License	–
6	Microattribution	Contribution type related to the dataset development
7	Project	–
8	Provenance	Including acquisition methods, equipment, etc.
9	Quality	Information on data limits, outliers, etc.
10	Reuse	Potential Value

b) Ensuring the quality of research data

The conceptual process of data publishing is shown in figure 4.1 (Kratz and Strasser, 2014). Authors store the dataset in a repository and submit a detailed description of the dataset as a data paper. After the prescribed procedure, the publisher checks the submitted dataset and data paper and assigns an identifier to both.

As mentioned above, data papers submitted to data journals go through the same third-party peer-review process as existing research papers; ensuring quality through peer-review improves interpretability. Note that the meaning of "quality" of a data paper differs from that of an existing research paper. In a data paper, a peer-reviewer does not evaluate "highly novel" but rather "accurately and richly described according to a certain descriptor."

c) Expansion of distribution channels

Since data publishing maintains the form of paper, data papers can be published on the same platform as existing research papers. Also, data papers will be given the same metadata as existing research papers. This metadata can improve research

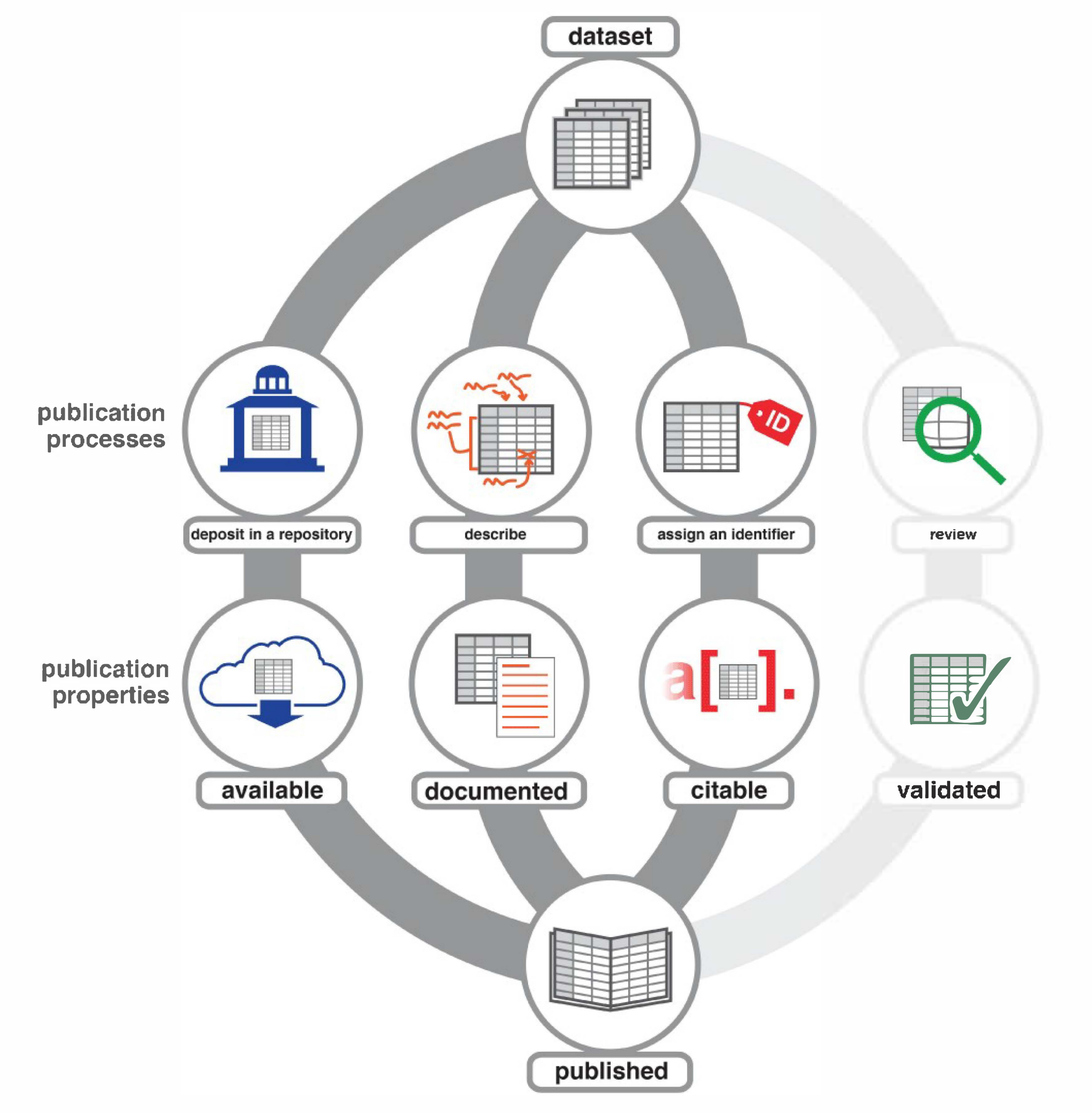


Figure 4.1: Conceptual process of data publishing. Source: Kratz J, Strasser C., 2014

data discoverability through searching data papers; the search route via data papers avoids complex interdisciplinary use of metadata problem across different fields (Willis et al., 2012). Moreover, from the data user's perspective, they can search the research data centralized with a familiar search platform. Publishing data paper leads to an expansion of the distribution channel.

d) Provide incentives for data curation

A high-quality description of research data will be recognized as a research contribution to evaluating the data paper's description by peers. This direction may motivate researchers to create and preserve higher-quality data. Also, research data have traditionally not been subject to citation (Robinson-García et al., 2015) so the data creators have not been fully recognized. In the data publishing norm, the data creators will be the author of the data paper. The data creators will evaluate their work as a research community' s contribution (Lawrence et al., 2011).

4.2.3 A reference model for data publishing

Traditional data deposition flow is accomplished in a manner tied to a repository, making it difficult to find a clear and consistent human-readable workflow representation for the repository (Austin et al., 2016). As one of the few examples of how this issue is being addressed, the RDA/WDS Publishing Data Workflows WG has analyzed 25 journals/repositories/guidelines and is developing standard workflows.

Figure 4.2 shows a framework for data publishing discussed in the RDA/WDS WG (Austin et al., 2016). No. 2-1 shows the traditional publication workflow model. In this model, both ensuring access and creating an explanation of the research data are left to the researchers. No. 2-2 shows the data publishing workflow model with "data publication." In this model, researchers should describe additional information such as methods, detailed descriptions, and related computer codes. No. 2-3 provides a detailed flow of the data publication and article publication process as defined in No. 2-2. "Data repository submission" shows the checking mechanism for mutual use and reuse of actual data. The "data article submission" shows the process of reviewing data papers. Data papers are published in scholarly journals using the same peer-review process as traditional research papers. The data paper contains link information about

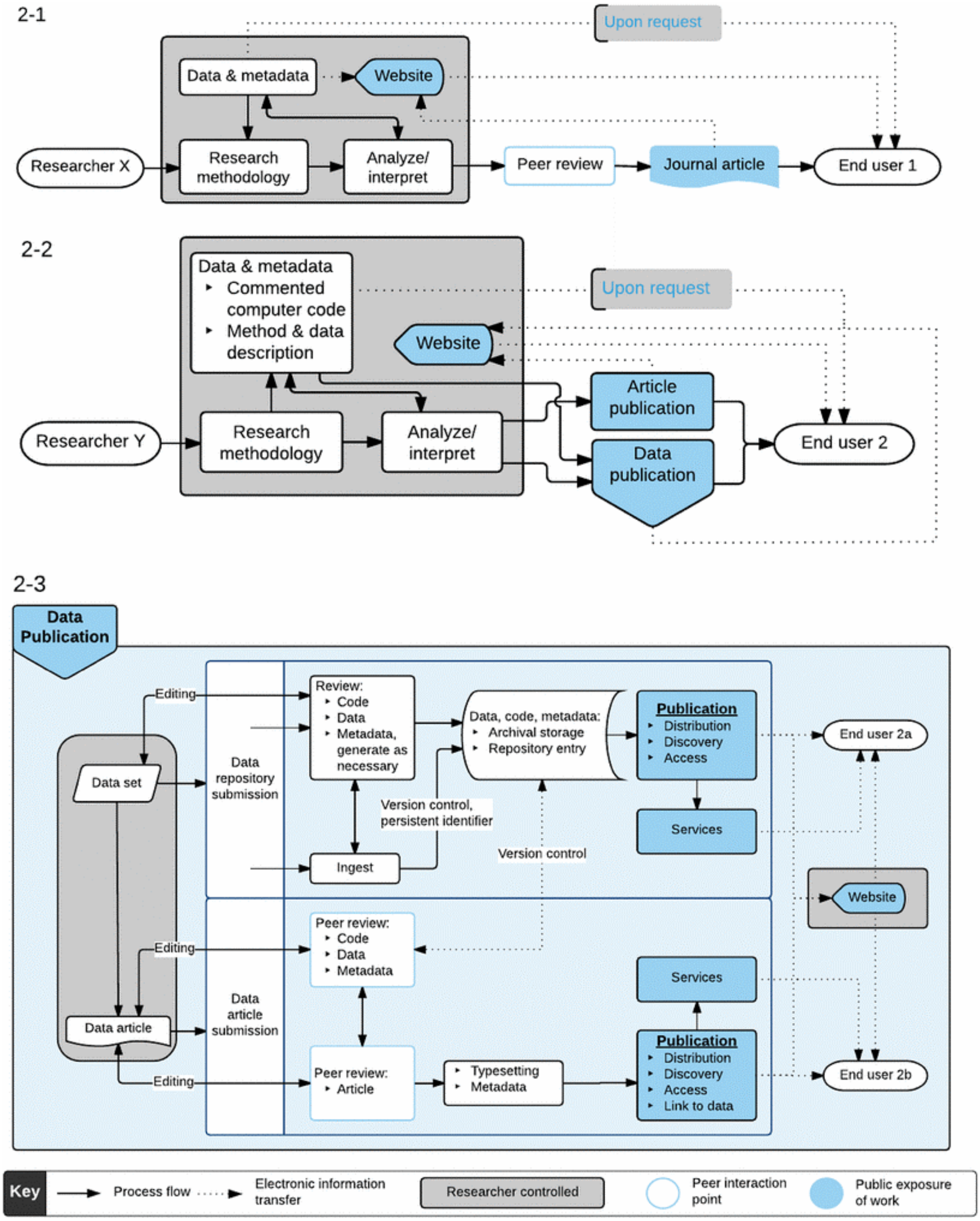


Figure 4.2: Research data publication workflows. Source: Austin, C.C. et al., 2016

the actual data and will be published on the journal platform in the same manner as the research paper.

"Data repository submission" and "Data article submission" can be evaluated as flow representations of the quality assurance process and the peer-review process formalized in Chapter 3. We have positioned this model as an effort consistent with a practical interpretation of the formalized evaluation process.

4.3 Case study

As detailed in Figure 4.2, the data publishing flow for data paper combines a quality assurance process and a peer-review process. Separating the two evaluation process will help make the system lighter and more flexible. However, there are few accumulated examples of application designs that follow this reference model. Therefore, we planned to implement some data curation support systems with each process as an application boundary. In this section, we introduce a case study in polar science field for referring to this reference model.

4.3.1 Context

The NIPR has acquired a wide variety of polar scientific data through its historical Antarctic and Arctic Research program. The acquired data are shared internationally, contributing significantly to the development of the field. The NIPR has published two reports: JARE Data Reports and NIPR Arctic Data Reports for international data sharing (Matsuzato, 1992).

Since the 2000s, NIPR started to operate some data repositories: the "NIPR Science Database" in 2007 and the "Arctic Data Archive System" in 2012. These data repositories realize the distribution of research data regardless of media format, but they have similar operational issues as mentioned in Section 4.1. Especially, the quality assurance of research data to be published under the responsibility of the NIPR was a critical issue for NIPR data curators. To address this issue, the data journal was proposed as a mechanism for both data distribution and quality assessment. The NIPR launched the Polar Data Journal (https://pdr.repo.nii.ac.jp/), the first data journal in a Japanese academic institution in January 2017. 32 data papers have been published as of March

2022.

4.3.2 Policy design

Since data journals are the same media format as existing journals, there are no significant differences in policy design considerations. Polar Data Journal is published as a free-access online journal, inheriting the characteristics of a general data journal. The license granted to the data paper is either the Creative Commons Attribution 4.0 International License (CC-BY 4.0) or the Creative Commons Attribution - Non-Commercial - No Derivative Works 4.0 International License (CC-BY-NC-ND 4.0). Also, DOIs (digital object identifiers) are assigned to data papers to enhance reusability.

4.3.3 Establishment of data publishing flow

According to the RDA/WDS reference model, research data and data papers need to be managed in parallel under version control. To achieve parallel flow management with multiple stakeholders, we represented each stakeholder’ s data curation process in a flow diagram. The parties involved in the workflow are as follows: the author who submit data papers, the referee who scrutinizes the papers, the editor and officer who perform the editing, and the data repository manager. Figure 4.3 shows an overview of the editorial process for the Polar Data Journal. The adopted system is described in detail in Section 4.3.4.

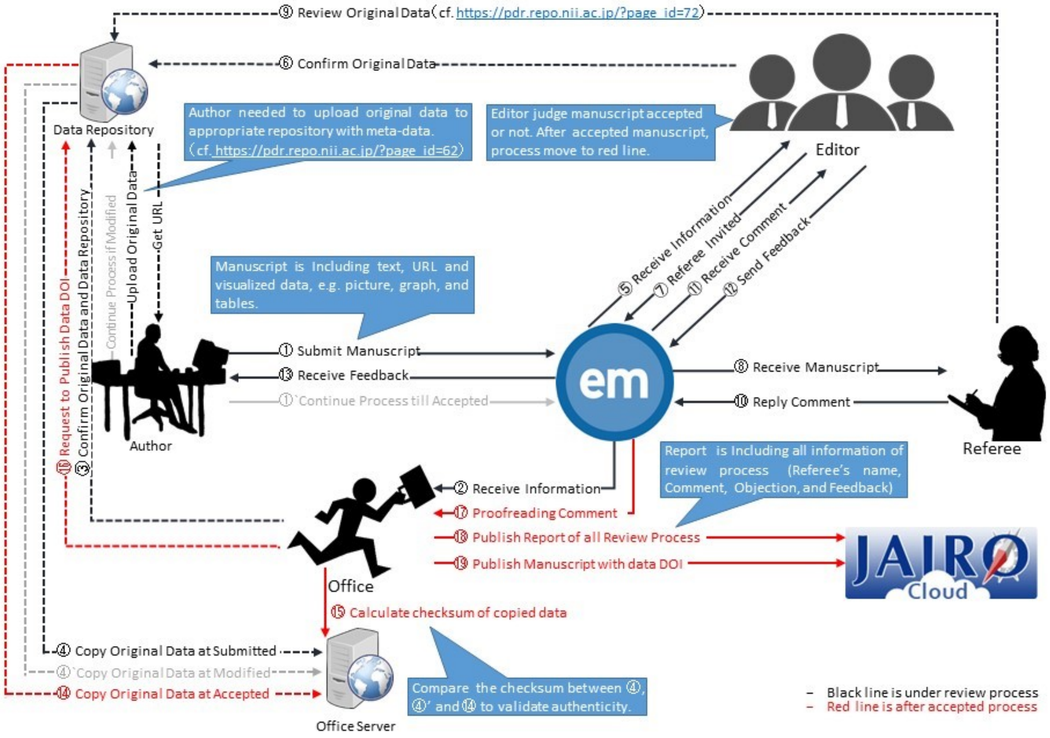
The description of each process is as follows:

(1) Authors submit a manuscript and associated files to EM.

(2) The information entered at the time of submission is sent to the editorial office.

(3) The editorial office confirmed no missing information and that the link to access the actual data is valid.

(4) The actual data in a data repository is copied to the office server; the officer calculates the checksum of the actual data.



⑨ Review Original Data（cf. https://pdr.repo.nii.ac.jp/?page_id=72）
⑥ Confirm Original Data
Data Repository
Editor
Author
Referee
Office
Office Server
em
JAIRO Cloud
⑯ Request to Publish Data DOI
③ Confirm Original Data and Data Repository
Continue Process if Modified
Upload Original Data
Get URL
① Submit Manuscript
⑬ Receive Feedback
①`Continue Process till Accepted
⑤ Receive Information
⑦ Referee Invited
⑪ Receive Comment
⑫ Send Feedback
⑧ Receive Manuscript
⑩ Reply Comment
② Receive Information
⑰ Proofreading Comment
⑱ Publish Report of all Review Process
⑲ Publish Manuscript with data DOI
⑮ Calculate checksum of copied data
④ Copy Original Data at Submitted
④`Copy Original Data at Modified
⑭ Copy Original Data at Accepted
– Black line is under review process
– Red line is after accepted process

Author needed to upload original data to appropriate repository with meta-data.（cf. https://pdr.repo.nii.ac.jp/?page_id=62）

Editor judge manuscript accepted or not. After accepted manuscript, process move to red line.

Manuscript is Including text, URL and visualized data, e.g. picture, graph, and tables.

Report is Including all information of review process (Referee's name, Comment, Objection, and Feedback)

Compare the checksum between ④, ④' and ⑭ to validate authenticity.

Figure 4.3: The editorial process for the Polar Data Journal.

(5) When the above process (2), (3), and (4) have been completed without any problems, the manuscript is sent to an editor.

(6) The editor checks the manuscript and the actual data.

(7) The editor selects two referees who are experts in the relevant fields.

(8) When they accept the review request, the selected referees become official referees and receive the manuscript and the link to the actual data.

(9) Referees scrutinize the content of the manuscript and prepare a review comment on the logicality of the manuscript, consistency with linked actual data, and the quality of the actual data.

(10) The referees reply to the editor with the acceptance or rejection of the manuscript with the referees’ comments.

(11) The editor confirms the referees’ comments and decides whether to accept or reject the manuscript as an editorial board member.

(12) The editor sends the decision results to the authors with the proofreading comments.

(13) Authors receive the decision results with the proofreading comments. If there are instructions for revision, authors revise the manuscript based on the comments and resubmit, and the set of processes is repeated.

(14) If accepted, the officer copies the actual data to the officer server again.

(15) The officer re-calculates the checksum of the actual data.

(16) After the checksum matches are confirmed before and after peer-review, the officer requests the data repository to issue a DOI for the actual data through the authors. The DOIs of the actual data will be included in the accepted data papers.

(17) The officer compiles all comments made by the editor and referees during the peer-review process, as well as corrections and rebuttals by the authors.

(18) Comments compiled by the officer will be published with the accepted data paper as a peer-reviewed report.

(19) The data paper will be published on the JAIRO Cloud.

Since authors can modify actual data, there remains the possibility that only the actual data may be changed during/after the peer-review period. If a version change of the actual data occurs during peer-review process, there will be a discrepancy between the description of the data paper and the actual state of the stored data. To avoid this situation, the Polar Data Journal requires authors to store their data statically from submission to acceptance. If authors wish to make additions or changes to the actual data submitted to the Polar Data Journal, they need to submit information about the differences. The mechanism for detecting irregular version changes is described in detail in Section 4.4.

4.3.4 System requirements

Next, we discussed system requirements to support each process. Since the data publishing system is expected to work with data repositories in multiple fields, the system must be designed to be flexible. We selected a mainly open-source system that allows flexible development. Conversely, since the peer-review flow is the same process as for existing research papers, we selected systems emphasizing the sustainability of development. As a result, we set the following three system requirements.

Peer-review process system

For the peer-review process system, it handles functions such as version control of data articles, assignment of peer-reviewers, management of peer-review schedules and comments, and notification of acceptance/rejection decisions that arise in the course of the peer-review process in addition to data papers and metadata associated with the papers. We adopted Editorial Manager, which is the proven online editorial workflow management system.

Quality assurance process system

For the quality assurance process system, it handles functions such as the management of research data and associated metadata, as well as functions such as issuing DOIs to research data and temporary data viewing during the peer-review period. We set the following three criteria for the quality assurance system: 1) Assignment of a persistent identifier (e.g., DOI), 2) Free access, and 3) Open license.

Unlike the previous peer-review process system, we did not adopt our own quality assurance process system and defined only the criteria on the system side. This is due to the following two reasons: a) research data management methods vary depending on the institution and/or field to which the authors belong, and b) that multiple data repositories have already been established within the field. Also, the previous three points must be clearly stated as operational policies.

The editorial office will check the link information provided by the authors and pre-screen the data repository. If the data repository does not meet the above criteria, the authors will be guided to resubmit the research data to the appropriate data repository. Note that this link need not be entirely public during the submission process; it only needs to be accessible to referees.

Data publishing process system

For the data publishing process system, it handles functions such as issuing DOIs to data papers that have been selected for acceptance, publishing peer-review reports, and backing up eligible research data. These requirements correspond to the processes in the Data Publishing category formalized in Chapter 3. We adopted the JAIRO Cloud service (Hayashi et al., 2021), developed by the National Institute of Informatics (NII).

4.4 Discussion

In this section, we discuss some issues that emerged during applying the model between multiple systems environments, including the possibility of inappropriate data manipulation in peer-reviews and how to prevent it.

Verification measure during peer-review process

We discuss inappropriate data manipulation that may occur during the peer-review period. Suppose the actual data in the data repository is altered or modified somehow, and the content differs from the actual data at the time of peer-review process. In that case, the reliability of the content described in the data paper cannot be certified. To ensure reliability, we added three steps in the original model: (4), (14), and (15) shown in Figure 4.4. The editorial office calculates a checksum of the actual data in each phase and compares both. By comparing the checksums, it is possible to verify that no modification or loss has occurred to the actual data in the data repository during the peer-review period.

Verification measure after peer-review process

We also discuss inappropriate data manipulation that may occur after the peer-review period. Any manipulations performed during peer-review must also be verifiable by a third party as necessary. Third parties cannot determine whether the data is reliable or not without accurate records; it is crucial for reproducibility to record the data curation process. We added two steps in the original model: (17) and (18) shown in Figure 4.4. The editorial office aggregates the history of each preceding process: recording the data curation process, including the start and end dates and times of each review process, the duration of the review, comments by the editor and referees, corrections and rebuttals by the authors, and the results of the checksum calculations. By publishing the peer-review report, anyone can check whether the peer-review process of the data paper was conducted correctly, which is expected to improve reliability.

4.5 Results and remaining issues in chapter 4

This chapter focused on publishing mechanisms for data papers that divert the journal peer-review. First, we evaluated the reference model for data publishing from a process perspective, positioning the model as a combination of existing quality assurance process and peer-review process. Furthermore, we conduct a technical implementation by interpreting RDA/WDS model as a case study. In applying the model, we clarified possible inappropriate activity between multiple systems environments in the data evaluation process. We developed two verification measures to deal with them. Through this study, we interpreted practically the relationship and boundaries between the existing quality assurance process and the peer-review process. This effort can also be positioned as an application-level interpretation of the processes in the Data Evaluation category formalized in Chapter 3.

We point out a limitation of this study regarding the processes in the Appraisal category formalized in Chapter 3. The tasks and procedures addressed in the data evaluation process are closely related to the risk management process and rights management process performed in the Appraisal category. Since the model does not explicitly mention the risk management process and rights management process, it can be assumed that it is under the researcher' s control. However, both processes are assumed to vary greatly depending on the nature of the handled research data. In the case of sensitive data, data curators may require involvement in the early stage. In order to apply this model to different fields with more diverse characteristics, it will be necessary to clarify the relationship between the processes in the Appraisal category and the Data Evaluation category.

5

Reframing the interdisciplinary appraisal processes by analyzing conditions of use of research data

In Chapter 5, we deal with the processes in the Appraisal category formalized in Chapter 3, focusing on the processes related to setting conditions of use. These processes have different variations due to the differences in legal restrictions and disciplinary norms by jurisdiction. Also, the conditions of use granted by data providers are more diverse. As a result, it is difficult for data re-users to interpret the results of the process accurately. This chapter investigates the actual processing status related to setting conditions of use in different fields and clarifies the correspondence of information tied to each formalized process. Furthermore, we conduct a practical implementation for stepwise interpretation of these processing results. This study reframes the processes in the Appraisal category and complements Chapter 4 practices each other.

5.1 Introduction

In Chapter 5, we deal with the processes in the Appraisal category formalized in Chapter 3, focusing on the processes related to setting conditions of use. These processes have different variations due to the differences in legal restrictions and disciplinary norms by jurisdiction (Ball, 2014). Also, the conditions of use granted by data providers are more diverse (Kindling et al., 2017). As a result, it is difficult for data re-users to interpret the results of the process accurately. Research data reuse based on ambiguous interpretation of processing results often leads to unintended reuse for the data providers. The conditions of use must be classified when research data are published in different fields (OECD, 2015a).

To accurately interpret the processing results involved in setting the conditions of use, it is necessary for data re-users to understand the actual status of processing from an interdisciplinary perspective. In recent years, there has been a move toward making data generated from public funds openly available in principle (ARDC, 2021; CODATA, 2019; Tsoukala et al., 2016). However, non-public funded data are not subject to policy implications. Standardization efforts based on the actual situation for conditions of use for data publishing has not been substantially explored.

This chapter investigates the actual processing status related to setting conditions of use in different fields and clarifies the correspondence of information tied to each formalized process. Furthermore, we conduct a technical investigation for stepwise interpretation of these processing results. This study is positioned as an effort to complement chapter 4 practices each other and support the building of an interdisciplinary infrastructure from a rule perspective.

5.2 Literature review

Following factors have been identified as occurring risks in research data sharing: lack of trust for proper use by data re-users, overly restrictive policies and unclear guidelines on research data sharing, and confusion over the ownership of data (Van Panhuis et al., 2014). The data curation activity involved in addressing these factors can be divided into two main processes: One is Risk Management, the process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information. This process also includes taking actions to reject or facilitate remediation when necessary. Another one is Rights Management, the process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse. As a result of the execution of these two processes, conditions of use are set.

Among the setting conditions of use, unclear copyrights and licenses are often a high barrier among the list of factors that hinder data reuse (Mayernik et al., 2020). A large international survey found that unsure about copyright and licensing (37%) are the second most crucial factor as the obstacles to publishing research data after organizing the data in a useful manner (46%) (Stuart et al., 2018). As a complementary report to the difficulty of interpreting conditions of use, the most widely used type of condition of use registered in re3data.org is "Other" at 57.2% and "Copyright" at 38.6%, with setting its own conditions of use or copyright notices being the most used situation (Kindling et al., 2017). The use of standardized tools, such as Creative Commons (CC) licenses, is limited to 21.8% at most. In some cases, CC licenses are granted for "non-copyrighted" works. The conditions of use for research data required by data providers are very diverse; hence, the interpretive costs for reuse are significant.

Initiatives for setting appropriate conditions of use have already been set up. The Digital Curation Centre mentioned early on the importance of promoting licensing as a way of maximizing the economic and social impact of data publishing (Ball, 2014). A survey conducted in 2017-2018 reported that 20 initiatives had been set to address ethical or legal issues (Grabus and Greenberg, 2019). Among these initiatives, the RDA/CODATA Legal Interoperability Interest Group (IG) published the article "Legal interoperability of research data: principles and implementation guidelines" (RDA-CODATA, 2016). These guidelines are primarily for data produced in or funded by the public sector and focus on legal interoperability to address misunderstanding and lack of knowledge and guidance on the legal issues generally related to research data.

In some fields, data that are not funded by public funds are widely used in research. In the case of non-public-funded data, the data are held by an individual or a company, and the conditions of use are determined by the data provider, except in cases provided by laws or regulations. A number of existing licensing tools can be applied to these data.

