Article

SAGE Open
July-September 2017: 1–13
© The Author(s) 2017
DOI: 10.1177/2158244017717301
journals.sagepub.com/home/sgo
SAGE

Identifying the Complex Position of Research Data and Data Sharing Among Researchers in Natural Science

Keiko Kurata[1], Mamiko Matsubayashi[2], and Shinji Mine[3]

Abstract

This article aims to provide an overview of researchers' practices and perceptions on data use and sharing. Semistructured interviews were conducted with 23 Japanese researchers in the natural sciences to identify their research practices and data use, including data sharing. We divided the interview scripts into meaningful phrases as a unit of analysis. Next, we focused on 406 statements on research data and reanalyzed them based on four aspects: stance on research data, practices and perceptions of data use, range of data sharing, and data type. A cluster analysis identified 14 clusters, which were divided into five groups: open access for data, restricted access for data, data interpretation, data processing and preservation, and data infrastructure. Our results reveal the complexity and diversity of the relationship between data and research practices. That is, the practice of research data sharing is heterogeneous, with no "one size fits all" between and among researchers.

Keywords

research data, research practices, data sharing, collaboration, contextuality, data ambiguity, data-sharing perspectives

Introduction

Science as a whole is making a paradigm shift toward data-driven research. The increasing level and development of data-driven research has raised an awareness of the importance of public access to data for all stakeholders in scholarly communication. Funding agencies in the United States and Europe—such as the National Science Foundation, Research Councils UK, and the European Commission—have begun to mandate data management plans as part of grant applications. One desired outcome of this data management is to increase the number of datasets shared among researchers and with the public. Moreover, data management plans may eventually evolve into a requirement to share data resulting from taxpayer-funded research.

The most successful example of this type of sharing involves the genetic sequence data deciphered by the Bermuda Principles in 1996 and the Human Genome Project in 2003, which stated that genetic sequence data should be released without delay, and public access to the data should be guaranteed. Far from becoming the scientific norm, however, the practice of data sharing does not appear to be prevalent in many fields. Although the significance of data sharing has been widely published and discussed, it has also raised some skepticism. In particular, many have expressed doubts and concerns about the effectiveness of data sharing, as we can see in expressions such as "empty archives" (Nelson, 2009) and "research parasites" (Longo & Drazen, 2016).

Nonetheless, data-sharing policies have increasingly been implemented by governments, funding agencies, universities, and journals worldwide (Jones, 2012; Stodden, Guo, & Ma, 2013). For example, the National Science Foundation makes data management and sharing plans mandatory for grant applications. Public Library of Science (*PLOS*) also requires authors to make their data available online upon publication. However, it is unclear whether data sharing can actually be considered prevalent throughout the scientific community. Considering the current, complicated situation regarding data and data sharing among researchers, asking researchers how appropriate it is to share data and whether they do so is insufficient. Research activities and data are closely intertwined, and exploring their contexts is a necessary step toward understanding research data and data sharing in the course of research activities.

However, few studies have focused on the manner by which researchers conduct and perceive the sharing of research data during research activities. Discussions on the

[1]Keio University, Tokyo, Japan
[2]University of Tsukuba, Tsukuba, Japan
[3]Mie University, Tsu, Japan

Corresponding Author:
Keiko Kurata, School of Library and Information Science, Faculty of Letters, Keio University, 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan.
Email: kei.kurata@keio.jp

subject typically concentrate on establishing a stance toward data sharing—either open or not. Therefore, it is necessary to examine researchers' individual situations, practices, and perceptions in greater depth and detail.

Literature Review

Numerous studies regarding data sharing among researchers have been published (Fecher, Friesike, & Hebing, 2015), and their primary focus has been the prevalence of actual data sharing (Tenopir et al., 2011; Tenopir et al., 2015) and the factors that influence the practice (Kim & Stanton, 2016; Kim & Zhang, 2015). However, there are many definitions of research data and data sharing, which makes interpreting and comparing the results of previous studies difficult. The reason for this confusion seems to be the lack of studies on the fundamental nature of the phenomenon; in other words, too few studies ask, "What is the definition of data and data sharing?" The dearth of studies regarding this topic is due to the fact that research data are highly ambiguous and context dependent. In the following section, previous studies are reviewed, focusing on (a) the ambiguity and (b) the contextuality of data and data sharing, and then the purpose of this study is presented.