Table 5.1 shows a comparison of the CC and Open Data Commons (ODC) licenses and the “Government of Japan Standard Terms of Use” used for Japanese government websites.

Table 5.1: License comparison chart.

	Permissions			Requirements			Prohibitions
License	Reproduction	Distribution	Derivative Works	Notice	Attribution	Share Alike	Non-Commercial
CC0	X	X	X				
CC-BY	X	X	X	X	X		
CC-BY-SA	X	X	X	X	X	X	
CC-BY-NC	X	X	X	X	X		X
CC-BY-ND	X	X		X	X		
CC-BY-NC-SA	X	X	X	X	X	X	X
CC-BY-NC-ND	X	X		X	X		X
ODC-PDDL	X	X	X				
ODC-BY	X	X	X	X	X		
ODC-ODbL	X	X	X	X	X	X	
Government of Japan Standard Terms of Use	X	X	X	X	X		

These licenses are designed to provide flexible conditions of use, considering the copyright law. However, the current laws and regulations that may apply to research data are more complex and do not fully reflect the actual conditions. In some cases, CC licenses are granted to data that are not subject to protection under the Japanese copyright law, such as numerical data, which may cause confusion about reuse. While there is no doubt that CC licenses are still a useful solution in many cases of data publishing, some challenges exist for handling non-copyrighted data. Some proposals have recently been pushed forward (e.g., UK Scholarly communication license (Baldwin and Pinfield, 2018) and Microsoft Open Use of Data Agreement (https://github.com/microsoft/Open-Use-of-Data-Agreement)). Meanwhile, the requirements for legal decisions on privacy and other matters slightly differ from country to country. For example, fair use in the U.S. have not been introduced in Japan, and alternatives are still under consideration. The process of setting conditions of use needs to be considered in the local context.

5.3 Survey

This study aims to investigate the actual processing status related to setting conditions of use in different fields and clarifies the correspondence of information tied to each formalized process. To achieve this purpose, we set the following research questions (RQs) herein.

RQ1: What are the limitations that arise in using the research data?

RQ2: What do the data providers desire or request when they publish research data?

RQ3: What support can be effective in promoting reuse in aspects of the data providers’ desire or data re-users’ request?

The conditions of use of research data are a complex combination of external constraints and the data provider’ s requests. We start clarifying the two boundaries based on the use-case of actual conditions of use in various fields.

5.3.1 Interview survey

First, we conducted the interview survey with data repository practitioners to organize the conditions of use that could be granted according to the types of research data. Since the data provider’s requests were assumed to be more diverse, we aimed to identify external constraints across fields as a first step. We also aimed to obtain clues for the subsequent questionnaire survey.

In selecting interviewees, we targeted data repositories with data submission policies and researchers knowledgeable about data policies. The reason was that data policies generally include items that stipulate the conditions of use, and we assumed that we could collect information on the external constraints behind the enactment. As a result, the survey included five experts, including data repository managers and researchers from space science, environmental sciences, social sciences, materials science, and humanities. Note that the interview study was limited to Japan, since we think that external constraints must be judged in the local context.

The survey period was from December 12th, 2017 to February 1st, 2018 and was conducted for approximately one hour each. In conducting the survey, we stated the research request document before the interview and obtained permission. We used a topic guide to share our interview outline with the interviewee. We set three major sections within the topic guide: "Sharing and publishing of research data," "Regarding granting licenses to research data," and "Licensing of research data and promotion of legal interoperability." The topic guide used for the actual questions is shown in Appendix 4. Table 5.2 shows the summary of the results.

We coded in terms of data characteristics, data holders, requests for users, penal regulations, and rights protections that may affect Risk Management Process. We also provided columns to describe issues and aspirations related to conditions of use.

Typical data holders are the researchers from whom the data were obtained, their affiliated institutions, research funders, and third-party data suppliers. In various cases, it is unclear who can claim rights because of the passage of time or the circumstances of the funding agency. Penalties for violations have not been established or strictly enforced. As the scope of data sharing becomes more expansive, it seems to be regarded as less effective. The demand for rights protection varies by research field. In a research field that deals with both constrained and unconstrained data, there seems to be that the less constrained datasets tend to be more commonly used. Challenges in data repository operations included national security, sensitive data, and fostering a culture to achieve data protection. They also indicated a demand for an interdisciplinary understanding of these issues.

5.3.2 Questionnaire survey

Next, we conducted a web questionnaire survey based on the data obtained in Section 5.3.1. This survey aims to take an exploratory look at the data provider's request. We provided following ten questions without mandatory items:

(1) Which of the following terms best describes your research field?

(2) Have you ever obtained or published any data, including the cases in which user

Table 5.2: Summary of the interview survey results.

Date	2017/12/12	2017/12/18	2017/12/20	2018/1/30	2018/2/1
Fields	Space Physics	Environmental Sciences	Social Sciences	Materials Science	Humanities
Characteristics	Mostly numerical data	Image data, numerical data, etc..	Survey data	Measurement data, calculation data, Material Informatics data, software	Image data, bibliographic data
Data holder (Representative)	The person(s) who is/are acquired the data	The funder(s)	The data provider	Institution of the data holder (with exception)	uncertain
Requests for users	(None)	Request for user name and purpose of use for searching metadata ; Primary data are negotiated on an individual basis	Only available by the researchers who belongs in the social sciences discipline ; Submission of a research proposal is mandatory ; Submit usage reports every year / Inclusion in the acknowledgments for using the data	Provide a provenance information of the data as well as literature citations	The metadata should be CC0 ; The conditions of use of the data should be clearly stated by data holder
Penal regulations	There are no publishing constraints from a scientific point of view	Under consideration	If the data is passed on to a third party without permission, the use of the data may be suspended	(None)	Considering the use of rightsstatements in case of publishing data
Rights protections	No rights protection is provided for the data to be published	There are two levels of access restrictions on data in the repository, depending on the contents	One-year and indefinite licenses are available according to the wishes of the data holder	Data marked as private will have restricted access	Non-public data will be considered for protection on an individual basis by contract. Among the public data, those with copyright properties will be subject to government of Japan standard terms of use and CC licenses
Issues	There are few explanation for national security related data (especially the description of disclosure period)	There are no institution to consult about research data rights in Japan	In addition to an organization that supports data management, data management personnel and personnel who can handle the technical aspects of metadata are needed ; The criteria for determining sensitive data change over time, so it cannot provide past data as it is	A point of contact is needed to receive inquiries about published data	Licensing standards for publishing non-copyrighted or obscure data ; How to develop a culture of data protection, who will bear the cost, and how to spread it The future discussion points are whether or not to do so
Aspirations related to conditions of use	Development of data utilization laws	Enhancement of the university's intellectual property department function	(None)	(None)	(None)

registration and fees are required?

(3) Are you familiar with the following license tools?

(4) Have you ever used any of the following license tools to publish your data?

(5) If you would like to publish your data, would you like to require the following to your users (those who use that data to publish results)?

(6) If the license is complied with, would you be willing to publish the data?

(7) If you are using public data for your part of the research, please choose the method of presentation that you think is appropriate.

(8) Do you have any requests or concerns about using your published data for commercial activities, patents, press, literature, art, etc.?

(9) Please select the initiatives that you believe are desirable to the use and publishing of data.

(10) Free description (any problems or requests regarding the use or publishing of data).

In designing the survey, we created questions without the external constraints suggested by the interview survey. The survey period was set from February 13, 2018 to March 20, 2018. We distributed our questionnaire form via some mailing lists and websites using Questant’ s questionnaire system. The survey targeted mainly researchers, data managers, and librarians. The final number of responses is 413, of which 409 are valid responses. It should be noted that two limitations of this survey are as follows: (a) This survey was not a random selection. (b) The respondents’ research fields were biased from Social Science (17.4%) to Astronomy (0.2%). The aggregated results of the 409 valid responses are presented in the order of the questions presented before. The data from the survey are publicly available (Ikeuchi and Minamiyama, 2020). The “n” in the charts indicates the number of respondents.

(1) Property of respondents

Table 5.3 shows the research fields of the respondents. Social sciences (17.4%), earth sciences (12.5%), and humanities (10.3%) were well represented among respondents, while mathematics and astronomy were not (both 0.2%). Other responses included library and information science, nursing, nutrition, and so on. In some cases, affiliation information was recorded for library staff and private companies rather than research fields. Fifty-eight (14.2%) respondents selected, “I am not currently engaged in any research activities.”

Table 5.3: Research fields of respondents (n = 409).

Research Field	Number	Ratio
Social Sciences	71	17.4%
Earth Sciences	51	12.5%
Humanities	42	10.3%
Medicine	35	8.6%
Engineering	32	7.8%
Computer Science	20	4.9%
Biological Science	19	4.6%
Agricultural Science	18	4.4%
Psychology	16	3.9%
Physics	8	2.0%
Chemistry	2	0.5%
Mathematics	1	0.2%
Astronomy	1	0.2%
Other	35	8.6%
I am not currently engaged in any research activities	58	14.2%
Total	409	100.0%

(2) Experience in obtaining published data and publishing data by themselves

In this question, we asked for experience in obtaining published data and publishing data by themselves from the nine sources. The respondent’ s choices are “Obtain,” “Publish,” and “None” and set as follows: “Obtain” and “Publish” are multiple selections, and “None” cannot be selected when “Obtain” or “Publish” is selected. Table 5.4 shows the aggregate results.

Highly selected answers as regards “where to obtain” are institutional repositories/data archives (62.3%), government repositories/data archives (48.4%), and personal/research lab websites and blogs (47.9%). Highly selected answers about “where to publish” are institutional repositories/data archives (25.7%), personal/research

Table 5.4: Experience in obtaining published data and publishing data by themselves (n = 409).

Sources	Obtain	Publish	None	No Answer
Institutional repositories/data archives	62.3%	25.7%	29.1%	1.5%
Government repositories/data archives	48.4%	1.7%	46.0%	4.6%
Personal/research lab websites or blogs	47.9%	23.5%	41.8%	2.2%
Supplementary materials (in research paper)	36.7%	9.3%	54.3%	6.8%
Academic SNS services (e.g. Mendeley, ResearchGate)	32.0%	11.5%	58.2%	6.6%
Data repositories/archives in specific field	28.6%	8.3%	64.3%	5.4%
Code sharing services (e.g. GitHub)	24.4%	8.1%	69.2%	5.1%
Repositories/data archives by Commercial company	18.1%	1.5%	73.6%	7.1%
Other data publishing services (e.g. figshare, zenodo)	12.7%	3.7%	79.2%	6.8%

lab websites and blogs (23.5%), and academic SNS services (11.5%). Compared to the experience of obtaining data, the proportion of respondents with experience in publishing data is lower.

Table 5.5 presents the results of obtaining public data and having experience in releasing data. Respondents who selected "Yes" for one or more of the items in Table 5.4 are tabulated as having "Yes" experience in obtaining and publishing. Consequently, 84.1% of the respondents had experience in obtaining data, and 46.5% of the respondents had experience in publishing data. One respondent did not respond at all.

Table 5.5: Experience in obtaining published data and publishing data.

	Yes	No/No response	Total
Obtain	344 (84.1%)	65 (15.9%)	409 (100%)
Publish	190 (46.5%)	219 (53.5%)	409 (100%)

(3) Awareness of existing licenses

We asked for the awareness of three licenses, which are well known in Japan to identify the extent to which existing licenses are recognized. To eliminate answers based on fuzzy memories, we also set a link to the license or a page explaining the license in this question form. Figure 5.1 shows the aggregate results.

The highest recognition is for CC license, but less than half (46.9%) of the respondents are aware of it. ODC (19.3%) and Government Standard Terms of Use (15.9%) follow, and both are less than two in ten. The survey respondents would be expected to

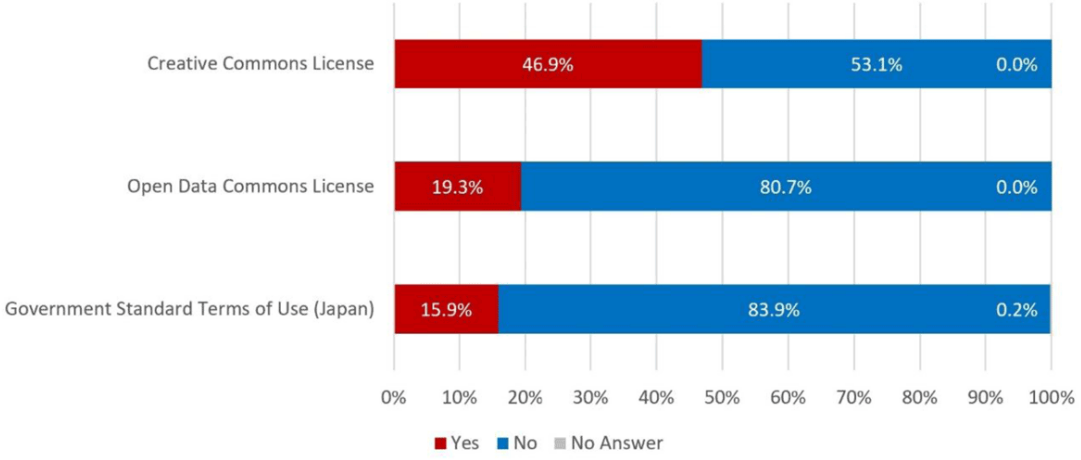


Figure 5.1: Awareness of existing licenses (n = 409).

have some level of interest in licensing their research data, but awareness of existing licenses was not high.

(4) Usage of existing licenses

To ascertain the use of the licenses listed in the previous question (3), we asked for respondents who are aware of each license about their experience in using each one. Figure 5.2 shows the aggregate results.

Fifty-nine respondents (30.7%) had used the CC license, which was the highest proportion, as was the case with (3). Only four (5.1%) and six (9.2%) respondents had experience using ODC and Government Standard Terms of Use.

(5) Desired condition of use when respondents publish their research data

We asked respondents to select their desired conditions of use from a list to quantify the extent of the requests they would make. We listed ten items to the possible requests; five items obtained during the interview survey and additional five items included in the CC license elements. Figure 5.3 depicts the aggregate results, with the following order: the sum of “Yes” and “It depends on cases” is the highest.

The aggregate results revealed a diverse reality; all categories were used to some degree. However, a certain degree of difference between preferred and non-preferred

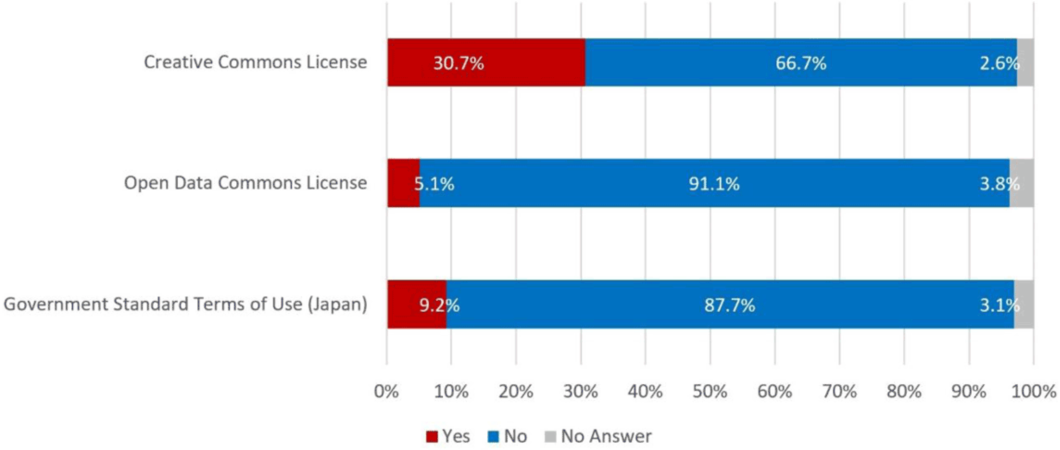


Figure 5.2: Usage of existing licenses (n = 409).

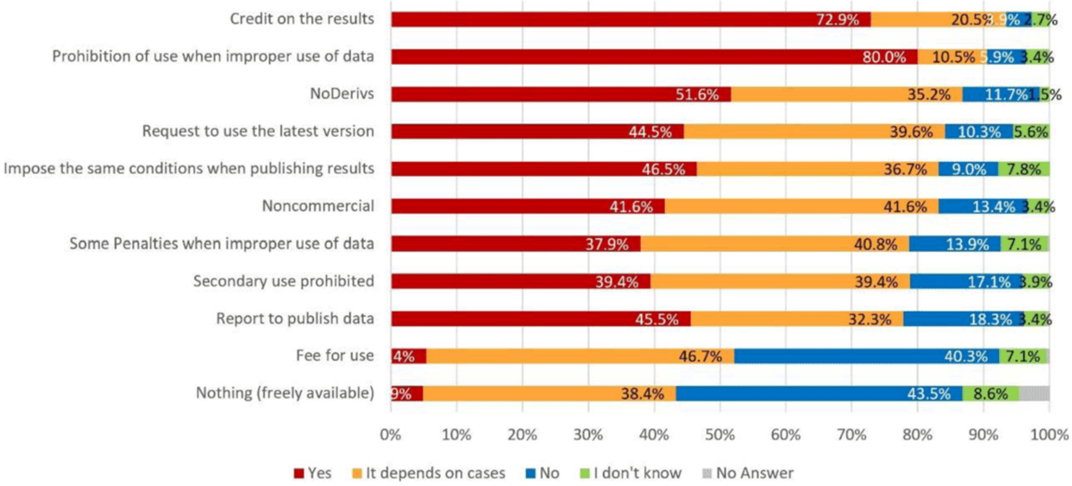


Figure 5.3: The desired conditions of use when respondents publish their research data (n = 409).

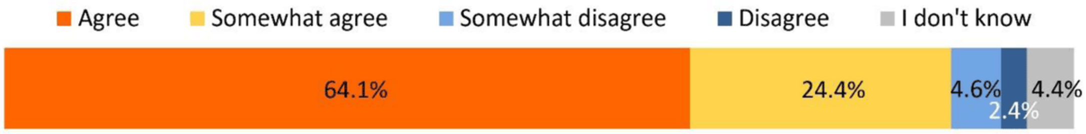


Figure 5.4: License compliance and willingness to publish data (n = 409).

categories can be observed. The highest percentage of "Yes" and "It depends on cases" is for "Credit on the results" (93.4%). The "Yes" percentage was higher than the credit indication (80.0%) for the "Prohibition of use when improper use of data." The total of "It depends on cases" was 90.5%. The items for which the total of "Yes" and "It depends on cases" exceeded 80% were "Request to use the latest version" (84.1%), "Impose the same conditions when publishing results" (83.1%), and "Noncommercial" (83.1%).

On the contrary, 43.5% of the respondents selected "No" when they asked for "Nothing (freely available)." In other words, just over 40% of respondents wanted to set some kinds of conditions to release their data. In addition, 40.3% of the respondents answered "No" to the question of "Fee for use," indicating that a certain number of respondents do not want to be compensated for data publishing. Note that there is an error of 0.1% between the Figure 5.3 and the main text due to rounding off numbers after the decimal point.

(6) License compliance and willingness to publish data

We asked whether they would be willing to publish their own data if the conditions of use listed in (5) were complied with. This question was asked to all respondents, including those whose data had already been exposed. Figure 5.4 depicts the aggregate results.

Consequently, 64.1% of the respondents said that they were "Agree," and 24.4% said they were "Somewhat agree" (total: 88.5%), exceeding "Somewhat disagree" (4.6%) and "Disagree" (2.4%).

(7) An appropriate method of displaying the use of published data

We asked the respondents about the display method they thought was appropriate for using published data by a third party. This question allows multiple choices. Table 5.6 presents the number and percentage of respondents who chose each option. Note that three respondents, who did not select any of the options, were excluded from the tabulation.

Table 5.6: Appropriate methods of displaying the use of published data (n = 406).

Choices	Numbers	Rates
Cite the source of the data in a paper (include it in the bibliography)	367	90.4%
Include source of the data information in the main text	224	55.2%
Include source of the data information in the acknowledgment	99	24.4%
Add the data provider name as a co-author	47	11.6%
It is not necessary to describe the data in a paper	0	0.0%

The highest selection rate was “Cite the source of the data in a paper (include it in the bibliography)” (90.4%). Most of the respondents judged that it would be appropriate to cite the data and the paper if the data were used. 55.2% of the respondents selected “Include source of the data information in the main text” as their next choice. None of the respondents selected, “It is not necessary to describe the data in a paper.”

(8) Requests and concerns about data reuse

We asked the respondents, “Do you have any requests or concerns if the data you’ ve published will be used for commercial activities, patents, press, literature, art, etc.?” in an additional comment space. This question aims to identify any other requests or concerns not raised in the literature or interview survey. As of 197 respondents’ major concerns were as follows: citation or indication of authorship (99 respondents), concern about misuse or inappropriate use (35 respondents), and concern about commercial use (14 respondents).

(9) Desired approach to data use and publishing

We asked the respondents about their preferred approach to data use and publishing. This question allows multiple choices. The choices were made with reference to the interview survey results. Table 5.7 presents the number and rate of respondents who chose each option. Note that five respondents who did not select any of the options were excluded.

Table 5.7: Desired approach for data use and publishing (n = 404).

Choices	Numbers	Rates
Establishment of standard data licenses (conditions of use)	312	77.2%
Development of appropriate guidelines for data licensing	285	70.5%
Establishment of a data licensing consultation, support, and management department (organization)	168	41.6%
Enabling a license to be specified in the data retrieval system	155	38.4%
Development of data rights legislation	148	36.6%
Establishment of a governing body for data licensing (external organization)	95	23.5%
Nothing in particular	21	5.2%

The highest rate was "Establishment of standard data licenses (conditions of use)" (77.2%), followed by "Establishment of guidelines for data licenses" (70.5%). Moreover, "establishment of a data licensing consultation, support, and management department (organization)" (41.6%) was selected higher than "establishment of a data licensing management organization (external organization)" (23.5%) as a contact point of data licensing issues.

(10) Free comments

A total of 84 respondents described the situation in the free comments. Regarding data publishing in general, various issues were pointed out, including inadequate systems, infrastructure, and technical difficulties and concerns about data publishing.

5.4 Analysis and design

This section discusses the correspondence between formalized processes and conditions of use based on two surveys in Section 5.3. The two possible reasons for not publishing research data are as follows: 1) external constraints, such as legal restrictions or disciplinary norms, and 2) data provider's request. We clearly separate the two and discuss the Risk Management process and the Rights Management process in relation to each. We also discuss the feasibility of these interpretations by the data re-user.

5.4.1 External constraints

In this section, we organize the external constraints regarding when to publish research data from the survey results in Section 5.3.1. Since the Risk Management process has been conducted on a field-by-field basis, individual risk assessments have been conducted for each keyword, such as personal information and intellectual property rights. We re-summarized these keywords into five categories from an interdisciplinary perspective. This categorization allows for a comprehensive understanding of what is considered in the Risk Management Process. Table 5.8 shows its category, definition, constraint subject matter, and some examples. Note that the examples described are not exhaustive.

Table 5.8: List of external constraints.

Category	Definition	Subject	Example
Discipline agreement and international treaties	Practices and standards in a specific discipline or research community that limit the data publishing. In some cases this is stated as an international treaty, but in others it is not always explicitly stated.	Disciplines and Norms	Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)
			Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property 1970

Continued on next page

Continued from previous page

Category	Definition	Subject	Example
			Convention Concerning the Protection of the World Cultural and Natural Heritage The Convention on the Protection and Promotion of the Diversity of Cultural Expressions The Nagoya Protocol on Access and Benefit-sharing Recommendation on the Safeguarding of Traditional Culture and Folklore Bereaved family' s request
Personal Information	It stipulates the handling of data that can identify individuals. It includes guidelines that define individual policies on anonymization and information disclosure.	Societies	The Personal Information Protection Commission, Government of Japan. "Laws and guidelines" (only in Japanese)

Continued on next page

Continued from previous page

Category	Definition	Subject	Example
			Japan External Trade Organization(JETRO). “About General Data Protection Regulation (GDPR)” (only in Japanese) Ministry of Health, Labor and Welfare (Japan). “About research guidelines” (only in Japanese)
Diplomatic / National security	Research data pertaining to national security. Data related to the development of weapons of mass destruction, etc. (as defined in the Foreign Exchange and Foreign Trade Act) and defense secrets (the Self-Defense Forces). law), important data that may affect national life (e.g., domestic energy (e.g., location of resources, blueprints for critical equipment, etc.).	State	Japan Society for Intellectual Production. “Security Trade Control Guidelines for Researchers in universities and other institutions of higher education. Revised 2nd ed”

Continued on next page

Continued from previous page

Category	Definition	Subject	Example
Agreements, contracts, Intellectual Property rights	An agreement with a research partner, contractor, etc. that restricts the data publishing in joint research or contract research.	Companies, etc.	Ministry of Economy, Trade and Industry (Japan). "Operation guidelines for data management in contract research and development" (only in Japanese)
			Ministry of Economy, Trade and Industry (Japan). "Contract Guidelines on Utilization of AI and Data. Data Section"
Data Policy	Where the research funder has a policy on limited data sharing for the research to be funded, or where a strategic business decision restricts the data publishing relating to pending industrial property rights or research data where the commercialization of the research results is envisaged.	Institutions	National Institute for Environmental Studies (Japan). "NIES Data Policy" (only in Japanese)

Continued on next page

Continued from previous page

Category	Definition	Subject	Example
			Teikyo University (Japan). "Intellectual Property policy in Teikyo University" (only in Japanese)
			Japan Agency for Medical Research and Development (Japan). "Data sharing policy for realization of genomic medicine" (only in Japanese)

End of table

1) Discipline agreement and international treaties

Data publishing is sometimes restricted by the discipline agreement of the field or research community, such as cases in which the data publishing causes harm to the research subject or cases in which the subsequent research activities themselves are severely affected. Although protection policies are established as international treaties in many cases, these policies are not always clearly defined, known, nor applied.

2) Personal Information

Data publishing is sometimes restricted to protect personal information by social demand. Its cases include disclosure, transfer, and anonymization restriction by relevant local laws, cross-border rules such as GDPR, and specific globalized guidelines.

3) Diplomatic/national security

Data publishing is restricted if the data is related to national security or international relations. These data are strictly operated in the global context, including the conditions of use.

4) Agreements, contracts, and intellectual property rights

Data publishing is sometimes restricted by agreements or contracts. In general, the parties' agreement determines jurisdiction over contracts and intellectual property rights; sometimes the conditions and the embargo period for data publishing are not uniformly determined because it is not a direct matter of concern.

5) Data Policy

Data publishing is restricted when there is a conflict with the policies of the data provider's organization or funding agency. There are many possible reasons for data policy restrictions; for example, a funding agency has a policy limiting data publishing for the research to be funded when a patent has been applied or the commercialization of research results is expected. These cases are restricted as an individual strategic decision and similar to previous data disclosure through contracts. The difference is that it is a management decision based on an "Open and Closed Strategy", which is a strategy to handle data by separating what should be released and what should be protected based on the characteristics of the data.

5.4.2 Data provider's request

This section discusses the data provider's requests subject to the Rights Management process. In Section 5.3.2, we observed quantitative differences among categories preferred and not preferred by the data providers. Therefore, we qualitatively analyzed the ten categories used in the questionnaire survey to understand the data provider's request diversity. Table 5.9 shows each condition of use analyzed from the data publishing perspective. This categorization will allow us to better understand the data provider's requests and give some direction to the Rights Management Process.

Note that items observed in the questionnaire survey that belong to the "Requests" category are not included in Table 5.9. "Requests" for the public are not legal contracts, but mainly moral matters; hence, no-obligation, prohibition, nor permission easily occur. The violation does not immediately imply termination of use, but data

Table 5.9: List of each condition of use analyzed from the perspective of expected users, duties, and constraints.

Condition of use	Expected user	Type of duty	Target of constraint	Compatible with CC licenses	Suggested categories
1) Waiver	Public	-	-	CC0	Preferable
2) Credit on the results (CC license term: Attribution)	Public	Obligation	Redistribution	BY	
3) Impose the same conditions when publishing results (CC license term: ShareAlike)	Public	Obligation	Redistribution and combination	SA (only for redistribution)	Available
4) Noncommercial	Public	Prohibition	Redistribution and data processing	NC (only for redistribution)	
5) NoDelivs	Public	Prohibition	Redistribution and data processing	ND	
6) Improper use of data	Public/Specific	Prohibition	Data processing	–	
7) Reporting	Public/Specific	Obligation	Continue to use	–	
8) Secondary use prohibited	Specific	Prohibition	Redistribution	–	Not Preferable
9) Request to use the latest version	Public (latest version only)	Obligation	Redistribution	–	
10) Fee for use	Specific	Obligation	Redistribution	–	

providers ask data re-users to comply as much as possible with the data provider’s request for the appropriate use of their data. A further study is needed.

Basis of categorization

This section discusses three categories shown in Table 5.9. From the data publishing perspective, an unspecified number of users must be given access, even if under some limited conditions. We categorized the targets assumed by each condition of use as “public,” “specific,” or both. If the conditions of use are only “specific,” we categorized these conditions of use as “Not Preferable” for data publishing.

We also analyze what obligations are imposed in the conditions of use and summarize the types and targets of these obligations. Consequently, we found two cases in which restrictions were placed on the “Redistribution of data” and on the “Data processing or related processes.” Restrictions on data processing or related processes are not desirable for research data reuse. We categorized the rest conditions of use as "Available" if they have restrictions on data processing or related processes. Conditions of use with no data constraints or only redistribution constraints were categorized as "Preferable." We also discuss the individual conditions of use in more detail below.

Preferable

1) Waiver

This declaration waives copyright and all other related rights, including any indication of the data provider's name; it can be evaluated to be definitely intended for the public. The cost to the data re-user is minimized because it is consistent with only the legal requirements. However, there seems not to include any incentives for data publishing in this declaration. Moreover, changing the conditions of use in a manner that makes them more stringent is extremely difficult, even when there is a request to prevent unwanted use due to changes in circumstances, such as increased property values of data. Although this declaration is ideal from the viewpoint of research data reuse, the number of data actually published may be limited.

2) Credit on the results (CC license term: Attribution)

This condition of use requires displaying the data provider' s name and data URL information. The cost to data re-users can be assessed as a negligible level because it remains a "minimal constraint" found in the openness debate (Open Knowledge, nd). We can evaluate this condition of use aiming for the public.

Relatedly, we observed from the questionnaire survey that the data citation expectation is particularly high. Crediting names could be a certain incentive for data providers; although the past practice has mentioned using the data in the acknowledgments or the text, data citation is meant to treat research data as an independent academic artifact.

Whereas, as data-specific concerns, it may be too costly and impractical for the data re-user to deal with many different data sources as the source data for data-intensive science (e.g., machine learning). Describing all credits in the presence of multiple dataset takes time and effort. One of the commenters stated that this should be resolved as a problem with citation and notation methods. Although it is out of the scope of this study, more discussions are underway at the Data Citation Synthesis Group in FORCE11 (FORCE11, 2014) and elsewhere.

3) Impose the same conditions when publishing results (CC license term: Share Alike)

This condition of use requires the same condition of use under redistribution and combination of multiple pieces of data. Unlike "Credit on the results," it may prevent them from combining the data with another dataset that have an incompatible license. Although it remains the "minimum constraint" in the openness debate for redistribution, it should be used with caution in research data reuse.

4) Noncommercial

This condition of use requires the "non-commercial" use of data. The habit of prohibiting commercial use is deeply rooted in the academic community, and it seems unavoidable given the significance of academia' s freedom from the society. There is another point of contention with this condition of use; the criteria for judgment fluctuate depending on people because "commercial use" is not clearly defined. For example, some people will have different interpretations of commercial use when selling visualizations derived from the data. The ambiguity of commercial use was also pointed out in discussions on copyrighted materials (Creative Commons, 2009). It should be used with a more clarifying scope of commercial use.

Available

5) NoDerivs (No Derivatives)

This condition of use prohibits data publishing after any modification. Although opening to the public is not restricted, the cost to the data re-user is relatively high due to the permitting process. In general, the data will be published for new knowledge through processing or combination. Furthermore, the effective case in which the prohibition of modification largely depends on the type of data and the manner of use (e.g., image data that are practically a work of art). From the viewpoint of research data reuse, it should be used in a limited manner.

6) Improper use of data

This condition of use prohibits “Improper use” of data in research data reuse. This condition of use has the problem that the meaning of "Improper use" is ambiguous. The inappropriate use in a legal context is limited in scope by the laws of each country. The inappropriate use of field conventions may exist, but sometimes, tacit knowledge remains for data re-users in different fields. As a result, this condition of use is difficult to be the basis for any restriction. Unclear conditions of use lead to contraction of usage; it should be used with more specific provisions.

7) Reporting

This condition of use requires post-reuse reporting, suggesting that the objective is to know detailed usage practices rather than mechanical access statistics. Although the data can be reused by both public and specific situations, the condition of use is stricter because of the proactive obligation. This usage condition will work well when the data re-user can be identified; in situations where the data re-user cannot be identified, it will remain at the same level of efficacy as a "request."

Not Preferable

8) Secondary use prohibited

This condition of use prohibits the secondary use of data. It clearly prohibits data redistribution, translation, or adaption and is intended for one-on-one use of its original form. The background for setting this condition of use is a mix of concerns about misuse/responsibility for quality and a desire to understand the data re-users accurately. Although the data processing or combination is not restricted, they cannot be re-distributed and have to be excluded from the definition of data publishing.

9) Request to use the latest version

This condition of use is used to limit the research data reuse to the latest one. Data in the past cannot be reused under this condition of use; thus, a large amount of data will be replaced when the latest data are published, resulting in marked costs of data usage. Also, it is impossible to know in advance when the condition will be violated, and the manner to notify the version update is very limited. The constraints for research data reuse are too strict to be included in the "Not Preferable" category.

10) Fee for use

This condition of use requires some fee for data reuse. The requirement of a user fee before data reuse is considered to be out of the scope of a condition of use that assumes that the data will be open to the public. The survey results also suggest that approximately half of the respondents are still uncomfortable with the act of monetizing data. However, given the sustainability of the data repository, monetization may be a major challenge in the future. The fee could be obtained in various ways, including shareware on a request basis, charging through a freemium model, download speed limits, and whether or not ads are displayed. In cases where the data provider itself requires a user fee, under what conditions the fee will be incurred must be clarified.

5.4.3 Feasibility of interpretation by the data re-user

The data re-user is required to understand the external constraints behind the granted conditions of use. From the data provider’ s perspective, compliance with the granted conditions of use leads to safe data publishing. However, as observed in Section 5.4.1, data re-user needs interdisciplinary and specialized knowledge to interpret these constraints or requests. In fact, we can see many concern on the topic of citation, misuse/inappropriate use, and commercial use in Section 5.3.2. The feasibility of interpretation by the data re-user remains an issue.

5.4.4 Section summary

This section discussed the correspondence between formalized processes and conditions of use based on two surveys in Section 5.3. In terms of the reasons preventing the data publishing, conditions of use can be divided into two categories: those resulting from external constraints and those resulting from the the data provider’s request.

The Risk Management process corresponds to setting conditions of use resulting from external constraints. We re-summarized external constraints into five categories from an interdisciplinary perspective. It highlighted constraints that should be judged locally and constraints that can be judged globally. This categorization helps to provide a comprehensive understanding of the matters covered by the Risk Management Process.

The Rights Management process corresponds to the setting of usage conditions resulting from the data provider's request. We observed quantitative differences among categories preferred and not preferred by the data providers. Therefore, we qualitatively analyzed the ten categories used in the questionnaire survey to understand the data provider's request diversity. This categorization will allow us to better understand the data provider's requests and give some direction to the Rights Management Process.

We also discussed the feasibility of these interpretations by the data re-user. Data re-users need to interpret interdisciplinary and specialized knowledge that differs from each field' s practice. This suggests that even if a data provider in one field accurately describes the conditions of use, the description may not be easy to understand for data re-users in different fields. Publishing research data may not progress without a mechanism to bridge this knowledge gap between the data providers and the data re-users.

5.5 Practical implementation

This section discusses practical implementation in promoting research data reuse in aspects of the data provider' s requests or data re-users' intention. There is a knowledge gap between data providers and data re-users in understanding the conditions of use, so we need to bridge this gap by implementing some mechanisms. To address this issue, we formalized the data publishing flow with licensing scenarios, including an expert consultation process. Furthermore, we also developed "Guidelines for specifying conditions of use in research data publishing" (Minamiyama et al., 2020) for a common understanding of this flow. A tentative English translation of the document is shown in Appendix 5.

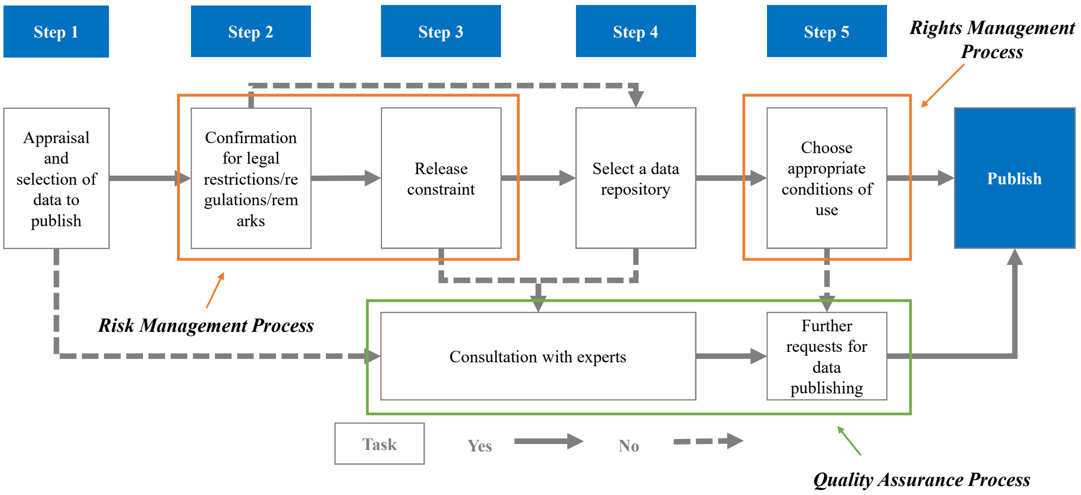

Step 1
Step 2
Step 3
Step 4
Step 5
Rights Management Process
Appraisal and selection of data to publish
Confirmation for legal restrictions/regulations/remarks
Release constraint
Select a data repository
Choose appropriate conditions of use
Publish
Risk Management Process
Consultation with experts
Further requests for data publishing
Task
Yes
No
Quality Assurance Process

Figure 5.5: Data Publishing flow with licensing scenarios.

5.5.1 Formalization of data publishing flow with licensing scenarios

From the analysis in Section 5.4, we mapped the Risk Management process and Rights Management process to the appropriate setting of conditions of use. Based on this understanding, we formalized a data publishing flow with licensing scenarios shown in Figure 5.5.

The flow consists of five steps. First, the data provider identifies the data to be published in Step 1. Next, the data provider confirms the external constraints from an interdisciplinary perspective in Step 2. In Step 3, for the external constraints identified in Step 2, the data provider set the necessary conditions of use for releasing constraint. Step 2 and step 3 are positioned as Risk Management processes. In Step 4, the data provider selects an appropriate data repository for the data judged to be open to the public. Steps 3 and 4 clearly state that expert consultation will be held because expert knowledge with an interdisciplinary perspective may be required for judgment. This process can also be understood as a part of the Quality Assurance process discussed in Chapter 4. Finally, the data provider chooses appropriate conditions of use in Step 5. Step 5 is positioned as Rights Management process. The details for each step are shown below.

Step 1: Appraisal and selection of data to publish

In this step, the data provider identifies various data used in the research, which can be curated and made available to the public. There are various types of data publishing motivations; mandated by publishers, funders, institutional policies, and researchers' intention. Also, some data types have already established methods of publication. The data provider is required to objectively understand their data as a prerequisite for the subsequent steps in this step.

Step 2: Confirmation for legal restrictions/regulations/remarks

In this step, the data provider considers whether or not the data identified in Step 1 falls under the five categories of external constraints shown in Section 5.4.1: 1) Disciplinary customs, including international treaty, 2) Personal information, 3) Diplomatic/national security, 4) Agreements, contracts, and intellectual property rights, and 5) Data policy.

The data provider extracts possible legal restrictions/regulations/remarks by category. If the target research data has not any category of concern, Step 3 can be skipped.

Step 3: Release constraint

In this step, the data provider sets the conditions of use for releasing constraints identified in Step 2. It includes limiting the number of users and setting an embargo period. The terms or periods set out here will be written as special conditions.

Step 3 also provides a route to consultation with experts. Even in cases where legal or disciplinary restrictions are imposed, restrictions may be lifted with appropriate data processing, such as anonymization or data release restrictions. Also, specific decisions on releases often require interdisciplinary expertise. Therefore, the involvement of experts should be envisaged from this step as a part of the Quality Assurance process discussed in Chapter 4.

There is another meaning to the involvement of experts in this step. There is no legal provision for the termination of the protection period for data as there is for copyrighted works. For example, even if the term of the collaboration agreement has expired, the data are not open to the public after any length of time unless the term is clearly defined. To prevent these unnecessary restrictions, this flow provides explicit steps for lifting the restrictions and encourages data providers to keep them to a minimum.

Step 4: Select a data repository

In this step, the data provider selects the repository where they want to publish the data. Well-known repositories/archives are likely to be the first candidate. If there are no representative repositories in each field, then the data provider’ s institutional repository or a generic repository such as Zenodo (https://zenodo.org/), figshare (https://figshare.com/) would be the following candidates. What should be checked in this step is whether the selected repository does not fall under the external constraints identified in Step 2. To complement Step 3, we describe a route to consult an expert in this step.

Step 5: Choose appropriate conditions of use

In this step, the data provider selects the appropriate conditions of use for data re-users and completes the data use covenants that set out conditions. The requests are more diverse than those for copyrighted works, and the situation has not yet been systematically organized. We propose data use covenants to be consistent with those protected by copyright law.

Note that there remains some concern regarding the complexity of conditions of use. Data use covenants can standardize the description of conditions of use to some extent, but they are not always simple. There are high demands for standardization and simple explanations. But having said that, the simple recommendation of open licenses avoiding the copyright problem will not promote research data use. If the data provider has further requests compared with copyrighted works, the data provider should choose a route to determine validity in consultation with experts.

5.5.2 Developing the guidelines for a common understanding of formalized flow

This section introduces the development of guidelines for a common understanding of formalized flow. In order to bridge the knowledge gap between data providers and data re-users, not only must the process be formalized, but the knowledge used in the process must also be shared. We developed the “Guidelines for specifying conditions of use in research data publishing” (Minamiyama et al., 2020) to help researchers and stakeholders understand and make appropriate publication decisions.

These guidelines provide necessary information and examples to consider when publishing research data with an interdisciplinary perspective. A critical feature of these guidelines are the inclusion of local implementation perspectives from experts. It should be possible to fill the knowledge gap by focusing the description on the local realities faced by data users. Therefore, some limitations exist in the versatility of the description. The main examples are listed below:

Scope of "research data" In Step 1, the data provider must identify their "research data" to publish. Although the scope of “research data” differs depending on the field of expertise, we defined “research data” as data that can be managed by digital means and released as research results and does not include physical objects such as samples, specimens, and recording media. In addition, although research articles and software can be treated as research data, this flow does not change or override the established methods for publishing in each content area. For example, CC licenses for paper publishing and GPL or other software licenses for software publishing are not an assumed target of the data handled by these guidelines. If a researcher has received research funding, they should follow the rule of the treatment of research data defined by the funding agency. These guidelines do not apply to such data; hence, their rules should be applied.

"well-known" repositories In Step 4, we recommend registration in well-known repositories. However, practical differences arise in this phase; "well-known" repositories differ by field and country. Although there are some famous registry sites such as re3data.org (https://www.re3data.org/) and FAIRsharing (https://fairsharing.org/), only a limited number of registrations are available in Japan. In light of this background situation, the guidelines provide both these registry sites and the list of recommended domestic data repositories in cooperation with the Japan Data Repository Network subcommittee under the RDUF (https://japanlinkcenter.org/rduf/).

Awareness of existing licensed tools In Step 5, the data provider selects the appropriate conditions of use. The guidelines prioritize practical use and present the preferable requirements consistent with existing licensed tools. Therefore, the additional categories of conditions of use revealed in the questionnaire survey analysis are not included. For example, we identify the “Preferable” conditions of use in Section 5.4. Still, we added “Impose the same conditions when publishing results,” “Noncommercial,” and “Noderivs” to the preferable requirements in the guidelines as a practical consideration. This decision corresponds to some contents that are difficult to distinguish from copyrighted works when giving conditions for data usage. A fundamental solution would be for research data to be managed under clear conditions of use from when it is created.

5.6 Results in chapter 5

In Chapter 5, we investigated the actual processing status related to setting conditions of use in different fields and clarified the correspondence of information tied to each formalized process. First, we conducted two surveys to investigate the actual processing status related to setting conditions of use in different fields. The conditions of use of research data are a complex combination of external constraints and the data provider’ s requests. We started clarifying the two boundaries based on the use-case of actual conditions of use in various fields. In the interview survey, we found the Risk Management process has been conducted on a field-by-field basis. Individual risk assessments have been conducted for each keyword, such as personal information and intellectual property rights. In the questionnaire survey, we observed differences among categories preferred and not preferred by the data providers quantitatively.

Through the analysis of the results of the two surveys, we discussed the correspondence between formalized processes and conditions of use. We clearly separated the external constraints and the data provider’s request, and we discussed the Risk Management process and the Rights Management process in relation to each. Regarding the Risk Management process, we re-summarized external constraints into five categories from an interdisciplinary perspective. It highlighted constraints that should be judged locally and constraints that can be judged globally. Regarding the Rights Management process, we qualitatively analyzed the ten categories used in the questionnaire survey to understand the data provider’s request diversity. This categorization will allow us to better understand the data provider’s requests and give some direction to the Rights Management Process. We also discussed the feasibility of these interpretations by the data re-user. Data re-users need to interpret interdisciplinary and specialized knowledge that differs from each field’ s practice.

Finally, we conducted a technical investigation for stepwise interpretation of these processing results. There is a knowledge gap between data providers and data re-users in understanding the conditions of use, so we need to bridge this gap by implementing some mechanisms. As a practical implementation, we developed a data publishing flow with licensing scenarios and guidelines based on our understanding of the formalized process. During the flow development process, we clarified the relationship to the Quality Assurance process discussed in Chapter 4. This study reframes the processes in the Appraisal category and complements Chapter 4 practices each other.

6
Conclusion

In this chapter, we discuss the results of this study and prospects.

6.1 Conclusion

This study aimed to interpret the data curation process in various fields from an interdisciplinary perspective. In chapter 1, we set the following two objectives:

Objective 1: Analysis and formalization of knowledge representing interdisciplinary data curation process

Objective 2: Practical studies to interpret the interdisciplinary data curation process

For Objective 1, we addressed this issue in chapter 3. As the first step to interpret the tasks and procedures performed in different fields at the same granularity, we investigated the practices of data curation conducted in each field. As a result, we found that about 87.2% of the processes are interpretable across multiple fields. Also, we realized that there needs a suitable model to describe the structure such as the relationships among Input-Output objects, processes, and staffing to accurately represent the data curation activity’ s structure in different fields.

Based on the preliminary analysis and survey results, we developed the Data Curation Process Ontology to formalize the data curation activity’ s structure in different fields. To verify the usefulness and validity of this ontology, we described and compared the several actual data curation activity’ s structures. Users can visualize the data curation activity’ s structure and the provenance information of the research data by using the format presented in Section 3.6. It is the important contribution of this study to compare the activity’ s structure of eight diverse repositories in a single model. Also, we showed that the ontology can use the specification of data curation process support functions by systems. This ontology can also help to allocate data curation process among multiple systems and improve the data curation process during data integration in different fields. With the formalized method enabled by the ontology, we can expect to expand the data curation process in various fields from an interdisciplinary perspective. Thus this study contributed to building a knowledge framework for a common understanding of the data curation process in different fields.

For Objective 2, we addressed this issue in chapter 4 and chapter 5. In chapter 4, we focused on publishing mechanisms for data papers that divert the journal peer-review. First, we evaluated the reference model for data publishing from a process perspective, positioning the model as a combination of existing quality assurance process and peer-review process. Furthermore, we conduct a technical investigation by implementing RDA/WDS model as a case study. In applying the model, we clarified possible inappropriate activity between multiple systems environments in the data evaluation process. We developed two verification measures to deal with them. Through this study, we examined practically the relationship and boundaries between the existing quality assurance process and the peer-review process in the Data Evaluation category. This effort can be positioned as an application-level representation of the processes in Data Evaluation category formalized in Chapter 3.

We point out a limitation of this study regarding the processes in the Appraisal category formalized in Chapter 3. The tasks and procedures addressed in the data evaluation process are closely related to the risk management process and rights management process performed in the Appraisal category. Since the model does not explicitly mention the risk management process and rights management process, it can be assumed that it is under the researcher’ s control. However, both processes are assumed to vary greatly depending on the nature of the handled research data. In the case of sensitive data, data curators may require involvement in the early stage. In order to apply this model to different fields with more diverse characteristics, it will be necessary to clarify the relationship between the processes in the Appraisal category and the Data Evaluation category.

In Chapter 5, we investigated the actual processing status related to setting conditions of use in different fields and clarified the correspondence of information tied to each formalized process. First, we conducted two surveys to investigate the actual processing status related to setting conditions of use in different fields. The conditions of use of research data are a complex combination of external constraints and the data provider’ s requests. We started clarifying the two boundaries based on the use-case of actual conditions of use in various fields. In the interview survey, we found the Risk Management process has been conducted on a field-by-field basis. Individual risk assessments have been conducted for each keyword, such as personal information and intellectual property rights. In the questionnaire survey, we observed differences among categories preferred and not preferred by the data providers quantitatively.

Through the analysis of the results of the two surveys, we discussed the correspondence between formalized processes and conditions of use. We clearly separated the external constraints and the data provider’s request, and we discussed the Risk Management process and the Rights Management process in relation to each. Regarding the Risk Management process, we re-summarized external constraints into five categories from an interdisciplinary perspective. It highlighted constraints that should be judged locally and constraints that can be judged globally. Regarding the Rights Management process, we qualitatively analyzed the ten categories used in the questionnaire survey to understand the data provider’s request diversity. This categorization will allow us to better understand the data provider’s requests and give some direction to the Rights Management Process. We also discussed the feasibility of these interpretations by the data re-user. Data re-users need to interpret interdisciplinary and specialized knowledge that differs from each field’ s practice.

Finally, we conducted a technical investigation for stepwise interpretation of these processing results. There is a knowledge gap between data providers and data re-users in understanding the conditions of use, so we need to bridge this gap by implementing some mechanisms. As a practical implementation, we developed a data publishing flow with licensing scenarios and guidelines based on our understanding of the formalized process. During the flow development process, we clarified the relationship to the Quality Assurance process discussed in Chapter 4. This study reframes the processes in the Appraisal category and complements Chapter 4 practices each other.

Through this study, we provided our framework for understanding data curation activities as processes from an interdisciplinary perspective. This framework allows the data curation process to be interpreted in a decoupled way of the original research context. Data re-users will be able to formally assess the increased interpretability by verifying that the processes included in the framework have been properly executed. Furthermore, we have demonstrated some practical implementations through our problem-solving approach as a stepwise formalization to the level of interpretation. We believe that as we move forward with these formalized efforts, we can improve the interpretability of research data and thereby contribute to the realization of open science.

6.2 Future works

One aspect of our approach is to reveal the bounded context of each field by intentionally providing a different perspective than the original contextual understanding. Even in the same data format, there are parts that can be understood in common and parts that cannot, and these unintelligible parts are the boundaries between fields. We hope to clarify the boundaries of expertise in the field by applying our frameworks. Also, in parallel with efforts in the leading fields, we would like to explore the possibility of using our frameworks outside the designated community, such as in different fields and industries. Through these efforts, we hope our framework will evolve beyond an interdisciplinary perspective to a transdisciplinary one in the future.

Bibliography

ARDC (2021). Sensitive data. https://ardc.edu.au/resources/working-with-data/sensitive-data/, (accessed 2022-5-11).

Assante, M., L. Candela, D. Castelli, and A. Tani (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal 15*.

Austin, C. C., T. Bloom, S. Dallmeier-Tiessen, V. K. Khodiyar, F. Murphy, A. Nurnberger, L. Raymond, M. Stockhause, J. Tedds, M. Vardigan, and A. Whyte (2016). Key components of data publishing: using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries 18*(2), 77–92.

Bahim, C., C. Casorrán-Amilburu, M. Dekkers, E. Herczog, N. Loozen, K. Repanas, K. Russell, and S. Stall (2020). The fair data maturity model: An approach to harmonise fair assessments. *Data Science Journal 19*.

Baker, K. S. and L. Yarmey (2009). Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation 4*(2), 12–27.

Baldwin, J. and S. Pinfield (2018). The uk scholarly communication licence: Attempting to cut through the gordian knot of the complexities of funder mandates, publisher embargoes and researcher caution in achieving open access. *Publications 6*(3).

Ball, A. (2010). Review of the state of the art of the digital curation of research data (version 1.2). https://purehost.bath.ac.uk/ws/portalfiles/portal/293012/erim1rep091103ab12.pdf, (accessed 2022-5-11).

Ball, A. (2012). *Review of Data Management Lifecycle Models.* UK United Kingdom: University of Bath. https://researchportal.bath.ac.uk/en/publications/review-of-data-management-lifecycle-models, (accessed 2022-5-11).

Ball, A. (2014). How to license research data. http://www.dcc.ac.uk/resources/how-guides, (accessed 2022-5-11).

Beagrie, N. (2008). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation 1*, 3–16.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet.* MIT Press.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World.* MIT Press.

Buneman, P., A. Chapman, and J. Cheney (2006). *A Provenance Model for Manually Curated Data.* Springer Berlin.

Cabinet Office, G. o. J. (2016). 5th science and technology basic plan. https://www8.cao.go.jp/cstp/kihonkeikaku/index5.html, (accessed 2022-5-11).

Cabinet Office, G. o. J. (2021). 6th science, technology, and innovation basic plan. https://www8.cao.go.jp/cstp/kihonkeikaku/index6.html, (accessed 2022-5-11).

Cabinet Office, Expert Panel on Open Science, G. o. J. (2015). Promoting open science in japan opening up a new era for the advancement of science. https://www8.cao.go.jp/cstp/sonota/openscience/, (accessed 2022-5-11).

Candela, L., D. Castelli, P. Manghi, and A. Tani (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology 66*(9), 1747–1762.

CASRAI (2019a). Data curation. https://codata.org/rdm-terminology/data-curation/, (accessed 2022-8-18).