Ambiguity in the Definition of Data and Data Sharing

"Data" is an umbrella term that covers a broad range of meanings and is thus difficult to define (Borgman, 2012, 2015). Many studies have failed to provide any definition of data or data sharing. The National Research Council's definition of data as "any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." (National Science Board, 2005, p. 8) has often been referenced in previous studies. However, we note that this definition only lists examples of possible forms of data.

At the opposite end of the spectrum, Hilgartner and Brandt-Rauf's (1994) "data stream" model provided a unique definition for data, and includes a variety of extremely heterogeneous entities in terms of both the input and output of scientific activities. In their model, data are described as "the result of experiments or surveys, biological materials and other samples, software, laboratory techniques, access to research sites, craft knowledge, and a wide variety of other forms of information and know-how" (Hilgartner & Brandt-Rauf, 1994, p. 359).

In many previous studies, data sharing simply means to "make data available to others" (Enke et al., 2012; Fischer & Zigmond, 2010; Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). This definition is just a paraphrase and is not sufficiently detailed to cover the diversity of data and data-sharing practices. For example, Tenopir includes the "deposition and preservation of data" as a part of data sharing (Tenopir et al., 2011), whereas Enke focuses on reuse and reanalysis by others and describes data sharing as "the practice of making one's data available to others or reusing it again for subsequent analyses" (Enke et al., 2012, p. 25). We note that the definitions of data sharing in both studies include "making data available to others." These two definitions, however, are markedly different from each other; indeed, data sharing has a great variety of definitions, and this variety is influenced by several factors. These factors include the following: (a) the methods involved in making data available to others, which range from "the attached datasets to published articles" (Wallis, Rolando, & Borgman, 2013, p. 2) to making data available to others through "their organization's website" or "national network" (Tenopir et al., 2011, p. 9); (b) the targets of data sharing, such as raw data or analyzed data; (c) the timing of data sharing; and (d) the range of data sharing, such as whether it is shared among members of a specific project, among all researchers in a given field, or among the general public.

The proportion of those who share data also varies depending on how questions about data sharing are asked and what types of data are actually being shared. For example, in Huang et al. (2012), 85.2% of biodiversity respondents reported having shared article-related data. By contrast, Wicherts, Bakker, and Molenaar (2011) reported that only 42.9% of respondents in psychology sent datasets that they used in their journal articles for reanalysis through a personal contact. We cannot identify the reason for these different percentages because the variety of methods to make data shareable is extremely diverse.

Data Contextuality

Data "exist in a context" and "have no value or meaning in isolation" (Borgman, 2015, p. 18, p. 4). It is not until contextualization occurs during the research process that data takes on its own meaning (Borgman, 2015; Stvilia et al., 2014; Swanson & Rinehart, 2016, p. 8). The significance of context for research data has long been demonstrated in previous studies (Birnholtz & Bietz, 2003; Borgman, 2015; Pryor, 2009), and the discussion of data contextuality can be summarized by two types of context. One is "explicit context," such as metadata, whereas the other is "implicit context," which includes tacit knowledge (Kowalczyk & Shankar, 2013).

First, explicitly describable context typically refers to metadata that have been produced or collected through research data. Representative examples include the name of the research instrument used, setup methods, properties, analytical methods, protocols, work flows, and the subjects of data collection. These metadata must be shared over places, organizations, and periods of time. Although researchers have difficulty in selecting which metadata can and should be provided (Borgman, 2012), the standardization of metadata for data sharing has begun to develop. For example, general data repositories, such as Dryad and Figshare, commonly require users to provide information describing the data's title, author, description, license, and subject keywords.