CASRAI (2019b). Research data. https://codata.org/rdm-terminology/research-data/, (accessed 2022-8-18).

Chao, T. C., M. H. Cragin, and C. L. Palmer (2014). Data practices and curation vocabulary (dpcvocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology 66*(3), 616–633.

Chin, G. and C. S. Lansing (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, New York, NY, USA, pp. 409–418. Association for Computing Machinery.

CHORUS (2021). Publisher data availability policies index. https://www.chorusaccess.org/resources/chorus-for-publishers/publisher-data-availability-policies-index/, (accessed 2022-5-11).

Choudhury, S., C. Huang, and C. L. Palmer (2020). Updating the dcc curation lifecycle model. *International Journal of Digital Curation 15*(1), 12.

Claerbout, J. (2010). Reproducible computational research: A history of hurdles, mostly overcome. http://sepwww.stanford.edu/sep/jon/reproducible.html, (accessed 2022-5-11).

Clarke, D. J., L. Wang, A. Jones, M. L. Wojciechowicz, D. Torre, K. M. Jagodnik, S. L. Jenkins, P. McQuilton, Z. Flamholz, M. C. Silverstein, B. M. Schilder, K. Robasky, C. Castillo, R. Idaszak, S. C. Ahalt, J. Williams, S. Schurer, D. J. Cooper, R. de Miranda Azevedo, J. A. Klenk, M. A. Haendel, J. Nedzel, P. Avillach, M. E. Shimoyama, R. M. Harris, M. Gamble, R. Poten, A. L. Charbonneau, J. Larkin, C. T. Brown, V. R. Bonazzi, M. J. Dumontier, S.-A. Sansone, and A. Ma' ayan (2019). Fairshake: Toolkit to evaluate the fairness of research digital resources. *Cell Systems 9*(5), 417–421.

CODATA (2019). The beijing declaration on research data. https://zenodo.org/record/3552330, (accessed 2022-5-11).

Commission, E. (2016). Guidelines on fair data management in horizon 2020 ver.3.0. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, (accessed 2022-5-11).

Consultative Committee for Space Data Systems, C. (2012). Rerence model for an open archival information system (oais): Magenta book. https://public.ccsds.org/Pubs/650x0m2.pdf, (accessed 2022-5-11).

Consultative Committee for Space Data Systems, C. (n.d.). Oais reference model (iso 14721) (website). http://www.oais.info/, (accessed 2022-5-11).

Creative Commons, U. (2009). Defining "noncommercial" : A study of how the online population understands "noncommercial use" . https://mirrors.creativecommons.org/defining-noncommercial/Defining_Noncommercial_fullreport.pdf, (accessed 2022-5-11).

DataONE, P. P. i. S. R. W. G. (2013). Data management guide for public participation in scientific research. https://old.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf, (accessed 2022-5-11).

Edwards, P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science 41*(5), 667–690.

Faniel, I., E. Kansa, S. Whitcher Kansa, J. Barrera-Gomez, and E. Yakel (2013). The challenges of digging data: A study of context in archaeological data reuse. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pp. 295–304. Association for Computing Machinery.

Faniel, I. M. and T. E. Jacobsen (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW) 19*(3–4), 355–375.

Faniel, I. M., A. Kriesberg, and E. Yakel (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology 67*(6), 1404–1416.

Faundeen, J. L., T. E. Burley, J. A. Carlino, D. L. Govoni, H. S. Henkel, S. L. Holl, V. B. Hutchison, E. Martín, E. T. Montgomery, C. Ladino, S. Tessler, and L. S. Zolly (2014). The united states geological survey science data lifecycle model. https://doi.org/10.3133/ofr20131265, (accessed 2022-5-11).

Fecher, B., S. Friesike, and M. Hebing (2015). What drives academic data sharing? *PLOS ONE 10*(2), 1–25.

FORCE11 (2014). *Joint Declaration of Data Citation Principles.* https://www.force11.org/group/joint-declaration-data-citation-principles-final, (accessed 2022-5-11).

G8 (2013). G8 science ministers statement. https://www.gov.uk/government/news/g8-science-ministers-statement, (accessed 2022-5-11).

Grabus, S. and J. Greenberg (2019). The landscape of rights and licensing initiatives for data sharing. *Data Science Journal 18*(1), 29.

Gray, J., A. S. Szalay, A. R. Thakar, C. Stoughton, and J. vandenBerg (2002). Online scientific data curation, publication, and archiving. In A. S. Szalay (Ed.), *Virtual Observatories*, Volume 4846, pp. 103 – 107. International Society for Optics and Photonics: SPIE.

Griffin, P., J. Khadake, K. LeMay, S. Lewis, S. Orchard, A. Pask, B. Pope, U. Roessner, K. Russell, T. Seemann, A. Treloar, S. Tyagi, J. Christiansen, S. Dayalan, S. Gladman, S. Hangartner, H. Hayden, W. Ho, G. Keeble-Gagnere, P. Korhonen, P. Neish, P. Prestes, M. Richardson, N. Watson-Haigh, K. Wyres, N. Young, and M. Schneider (2018). Best practice data life cycle approaches for the life sciences [version 2; peer review: 2 approved]. *F1000Research 6*(1618).

Grootveld, M. and J. van Egmond (2012). Peer-reviewed open research data: Results of a pilot. *International Journal of Digital Curation 7*.

Hayashi, M., Y. Hayashi, M. Asaoka, M. Kawai, Y. Minamiyama, and K. Yamaji (2021). Development of national-level institutional repository cloud service for open science. https://doi.org/10.5281/zenodo.5442279, (accessed 2022-5-11).

Hemphill, L., A. Pienta, S. Lafia, D. Akmon, and D. A. Bleckley (2022). How do properties of data, their curation, and their funding relate to reuse? *Journal of the Association for Information Science and Technology*, 1–13.

Higgins, S. (2008). The dcc curation lifecycle model. *International Journal of Digital Curation 3*(1), 134–140.

Higgins, S. (2011). Digital curation: The emergence of a new discipline. *International Journal of Digital Curation 6*(2), 78–88.

Higgins, S. (2018). Digital curation: the development of a discipline within information science. *Journal of Documentation 74*(6), 1318–1338.

Howard, T., M. Darlington, A. Ball, S. Cully, and C. McMahon (2010). *Opportunities for and Barriers to Engineering Research Data Re-use.* Number ERIM Project Document erim3rep100805tjh10 in ERIM Project Document. UK United Kingdom: University of Bath.

Hrynaszkiewicz, I. and Y. Shintani (2014). Scientific data : An open access and open data publication to facilitate reproducible research. *Journal of Information Processing and Management 57*(9), 629–640.

Hudson-Vitale, C., I. Heidi, J. R. Lisa, C. Jake, K. Wendy, O. Robert, and S. Claire (2017). *Data Curation. SPEC Kit 354.* Association of Research Libraries.

Humphrey, C. (2006). *e-Science and the Life Cycle of Research.* https://era.library.ualberta.ca/items/3334684b-fa6a-4c9d-a74b-559fecd42f9f, (accessed 2022-5-11).

ICPSR (2021). Guide to social science data preparation and archiving : Best practice throughout the data life cycle 6th edition. https://www.icpsr.umich.edu/web/pages/deposit/guide/, (accessed 2022-5-11).

Ikeuchi, U. and Y. Minamiyama (2020). Data supporting “investigation and development of the workflow to clarify conditions of use for research data publishing in japan” version 1.0. https://doi.org/10.20736/00001468, (accessed 2022-5-11).

ISO (2012). Iso 14721:2012 space data and information transfer systems — open archival information system (oais) — reference model. https://www.iso.org/standard/57284.html, (accessed 2022-5-11).

Johnson, W. G. (2008). The icpsr and social science research. *Behavioral amp; Social Sciences Librarian 27*(3–4), 140–157.

Johnston, L. R. (Ed.) (2017). *Curating Research Data. Volume Two: A Handbook of Current Practice.* Association of College Research Libraries.

Johnston, L. R., J. Carlson, C. Hudson-Vitale, H. Imker, W. Kozlowski, R. Olendorf, and C. Stewart (2016). Definitions of data curation activities used by the data curation network. https://hdl.handle.net/11299/188638, (accessed 2022-5-11).

Jones, M. (2009). The cedars project. *Library and Information Research 26*(84), 11–16.

Kindling, M., H. Pampel, S. van de Sandt, J. Rücknagel, P. Vierkant, G. Kloska, M. Witt, P. Schirmbacher, R. Bertelmann, and F. Scholze (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine 23*(3/4).

Klump, J., R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Höck, M. Lautenschlager, U. Schindler, I. Sens, and J. Wächter (2006). Data publication in the open access initiative. *Data Science Journal 5*, 79–83.

Koso, A. (2013). Data sharing in the life sciences: How far we have come and where we must go. *Journal of Information Processing and Management 56*(5), 294–301.

Kowalczyk, S. and K. Shankar (2011). Data sharing in the sciences. *Annual Review of Information Science and Technology 45*(1), 247–294.

Kowalczyk, S. T. (2018). Modelling the research data lifecycle. *International Journal of Digital Curation 12*(2), 331–361.

Kratz, J. and C. Strasser (2014). Data publication consensus and controversies [version 3; peer review: 3 approved]. *F1000Research 3*(94).

Lawrence, B., C. Jones, B. Matthews, S. Pepler, and S. Callaghan (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation 6*(2), 4–37.

Lebo, T., S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao (2013). Prov-o: The prov ontology. https://www.w3.org/TR/prov-o/, (accessed 2022-5-11).

Lee, C. A. and H. R. Tibbo (2007). Digital curation and trusted repositories: Steps toward success. *Journal of Digital Information 8*(2).

Lord, P. and A. Macdonald (2003). e-science curation report: Data curation for e-science in the uk: an audit to establish requirements for future curation and provision. https://digitalpreservation.gov/news/2004/e-ScienceReportFinal.pdf, (accessed 2022-5-11).

Matsuzato, F. (1992). A brief history of the national institute of polar research library (in japanese). *2*(2).

Mayernik, M. S., K. Breseman, R. R. Downs, R. Duerr, A. Garretson, and C.-Y. S. Hou (2020). Risk assessment for scientific data. *Data Science Journal 19*.

Mayernik, M. S., S. Callaghan, R. Leigh, J. Tedds, and S. Worley (2015). Peer review of datasets: When, why, and how. *Bulletin of the American Meteorological Society 96*(2).

Mayernik, M. S., T. DiLauro, R. Duerr, E. Metsger, A. E. Thessen, and G. S. Choudhury (2013). Data conservancy provenance, context, and lineage services: Key components for data preservation and curation. *Data Science Journal 12*.

Minamiyama, Y., U. Ikeuchi, K. Ueshima, M. Suto, N. Okayama, I. Yamada, K. Ebisawa, H. Nakanishi, and Y. Kumazaki (2020). Guidelines for specifying conditions of use in research data publishing, ver. 1.0. https://doi.org/10.11502/rduf_license_guideline, (accessed 2022-5-11).

Murayama, Y. and K. Hayashi (2014). New trends in open science (part 1): International trends in science and technology and academic information sharing frameworks and open research data (in japanese). *Science Technology Trends* (146).

NISTEP (2015). New trends in open science (part 3): Trends in research data publishing and the promotion for sharing basis data of research papers (in japanese). *Science Technology Trends* (148).

Noy, N. and D. McGuinness (2001). Ontology development 101: A guide to creating your first ontology. *Technical Report SMI-2001-0880*.

OECD (2015a). Enquiries into intellectual property' s economic impact, chapter 7: Legal aspects of open access to publicly funded research in enquiries into intellectual property' s economic impact. https://www.oecd.org/sti/ieconomy/KBC2-IP.Final.pdf, (accessed 2022-5-11).

OECD (2015b). Making open science a reality. *OECD Science, Technology and Industry Policy Papers*. http://dx.doi.org/10.1787/5jrs2f963zs1-en, (accessed 2022-5-11).

OECD (2017). Business models for sustainable research data repositories. http://dx.doi.org/10.1787/302b12bb-en, (accessed 2022-5-11).

of Illinois Urbana-Champaign, U. (2017). Foundations of data curation. https://ischool.illinois.edu/degrees-programs/courses/is547, (accessed 2022-6-30).

Oostdijk, N., H. Van den Heuvel, and M. Treurniet (2013). The clarin-nl data curation service: Bringing data to the foreground. *International Journal of Digital Curation 8*(2), 134–145.

Open Knowledge, F. (n.d.). Open definition. https://opendefinition.org/, (accessed 2022-7-8).

Pampel, H., H. Pfeiffenberger, A. Schäfer, E. Smit, S. Pröll, and C. Bruch (2012). Report on peer review of research data in scholarly communication (part a of d 33.1. *Trends in Biotechnology*.

Peer, L., A. Green, and E. Stephenson (2014). Committing to data quality review. *International Journal of Digital Curation 9*(1), 263–291.

Piwowar, H. A. (2011). Who shares? who doesn't? factors associated with openly archiving raw research data. *PLOS ONE 6*(7), 1–13.

RDA-CODATA, L. I. I. G. (2016). *Legal Interoperability Of Research Data: Principles And Implementation Guidelines*. https://zenodo.org/record/162241, (accessed 2022-5-11).

Robinson-García, N., E. Jiménez-Contreras, and D. Torres-Salinas (2015). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology 67*(12), 2964–2975.

Rolland, B. and C. P. Lee (2013). Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pp. 435–444. Association for Computing Machinery.

Ruggles, S. (2017). The importance of data curation. *The Palgrave Handbook of Survey Research*, 303–308.

Sands, A., C. L. Borgman, L. Wynholds, and S. Traweek (2012). Follow the data: How astronomers use and reuse data. *Proceedings of the American Society for Information Science and Technology 49*(1), 1–3.

Spier, R. (2002). The history of the peer-review process. *Trends in Biotechnology 20.*

Stuart, D., G. Baynes, I. Hrynaszkiewicz, K. Allin, D. Penny, M. Lucraft, and M. Astell (2018). Whitepaper: Practical challenges for researchers in data sharing. *figshare.* https://figshare.com/articles/Whitepaper_Practical_challenges_for_researchers_in_data_sharing/5975011/1, (accessed 2022-5-11).

Sun, G. and C. S. G. Khoo (2017). Social science research data curation: issues of reuse. *Libellarium: journal for the research of writing, books, and cultural heritage institutions 9*(2).

Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE 6*(6), 1–21.

Tenopir, C., N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf, and R. J. Sandusky (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE 15*(3), 1–26.

Tibbo, H. R. (2012). Placing the horse before the cart: Conceptual and technical dimensions of digital curation. *Historical Social Research / Historische Sozialforschung 37*(3 (141)), 187–200.

Tsoukala, V., M. Angelaki, V. Kalaitzi, B. Wessels, L. Price, M. J. Taylor, R. Smallwood, P. Linde, J. Sondervan, S. Reilly, M. Noorman, S. Wyatt, L. Bigagli, R. Finn, T. Sveinsdottir, and K. Wadhwa (2016). Recode: Policy recommendations for open access to research data. https://zenodo.org/record/50863, (accessed 2022-5-11).

UK Data Archive, U. (n.d.a). Curation process (webiste). https://www.data-archive.ac.uk/managing-data/digital-curation-and-data-publishing/curation-process/, (accessed 2022-5-11).

UK Data Archive, U. (n.d.b). Quality control (webiste). https://www.data-archive.ac.uk/managing-data/digital-curation-and-data-publishing/quality-control/, (accessed 2022-7-8).

UNESCO (2021). Unesco recommendation on open science. pp. 1–34. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en, (accessed 2022-5-11).

Uschold, M. and M. Gruninger (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review 11*(2), 93–136.

Van de Sompel, H., S. Payette, J. Erickson, C. Lagoze, and S. Warner (2004). Rethinking scholarly communication. *D-Lib Magazine 10*(9).

Van Panhuis, W. G., P. Paul, C. Emerson, J. Grefenstette, R. Wilder, A. J. Herbst, D. Heymann, and D. S. Burke (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health 14*(1).

Venkatesan, A., N. Karamanis, M. Ide-Smith, J. Hickford, and J. McEntyre (2019). Understanding life sciences data curation practices via user research [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research 8*(1622).

Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison (2014). The availability of research data declines rapidly with article age. *Current Biology 24*(1), 94–97.

Wallis, J. C., C. L. Borgman, M. S. Mayernik, and A. Pepe (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation 3*(1), 114–126.

Wallis, J. C., E. Rolando, and C. L. Borgman (2013). If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology. *PLOS ONE 8*(7), 1–17.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. ’ t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data 3*(1).

Wilkinson, M. D., M. Dumontier, S.-A. Sansone, L. O. Bonino da Silva Santos, M. Prieto, D. Batista, P. McQuilton, T. Kuhn, P. Rocca-Serra, M. Crosas, and E. Schultes (2019). Evaluating fair maturity through a scalable, automated, community-governed framework. *Scientific Data 6*(1).

Willis, C., J. Greenberg, and H. White (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology 63*(8), 1505–1520.

Yoon, A. (2014). “making a square fit into a circle”: Researchers’ experiences reusing qualitative data. *Proceedings of the American Society for Information Science and Technology 51*(1), 1–4.

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology 53*(1), 1–6.

Zimmerman, A. S. (2008). New knowledge from old data. *Science, Technology, amp; Human Values 33*(5), 631–652.

Acknowledgments

I would like to express my sincere gratitude to my advisor Prof. Hideaki Takeda for his guidance. I got a lot of learning and awareness through daily discussions. I cannot count the times of progressing in my research with his rich knowledge and deep insights. This research would not have been possible without him.

I would like to show my deepest appreciation to Prof. Kazutsuna Yamaji for his dedicated support. He allowed me to balance research and work in my job, and I was able to lead a research life in a highly favorable environment. I also learned how to deal with my research as a researcher.

I would like to offer my special thanks to Prof. Ikki Ohmukai for his kindness. He paved the way for me to enter the SOKENDAI doctoral program, even though I have been conducting research in a different field from informatics. Furthermore, he provided valuable suggestions for defining the data curation process as the evaluation committee member.

I would like to express my gratitude to Prof. Asanobu Kitamoto, Prof. Hitoshi Okada, and Prof. Masao Takaku for serving on the evaluation committee for this thesis. Prof. Kitamoto especially commented on the structure and direction of this research. Prof. Okada especially commented on the consistency of this research with the multiple fields behind it. Prof. Takaku commented on the methodology of this research and the perspective of writing the thesis. All of their suggestions are insightful and enhance the completeness of this thesis.

I appreciate my two lab members, especially Prof. Yusuke Komiyama, Dr. Sungmin Joo, Dr. Phuc Nguyen, Dr. Masaharu Hayashi, Mr. Makoto Asaoka, and Mr. Koichi Ojiro, for their insightful discussions. Daily discussions with them helped me to deepen my research.

I appreciate Dr. Yukiko Sakai for her kindness and support. Not only did she give the opportunity to pursue my research in earnest, but she also encouraged me to enter the doctoral program.

I gratefully acknowledge my parents and family, who together decided to enter the doctoral program and have always been by my side to encourage me. Without their support, this challenge would not have been possible.

Last but not least, this research was made possible by collaborating with many people whose names could not be mentioned here. I would like to thank them all.

List of Publications

Journal Papers

(1) Yasuyuki Minamiyama, Hideaki Takeda, Masaharu Hayashi, Makoto Asaoka, and Kazutsuna Yamaji. A study on formalizing data curation activities across different fields. (submitted to Information Processing and Management).

(2) Yasuyuki Minamiyama, Ui Ikeuchi, Kunihiko Ueshima, Nobuya Okayama, and Hideaki Takeda (2020). Investigation and Development of the Workflow to Clarify Conditions of Use for Research Data Publishing in Japan. Data Science Journal, 19(1), p.53. http://doi.org/10.5334/dsj-2020-053

(3) Yasuyuki Minamiyama, Takeshi Terui, Yasuhiro Murayama, Hironori Yabuki, Kazutsuna Yamaji, and Masaki Kanao (2017). Launching a new data journal "Polar Data Journal": Toward a new data publishing framework for polar science. Journal of Information Processing and Management. 60(3) p.147-156. https://doi.org/10.1241/johokanri.60.147 (in Japanese)

Technical and Poster Papers

(1) Yasuyuki Minamiyama, Hideaki Takeda, Masaharu Hayashi, Makoto Asaoka, Kazutsuna Yamaji (2022). Development and utilization of data curation process ontology in 17th International Digital Curation Conference, (Poster).

(2) Y. Minamiyama, H. Takeda, M. Hayashi, M. Asaoka, K. Ojiro and K. Yamaji (2021). Structuring data curation activities by using ontology-based data modeling in

16th International Digital Curation Conference, (Poster).

(3) Yasuyuki Minamiyama, Akira Kadokura, Masaki Kanao, Takeshi Terui, Hironori Yabuki, Kazutsuna Yamaji (2017). "Polar Data Journal" ; A new data publishing platform for polar science in International Workshop on Sharing, Citation and Publication of Scientific Data across Disciplines, Joint Support-Center for Data Science Research (DS), (Poster).

(4) Yasuyuki Minamiyama, Takeshi Terui, Akira Kadokura, Masaki Kanao, Hironori Yabuki, Kazutsuna Yamaji (2017). Polar Data Journal by National Institute of Polar Research in JpGU-AGU Joint Meeting 2017, (Poster).

(5) Yasuyuki Minamiyama (2015). Data journal : A new approach to research data management. Current-Awareness. 325, pp.19-22. https://doi.org/10.11501/9497651 (in Japanese), (technical report).

Other works

(1) Yasuyuki Minamiyama (2022). Data Curation Process Ontology ver. 1.0. https://purl.archive.org/curation-ontology, (Dataset).

(2) Yasuyuki Minamiyama, Ui Ikeuchi, Kunihiko Ueshima, Misaki Suto, Nobuya Okayama, Issaku Yamada, Ken Ebisawa, Hodaka Nakanishi, and Yui Kumazaki (2020). Guideline for specifying conditions of use in research data publishing, ver. 1.0., https://doi.org/10.11502/rduf_license_guideline (Other).

Appendix 1. Topic guide of "Questions related to data curation activities"

We used a topic guide in Section 3.4.2 to share the specific phase of data curation activities with the interviewee. In the topic guide, we set the following nine questions referred to the previous study categories (Johnston, 2017). Note that the topic guide sent to interviewees was prepared in Japanese.

List of questions

1) Ingest
In this section, we will ask you about your data ingestion process details. This includes the identification of datasets available for registration, the procedure for receiving a deposit agreement, available media, and procedures for obtaining metadata and documentation.

2) Appraisal and Selection
In this section, we will ask you about your data appraisal and selection process details. This includes the identification of legal risks, such as the handling of personal information, arising from the characteristics of the data to be accepted, and the operation of the collection policies established by each repository.

3) Data Processing
In this section, we will ask you about your data processing process details. This

includes the storage and work space environment for processing data files before publishing, the scope of work logging and who created it, matters related to the software used to handle the data, and other data processing policies.

4) Data Storing
In this section, we will ask you about the system for storing data that has been processed. This includes the workflow up to preservation and the history management in each phase.

5) Metadata generation
In this section, we will ask you about handling metadata details submitted by data providers. This includes the validation, modifications, and additions that the data curator will make to the submitted metadata, distribution of the metadata envisioned, and the handling of metadata tied to a specific study.

6) Access level
In this section, we will ask you about the status of your repository’ s data access levels. This includes the options of access restrictions (e.g., affiliation, IP range, specific circumstances) and the types of conditions of use that are granted with description examples.

7) Long-term preservation
In this section, we will ask you about efforts for long-term preservation. This includes statements on preferred file formats, migration, format standardization/restructuring, emulation support, etc.

8) Re-evaluation and disposal
In this section, we will ask you about the metrics used to evaluate published data and how they are operationalized. This includes metrics information used as evaluation criteria (e.g., number of accesses, downloads, citations, papers linked to published data, etc.) and how the collected information is used fall under this category.

9) Other

In this section, we will ask you about the operation for conducting data curation activities. This includes the actual number of people involved in data curation activities and the allocation of specialized personnel, in addition to the number of cases processed per year and the time each process takes.

Appendix 2. List of data curation process description rationale

In Section 3.4.2, we read and referred each organization’ s data curation process manuals and related documents for the rationale for the activities to ensure consistency with the interview results. We mapped the specific description of the activities and the data curators’ information onto the working framework for those activities for which we were able to identify a description of the rationale for the activities. The following tables show the description of the rationale for the activities. We classified each data curation process mapped to the working framework into three levels: 1. Implemented, 2. Partially implemented, and 3. Not implemented. The statements on which the interviewee relies are masked: recorded as "Survey participant". Note that two of the MDR’ s processes have undergone changes involving category revisions since the time of the interview. We changed the mappings to match the currently available sources of information.

Table 1: List of data curation process description rationale (khirin).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	Survey participant	–
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	3	Survey participant	–
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text “Readme” file).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	1	Survey participant	–
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	Survey participant	–
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	Survey participant	–
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	–
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	1	Survey participant	–
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	3	Survey participant	–
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	Survey participant	–
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	1	Survey participant	–
16	Deidentific-ation	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	1	Survey participant	–
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	1	Survey participant	–
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	1	Survey participant	–
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	–
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of “null” and “blank” values, or unclear acronyms.	1	Survey participant	–
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	3	Survey participant	–
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	Survey participant	–
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	1	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	3	Survey participant	–
32	File download	Allow access to the data materials by authorized third parties.	3	Survey participant	–
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	3	Survey participant	–
34	Metadata brokerage	Active dissemination of a data set’ s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	3	Survey participant	–
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Survey participant	–
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	2	Survey participant	–
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	3	Survey participant	–
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	3	Survey participant	–
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	2	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	1	Survey participant	–

End of table

Table 2: List of data curation process description rationale (DiPLAS).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	Survey participant	Internal manual

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Survey participant	Internal manual
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	3	The National Museum of Ethnology will sign a memorandum of license agreement for each photographer, copyright holder, and owner of photographic materials.	http://diplas.jp/pdf/requirements.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	The members of the publicly solicited project will use this database to successively add to the available information and carry out the research plan for the Grant-in-Aid for Scientific Research. The database will contain the following items: (1) ID, (2) photographic image, (3) photographer, copyright holder, (4) time of photographing, (5) region of photographing (country and local name at the time of photographing), (6) ethnic name (if identifiable), (7) date of photographing, (8) image content tag, (9) related information (references, etc.), and (10) free text (Japanese and English in principle).	http://diplas.jp/pdf/requirements.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	2	Survey participant	–
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	We will also convert the basic information attached to the photographic materials into data. For data that has already been digitized, such as photographs taken with a digital camera, we will assign an ID (if necessary, we will rename the file and retrieve the basic information).	http://diplas.jp/pdf/requirements.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	2. Copyright of the academic materials themselves. ; 3. copyright of the image data (still and moving images) taken of the academic materials.	http://diplas.jp/pdf/outline_Guideline.pdf
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	4. Rights relating to persons and acts recorded in academic information.	http://diplas.jp/pdf/outline_Guideline.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	"8. Acceptance or rejection: (1) The review will be conducted by the Open Call Project Review Committee, which will be placed under the Platform Committee that oversees the entire ""Digital Picture Library for Area Studies"", and will decide which proposals to adopt."	http://diplas.jp/pdf/requirements.html
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	Internal manual

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	1	Survey participant	–
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	1	Survey participant	–
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	3	3. Digitization of images and registration of basic information: Serial numbers (IDs) are assigned to materials, and images are digitized and registered in the database.	http://diplas.jp/outline.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	Survey participant	Internal manual
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	1	Survey participant	–
16	Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3	Survey participant	Internal manual
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3	Survey participant	Internal manual
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	3	Survey participant	Internal manual
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	Internal manual
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	Survey participant	Internal manual

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.	3	Survey participant	Internal manual
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	1	Survey participant	–
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	Survey participant	–
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	1	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
32	File download	Allow access to the data materials by authorized third parties.	3	"The content entered into the database will be made available to the public if possible, after confirming the scope of use and rights in accordance with relevant laws and regulations, the National Museum of Ethnology's ""Guidelines for the Disclosure of Academic Information on the Internet"" and its ""Guidelines,"" based on the copyright holder's decision to make each image public or private, and the possibility that a photograph of a person may lead to a violation of rights or damage to dignity."	http://diplas.jp/outline.html
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
34	Metadata brokerage	Active dissemination of a data set' s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	3	Survey participant	Internal manual
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	3	"The content entered into the database will be made available to the public if possible, after confirming the scope of use and rights in accordance with relevant laws and regulations, the National Museum of Ethnology's ""Guidelines for the Disclosure of Academic Information on the Internet"" and its ""Guidelines,"" based on the copyright holder's decision to make each image public or private, and the possibility that a photograph of a person may lead to a violation of rights or damage to dignity."	http://diplas.jp/outline.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Survey participant	–
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Survey participant	–
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	3	Survey participant	–
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern webbrowsers)	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	1	Survey participant	–
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	2	Survey participant	–
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	1	Survey participant	–
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	2	Survey participant	–
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	1	Survey participant	–

End of table

Table 3: List of data curation process description rationale (Materials Data Repository).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	Users who already have a NIMS account (the account that you use to log in to desknet's, Denbun, and other NIMS systems) can use MDR using the same account.	https://dice.nims.go.jp/services/MDR/manual/html/login.html
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	3	Click the checkbox to indicate that you have understood and agree to the Deposit Agreement. If the box is checked, one of the items in the "Requirements" at the top right turns from a red "!" to a green check, indicating that you satisfied one of the requirements(Fig. 13).	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	Enter your metadata (title, creators, etc.) in the Descriptions tab. The following fields must be filled out if they are applicable.	https://dice.nims.go.jp/services/MDR/manual/html/metadata-data.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	2	Survey participant	–
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	Enter your metadata (title, creators, etc.) in the Descriptions tab. The following fields must be filled out if they are applicable.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	Copyright license or a statement of the condition that re-users of this work must adhere to (how others may reuse your work). For further details see the entry for license. If the work has already been released under a license elsewhere (e.g. from the publisher), select the one that applies.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	Enter the URL where we can confirm that your work is already public. (E.g. conference website, external database, etc.)	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	After you click on SUBMIT, MDR staff will review your deposit. If it has no issues regarding the terms of the deposit agreement and copyright policies, your work will be made public.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	–
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	Your work only contains the article: Publication ; Your primary intention is to make your paper available, with data accompanying the paper: Publication ; Your primary intention is to make your data/software available, with your paper documenting your data/software: Dataset ; You wish to describe your work using detailed metadata: Dataset	https://dice.nims.go.jp/services/MDR/manual/html/faq.html
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	Survey participant	–
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	1	Survey participant	–
16	Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	1	Survey participant	–
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	1	Survey participant	–
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	1	Survey participant	–
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	–
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	3	Your work will be assigned a new DOI ,and will be made public.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.	3	After you click on SUBMIT, MDR staff will review your deposit. If it has no issues regarding the terms of the deposit agreement and copyright policies, your work will be made public. In some circumstances, an admin may request changes before accepting your work.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	1	Survey participant	–
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Instruments used for the work.	https://dice.nims.go.jp/services/MDR/manual/html/metadata-data.html#instruments

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	Survey participant	–
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	3	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	1	Survey participant	–
32	File download	Allow access to the data materials by authorized third parties.	3	Open the Files tab from the top of the form and upload your files. You can use any of the following methods:	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	3	Users can discover publications and datasets using metadata tailored for materials or by a full-text search, and can view and download them.	https://mdr.nims.go.jp/about?locale=en

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
34	Metadata brokerage	Active dissemination of a data set’ s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	1→3	Datasets and software resource types on MDR are indexed by Clarivate Data Citation Index.	https://dice.nims.go.jp/services/MDR/manual/html/functions.html
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	1	You do not need to change this from MDR Open.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Embargoes are not available in MDR. Please deposit your work only after it is ready to be viewed without restrictions.	https://dice.nims.go.jp/services/MDR/manual/html/faq.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Copyright license or a statement of the condition that re-users of this work must adhere to (how others may reuse your work). For further details see the entry for license. If the work has already been released under a license elsewhere (e.g. from the publisher), select the one that applies.	https://dice.nims.go.jp/services/MDR/manual/html/deposit.html
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	2	Survey participant	–
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	1	Survey participant	–
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern webbrowsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	1	Survey participant	–
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	3	Survey participant	–
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	1→3	"We recommend uploading your new version as a new work and refer to your older version in ""Related item"" metadata. Select ""is new version of"" for the relationship and enter the old version's URL and link title."	https://dice.nims.go.jp/services/MDR/manual/html/faq.html

End of table

Table 4: List of data curation process description rationale (DARWIN).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	Survey participant	–
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	3	4.1. Before the voyage "Pre-voyage preparations for the handling data and samples are listed in Table 4.1. The forms for the pledge/agreement, metadata sheets and reporting documents will be sent by the management department prior to the voyage."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	3.2. Summary of Submissions "The types and summary of reporting documents, metadata, data, and samples to be submitted are listed in Table 3.2."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	2	3.3 Exceptions to the Data or Sample to be Submitted "In the event of any of the reasons listed in Table 3.3, the principal investigator may decide that submission is not required."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	3.2. Summary of Submissions "The types and summary of reporting documents, metadata, data, and samples to be submitted are listed in Table 3.2."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	Survey participant	–
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	Survey participant	–
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	2.2 Confirmation of data and sample handling regulations	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	–
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	3	Survey participant	–
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert “fixed” data (e.g., PDF formats) into machine-readable formats.	1	Survey participant	–
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	1. Introduction "This system has functions for inputting and outputting information via a Web browser, and enables management of metadata data, assistance with data publication tasks, and searching for past research voyages."	https://doi.org/10.5918/jamstecr.18.53
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
16	Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	3	Survey participant	–
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3	Survey participant	–
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3	Survey participant	–
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	3	Survey participant	–
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	–
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	3	Minting DOIs for Information of Research Cruises Disseminated through DARWIN	http://www.jamstec.go.jp/e/database/darwin_doi.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of “null” and “blank” values, or unclear acronyms.	3	Survey participant	–
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	3	Survey participant	–
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	6.1 Information for Proposal Applicants "The researcher who adopted the proposal will contact the department in charge whenever there are matters listed in Table 6.1.; (Metadata sheet content changes)"	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	3	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	3	5.3. Routine data reporting "We report the acquired data and observation information based on laws and regulations and agency collaboration, as listed in Table 5.3."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf
32	File download	Allow access to the data materials by authorized third parties.	3	3.1 Submission period and embargo period "The Principal Investigator or Project Leader will compile the data acquired during the voyage and the samples for submission and submit them to the responsible department."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
34	Metadata brokerage	Active dissemination of a data set' s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	3	Survey participant	–
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	3	Survey participant	–
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	3	3.1 Submission period and embargo period "The Principal Investigator or Project Leader will compile the data acquired during the voyage and the samples for submission and submit them to the responsible department."	https://www.jamstec.go.jp/ceist/e/datasample/JAM_DS_Handbook.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Method of Provision of Data and Samples "JAMSTEC will provide the information on the Data and Samples in an easily accessible manner, and will try to satisfy the needs and requirements of users." ; Pricing Policies of Data and Samples "The Data and Samples belonging to JAMSTEC will be available free of charge for scientific and educational uses in principle except for the actual providing costs. Industrial uses of Data and Samples will be subject to be charged appropriately in principle and they will be depending on the nature of the use."	http://www.jamstec.go.jp/e/database/data_policy.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	2	Survey participant	–
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	3	Survey participant	–
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	3	Survey participant	–
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	3	Survey participant	–

End of table

Table 5: List of data curation process description rationale (Global Environmental Database).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Survey participant	–
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	Survey participant	–
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	Survey participant	–
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	Survey participant	–
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	–
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	1	Survey participant	–
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	1	Survey participant	–
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	Survey participant	–
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	3	Survey participant	–
16	Deidentific-ation	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3	Survey participant	–
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3	Survey participant	–
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	3	Survey participant	–
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	–
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.	3	Survey participant	–
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	3	Survey participant	–
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	Survey participant	–
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	3	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	3	Survey participant	–
32	File download	Allow access to the data materials by authorized third parties.	3	Survey participant	–
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	1	Survey participant	–
34	Metadata brokerage	Active dissemination of a data set' s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	3	Survey participant	–
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Survey participant	–
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	2	Survey participant	–
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	1	Survey participant	–
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	1	Survey participant	–
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	3	Survey participant	–

End of table

Table 6: List of data curation process description rationale (Rikkyo University Data Archive).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	1. Contact about depositing the data "In the first instance, please contact RUDA if you are considering depositing data. Details of the deposition process will be explained by the relevant staff (In addition, the staff may ask you some questions regarding the data)."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Record of the work	Internal manual
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	3	4. Concluding the Deposit Agreement "Once RUDA receives the materials to be deposited, it will forward to you two copies of the "Deposit Agreement." After you carefully read the form, please sign and stamp (if possible) both forms and return one of them to RUDA via post (Please keep the other as a receipt)."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	3. Transfer of deposition items "On the basis of the content written in the "Deposition Checklist," RUDA will determine whether or not to accept the data. If it is deemed acceptable, RUDA will ask the depositor for basic information about the survey by forwarding a "Metadata Sheet" The information provided in the "Metadata Sheet" will be publicized on our website after the data has been deposited. After filling the "Metadata Sheet," please send it to RUDA via either post or email. After receiving the "Metadata Sheet," our staff will provide the details for transferring the data and related material."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	3	How to transfer with Proself	Internal manual

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	3. Transfer of deposition items "On the basis of the content written in the "Deposition Checklist," RUDA will determine whether or not to accept the data. If it is deemed acceptable, RUDA will ask the depositor for basic information about the survey by forwarding a "Metadata Sheet" The information provided in the "Metadata Sheet" will be publicized on our website after the data has been deposited. After filling the "Metadata Sheet," please send it to RUDA via either post or email. After receiving the "Metadata Sheet," our staff will provide the details for transferring the data and related material."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	Prior to depositing your data. "Before proceeding to deposit the data, please confirm the following requirements: The depositor holds the copyright, ownership rights, and the license of the data. ; No other party' s right are violated by depositing your data. ; Unless these requirements are satisfied, RUDA may refuse any offers to deposit the data."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	Prior to depositing your data. "Before proceeding to deposit the data, please confirm the following requirements: The depositor holds the copyright, ownership rights, and the license of the data. ; No other party' s right are violated by depositing your data. ; Unless these requirements are satisfied, RUDA may refuse any offers to deposit the data."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	3. Transfer of deposition items "On the basis of the content written in the “Deposition Checklist,” RUDA will determine whether or not to accept the data. If it is deemed acceptable, RUDA will ask the depositor for basic information about the survey by forwarding a “Metadata Sheet” The information provided in the “Metadata Sheet” will be publicized on our website after the data has been deposited. After filling the “Metadata Sheet,” please send it to RUDA via either post or email. After receiving the “Metadata Sheet,” our staff will provide the details for transferring the data and related material."	https://spirit.rikkyo.ac.jp/csi/RUDA/SitePages/index.aspx

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	Internal manual
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	3	3 Check with published materials	Internal manual
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	3	Cases of refusal (did) / consider (did)	Internal manual
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	Record of the work	Internal manual
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	3	Cleaning manual	Internal manual
16	Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	3	0.8 Delete unnecessary variables	Internal manual

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3	2. After data cleaning	Internal manual
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3	Double check	Internal manual
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	3	Survey participant	–
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	–
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	6.3 Creating and Writing a Modified Syntax	Internal manual

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	3	7 Logical check	Internal manual
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.	3	3 Check with published materials	Internal manual
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	3	1.5 Reordering Variables	Internal manual
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	Survey participant	–
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	3	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	3	Survey participant	–
32	File download	Allow access to the data materials by authorized third parties.	3	Guide to Downloading Data	https://spirit.rikkyo.ac.jp/csi/RUDA/userguidance/english/datadownload/Home.aspx
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
34	Metadata brokerage	Active dissemination of a data set' s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	3	Survey participant	–
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	3	Article 7. Prohibition of Use of User Account by Third Party "The User shall not cause or allow a third party to use the User account issued in response to his/her User registration. Further, the User shall make every effort to take necessary measures (such as password management) for preventing such use by a third party. "	https://spirit.rikkyo.ac.jp/csi/RUDA/userguidance/english/download/doc/agreement_e.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Article 11. Data Access Permit Period "1 The User shall promptly erase any Data for which the data access permit period has elapsed, and thereafter shall not access such Data." "2 If the User desires to continue accessing the Data, the User shall submit a new application for such access. "	https://spirit.rikkyo.ac.jp/csi/RUDA/userguidance/english/download/doc/agreement_e.pdf
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Article 3. Agreement to Terms of Use "The User shall be deemed to apply for access to the services provided by the Center and access the same after giving consent to the terms and conditions of these Terms of Use. "	https://spirit.rikkyo.ac.jp/csi/RUDA/userguidance/english/download/doc/agreement_e.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	2	Usage report	https://spirit.rikkyo.ac.jp/csi/RUDA/userguidance/english/report/Home.aspx
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	1	Survey participant	–
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	1	Survey participant	–
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	3	Survey participant	–

End of table

Table 7: List of data curation process description rationale (IUGONET).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	If you are interested, please contact with the IUGONET members.	http://www.iugonet.org/product/metadata.jsp?lang=en

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Survey participant	–
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	"When you first contact us, we will ask you the following questions regarding the data you plan to register: The name of the data set you are registering, A brief description of the data set, Name and affiliation of the Principal Investigator, Name and affiliation of the person responsible for creating the metadata"	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	2	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	The IUGONET Common Metadata Format (http://www.iugonet.org/mdformat.html) allows for the creation of metadata describing not only a single dataset, but also instruments, observation sites, human resources, and databases of real data. In addition to metadata describing a single data set, metadata describing instruments, observation sites, human resources, and databases of actual data are created independently and linked to each other.	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	Acknowledgment statement required when using the data set.	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	Access policy for the actual dataset (open to the public, restricted access, closed, etc.).	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	IUGONET welcomes your registration of the metadata of your observation data to our metadata database.	http://www.iugonet.org/product/metadata.jsp?lang=en

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	Survey participant	–
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	The IUGONET Common Metadata Format (http://www.iugonet.org/mdformat.html) allows for the creation of metadata describing not only a single dataset, but also instruments, observation sites, human resources, and databases of real data. In addition to metadata describing a single data set, metadata describing instruments, observation sites, human resources, and databases of actual data are created independently and linked to each other.	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	1	Survey participant	–
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	1	Survey participant	–
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	3	Survey participant	–
16	Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3	Survey participant	–
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3	Survey participant	–
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	3	Survey participant	–
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	3 How to submit your created metadata and confirm registration	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.	3	The sent metadata will be run through a check script on the IUGONET server.	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	1	Survey participant	–
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	Survey participant	–
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	1	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	1	Survey participant	–
32	File download	Allow access to the data materials by authorized third parties.	3	3 How to submit your created metadata and confirm registration	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
34	Metadata brokerage	Active dissemination of a data set' s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	1	Survey participant	–
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	1	Survey participant	–
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	Acknowledgment statement required when using the data set.	http://www.iugonet.org/data/manual/IUGONET_metadata_manual_v2_20170418.pdf
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	2	Survey participant	–
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	1	Survey participant	–
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	1	Survey participant	–
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	2	Survey participant	–
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	3	Survey participant	–

End of table

Table 8: List of data curation process description rationale (NBDC Archive).

No	Activity	Definition	Imp. levels	Rationale description	Link
1	Authentication	The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.	3	If you have any question or request about database archive and database deposit, please contact the follwing:	https://dbarchive.biosciencedbc.jp/contents-en/contact/contact.html
2	Chain of custody	Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.	3	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
3	Deposit agreement	The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.	3	The NBDC will check the archived data, metadata, and the wording of the license agreement, and approve the release of the archive.	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
4	Documentation	Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).	3	7 Metadata creation	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
5	File Validation	A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.	1	Survey participant	–
6	Metadata	Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).	3	7 Metadata creation	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
7	Rights management	The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).	3	8 License decision	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
8	Risk management	The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.	3	3.1.2 Databases that are not accepted for archiving	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
9	Selection	The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.	3	3.1 Archive Acceptance Policy	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
10	Arrangement and description	The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).	3	5.1.2 Designing the File Structure	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
11	Code review	Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.	3	Survey participant	–
12	Contextualize	Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.	3	Create detailed information (metadata) about the contents of the database and each item in the archive file.	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
13	Conversion (Analog)	In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.	1	Survey participant	–
14	Curation log	A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.	3	Survey participant	–
15	Data cleaning	A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.	3	Survey participant	–
16	Deidentification	Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.	3	3.1 Archive Acceptance Policy	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
17	File format transformations	Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.	3	Convert the original data into the archive file designed in (1).	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
18	Transcoding	With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
19	File inventory or manifest	The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.	3	Survey participant	–
20	File renaming	To rename files in a dataset, often to standardize and/or reflect important metadata.	3	5.2.2 Naming conventions for archive files	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
21	Indexing	Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.	3	Survey participant	–
22	Interoperability	Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.	3	5.1.3 Determination of data items	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
23	Peer-review	The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.	1	Survey participant	–
24	Persistent Identifier	A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.	3	DOI in Life Science Database Archive	https://dbarchive.biosciencedbc.jp/contents-en/doi/list.html

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
25	Quality assurance	Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of “null” and “blank” values, or unclear acronyms.	3	The NBDC will check the archived data, metadata, and the wording of the license agreement, and approve the release of the archive.	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
26	Restructure	Organize and/or reformate poorly structured data files to clarify their meaning and importance.	3	"5.1.3 Determination of data items ; 5.1.3.4 Integration and division of items"	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
27	Software registry	Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
28	Contact information	Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.	3	7.1 Database Metadata	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
29	Data citation	Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.	1	Survey participant	–
30	Data visualization	The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
31	Discovery Services	Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).	3	6 Create a simple search site (optional)	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
32	File download	Allow access to the data materials by authorized third parties.	3	About the Life Science Database Archive	https://dbarchive.biosciencedbc.jp/contents-en/about/about.html
33	Full-text indexing	Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
34	Metadata brokerage	Active dissemination of a data set' s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.	3	6 Create a simple search site (optional)	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
35	Restricted access	In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.	3	3.1.2 Databases that are not accepted for archiving	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
36	Embargo	To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
37	Terms of use	Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).	3	8 License decision	https://dbarchive.biosciencedbc.jp/files/nbdc_dbarchive_guidelines.pdf
38	Use analytics	Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.	3	Survey participant	–
39	Cease data curation	Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.	1	Survey participant	–
40	Migration	Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
41	Emulation	Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)	1	Survey participant	–
42	Secure storage	Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.	3	Survey participant	–
43	File audit	Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.	1	Survey participant	–

Continued on next page

Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
44	Repository certification	The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).	1	Survey participant	–
45	Succession planning	Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.	3	Survey participant	–
46	Technology monitoring and Refresh	Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.	3	Survey participant	–

Continued on next page

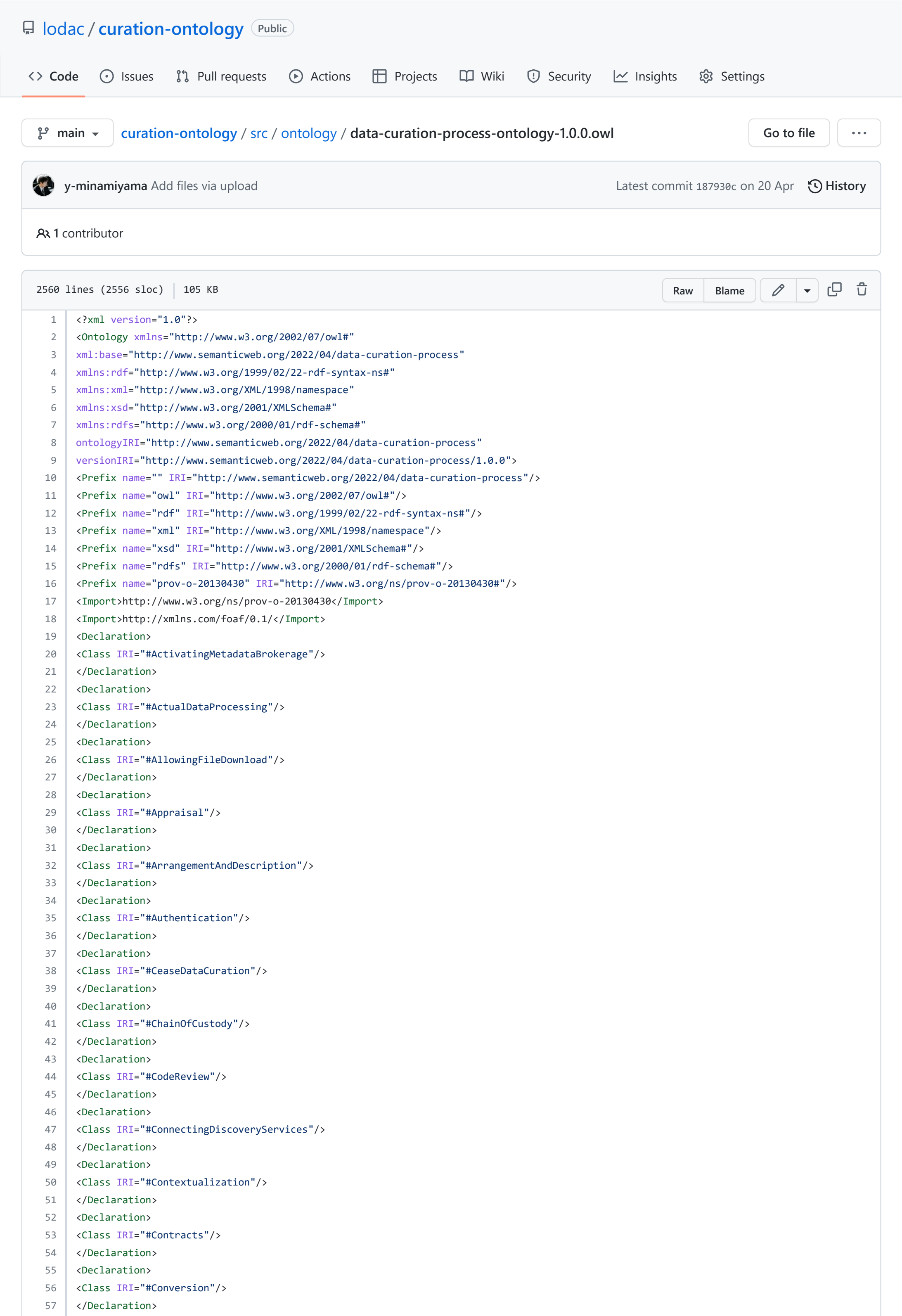
Continued from previous page

No	Activity	Definition	Imp. levels	Rationale description	Link
47	Versioning	Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.	3	Survey participant	–

End of table

Appendix 3. Data Curation Process Ontology

In Section 3.5, we developed the Data Curation Process Ontology. The ontology aims to represent commonalities and differences in data curation activity’ s structure. The domain to be covered by this ontology is that of data curation. Providing the structured data curation activity in a machine-readable format can support the knowledge-sharing process between humans and information systems in a scalable manner. It would be desirable to maintain the ontology through the collaboration of data curators in each field and ontologists who deal with knowledge sharing in information systems. The description of this ontology is based on the vocabulary of the PROV ontology (https://www.w3.org/TR/prov-o/) endorsed by W3C. The described activities were defined by the Data Curation Network, with some additions. The ontology is available at the following URL (https://purl.archive.org/curation-ontology). To open the ontology the program Protégé (https://protege.stanford.edu/) is recommended.