Second, research activities depend on informal knowledge, which is not easily documented, largely because it is often shared as implicit knowledge among various levels of research communities (Collins, 1983; Latour & Woolgar, 1986). Whether researchers can interpret the data produced by others depends on the availability of these implicit contexts. Although the simple act of sharing data is often supposed to be enough to ensure its uptake and impact (Davies & Edwards, 2012), researchers have repeatedly expressed concern that a lack of context would prevent them from sharing and reusing data as producers and reusers. Situations such as these can result in the misuse and misinterpretation of the data (Cragin, Palmer, Carlson, & Witt, 2010; Sayogo & Pardo, 2013). When making data available to others, a researcher cannot easily communicate the entirety of the context in which that data was produced, and it is difficult to interpret and reuse others' data without this contextual information. Although the importance of context in data sharing continues to be discussed, how and what type of implicit contexts are required in research activities is still unclear.

Purpose of This Study

As explained above, the concepts of research data and data sharing are ambiguous and context dependent, and researchers can take different meanings from data in different contexts. Research data are inextricable from the research process in which they are produced and are "an indispensable element of scientific research" (Tenopir et al., 2011, p. 2). To better understand data-sharing practices, we need to examine what researchers consider to be data and how they conduct and perceive data sharing during the process of research activities. We must also recognize the realities of data ambiguity and contextuality. The primary aim of this study is to examine researchers' perceptions and practices of data use and the manner by which they process, analyze, and store data. We accomplish this aim through interviewing researchers in diverse natural science fields.

Method

This study employed a combination of qualitative methods (i.e., interviews with researchers and content analysis of the statements obtained from the interviewees) and quantitative methods (i.e., cluster analysis on the results of the content analysis). Both approaches were utilized in this study because each provides unique insights (Guest, MacQueen, & Namey, 2012). Content analysis is appropriate for understanding the nuances of interviewees' statements, whereas cluster analysis helps marshal complex practices and perceptions of research data and presents their overview picture.

Table 1. The Profiles of Interviewees.

Institution	Subject	Position
National university	Information engineering	Associate professor
National university	Neutron science	Professor
National university	Quantum information science	Assistant professor
National university	Quantum electronics	Professor
National research institution	Nuclear physics	Professor
National university	Computer science	Associate professor
National university	Crystallography	Professor
National university	Plasma science	Assistant professor
National university	Molecular cell biology	Professor
National university	Space robotics	Professor
National university	Genetic disorder (pediatrics)	Professor
National university	Public health	Professor
National university	Space information science	Assistant professor
National research institution	Brain science	Professor
Private university	Regenerative medicine	Professor
National university	Biophysics	Professor
National university	Neuroscience	Professor
National university	Orthopedics	Associate professor
National university	Brain physiology	Assistant professor
National university	Physics and chemistry of solids	Professor
Private university	Metabolomic analysis	Assistant professor
National university	Particle physics	Associate professor
Private university	Algebraic analysis	Professor

Interviews With Researchers

Using snowball sampling, 23 Japanese active senior researchers were selected from various fields of the natural sciences, which is a discipline that exemplifies diversity among research practices. To be more precise, first we recruited some researchers whom our collaboration project members knew as interviewees before, and then we asked them to introduce some other interviewees. As a result, the majority of the interviewees were senior professors at relatively large universities. Profiles of the interviewees are shown in Table 1.

We conducted semistructured interviews in Japanese, lasting 60 to 90 min during which we asked the researchers to describe three topics: (a) the purpose and various methods of their research practices, especially how they collaborate with other researchers; (b) the relationship between their research practice and research data, that is, the positioning of research data within their research and their methods for processing and preserving research data; and (c) whether they share and reuse research data. Each interviewee was asked a set of questions customized for their situations to explore these three topics in detail. We did not ask any questions about specified issues on data sharing or open research data such as the preparation of metadata, data-sharing policies by funding agencies or journal publishers, and perceived effort in order not to induce the interviewees to mention the issues we think are important. We would like to elicit a diversity of opinion or views on research data and the relationship between their research practice and research data. We recorded and transcribed all the interviews.