lodac / curation-ontology Public

curation-ontology / src / ontology / data-curation-process-ontology-1.0.0.owl

y-minamiyama Add files via upload

Latest commit 187930c on 20 Apr

1 contributor

2560 lines (2556 sloc) | 105 KB

```
<?xml version="1.0"?>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
xml:base="http://www.semanticweb.org/2022/04/data-curation-process"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
ontologyIRI="http://www.semanticweb.org/2022/04/data-curation-process"
versionIRI="http://www.semanticweb.org/2022/04/data-curation-process/1.0.0">
<Prefix name="" IRI="http://www.semanticweb.org/2022/04/data-curation-process"/>
<Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#"/>
<Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
<Prefix name="xml" IRI="http://www.w3.org/XML/1998/namespace"/>
<Prefix name="xsd" IRI="http://www.w3.org/2001/XMLSchema#"/>
<Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
<Prefix name="prov-o-20130430" IRI="http://www.w3.org/ns/prov-o-20130430#"/>
<Import>http://www.w3.org/ns/prov-o-20130430</Import>
<Import>http://xmlns.com/foaf/0.1/</Import>
<Declaration>
<Class IRI="#ActivatingMetadataBrokerage"/>
</Declaration>
<Declaration>
<Class IRI="#ActualDataProcessing"/>
</Declaration>
<Declaration>
<Class IRI="#AllowingFileDownload"/>
</Declaration>
<Declaration>
<Class IRI="#Appraisal"/>
</Declaration>
<Declaration>
<Class IRI="#ArrangementAndDescription"/>
</Declaration>
<Declaration>
<Class IRI="#Authentication"/>
</Declaration>
<Declaration>
<Class IRI="#CeaseDataCuration"/>
</Declaration>
<Declaration>
<Class IRI="#ChainOfCustody"/>
</Declaration>
<Declaration>
<Class IRI="#CodeReview"/>
</Declaration>
<Declaration>
<Class IRI="#ConnectingDiscoveryServices"/>
</Declaration>
<Declaration>
<Class IRI="#Contextualization"/>
</Declaration>
<Declaration>
<Class IRI="#Contracts"/>
</Declaration>
<Declaration>
<Class IRI="#Conversion"/>
</Declaration>
```

```xml
<Declaration>
<Class IRI="#Copyright"/>
</Declaration>
<Declaration>
<Class IRI="#CreatingLandingPage"/>
</Declaration>
<Declaration>
<Class IRI="#DataCleaning"/>
</Declaration>
<Declaration>
<Class IRI="#DataCurationActivities"/>
</Declaration>
<Declaration>
<Class IRI="#DataEvaluation"/>
</Declaration>
<Declaration>
<Class IRI="#DataLifecycleActivities"/>
</Declaration>
<Declaration>
<Class IRI="#DataPolicy"/>
</Declaration>
<Declaration>
<Class IRI="#DataPreservationActivities"/>
</Declaration>
<Declaration>
<Class IRI="#DataProcessing"/>
</Declaration>
<Declaration>
<Class IRI="#DataPublishing"/>
</Declaration>
<Declaration>
<Class IRI="#DataVisualization"/>
</Declaration>
<Declaration>
<Class IRI="#Deidentification"/>
</Declaration>
<Declaration>
<Class IRI="#DepositAgreement"/>
</Declaration>
<Declaration>
<Class IRI="#Digitization"/>
</Declaration>
<Declaration>
<Class IRI="#DiplomaticOrNationalSecurity"/>
</Declaration>
<Declaration>
<Class IRI="#DisciplinaryCustoms"/>
</Declaration>
<Declaration>
<Class IRI="#DisplayingDataCitation"/>
</Declaration>
<Declaration>
<Class IRI="#Documentation"/>
</Declaration>
<Declaration>
<Class IRI="#Embargo"/>
</Declaration>
<Declaration>
<Class IRI="#Emulation"/>
</Declaration>
<Declaration>
<Class IRI="#FileAuditing"/>
</Declaration>
<Declaration>
<Class IRI="#FileFormatTransformation"/>
</Declaration>
<Declaration>
<Class IRI="#FileInventoryOrManifest"/>
</Declaration>
<Declaration>
<Class IRI="#FileRenaming"/>
</Declaration>
<Declaration>
<Class IRI="#FileValidation"/>
```

```
</Declaration>
<Declaration>
<Class IRI="#FormulateSuccessionPlanning"/>
</Declaration>
<Declaration>
<Class IRI="#GeneratingFulltextIndexing"/>
</Declaration>
<Declaration>
<Class IRI="#IPR"/>
</Declaration>
<Declaration>
<Class IRI="#Indexing"/>
</Declaration>
<Declaration>
<Class IRI="#InformedConsent"/>
</Declaration>
<Declaration>
<Class IRI="#Ingest"/>
</Declaration>
<Declaration>
<Class IRI="#Interoperability"/>
</Declaration>
<Declaration>
<Class IRI="#MaintainingContactInformation"/>
</Declaration>
<Declaration>
<Class IRI="#MetadataGeneration"/>
</Declaration>
<Declaration>
<Class IRI="#MetadataProcessing"/>
</Declaration>
<Declaration>
<Class IRI="#Migration"/>
</Declaration>
<Declaration>
<Class IRI="#MintingPersistentIdentifier"/>
</Declaration>
<Declaration>
<Class IRI="#PeerReview"/>
</Declaration>
<Declaration>
<Class IRI="#PersonalInformation"/>
</Declaration>
<Declaration>
<Class IRI="#ProvidingRestrictedAccess"/>
</Declaration>
<Declaration>
<Class IRI="#QualityAssurance"/>
</Declaration>
<Declaration>
<Class IRI="#RegisteringSoftware"/>
</Declaration>
<Declaration>
<Class IRI="#Restructure"/>
</Declaration>
<Declaration>
<Class IRI="#RightsManagement"/>
</Declaration>
<Declaration>
<Class IRI="#RiskManagement"/>
</Declaration>
<Declaration>
<Class IRI="#SecuringStorage"/>
</Declaration>
<Declaration>
<Class IRI="#Selection"/>
</Declaration>
<Declaration>
<Class IRI="#SensitiveData"/>
</Declaration>
<Declaration>
<Class IRI="#SettingTermsOfUse"/>
</Declaration>
<Declaration>
```

```
<Class IRI="#SubmitData"/>
</Declaration>
<Declaration>
<Class IRI="#TechnologyMonitoringAndRefreshing"/>
</Declaration>
<Declaration>
<Class IRI="#TrackingUseAnalytics"/>
</Declaration>
<Declaration>
<Class IRI="#Transcoding"/>
</Declaration>
<Declaration>
<Class IRI="#TranscribingAndTranslatingData"/>
</Declaration>
<Declaration>
<Class IRI="#Versioning"/>
</Declaration>
<Declaration>
<Class IRI="#accessRestriction"/>
</Declaration>
<Declaration>
<Class IRI="#administrator"/>
</Declaration>
<Declaration>
<Class IRI="#agreement"/>
</Declaration>
<Declaration>
<Class IRI="#appropriateRepository"/>
</Declaration>
<Declaration>
<Class IRI="#authenticationResults"/>
</Declaration>
<Declaration>
<Class IRI="#backupData"/>
</Declaration>
<Declaration>
<Class IRI="#contactInformation"/>
</Declaration>
<Declaration>
<Class IRI="#costAndLeadTime"/>
</Declaration>
<Declaration>
<Class IRI="#creditOnTheResults"/>
</Declaration>
<Declaration>
<Class IRI="#curatedData"/>
</Declaration>
<Declaration>
<Class IRI="#curationLog"/>
</Declaration>
<Declaration>
<Class IRI="#curationRecord"/>
</Declaration>
<Declaration>
<Class IRI="#dataCurator"/>
</Declaration>
<Declaration>
<Class IRI="#dataDepositer"/>
</Declaration>
<Declaration>
<Class IRI="#dataDocument"/>
</Declaration>
<Declaration>
<Class IRI="#dataProcessingPolicy"/>
</Declaration>
<Declaration>
<Class IRI="#dataUser"/>
</Declaration>
<Declaration>
<Class IRI="#documentationPolicy"/>
</Declaration>
<Declaration>
<Class IRI="#evaluationResults"/>
</Declaration>
```

```
<Declaration>
<Class IRI="#expertiseNecessity"/>
</Declaration>
<Declaration>
<Class IRI="#externalConstraints"/>
</Declaration>
<Declaration>
<Class IRI="#externalServiceProvider"/>
</Declaration>
<Declaration>
<Class IRI="#feasibility"/>
</Declaration>
<Declaration>
<Class IRI="#feeForUse"/>
</Declaration>
<Declaration>
<Class IRI="#fileLocation"/>
</Declaration>
<Declaration>
<Class IRI="#fileValidationResults"/>
</Declaration>
<Declaration>
<Class IRI="#fullTextInformation"/>
</Declaration>
<Declaration>
<Class IRI="#imposeTheSameConditions"/>
</Declaration>
<Declaration>
<Class IRI="#improperUse"/>
</Declaration>
<Declaration>
<Class IRI="#indexingInformation"/>
</Declaration>
<Declaration>
<Class IRI="#landingPage"/>
</Declaration>
<Declaration>
<Class IRI="#linkingInformation"/>
</Declaration>
<Declaration>
<Class IRI="#metadata"/>
</Declaration>
<Declaration>
<Class IRI="#metadataCurator"/>
</Declaration>
<Declaration>
<Class IRI="#metadataSchema"/>
</Declaration>
<Declaration>
<Class IRI="#noDelivs"/>
</Declaration>
<Declaration>
<Class IRI="#nonCommercial"/>
</Declaration>
<Declaration>
<Class IRI="#peerReviewer"/>
</Declaration>
<Declaration>
<Class IRI="#persistentIdentifier"/>
</Declaration>
<Declaration>
<Class IRI="#policy"/>
</Declaration>
<Declaration>
<Class IRI="#preservationPolicy"/>
</Declaration>
<Declaration>
<Class IRI="#quality"/>
</Declaration>
<Declaration>
<Class IRI="#reporting"/>
</Declaration>
<Declaration>
<Class IRI="#repositorySystem"/>
```

```xml
</Declaration>
<Declaration>
<Class IRI="#researchData"/>
</Declaration>
<Declaration>
<Class IRI="#retrievalMetadata"/>
</Declaration>
<Declaration>
<Class IRI="#reusability"/>
</Declaration>
<Declaration>
<Class IRI="#secondaryUseProhibited"/>
</Declaration>
<Declaration>
<Class IRI="#selectionPolicy"/>
</Declaration>
<Declaration>
<Class IRI="#selectionResults"/>
</Declaration>
<Declaration>
<Class IRI="#size"/>
</Declaration>
<Declaration>
<Class IRI="#softwareRegistry"/>
</Declaration>
<Declaration>
<Class IRI="#sourceCode"/>
</Declaration>
<Declaration>
<Class IRI="#submittedData"/>
</Declaration>
<Declaration>
<Class IRI="#successionPlan"/>
</Declaration>
<Declaration>
<Class IRI="#systemAdministrator"/>
</Declaration>
<Declaration>
<Class IRI="#technicalInformation"/>
</Declaration>
<Declaration>
<Class IRI="#termsOfUse"/>
</Declaration>
<Declaration>
<Class IRI="#typeOfResource"/>
</Declaration>
<Declaration>
<Class IRI="#usageResults"/>
</Declaration>
<Declaration>
<Class IRI="#useLatestVersion"/>
</Declaration>
<Declaration>
<Class IRI="#versionInformation"/>
</Declaration>
<Declaration>
<Class IRI="#visualizedData"/>
</Declaration>
<Declaration>
<Class IRI="#waiver"/>
</Declaration>
<Declaration>
<ObjectProperty IRI="http://www.w3.org/ns/prov#Revision"/>
</Declaration>
<Declaration>
<AnnotationProperty IRI="http://www.w3.org/2004/02/skos/core#altLabel"/>
</Declaration>
<SubClassOf>
<Class IRI="#ActivatingMetadataBrokerage"/>
<Class IRI="#DataPublishing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActivatingMetadataBrokerage"/>
<ObjectSomeValuesFrom>
```

```
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#indexingInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActivatingMetadataBrokerage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#landingPage"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActivatingMetadataBrokerage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#externalServiceProvider"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<Class IRI="#DataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#visualizedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#accessRestriction"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataProcessingPolicy"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#curatedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#submittedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#ActualDataProcessing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataCurator"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<Class IRI="#DataPublishing"/>
```

```
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#versionInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#accessRestriction"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#sourceCode"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#AllowingFileDownload"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#fileLocation"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Appraisal"/>
<Class IRI="#DataCurationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Appraisal"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#administrator"/>
<Class IRI="#dataCurator"/>
<Class IRI="#repositorySystem"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#ArrangementAndDescription"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Authentication"/>
<Class IRI="#Appraisal"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Authentication"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#authenticationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
```

```
<Class IRI="#Authentication"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#contactInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Authentication"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDepositer"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#CeaseDataCuration"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#CeaseDataCuration"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#preservationPolicy"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CeaseDataCuration"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataCurator"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ChainOfCustody"/>
<Class IRI="#DataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#ChainOfCustody"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#curationLog"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#ChainOfCustody"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#submittedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#CodeReview"/>
<Class IRI="#DataEvaluation"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#CodeReview"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataDepositer"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CodeReview"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#sourceCode"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#ConnectingDiscoveryServices"/>
<Class IRI="#DataPublishing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#ConnectingDiscoveryServices"/>
<ObjectSomeValuesFrom>
```

```
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#fullTextInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ConnectingDiscoveryServices"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#externalServiceProvider"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#ConnectingDiscoveryServices"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#indexingInformation"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Contextualization"/>
<Class IRI="#MetadataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Contextualization"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#linkingInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Contracts"/>
<Class IRI="#externalConstraints"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Conversion"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Copyright"/>
<Class IRI="#Contracts"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Copyright"/>
<ObjectExactCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#qualifiedUsage"/>
<Class IRI="#termsOfUse"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<Class IRI="#DataPublishing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#agreement"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#contactInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#fileLocation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
```

```
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#indexingInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#persistentIdentifier"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#selectionResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#versionInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#repositorySystem"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#CreatingLandingPage"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#landingPage"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataCleaning"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataCurationActivities"/>
<Class IRI="#DataLifecycleActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataCurationActivities"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#curationRecord"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataEvaluation"/>
<Class IRI="#DataCurationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataEvaluation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#evaluationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataEvaluation"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#dataCurator"/>
<Class IRI="#peerReviewer"/>
```

```
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataLifecycleActivities"/>
<Class IRI="http://www.w3.org/ns/prov#Activity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataPolicy"/>
<Class IRI="#externalConstraints"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataPreservationActivities"/>
<Class IRI="#DataLifecycleActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataProcessing"/>
<Class IRI="#DataCurationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataProcessing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#dataCurator"/>
<Class IRI="#dataDepositer"/>
<Class IRI="#metadataCurator"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataPublishing"/>
<Class IRI="#DataCurationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataPublishing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#evaluationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataPublishing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#dataCurator"/>
<Class IRI="#repositorySystem"/>
<Class IRI="#systemAdministrator"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#DataVisualization"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Deidentification"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DepositAgreement"/>
<Class IRI="#Ingest"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DepositAgreement"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#selectionPolicy"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#DepositAgreement"/>
<ObjectExactCardinality cardinality="1">
```

```xml
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#agreement"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Digitization"/>
<Class IRI="#Conversion"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DiplomaticOrNationalSecurity"/>
<Class IRI="#externalConstraints"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DisciplinaryCustoms"/>
<Class IRI="#externalConstraints"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#DisplayingDataCitation"/>
<Class IRI="#CreatingLandingPage"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Documentation"/>
<Class IRI="#Ingest"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Documentation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#documentationPolicy"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Documentation"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#dataDocument"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Embargo"/>
<Class IRI="#AllowingFileDownload"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Emulation"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Emulation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#softwareRegistry"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Emulation"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#systemAdministrator"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileAuditing"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileAuditing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#fileValidationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileAuditing"/>
<ObjectMinCardinality cardinality="1">
```

```
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curatedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileAuditing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#systemAdministrator"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileAuditing"/>
<ObjectExactCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasInformedBy"/>
<Class IRI="#FormulateSuccessionPlanning"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileFormatTransformation"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileInventoryOrManifest"/>
<Class IRI="#DataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileInventoryOrManifest"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#curationLog"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileInventoryOrManifest"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileRenaming"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileValidation"/>
<Class IRI="#Appraisal"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileValidation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#fileValidationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileValidation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#sourceCode"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FileValidation"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#submittedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#FormulateSuccessionPlanning"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
```

```xml
<Class IRI="#FormulateSuccessionPlanning"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#successionPlan"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FormulateSuccessionPlanning"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#preservationPolicy"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#FormulateSuccessionPlanning"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#administrator"/>
<Class IRI="#dataCurator"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#GeneratingFulltextIndexing"/>
<Class IRI="#DataPublishing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#GeneratingFulltextIndexing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curatedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#GeneratingFulltextIndexing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#GeneratingFulltextIndexing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#submittedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#GeneratingFulltextIndexing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#fullTextInformation"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#IPR"/>
<Class IRI="#Contracts"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Indexing"/>
<Class IRI="#DataPublishing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Indexing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Indexing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
```

```
<Class IRI="#linkingInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Indexing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#indexingInformation"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Indexing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#retrievalMetadata"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#InformedConsent"/>
<Class IRI="#Contracts"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Ingest"/>
<Class IRI="#DataCurationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Ingest"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataDepositer"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Interoperability"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#MaintainingContactInformation"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#MaintainingContactInformation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#contactInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MaintainingContactInformation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataCurator"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MaintainingContactInformation"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAttributedTo"/>
<Class IRI="#dataDepositer"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MaintainingContactInformation"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#contactInformation"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataGeneration"/>
<Class IRI="#MetadataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataGeneration"/>
```

```
<ObjectExactCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#retrievalMetadata"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataProcessing"/>
<Class IRI="#DataProcessing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#metadataSchema"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MetadataProcessing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#dataDepositer"/>
<Class IRI="#metadataCurator"/>
</ObjectUnionOf>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Migration"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Migration"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#curatedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Migration"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curatedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Migration"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<ObjectUnionOf>
<Class IRI="#dataCurator"/>
<Class IRI="#repositorySystem"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#MintingPersistentIdentifier"/>
<Class IRI="#DataPublishing"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#MintingPersistentIdentifier"/>
```

```
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#persistentIdentifier"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MintingPersistentIdentifier"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#indexingInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#MintingPersistentIdentifier"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<Class IRI="#DataEvaluation"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curationLog"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#metadata"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#sourceCode"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curatedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#PeerReview"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#peerReviewer"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#PersonalInformation"/>
<Class IRI="#externalConstraints"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#ProvidingRestrictedAccess"/>
<Class IRI="#AllowingFileDownload"/>
</SubClassOf>
<SubClassOf>
```

```
<Class IRI="#QualityAssurance"/>
<Class IRI="#DataEvaluation"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#QualityAssurance"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curatedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#QualityAssurance"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#QualityAssurance"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#metadata"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#QualityAssurance"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#sourceCode"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#QualityAssurance"/>
<ObjectAllValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataCurator"/>
</ObjectAllValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RegisteringSoftware"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#RegisteringSoftware"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#softwareRegistry"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RegisteringSoftware"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RegisteringSoftware"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#retrievalMetadata"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RegisteringSoftware"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#dataCurator"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Restructure"/>
<Class IRI="#ActualDataProcessing"/>
</SubClassOf>
```

```
<SubClassOf>
<Class IRI="#RightsManagement"/>
<Class IRI="#Appraisal"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#RightsManagement"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#termsOfUse"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RightsManagement"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RiskManagement"/>
<Class IRI="#Appraisal"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#RiskManagement"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#accessRestriction"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RiskManagement"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#externalConstraints"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#RiskManagement"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<ObjectUnionOf>
<Class IRI="#dataDocument"/>
<Class IRI="#submittedData"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#SecuringStorage"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#SecuringStorage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#backupData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#SecuringStorage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#metadata"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#SecuringStorage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#SecuringStorage"/>
<ObjectMinCardinality cardinality="1">
```

```
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#systemAdministrator"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<Class IRI="#Appraisal"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#accessRestriction"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#authenticationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#fileValidationResults"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#selectionPolicy"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#termsOfUse"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Selection"/>
<ObjectAllValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#selectionResults"/>
</ObjectAllValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#SensitiveData"/>
<Class IRI="#PersonalInformation"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#SettingTermsOfUse"/>
<Class IRI="#CreatingLandingPage"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#SettingTermsOfUse"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#termsOfUse"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#SubmitData"/>
```

```
<Class IRI="#Ingest"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#SubmitData"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#sourceCode"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#SubmitData"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#submittedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#TechnologyMonitoringAndRefreshing"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#TechnologyMonitoringAndRefreshing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#technicalInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#TechnologyMonitoringAndRefreshing"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#technicalInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#TechnologyMonitoringAndRefreshing"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAssociatedWith"/>
<Class IRI="#systemAdministrator"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#TrackingUseAnalytics"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#TrackingUseAnalytics"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#dataDocument"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#TrackingUseAnalytics"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#landingPage"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#TrackingUseAnalytics"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#TrackingUseAnalytics"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#generated"/>
<Class IRI="#usageResults"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
```

```xml
<Class IRI="#Transcoding"/>
<Class IRI="#FileFormatTransformation"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#TranscribingAndTranslatingData"/>
<Class IRI="#Conversion"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Versioning"/>
<Class IRI="#DataPreservationActivities"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#Versioning"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#influenced"/>
<Class IRI="#versionInformation"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Versioning"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#curatedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Versioning"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#submittedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#Versioning"/>
<ObjectExactCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#versionInformation"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#accessRestriction"/>
<Class IRI="#fileLocation"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#administrator"/>
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#agreement"/>
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#appropriateRepository"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#authenticationResults"/>
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#backupData"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#contactInformation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#costAndLeadTime"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#creditOnTheResults"/>
<Class IRI="#termsOfUse"/>
```

```
</SubClassOf>
<SubClassOf>
<Class IRI="#curatedData"/>
<Class IRI="#researchData"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#curatedData"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasDerivedFrom"/>
<Class IRI="#submittedData"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#curationLog"/>
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#curationRecord"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#dataCurator"/>
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#dataDepositer"/>
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#dataDocument"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#dataProcessingPolicy"/>
<Class IRI="#policy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#dataUser"/>
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#documentationPolicy"/>
<Class IRI="#policy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#evaluationResults"/>
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#expertiseNecessity"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#externalConstraints"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#externalServiceProvider"/>
<Class IRI="http://www.w3.org/ns/prov#SoftwareAgent"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#feasibility"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#feeForUse"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#fileLocation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#fileValidationResults"/>
```

```xml
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#fullTextInformation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#imposeTheSameConditions"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#improperUse"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#indexingInformation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#landingPage"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#landingPage"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#visualizedData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#landingPage"/>
<ObjectAllValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasAttributedTo"/>
<Class IRI="#repositorySystem"/>
</ObjectAllValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#landingPage"/>
<ObjectAllValuesFrom>
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasStartedBy"/>
<Class IRI="#DataPublishing"/>
</ObjectAllValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#landingPage"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#used"/>
<Class IRI="#retrievalMetadata"/>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#landingPage"/>
<ObjectExactCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasEndedBy"/>
<Class IRI="#CeaseDataCuration"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#linkingInformation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#metadata"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#metadata"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://xmlns.com/foaf/0.1/primaryTopic"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#metadataCurator"/>
```

```
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#metadataSchema"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#noDelivs"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#nonCommercial"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#peerReviewer"/>
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#persistentIdentifier"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#policy"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#preservationPolicy"/>
<Class IRI="#policy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#quality"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#reporting"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#repositorySystem"/>
<Class IRI="http://www.w3.org/ns/prov#SoftwareAgent"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#researchData"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#retrievalMetadata"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#reusability"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#secondaryUseProhibited"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#selectionPolicy"/>
<Class IRI="#policy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#selectionResults"/>
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#size"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#softwareRegistry"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
```

```
<SubClassOf>
<Class IRI="#sourceCode"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#sourceCode"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="http://xmlns.com/foaf/0.1/primaryTopic"/>
<Class IRI="#researchData"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<SubClassOf>
<Class IRI="#submittedData"/>
<Class IRI="#researchData"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#successionPlan"/>
<Class IRI="#policy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#systemAdministrator"/>
<Class IRI="http://www.w3.org/ns/prov#Person"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#technicalInformation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#termsOfUse"/>
<Class IRI="http://www.w3.org/ns/prov#Entity"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#typeOfResource"/>
<Class IRI="#selectionPolicy"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#usageResults"/>
<Class IRI="#curationRecord"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#useLatestVersion"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#versionInformation"/>
<Class IRI="#metadata"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#visualizedData"/>
<Class IRI="#researchData"/>
</SubClassOf>
<SubClassOf>
<Class IRI="#visualizedData"/>
<ObjectMinCardinality cardinality="1">
<ObjectProperty IRI="http://www.w3.org/ns/prov#wasDerivedFrom"/>
<ObjectUnionOf>
<Class IRI="#curatedData"/>
<Class IRI="#submittedData"/>
</ObjectUnionOf>
</ObjectMinCardinality>
</SubClassOf>
<SubClassOf>
<Class IRI="#waiver"/>
<Class IRI="#termsOfUse"/>
</SubClassOf>
<DisjointClasses>
<Class IRI="#DataCurationActivities"/>
<Class IRI="#DataPreservationActivities"/>
</DisjointClasses>
<DisjointClasses>
<Class IRI="#RiskManagement"/>
<Class IRI="#Selection"/>
</DisjointClasses>
<AnnotationAssertion>
```

```
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#ActivatingMetadataBrokerage</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Active dissemination of a data set’s metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#AllowingFileDownload</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Allow access to the data materials by authorized third parties.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Appraisal</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Appraisal is the process whereby some records are selected for retention, others (the great majority) are deemed of insufficient value to justify permanent retention.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#ArrangementAndDescription</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Authentication</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#CeaseDataCuration</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#ChainOfCustody</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
```

```xml
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#CodeReview</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#ConnectingDiscoveryServices</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">To connect services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Contextualization</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">http://doi.org/10.5334/dsj-2020-053</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Contracts</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">An agreement with a research partner, contractor, etc. that restricts the data publishing in joint research or contract research.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Conversion</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#CreatingLandingPage</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">A web page associated with some dataset or identifier.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataCleaning</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">A process used to improve data quality by detecting and correcting (or removing) defects &amp; errors in data.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataCurationActivities</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The encompassing work and actions taken by curators of a data repository in order to provide meaningful and enduring access to data.</Literal>
</AnnotationAssertion>
```

```
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://casrai.org/term/evaluation/</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataEvaluation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Evaluation is a decision about significance, value, or quality of something, based on careful study of its good and bad features.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">http://doi.org/10.5334/dsj-2020-053</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataPolicy</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Where the research funder has a policy on limited data sharing for the research to be funded, or where a strategic business decision restricts the data publishing relating to pending industrial property rights or research data where the commercialization of the research results is envisaged.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://casrai.org/term/data-processing/</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataProcessing</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">A generic concept referring to all kinds of procedures being executed on data at any point in the data life cycle.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://casrai.org/term/data-publication/</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataPublishing</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DataVisualization</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Deidentification</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third-parties.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DepositAgreement</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the rep
```

```
ository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">http://doi.org/10.5334/dsj-2020-053</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DiplomaticOrNationalSecurity</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Research data pertaining to national security. Data related to the development of weapons of mass destruction, etc. (as defined in the Foreign Exchange and Foreign Trade Act) and defense secrets (the Self-Defense Forces). law), important data that may affect national life (e.g., domestic energy (e.g., location of resources, blueprints for critical equipment, etc.).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">http://doi.org/10.5334/dsj-2020-053</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DisciplinaryCustoms</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Practices and standards in a specific discipline or research community that limit the data publishing. In some cases this is stated as an international treaty, but in others it is not always explicitly stated.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#DisplayingDataCitation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Documentation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">To describe Information about any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text “Readme” file).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Embargo</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Emulation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Provide legacy system configurations in modern equipment in order to ensure long-term usability of data. (E.g., arcade games emulated on modern web-browsers)</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#FileAuditing</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.</Literal>
```

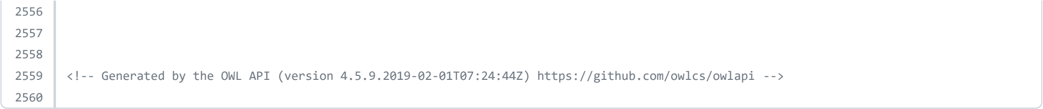
```
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#FileFormatTransformation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#FileInventoryOrManifest</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#FileRenaming</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">To rename files in a dataset, often to standardize and/or reflect important metadata.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#FileValidation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#FormulateSuccessionPlanning</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#GeneratingFulltextIndexing</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Indexing</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
```

```xml
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Ingest</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The process of turning a Submission Information Package (SIP) into an Archival Information Package (AIP), i.e. putting data into a digital archive (OAIS term)</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Interoperability</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#MaintainingContactInformation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#MetadataGeneration</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">A generation process of metadata which is the Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Migration</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#MintingPersistentIdentifier</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">To mint a permanent URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary. E.g., a Digital Object Identifier or DOI.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#PeerReview</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">http://doi.org/10.5334/dsj-2020-053</Literal>
</Annotation>
```

```xml
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#PersonalInformation</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">It stipulates the handling of data that can identify individuals. It includes guidelines that define individual policies on anonymization and information disclosure.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#ProvidingRestrictedAccess</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#QualityAssurance</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of “null” and “blank” values, or unclear acronyms.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#RegisteringSoftware</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Restructure</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Organize and/or reformate poorly structured data files to clarify their meaning and importance.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#RightsManagement</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#RiskManagement</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
```

```
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#SecuringStorage</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Selection</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">The process of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#SettingTermsOfUse</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#SubmitData</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">To submit data to start data curation activities.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#TechnologyMonitoringAndRefreshing</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#TrackingUseAnalytics</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Transcoding</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions. (E.g, Convert QuickTime files to MPEG4).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:isDefinedBy"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">https://hdl.handle.net/11299/188638</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#Versioning</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
```

```
<IRI>#creditOnTheResults</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use requires displaying the creator’s name and data URL information</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#feeForUse</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use requires some fee for data use.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#imposeTheSameConditions</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use requires the same condition of use under redistribution and combination of multiple pieces of data. Unlike “Attribution,” it may prevent them from combining the data with other sets that have an incompatible license.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#improperUse</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use prohibits “Improper use” of data in the data processing phase. Note that the definition of inappropriate use will probably depend on the conventions of the field.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#noDelivs</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use prohibits data publishing after any modification. The case in which the prohibition of modification is effective is presumed to largely depend on the type of data and the manner of use (e.g., image data that are practically a work of art).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#nonCommercial</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use requires the “non-commercial” use of data. Also, the limitation is on redistribution and data processing (e.g., selling visualizations derived from the data).</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#reporting</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use requires post-reuse reporting, suggesting that the objective is to know detailed usage practices rather than mechanical access statistics.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#secondaryUseProhibited</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use prohibits the secondary use of data. This license clearly prohibits data redistribution, translation, or adaption and is intended for one-on-one use of its original form.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<Literal datatypeIRI="http://www.w3.org/2001/XMLSchema#anyURI">http://doi.org/10.5334/dsj-2020-053</Literal>
</Annotation>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#termsOfUse</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">“Condition of use” generally means use permission/prohibition, obligations, and constraints based on the types of research data.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty IRI="http://www.w3.org/2004/02/skos/core#altLabel"/>
<IRI>#termsOfUse</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Conditions of Use</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#useLatestVersion</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">This condition of use is used to limit the use of data to the latest one. Data in the past cannot be reused.</Literal>
</AnnotationAssertion>
<AnnotationAssertion>
<AnnotationProperty abbreviatedIRI="rdfs:comment"/>
<IRI>#waiver</IRI>
<Literal datatypeIRI="http://www.w3.org/2000/01/rdf-schema#Literal">Waives copyright and all other related rights</Literal>
</AnnotationAssertion>
</Ontology>
```



```



<!-- Generated by the OWL API (version 4.5.9.2019-02-01T07:24:44Z) https://github.com/owlcs/owlapi -->

```

Appendix 4. Topic guide of "Basic research on licensing and legal interoperability of research data Questions"

We used a topic guide in Section 5.3.1 to share our interview outline with the interviewee. We set three major sections within the topic guide: "Sharing and publishing of research data," "Regarding granting licenses to research data," and "Licensing of research data and promotion of legal interoperability." Details of each section are as follows. Note that the topic guide sent to interviewees was prepared in Japanese.

List of sections

1. Sharing and publishing of research data

- Overview and characteristics of your data

e.g., ownership of data, public or commercial value, existence of personal information

- Current status of data sharing and publication in your research, your institution, and your research community

- Difficulties in sharing and publishing research data

- What you would like to ask/prohibit users from doing with the research data you have made public

  e.g., prohibition or suspension of use and penalties for license violations

2. Regarding granting licenses to research data

- The type of license used, the stipulations of the license granted, and information/guidelines referred to at the time of granting the license

- Difficulties in granting licenses to research data

3. Licensing of research data and promotion of legal interoperability

- The needs for licensing and legal interoperability of research data

- Personal views on existing licenses, guidelines, and existing discussions

- The degree of protection of rights to the research data sought and the basis for the request for protection of rights (and why such a request is made by the user)

- Expectations regarding licensing, rights management, data release support, etc.

4. Other comments

Appendix 5. Guidelines for specifying conditions of use in research data publishing

In Section 5.5 we introduced the “Guidelines for specifying conditions of use in research data publishing” (Minamiyama et al., 2020) to help researchers and stakeholders understand and make appropriate publication decisions. These guidelines provide necessary information and examples that should be considered when publishing research data with an interdisciplinary perspective. The original document was written in Japanese, and a tentative English translation of the document is shown below.

December 25, 2019

Guidelines for specifying conditions of use in research data publishing

Research Data License Subcommittee
under the Research Data Utilization Forum (RDUF)

Index

Introduction. Five Questions on the Research Data Publishing and the Specification of Conditions of Use

Purpose and objectives

These guidelines are for the research data publishing and the specification of conditions of use developed by the Research Data License Subcommittee under the Research Data Utilization Forum (RDUF)[1]. The basic policy of open science in Japan is to expand the utilization of research results funded by public research funds as much as possible[2], but some types of research data are exceptions. Therefore, these guidelines aim to enable data providers to publish research data under appropriate conditions of use by organizing information and examples that generally require attention when publishing research data, along with the decision-making process. It also expects to be used as a tool for data reusers to easily understand the background of the conditions of use required by the data provider.

(1) When the data provider specifies the conditions of use

To enable those who wish to publish research data (individual researchers, teams, and repository managers) to inform third parties of their conditions of use concisely. Possible scenarios are as follows: a) when publishing research data underlying a research paper and b) when publishing research data itself as research results. These guidelines prevent unauthorized reproduction, plagiarism, inappropriate processing, and trouble with interested parties in research data publishing.

(2) When the data reuser checks the existing conditions of use

When the researcher acquires and reuses published research data, they can easily understand the conditions of use required by the data provider.

How to use these guidelines

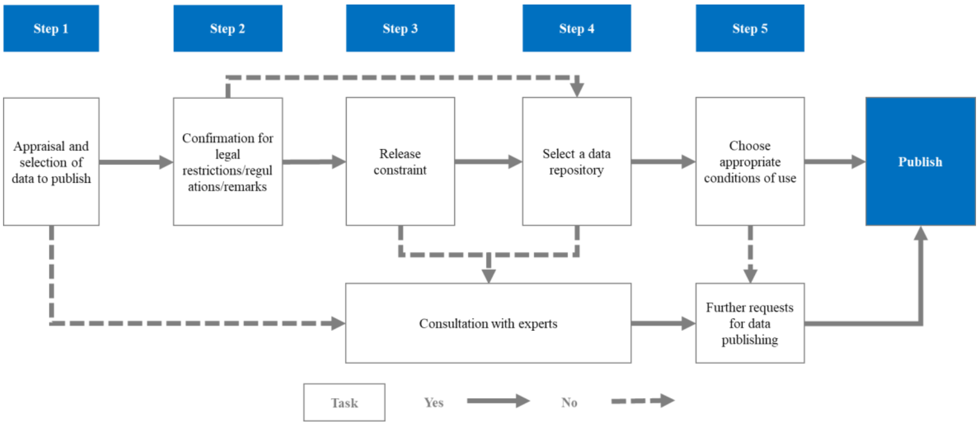
First, select your research data to be published in Q1. Next, confirm any external constraints as listed in Q2. In Q3, you can confirm the processes for enabling research data publishing for the external constraints identified in Step 2. Even if external constraints exist, it may be

[1] Research Data Utilization Forum (RDUF) was established in May 2016. The principal mission of the RDUF is to encourage the utilization of research data based on open science. The RDUF is willing to support the communication on research data management and open science in various inter-disciplinary and inter-sectional views.

[2] 5th Science and Technology Basic Plan. https://www8.cao.go.jp/cstp/kihonkeikaku/index5.html

possible to publish your research data by setting a certain embargo period, as described in Q3. If your research data can be made public, refer to Q4 to select a data repository. Finally, specify the conditions of use for your research data in Q5.

Publishing flow and licensing scenarios for research data


Step 1
Step 2
Step 3
Step 4
Step 5
Appraisal and selection of data to publish
Confirmation for legal restrictions/regulations/remarks
Release constraint
Select a data repository
Choose appropriate conditions of use
Publish
Consultation with experts
Further requests for data publishing
Task
Yes
No

Usage Notes

- These guidelines organize information and examples that generally need to be kept in mind when publishing research data and the decision-making process. We do not expect to cover all academic fields; When publishing research data, please check general guidelines and policies related to research ethics and the handling of research data in your own academic field.
- When the researchers share research data, even if one wants a particular publication method or conditions of use, the other researchers may not have the same intention. Please confirm their intent well in advance.
- The conditions of use recommended by these guidelines are proposed regarding the Creative Commons 4.0 International License. Note that we do not consider the compatibility with different versions.
- If you wish to specify conditions of use for databases and/or repositories, we recommend seeking advice from experts in the relevant legal systems. The same advice should be given when there is more than one interested party, such as in the case of research conducted through industry-academia-government collaboration. Also, if you wish to state your conditions of use for research data as an organization or institution, please consider developing your institution's data policy.

Conditions of use for this document/Disclaimers

- Copyright of the text and figures in these guidelines belongs to the Research Data License Subcommittee under the Research Data Utilization Forum (RDUF). Except where otherwise noted, all materials are available under the Creative Commons Attribution 4.0 International License (CC-BY) terms. When using the materials, please clearly indicate the source information and any modifications regarding the following license notice:

  Source: "Guidelines for specifying conditions of use in research data publishing ver.1.0". Research Data License Subcommittee under the Research Data Utilization Forum (RDUF), 2019, 32p. https://doi.org/10.11502/rduf_license_guideline, (accessed YYYY-MM-DD).

- We do not guarantee the accuracy, certainty, fitness for purpose, or other quality of the statements in your context. The responsibility for all actions using these guidelines rests entirely with the user. The users themselves should make decisions based on the information obtained.

Q1. Appraisal and selection of data to publish

First, select the research data to be published in the whole study. The scope of the "research data" term varies from field to field, so these guidelines limit the scope of the term to the extent that it can be managed by electronic means. In other words, the term “research data” does not include physical materials such as samples (specimens, samples) or recording media (paper, disks, etc.) in these guidelines.

- Definition of “research data” (*Subject of these guidelines)
  Digital data used as a source information for scientific research. It includes a variety of formats, such as numerical, textual, image, audio, and video. Various designations may be used depending on the context in which the data is used; e.g., evidence data, source data, and derived data.

  Example:
  1) Evidence data
     Data underlying a research paper or research results is called "evidence data.” Raw data may be published as the evidence data, or sometimes processed data are selected to publish.
  2) Source data
     The original data newly collected from the observation is called "source data (or primary data).” Researchers may observe and create the source data themselves, or it may be held by a third party (e.g., another researcher, a company, or a local government).
  3) Derived data
     Data created by derivation from source data is called "derived data.” If the creator of the source data and the creator of the derived data are different, the data citation method and rights attribution tends to be complicated. It is necessary to pay attention to the source data version information.

(cf: Data not included in the “research data”) (Outside the scope of these guidelines)

- Non research data
  Descriptions to explain the outline and status of the research data, as well as physical objects such as research notes, diaries, samples, etc.
  - Ex. Descriptions about research data (including metadata or meta-information)
  - Other records or logs (research notes, samples, and other physical objects)

- Copyrighted work
  For example:
  - ➢ Research papers
  - ➢ Books, derivative works by the authors, secondary works
  - ➢ Derivative works by third parties

- Research environment
  For example:
  - ➢ Databases
  - ➢ Software (e.g., Analysis and visualization programs, estimation models, machine learning algorithms)
  - ➢ Other source codes

When publishing research data is required

Publishing research data may be required by your funding agency, publisher, or institution to promote research data reuse.

[Ex. 1] by funding agency
JST Policy on Open Access to Research Publications and Research Data Management (April 1, 2017)
https://www.jst.go.jp/pr/intro/openscience/guideline_openscience.pdf

[Ex. 2] by publisher
Elsevier. Research Data Guidelines
https://www.elsevier.com/authors/author-resources/research-data/data-guidelines

[Ex. 3] by institution (project data)
JAXA. ISAS Data Policy (March 14, 2018)
http://www.isas.jaxa.jp/researchers/data-policy/

[Ex.4] by institution (evidence data)
National Institute of Polar Research. National Institute of Polar Research Open Access Policy (November 24, 2017)
https://www.nipr.ac.jp/outline/activity/oap.html

Considerations for handling data not included in "research data"

- Metadata plays an essential role in informing the existence of research data. Even if the source or derived data is kept private, the metadata should be widely available for the public to search, view, and retrieve.
- Data not included in ”research data” should be segregated and managed with the research data. If you treat past research notes or diaries as source data, be careful what you publish in digital form.
- The requirement of copyrighted works such as articles, papers, posters, slide materials, and projection materials differs from the research data to be published. For these works, consider publishing them in your institutional repository and applying standard licensing tools such as the Creative Commons License.
- The research environment and research data must be treated separately in the licensing context. You may refer to the following if you wish to specify a database license or a source code license.

  [Ref. 1] Open Data Commons. Open Data Common Open Database License (ODbL).
  https://opendatacommons.org/licenses/odbl/index.html

  [Ref. 2] Choosealicense.com. Licenses.
  https://choosealicense.com/licenses/

Q2. Confirmation for legal restrictions/regulations/remarks

There may be restrictions on publishing research data due to the sensitive content contained in the data (e.g., privacy information) or the research participant's request. Please confirm if your data falls into the descriptions contained in the following categories.

In cases of disciplinary customs restriction, including international treaty

- Individual disciplines and research communities may have conventions or standards regarding data release restrictions. The provisions of international treaties are indicated in some cases[3], but the provisions are not always explicitly stated[4].

  [Ex. 1] Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)
  https://www.cites.org/eng

  [Ex. 2] Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property
  http://portal.unesco.org/en/ev.php-URL_ID=13039&URL_DO=DO_TOPIC&URL_SECTION=201.html

In cases of containing personal information

- The laws of each country regulate the handling of personal information. In Japan, if the research entity is private, the Act on the Protection of Personal Information applies. If the research entity is an Incorporated Administrative Agency, including a National Research and Development Agency, the Act on the Protection of Personal Information Held by Incorporated Administrative Agencies, etc.
- As for discipline-specific regulations, for example, guidelines may be formulated for each field, with separate policies for anonymization and information disclosure.

[3] Other examples of publishing research data restrictions are mentioned within the guidelines prepared by RDA/CODATA, such as the protection of endangered species, cultural resources, sovereign genetic resources, and traditional knowledge.
Legal Interoperability of Research Data: Principles and Implementation Guidelines
https://doi.org/10.5281/zenodo.162241

[4] For example, materials may be withheld from the public due to the Bereaved family's request in literary research.

[Ref. 1] Personal Information Protection Commission, Government of Japan. “Laws and guidelines” (only in Japanese)
https://www.ppc.go.jp/personalinfo/legal/

[Ref. 2] Japan External Trade Organization(JETRO). “About General Data Protection Regulation (GDPR)” (only in Japanese)
https://www.jetro.go.jp/world/europe/eu/gdpr/

[Ref. 3] Ministry of Health, Labor and Welfare (Japan). “About research guidelines” (only in Japanese)
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hokabunya/kenkyujigyou/i-kenkyu/index.html

In cases of Diplomatic / National security restriction

- Publishing research data on national security is regulated by law and includes data related to the weapons development of mass destruction (as defined in the Foreign Exchange and Foreign Trade Law) and defense secrets (defined in the Self Defense Forces Law).
- In other cases, there are special legal measures for data that may affect the lives of the public (e.g., location of domestic energy resources, blueprints of important facilities, etc.).

[Ref. 1] Japan Society for Intellectual Production. “Security Trade Control Guidelines for Researchers in universities and other institutions of higher education. Revised 2nd ed”
http://j-sip.org/info/pdf/anzenhosho1-1_2.pdf

In cases of keeping agreements, contracts, Intellectual Property rights

- In joint or contract research, it is necessary to comply with agreements with research partners, contractors, etc. regarding publishing research data.
- If you have some agreements or contracts with a private company (including a commercial publisher) that restricts publishing your data, you must comply with the terms of the agreements or contracts.

[Ref. 1] Ministry of Economy, Trade and Industry (Japan). "Operation guidelines for data management in contract research and development" (only in Japanese)
https://www.meti.go.jp/press/2017/12/20171227001/20171227001-1.pdf

[Ref. 2] Ministry of Economy, Trade and Industry (Japan). "Contract Guidelines on Utilization of AI and Data. Data Section"
https://www.meti.go.jp/press/2019/12/20191209001/20191209001.html

In cases of complying data policy

- Your institution may have a restricted data sharing policy. If your institution has an intellectual property policy or data policy[5], you must confirm the scope or embargo period within the policy.
- In some cases, publishing research data on industrial property rights pending application or research data expected to be commercialized may be restricted as a management strategy decision[6]. You also need to confirm the target data attribution.

[Ex. 1] National Institute for Environmental Studies. "NIES Data Policy" (only in Japanese)
https://www.nies.go.jp/kihon/kitei/kt_datapolicy.pdf

[5] In Japan, all national research and development agencies must have a data policy by 2020.
Cabinet Office, Japan. "Integrated Innovation Strategy"
https://www8.cao.go.jp/cstp/tougosenryaku/index.html
Cabinet Office, Japan. "Guidelines for the Development of Data Policies in National Research and Development Agency"
https://www8.cao.go.jp/cstp/stsonota/datapolicy/datapolicy.html

[6] The "Open and Closed Strategy" is "a strategic choice to increase the company's profits by adopting an open model IP strategy of disclosing or licensing to other companies, in addition to a closed model IP strategy of keeping technologies and other information secret or implementing exclusive rights such as patent rights." The "Guidelines for the Formulation of Data Policies for National Research and Development Institutions" also calls for the formulation of policies based on this concept.
Japan Patent Office. "Open and Closed Strategy"
https://faq.inpit.go.jp/content/tradesecret/files/100578260.pdf

[Ex. 2] Teikyo University. “Intellectual Property policy in Teikyo University” (only in Japanese)
https://www.teikyo-u.ac.jp/affiliate/laboratory/tttc_center/policy.html

[Ex. 3] Japan Agency for Medical Research and Development. “Data sharing policy for realization of genomic medicine” (only in Japanese)
https://www.amed.go.jp/content/000023353.pdf
https://humandbs.biosciencedbc.jp/files/DAC/4th_meeting/2_ref_AMED_DSP.pdf

Q3. Release constraint

In most cases, even if there are some restrictions on publishing research data as described in Q2, it is possible to publish the data by applying appropriate data processing or allowing a certain embargo period[7]. Set the necessary conditions of use based on the following information[8] and show them to the data reuser:

In cases of disciplinary customs restriction, including international treaty

You need to confirm the data publishing procedures with the corresponding national law if the disciplinary customs restriction is explicitly stated in an international treaty. In the absence of a specified period for restricted publication, you must set an appropriate period that considers the disciplinary practice and/or the treaty's purpose. In addition to checking with the IP department of your institution, you can consult with an expert if necessary.

[Ref. 1] Ministry of Foreign Affairs of Japan. "Treaty Data Search"
https://www3.mofa.go.jp/mofaj/gaiko/treaty/

[Ref. 2] FAIRSharing. "FAIRSharing policies"
https://fairsharing.org/policies/
* Collection of field-specific policies/guidelines

[7] In principle, the legal protection of a copyrighted work expires 70 years after the author's death. Still, there is no corresponding provision for research data, so it is necessary to be more careful in setting an embargo period. From the viewpoint of protecting research papers, it is generally 12 months for science, engineering, and medicine, and typically 24 to 36 months for humanities and social sciences. However, in recent years, the evidence data tend to publish immediately.
[Ref.] U.S. Department of Health & Human Services. "NIH Public Access Policy Details" https://publicaccess.nih.gov/policy.htm
[Ref.] Wiley. "Wiley's Self-Archiving Policy" https://authorservices.wiley.com/author-resources/Journal-Authors/licensing/self-archiving.html
[Ref.] SHERPA/RoMEO. "Publisher copyright policies & self-archiving" http://sherpa.ac.uk/romeo/index.php

[8] There are some options for setting an embargo period, such as a timer, date/time-specified, and user-only limitation methods. These options will be used in some combination.

In cases of containing personal information

Even if your research data contains some personal information, it can be published when you anonymize your data in an appropriate method.

[Ref. 1] Personal Information Protection Commission, Government of Japan. “Guidelines for the Act on the Protection of Personal Information. Anonymized Information section”
https://www.ppc.go.jp/files/pdf/guidelines04.pdf

You can also refer to these particular guidelines to specify anonymizing methods in your fields:

[Ref. 2] Personal Information Protection Commission, Government of Japan. “Guidelines for Specific Fields”
https://www.ppc.go.jp/personalinfo/legal/guidelines/

[Ref. 3] Japan Pharmaceutical Manufacturers Association. “Institutional Review Board (IRB) overview and date of event”
http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/leaflet/009.html

In cases of Diplomatic / National security restriction

If your research data is restricted by export control, military diversion, or other treaties, you must follow established procedures to consider whether or not to disclose the data. Please consult with the department in charge and follow the disclosure procedures.

[Ref. 1] Ministry of Finance, Japan. “Overview of Foreign Exchange and Foreign Trade Act”
https://www.mof.go.jp/international_policy/gaitame_kawase/gaitame/index.html

[Ref. 2] Ministry of Economy, Trade and Industry, Japan. “Security Export Control System in Japan”
https://www.meti.go.jp/policy/anpo/

If your research data is categorized in public records, it is subject to the Public Records and Archives Management Act. You must be considered for disclosure following the enforcement order of the Act. As above, please consult with the department in charge and follow the disclosure procedures.

[Ref. 1] Public Records and Archives Management Act
https://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=421AC0000000066

[Ref. 2] Cabinet Office, Government of Japan. “Documents subject to management under the Public Records and Archives Management Act”
https://www8.cao.go.jp/chosei/koubun/about/bunsho/bunsho.html

[Ref. 3] Cabinet Secretariat, Japan. “Standard document retention period”
https://www.cas.go.jp/jp/koukai/hyoujunbunsho/anzenhosho.html

In cases of keeping agreements, contracts, Intellectual Property rights

Based on the agreement or contract, set the publication date and time after confirming and coordinating the cases in which publication is possible. In addition to the statements related to the publication of research results, check the agreements after the joint research period has ended. Even in cases where there is no explicit agreement or contract, you should check with your IP department before publishing your research data.

[Ref. 1] Joint Research and Development Agreement (in Japanese)

> 第〇条（研究成果の公表等）
> 甲又は乙は、本契約の有効期間中及び契約終了後〇年間は、本共同研究によって得られた研究成果を公表又は第三者に開示しようとする場合には、その内容、時期、方法等について、書面により事前に相手方の承諾を受けるものとする。
>
> Source: Ministry of Economy, Trade and Industry, Japan. “Handbook for Protection of Confidential Information - Toward Enhancing Corporate Value”
> https://www.meti.go.jp/policy/economy/chizai/chiteki/trade-secret.html#handbook

[Ref. 2] Ministry of Education, Culture, Sports, Science and Technology, Japan. “Sakura tool”
https://www.mext.go.jp/a_menu/shinkou/sangaku/1383777.htm
* Provides contract templates that can be used for joint research. The link provides the consortium type of contract and revised materials for individual versions.

In cases of complying data policy

Based on the applicable data policy, set the publication date and time to the end of the embargo period. If the policy does not specify the embargo period, it should be decided in consultation with the department in charge.

In cases of not available to be published

Even if your research data cannot be published at this time, it is necessary to leave a trail of evidence that the research data exists to support future research activities. Please document the decision-making process up to this point and store the research data in appropriate storage. Also, please publish it as metadata to the institution's platform if possible.

Q4. Select a data repository

When you have finished checking the constraints, you need to select an appropriate repository. You can publish your research data using file-sharing services or data management software. However, we recommend using data repositories in related fields or institutional repositories from the viewpoints of organization and preservation, data permanence assurance, user recognition, and security management[9]. Some examples of data repositories in Japan are as below. If you want to search for international data repositories, please see the "List of International Data Repositories.”

Disciplinary data repositories

- Social sciences
  - SSJDA (https://ssjda.iss.u-tokyo.ac.jp/Direct/)
    - Acceptable data type: Micro data with questionnaires used in various social and statistical surveys
  - RUDA (https://ruda.rikkyo.ac.jp/dspace/)
    - Acceptable data type: Social survey data (Economics, Business Administration, Sociology, Social Psychology, Political Science, Political Psychology, Law, Sociology of Law, Education, Sociology of Education, etc.)
- Life sciences
  - DDBJ (https://www.ddbj.nig.ac.jp/index.html)
    - Acceptable data type: Annotated/assembled sequences, Sequencing and alignment data from next-generation sequencing platforms, Functional genomics data, Research project, Biological sample, Human data requiring controlled-access
      * Further information:
      https://www.ddbj.nig.ac.jp/data-categories.html
  - NBDC human database (https://humandbs.biosciencedbc.jp/)
    - Acceptable data type: Human data produced from publicly funded research

[9] Even when research data is published as an Appendix or Supplement to a research paper, registering it in repositories and databases in related fields will further increase its discoverability and make it more likely to be used. It also facilitates maintenance of broken links.

  - jPOSTrepo (https://repository.jpostdb.org/)
    - Acceptable data type: ProteOme data in Japan
  - GlyTouCan (https://glytoucan.org/)
    - Acceptable data type: Glycan structures data
  - Life Science Database Archive (https://dbarchive.biosciencedbc.jp/index.html)
    - Acceptable data type: Datasets generated by domestic life science researchers
- Earth science
  - DIAS (http://www.diasjp.net/)
    - Acceptable data type: Earth and Environmental data
  - IUGONET (http://search.iugonet.org/list.jsp)
    - Acceptable data type: Solar-Terrestrial Science Observations data
  - Global Environment Database (http://db.cger.nies.go.jp/portal/)
    - Acceptable data type: Global Environmental Research Data
- Biological science
  - Biological Information System for Marine Life (BISMaL) (https://www.godac.jamstec.go.jp/bismal/j/)
    - Acceptable data type: Data on marine biotic occurrence records
  - Global Biodiversity Information Facility Japan Node (JBIF) (http://www.gbif.jp/v2/)
    - Acceptable data type: World's biodiversity data
    - Contact: http://www.gbif.jp/v2/regist/index.html
- Synchrotron radiation science
  - SPring-8 case studies & reports cross research (http://www.spring8.or.jp/ja/science/customsearch/)
    - Acceptable data type: Data on polymers, organic thin films, and green energy fields
  - SPring-8 BL14B2 XAFS Standard Sample Database (https://support.spring8.or.jp/xafs/standardDB/standardDB.html)
    - Acceptable data type: XAFS (X-ray absorption fine structure) data

– Other fields
  - List of international data repositories (https://www.re3data.org/)

Institutional repositories

- List of Japanese institutional repositories (https://www.nii.ac.jp/irp/list/)

Considerations

- In the event of data leakage or unauthorized use, you will warn the data reuser and request an injunction against the data reuse. Depending on the circumstances, you may consider injunctive relief, damages, or criminal legal proceedings. Protection may be available under the Copyright Act, or relief may be obtained under the Unfair Competition Prevention Act. You should consult with the department in charge and follow the appropriate procedures.
- When selecting a data repository, information such as whether it conforms to international standards certification and which country's laws it complies with may be helpful.

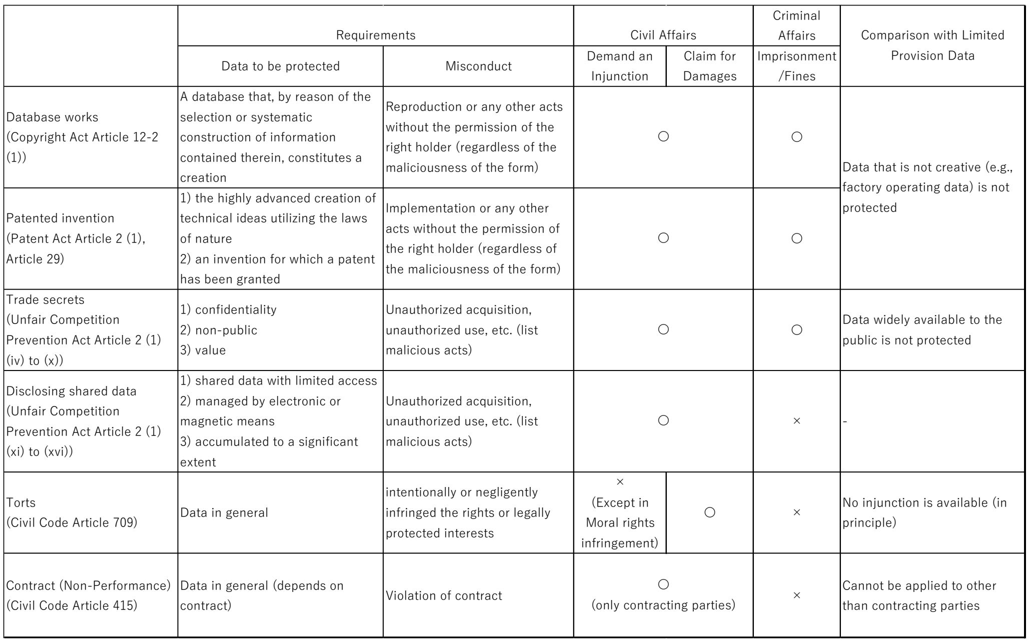
[Ref. 1] FAIR principles

- FAIR principles as a standard for data sharing (https://doi.org/10.18908/a.2018041901)
- Is the repository listed on the "FAIRsharing" website? (https://fairsharing.org/)
- Is the repository listed on the "Repository Finder," which complies with FAIR principles? (https://repositoryfinder.datacite.org/)

[Ref. 2] International Standards Certification

- From the re3data.org search page (https://www.re3data.org/search), select "Certificates" in the Filter to check.
- A list of recommended repositories may be provided by the publisher.
  Ex. Nature "Scientific Data. Recommended Data Repositories"
  https://www.nature.com/sdata/policies/repositories

List of Legal Protection of Data in Japan



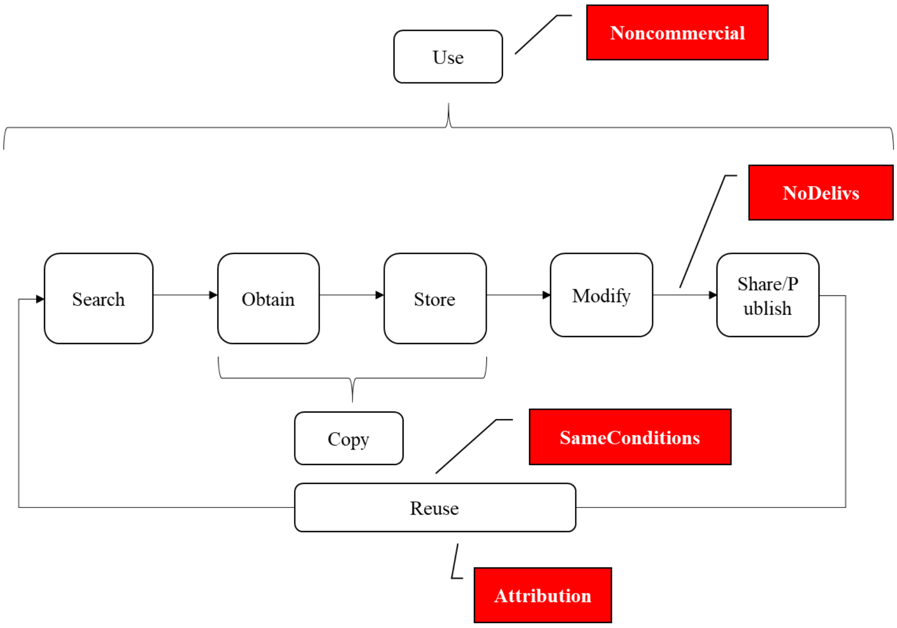
	Requirements		Civil Affairs		Criminal Affairs	Comparison with Limited Provision Data
	Data to be protected	Misconduct	Demand an Injunction	Claim for Damages	Imprisonment /Fines	
Database works (Copyright Act Article 12-2 (1))	A database that, by reason of the selection or systematic construction of information contained therein, constitutes a creation	Reproduction or any other acts without the permission of the right holder (regardless of the maliciousness of the form)	○		○	Data that is not creative (e.g., factory operating data) is not protected
Patented invention (Patent Act Article 2 (1), Article 29)	1) the highly advanced creation of technical ideas utilizing the laws of nature 2) an invention for which a patent has been granted	Implementation or any other acts without the permission of the right holder (regardless of the maliciousness of the form)	○		○	
Trade secrets (Unfair Competition Prevention Act Article 2 (1) (iv) to (x))	1) confidentiality 2) non-public 3) value	Unauthorized acquisition, unauthorized use, etc. (list malicious acts)	○		○	Data widely available to the public is not protected
Disclosing shared data (Unfair Competition Prevention Act Article 2 (1) (xi) to (xvi))	1) shared data with limited access 2) managed by electronic or magnetic means 3) accumulated to a significant extent	Unauthorized acquisition, unauthorized use, etc. (list malicious acts)	○		×	-
Torts (Civil Code Article 709)	Data in general	intentionally or negligently infringed the rights or legally protected interests	× (Except in Moral rights infringement)	○	×	No injunction is available (in principle)
Contract (Non-Performance) (Civil Code Article 415)	Data in general (depends on contract)	Violation of contract	○ (only contracting parties)		×	Cannot be applied to other than contracting parties

Ref. Ministry of Economy, Trade and Industry, Japan. Major legal systems against data misuse https://www.meti.go.jp/policy/economy/chizai/chiteki/H30nen_fukyohoshosai.pdf

Q5. Choose appropriate conditions of use

Once you have decided where to publish the data, you need to specify the conditions of use for your research data. When you have multiple datasets or are publishing them together with derived data, it is convenient to specify their conditions of use together. Specified conditions of use should be appropriately described as metadata[10] when registering to the repository.

Flow of research data reuse and actions subject to recommended conditions of use specification



- Please clearly indicate the following four points: attribution, same conditions, NoDelivs, and Noncommercial.
- “NoDelivs” means that the release of modified data is prohibited.

[10] To ensure transparency, we strongly recommend that a description of how data is acquired and modified be created in the metadata according to the conditions of use or a link be provided to a report or data paper describing the procedure.

Recommended conditions of use

In specifying the conditions of use, you can combine some conditions such as attribution and/or permission for modification. You can also choose to waive your rights. The description in square brackets for each conditional specification can be used when describing the conditions of use.

List of conditions of use combinations

Conditions of use	Description
Waiver	Freely available
Attribution	Clearly indicate the data source and credit information
Attribution - Noncommercial	Clearly indicate the data source and credit information; Commercial use prohibited
Attribution - NoDelivs	Clearly indicate the data source and credit information; the release of modified data prohibited
Attribution - Noncommercial - NoDelivs	Clearly indicate the data source and credit information; Commercial use prohibited; the release of modified data prohibited
Attribution - SameConditions	Clearly indicate the data source and credit information; the release of modified data granted different conditions of use prohibited
Attribution - SameConditions - Noncommercial	Clearly indicate the data source and credit information; the release of modified data granted different conditions of use prohibited; Commercial use prohibited
Other	Individual restrictions by contract (e.g., limited sharing)

- Although these guidelines aim to set appropriate conditions of use for a non-copyrighted data, there are many cases in which it is difficult to determine whether or not a copyrighted work. The conditions of use recommended in this section are compatible with the Creative Commons License (https://creativecommons.jp/licenses/) International 4.0, and you can use these conditions of use regardless of their copyrightability.
- Note that when research data are based on derived data (see p. 5), it is impossible to grant fewer conditions of use than the source data. For example, you cannot grant

“Waiver” if the original conditions of use are “Attribution,” even if the research data was published by yourself.

- “Waiver” can be understood as a declaration by the data provider that they waive the right to take legal action against copyright infringement. However, even if the data provider declares a "waiver," moral rights, privacy rights, and the right to prevent unfair competition remain. Trademark and patent rights are not also waived. Therefore, legal action against rights infringement other than copyright can be considered by the data provider, the institution to which the data belongs, or the repository manager.
  Cf. https://creativecommons.jp/sciencecommons/aboutcc0/
- “NoDelivs” prohibits the sharing or publication of modified data, so it does not prohibit private data reuse. Also, even under conditions of use that do not grant “NoDelivs,” the Unfair Competition Prevention Act may be applied in case of data falsification.

1） Waiver

- The data are freely available for commercial or non-commercial purposes. It is unnecessary to indicate the data source information or the modification methods.

[Description (Japanese)]

- 本データの利用に当たり、原則として、何らの制約はありません。

※本データが著作物である場合は、CC0（権利放棄）が付与されます。

[Notice]

- If you select a "waiver," you may not revoke or change your choice in the future. Please check carefully with your institution or professional before selection.
- This conditions of use will be interpreted that the data reuser does not require third parties to indicate the data source information or the same conditions of use as the original data.
- Even if the data provider has waived their rights, it may be necessary to indicate the data source information according to the someone's policy, such as journal policy. You need to cite data appropriately, taking into account research ethics regulations.

2） Attribution

- The data are freely available as long as the data source and credit information are clearly indicated.

[Description (Japanese)]

- ➢ 本データの公開に当たっては、出所を明示してください。
- ➢ 本データを改変した場合には、その手順を何らかの手段で明記してください。

※本データに著作権が発生する場合、クリエイティブ・コモンズ 表示 4.0 国際ライセンス（CC-BY）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示を条件に利用することが可能です。

[Notice]

- In displaying attribution for your research data, clearly indicate the credit information, including the version and date/time information on the landing page.
- There are several ways to specify the means of data modification: 1) mentioning when the data source is referred, 2) including it in the metadata, and 3) writing a report or data paper that describes the procedure in more detail. Choose a method appropriate to the degree of alteration.

3）Attribution - Noncommercial

- For noncommercial purposes, the data are freely available as long as the data source and credit information are clearly indicated.

[Description (Japanese)]

- 本データの公開に当たっては、出所を明示してください。
- 本データは、営利目的で利用することができません。
- 本データを改変した場合には、その手順を何らかの手段で明記してください。

※本データに著作権が発生する場合、クリエイティブ・コモンズ 表示-非営利 4.0 国際ライセンス（CC-BY-NC）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示及び非営利目的での利用を条件に利用することが可能です。

4） Attribution - NoDelivs

- The data are freely available as long as the data source and credit information are clearly indicated; The release of modified data is prohibited.

[Description (Japanese)]

- 本データの公開に当たっては、出所を明示してください。
- 本データを改変した場合、改変されたデータを公開することはできません。

※本データに著作権が発生する場合、クリエイティブ・コモンズ 表示-改変禁止 4.0 国際ライセンス（CC-BY-ND）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示及び改変されたデータを公開しないことを条件に利用することが可能です。

[Notice]

- It is common for data acquired by third parties to be modified in the reuse process, except when the data is only for observation, viewing, or browsing. If you have a particular modification method that you want to prohibit, clearly state it.
- The term “Modified“ includes partially altered from the source data and derived data. For example, this applies when tabular data are prepared based on individual data or estimates are organized based on observed data.

5）Attribution - Noncommercial - NoDelivs

- For noncommercial purposes, the data are freely available as long as the data source and credit information are clearly indicated; The release of modified data is prohibited.

[Description (Japanese)]

- 本データの公開に当たっては、出所を明示してください。
- 本データは、営利目的で利用することができません。
- 本データを改変した場合、改変されたデータを公開することはできません。

※本データに著作権が発生する場合、クリエイティブ・コモンズ 表示-非営利-改変禁止 4.0 国際ライセンス（CC-BY-NC-ND）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示、非営利目的での利用及び改変されたデータを公開しないことを条件に利用することが可能です。

6） Attribution - SameConditions

- The data are freely available as long as the data source and credit information are clearly indicated; The release of modified data granted different conditions of use is prohibited.

[Description (Japanese)]

- 本データの公開に当たっては、出所を明示してください。
- 本データを改変した場合には、本データと同じ利用条件で公開し、かつその手順を何らかの手段で明記してください。

※本データに著作権が発生する場合、クリエイティブ・コモンズ 表示-継承 4.0 国際ライセンス（CC-BY-SA）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示及び元データと同じ利用条件要素を付与することを条件に利用することが可能です。

7）Attribution - SameConditions - Noncommercial

- For noncommercial purposes, the data are freely available as long as the data source and credit information are clearly indicated; The release of modified data granted different conditions of use is prohibited.

[Description (Japanese)]

- 本データの公開に当たっては、出所を明示してください。
- 本データは、営利目的で利用することができません。
- 本データを改変した場合には、本データと同じ利用条件で公開し、かつその手順を何らかの手段で明記してください。

※本データに著作権が発生する場合、クリエイティブ・コモンズ 表示-継承-非営利 4.0 国際ライセンス（CC-BY-SA-NC）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示、非営利目的での利用及び元データと同じ利用条件要素を付与することを条件に利用することが可能です。

8) Additional conditions of use

In cases of some additional conditions of use caused by an institution's data policy or individual contracts, the detailed description will be more readily understood when displayed along with the usage notes. We recommend preparing a concise user guide with references to these policies or contracts. Note that any new conditions granted to the data will no longer make it compatible with the Creative Commons License.

Appendix. Terms of use statement

Since the research data assumed by these guidelines are not protected by copyright law, it is necessary to indicate the specified terms of use in the metadata and establish more detailed terms in advance to legally guarantee the specified conditions of use. Referring to the following sample format, please check your description, such as the data source and credit information, an example of how to modify the data, and a disclaimer. If there is missing information in the landing page, add the necessary information to the metadata and consider changing the data repository if necessary.

- Sample format: in cases for "Attribution" (Japanese)

  本データ及び付録資料に収録された情報(以下「本データ等」といいます)に関する一切の権利は、原則として、本データ等の作成に関与した研究者、研究機関又は当該データの提供者（以下「情報提供者」といいます）に帰属します。本データ等に関する権利は、我が国国内法及び国際条約により保護されており、情報提供者が指定する利用規約又はライセンス表示に従う場合を除いて、本データ等を無断で利用することはできません（使用、複製、頒布、上映、公衆送信、上演、出版、送信可能化、翻案、改変及び商用利用を含みますが、これらに限られません）。本データ等の利用に当たっては、情報提供者が指定する利用規約又はライセンス表示に同意したものとみなします。

  （利用条件）
  - 本データの公開に当たっては、出所を明示してください。
  - 本データを改変した場合には、その手順を何らかの手段で明記してください。

  ※本データ等に著作権が発生する場合、クリエイティブ・コモンズ 表示 4.0 国際ライセンス（CC-BY）の条件で利用することが可能です。著作権が発生しない場合でも、出所の明示を条件に利用することが可能です。

（例 1）出所：「本データ等の名称」（本データ等の作者名）（本データ等の URL）（バージョン表記などの日時情報）

（例 2）出所：「本データ等の名称」（本データ等の作者名）（本データ等の URL）をもとに（利用者名）が加工して作成

なお、本データ等に関しては、万全を期してはおりますが、正確性、確実性、目的適合性その他の品質を保証するものではありません。本データ等を用いて行うすべての行為に関して、その責任はすべて利用者自身に帰属します。

万が一、本データ等を用いたことによって利用者が何らかの損害を被った場合、その損害に関して情報提供者は一切の責任を負うものではありません。得られた情報に基づく決定は、本データ等の利用者ご自身でご判断いただきますようお願い申し上げます。

また、情報提供者は本データ等からアクセス可能な、第三者が権利を有する情報の正確性、信頼性、安全性を何ら保証するものではなく、第三者が権利を有する情報の利用により生じたいかなる損害に関しても、情報提供者は一切の責任を負うものではありません。

本データ等は、予告なく追加、変更、削除されることがありますので、あらかじめご了承ください。

（氏名）
（文書の公開年月日）

References

1) Shu Higashi. “Considering open data licenses (6)” (translated from Japanese) Open Data Commons, 2012.
http://okfn.jp/2012/10/24/opendata_license06/, (accessed 2019-12-20).

2) Masafumi Ono, Toshio Koike, Ryosuke Shibasaki. Survey for research data sharing in earth environmental information domain: Realities in research community. Journal of Information Processing and Management. 2016, Vol. 59, No. 8, p. 514-525.
https://doi.org/10.1241/johokanri.59.514, (accessed 2019-12-20).

3) Cabinet Office, Japan. “Government of Japan Standard Terms of Use (Version 2.0)”. 2015. https://www.kantei.go.jp/jp/singi/it2/densi/kettei/gl2_betten_1.pdf, (accessed 2019-12-20).

4) Liaison Committee of Relevant Ministries and Agencies and Working-Level Consultative Meeting on Collaboration of Digital Archives. “Guidelines for Building, Sharing, and Using Digital Archives” (translated from Japanese). 2017.
https://www.kantei.go.jp/jp/singi/titeki2/digitalarchive_kyougikai/guideline.pdf, (accessed 2019-12-20).

5) Shinnosuke Fukuoka and Hidetoshi Matsumura. Data laws and contracts. SHOJIHOMU. 2019, 440p.

6) Alex Ball. How to License Research Data. 2011. http://www.dcc.ac.uk/resources/how-guides/license-research-data, (accessed 2019-12-20).

7) FORCE11: Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. 2014. https://doi.org/10.25490/a97f-egyk, (accessed 2019-12-20).
(Research Data Utilization Forum (RDUF): Research Data Citation Subcommittee, Trans. https://doi.org/10.11502/rduf_rdc_jddcp_ja, (accessed 2019-12-20).)

8) RDA-CODATA: Legal Interoperability Interest Group. Legal Interoperability of Research Data. Principles and Implementation Guidelines. Zenodo, 2016.
https://doi.org/10.5281/zenodo.162241, (accessed 2019-12-20).

* All URLs in the text and footnotes are as of December 20, 2019.

Members (ver.1, as of December 2019)

Yasuyuki Minamiyama (National Institute of Informatics)
Ui Ikeuchi (Bunkyo University)
Kunihiko Ueshima (Japan Data Exchange, Inc.)
Misaki Suto (Formerly Mitsubishi UFJ Research and Consulting Co., Ltd.)
Nobuya Okayama (Hitachi Consulting)
Issaku Yamada (The Noguchi Institute)
Ken Ebisawa (Japan Aerospace Exploration Agency, Institute of Space and Astronautical Science)
Hodaka Nakanishi (Teikyo University)
Yui Kumazaki (Japan Atomic Energy Agency)

Glossary of "Guideline for specifying conditions of use in research data publishing"

No.	Term	Definition	Source/Reference
1	Falsification	Manipulating research materials, equipment, or processes to change data or results obtained from research activities.	Ministry of Education, Culture, Sports, Science and Technology, Japan. "Guidelines for Responding to Misconduct in Research" https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu12/houkoku/attach/1334660.htm, (accessed 2019-12-25).
2	Creative Commons	The name of a project or a non-profit organization that promotes the smooth distribution and reuse of copyrighted works by providing copyright holders with a means of indicating the conditions of use of their works with a simple mark. By declaring the conditions of use of texts, photos, videos, sounds, etc. on websites, etc., using the Creative Commons-defined marks in advance, the copyright holder can save users from having to go through the licensing procedure.	Japan Society of Library and Information Science, Dictionary of Terms Editorial Committee. Dictionary of Library and Information Science Terms. 4th edition. 2014.
3	Credit	Formal recognition of the contributions made by an individual or group to the research outputs.	RDA-CODATA: Legal Interoperability Interest Group. "Legal Interoperability of Research Data. Principles and Implementation Guidelines". Zenodo, 2016. https://doi.org/10.5281/zenodo.162241, (accessed 2019-12-25).
4	Research data	Digital data used as a source of information for scientific research. It includes a variety of formats, such as numerical, textual, image, audio, and video. In these guidelines, it does not include physical materials such as samples (specimens, samples) or recording media (paper, disks, etc.).	1) Cabinet Office, Japan. "Report of the Working Group on Research Data Infrastructure and Global Outreach - Strategies for the Development of Research Data Infrastructure and Global Outreach". 2019. https://www8.cao.go.jp/cstp/tyousak

			ai/kokusaiopen/houkokusho.pdf, (accessed 2019-12-25). 2) Japan Science and Technology Agency. “JST Policy on Open Access to Research Publications and Research Data Management”. 2017. https://www.jst.go.jp/pr/intro/openscience/policy_openscience.pdf, (accessed 2019-12-25). 3) OECD. OECD principles and guidelines for access to research data from public funding. 2007. https://doi.org/10.1787/9789264034020-en-fr, (accessed 2019-12-25).
5	Industrial Property rights	Four of the intellectual property rights, patent rights, utility model rights, design rights, and trademark rights, are referred to as industrial property rights. The purpose of a system of industrial property rights is to encourage and to motivate inventors of inventions and creators of designs, to protect their rights, and to instill confidence in the maintenance of business activities related to trademarks.	Japan Patent Office. “System of Industrial Property Rights”. https://www.jpo.go.jp/system/patent/gaiyo/seidogaiyo/chizai01.html, (accessed 2019-12-25).
6	Sample	A physical object that has substance, such as a specimen to be used for research.	Science Council of Japan. “Response: Improving Soundness in Scientific Research” (in Japanese). 2015. http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-23-k150306.pdf, (accessed 2019-12-25).
7	Information disclosure	A person with information shows it to another. This document distinguishes between "access to information," which is done by government agencies and academic institutions for the public. In other words, information disclosure may have some restrictions, such as limiting the scope and subject matter of disclosure, confidentiality	1) Ministry of Internal Affairs and Communications, Japan. “Access to Information System” (in Japanese). https://www.soumu.go.jp/main_sosiki/gyoukan/kanri/jyohokokai/index.html, (accessed 2019-12-25). 2) Strike Co., Ltd. “Glossary of M&A”

		obligations, and payment of compensation. When the transfer, lease, or assignment of rights to that information is involved, it is called "provision of information.”	https://www.strike.co.jp/maword/0331.html, (accessed 2019-12-25). 3) Cambridge Dictionary. https://dictionary.cambridge.org/ja/dictionary/english/disclosure, (accessed 2019-12-25).
8	Storage	External memory, one of the main devices that make up a computer, is used to store data for an unspecified period. This term is used to maintain its contents even when electricity is not supplied, such as hard disks, optical disks (CDs and DVDs), flash memory storage devices (USB memory sticks and memory cards), magnetic tapes, and so on.	1) Online Dictionary for Library and Information Science, https://www.abc-clio.com/ODLIS/odlis_s.aspx, (accessed 2019-12-25). 2) IDC Frontier Inc. “Glossary of Cloud / Data Center”. https://www.idcf.jp/words/storage.html, (accessed 2019-12-25).
9	Data sharing	Providing or disclosing data only to a limited number of subjects. “Data sharing” term is sometimes used in the sense of data publishing; these guidelines distinguish it from data publishing.	Cabinet Office, Japan. “Report of the Working Group on Research Data Infrastructure and Global Outreach - Strategies for the Development of Research Data Infrastructure and Global Outreach”. 2019. https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/houkokusho.pdf, (accessed 2019-12-25).
10	Data publishing	Publishing data to third parties on websites, repositories, or supplements to research papers and is accessible via the Internet.	1) Cabinet Office, Japan. “Report of the Working Group on Research Data Infrastructure and Global Outreach - Strategies for the Development of Research Data Infrastructure and Global Outreach”. 2019. https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/houkokusho.pdf, (accessed 2019-12-25). 2) G7 Science and Technology Ministers’ Meeting. “Tsukuba Communiqué”. 2016.

			https://www8.cao.go.jp/cstp/kokusaiteki/g7_2016/2016communique.html, (accessed 2019-12-25).
11	Data paper	A paper that describes the content, acquisition method, data format, access information, etc., regarding publicly available data such as observation data, measurement data, analysis data, and calculation simulation results. It does not include analysis, interpretation, or scientific conclusions.	1) Vishwas Chavan, Lyubomir Penev. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics Vol.12, S2, 2011. https://doi.org/10.1186/1471-2105-12-S15-S2, (accessed 2019-12-25). 2) Introduction of “Data Paper”: New Category for JAMSTEC-R Article, JAMSTEC Report of Research and Development, 2017, Vol. 24, P. 21-22. 2017, https://doi.org/10.5918/jamstecr.24.21, (accessed 2019-12-25).
12	Data policy	A statement of the data and information management processes that the organization has designed to support and protect the organization's research data assets. It is a set of high-level principles that establishes a guiding framework for data management. Data policies can be used to address strategic aspects such as data access, relevant legal matters, data management issues and storage operations, data acquisition, and other issues.	Research Data Canada and CASRAI. Trans-Disciplinary Glossary for Research Data Management. https://dictionary.casrai.org/Data_policy, (accessed 2019-12-25).
13	Anonymization	A manipulation that reduces the risk of personal identification by processing information that could directly/indirectly identify an individual. It is distinguished from “pseudonymization”; Pseudonymization is an operation that removes or separates information that can directly identify an individual by itself (e.g., name, mug shot, fingerprints, driver's license number, etc.) from other information. Anonymization involves	1) Personal Information Protection Commission, Government of Japan. “Guidelines for the Act on the Protection of Personal Information. Anonymized Information section”. 2017. https://www.ppc.go.jp/files/pdf/report_office.pdf, (accessed 2019-12-25).

		disambiguating or replacing age, gender, occupation, behavior logs, etc., so normal methods cannot recover the original information.	2) Jun Sakuma. Privacy Protection in Data Analysis: Machine Learning Professional Series. 2016.
14	Plagiarism	The appropriation of another person's ideas, processes, results, or words without giving appropriate credit.	U.S. Department of Health and Human Service, Office of Research Integrity. Definition of Research Misconduct. 2000. https://ori.hhs.gov/definition-misconduct, (accessed 2019-12-25).
15	Metadata	Data describing the characteristics of an information resource in order to effectively identify, describe, and explore it. Every academic community has its unique metadata tied to its information use practices, which are numerous and varied.	Japan Society of Library and Information Science, Dictionary of Terms Editorial Committee. Dictionary of Library and Information Science Terms. 4th edition. 2014.
16	License	To grant official permission for any action, use or possession. It also means an official document that sets forth the terms and conditions for a patented invention or the right to use intellectual property such as software. The Creative Commons License and the MIT License are well known in the academic community. In this document, the basic declaration of intent by the rights holder is called a "license indication" and is distinguished from the detailed "conditions of use" set forth in contracts or terms of use.	Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/licence, (accessed 2019-12-25).
17	Landing page	In a broad sense, it refers to the first page that a visitor sees through a website through a link or advertisement. It is distinguished from the top page, which is the entrance to a website. In the academic community, it is a page that describes the metadata and access methods for digital content in a repository.	1) NTTCom Online Marketing Solutions Corporation. Visionalist, "Glossary of Web Marketing". https://www.visionalist.com/glossaries/10_ra_001.html, (accessed 2019-12-25). 2) Basic Inc. ferret, "Web Marketing Dictionary". https://ferret-plus.com/words/1048, (accessed 2019-12-25). 3) Cambridge Dictionary.

			https://dictionary.cambridge.org/dictionary/english/landing-page, (accessed 2019-12-25). 4) Japan Link Center Joint Steering Committee. “Guidelines for Registering DOIs for Research Data”. https://doi.org/10.11502/rd_guideline_ja, (accessed 2019-12-25).
18	Repository (Digital repository)	Information systems that capture, store, manage, preserve, and provide access to digital content. Depending on the managing entity, there are different names for institutional repositories, discipline-specific repositories, government repositories, etc. The definition category also differs depending on the purpose of repository construction. For example, an open access repository is defined as a collection of full-text documents available in an online database on the Internet and characterized by free and immediate access.	1) Iris Xie, Krystyna K. Matusiak, Chapter 1 - Introduction to digital libraries, Discover Digital Libraries. 2016. https://doi.org/10.1016/B978-0-12-417112-1.00001-6, (accessed 2019-12-25). 2) Iris Xie, Krystyna K. Matusiak, Chapter 9 - Digital preservation, Discover Digital Libraries. 2016. https://doi.org/10.1016/B978-0-12-417112-1.00009-0, (accessed 2019-12-25).