Analysis Procedure

We conducted a content analysis in Japanese on the statements acquired from the interviews. Initially, each author coded the data without any constraints. Specifically, we divided each statement into small segments (some were phrases, others were a few sentences in length), and then for each segment, we assigned keywords that described the content discussed in each statement. We attached 158 keywords to the 5,086 segments acquired. Because the 5,086 items with these 158 keywords are heterogeneous, in this article, we analyzed only the 406 statements with keywords related to "research data." All 406 target statements are related to the features, interpretations, processing, and preservation of research data and data sharing. We numbered these 406 statements. In the following descriptions, the parenthesized numbers function as unique identifiers that correspond to the 406 items. In addition, we translated only some statements into English, which were illustrated as examples in this article. On translating the statements into English, we paid careful attention not to change their nuances of meaning.

We examined these 406 statements and uncovered four aspects that every statement had in common. Then, we recoded the 406 statements using these four "aspects," which are (a) stance on research data, (b) practices and perceptions of data use, (c) range of data sharing, and (d) data type. One reason we employed these aspects is that all four of them were present in every target statement, as we discovered through the results of the content analysis. The other reason is that all these aspects have been referenced in previous studies of research data and data sharing. We display the options within each of the four aspects in Table 2.

The first aspect, "stance on research data," has been addressed in previous studies of data sharing. We provided four options regarding this aspect from which the researchers could choose: (a) open, (b) restricted, (c) interpretation, and (d) N/A. Previous studies typically discussed two stances, usually one in favor of data sharing (such as the "open" option) and one that is against it (such as the "restricted" option). This article introduces a third option, "interpretation." We assigned this option to statements that describe researchers' complicated response to the question, "What are data?" We considered these statements to be distinct from a simple dichotomy between "open" and "restricted." To provide further specificity, we assigned statements in which researchers made no comments about their stance regarding data sharing to "N/A."

Table 2. Four Aspects for Reclassifying Statements in "Research Data."

Aspect	Option	Description
Stance on research data	Open	Do or should make research data open
	Restricted	Do not or cannot make research data open
	Interpretation	Do or need to interpret research data
	N/A	Neither
Practices and perceptions of data use	My data	Research data are mine
	For science and society	Research data are for science or society
	National or institutional policy	Influence of national or institution policy of research data
	Preservation	How to preserve research data
	Processing	How to operate data processing
	Issues on research data	Problem or issues or challenge in dealing with research data
	Specific opinion on data	Personal or unconventional or specific manner (way) of data use in research
Range of data sharing	Small-scale collaboration	Sharing among researchers in small-scale collaboration
	Large-scale research project	Sharing among researchers in large-scale project research
	Open to the public	Data sharing for all researchers or general public
Data type	Raw data	Raw data such as monitoring data, empirical data
	Analyzed data	Data after any analysis
	Material data	Target or object for test or experiment
	Statistical data	Public statistical data
	Clinical data	Clinical data in medicine
	N/A or multiple	Neither or multiple

We configured the practices and perceptions of data use into one aspect because these two factors (elements) were indistinct combinations in each statement. We determined the seven options for the second aspect, "practices and perceptions of data use," using a bottom-up analysis. The first two options, "my data" and "for science and society," refer to statements that explain why interviewees have a positive or negative stance on data sharing. Statements by interviewees who believe that their research data should be private were classified into "my data," whereas those made by interviewees who believe that data should be shared more broadly were included in the "for science and society" option. The third option, "national or institutional policy," includes statements by interviewees who feel that the choice of whether to share data is affected by national and institutional policies. Options 4 and 5—"preservation" and "processing," respectively—were assigned to statements in which interviewees claim to process and preserve research data prior to the actual act of data sharing. These two options are of a different type than the previous three because they engage with interviewees' actual research practices, whereas the other options describe interviewees' perceptions about research data. Option 6, "issues on research data," contains statements that address various problems with processing research data in research practices. The final option, "specific opinion on data," refers to statements that describe interviewees' opinions about the nature of data or the relationship between science and research data.

We offered three options—"small-scale collaboration," "large-scale research project," and "open to the public"—for the third aspect, "range of data sharing." The term "data sharing" may not typically imply sharing research data in small-scale collaboration but instead means making data resources available to the public. In this study, however, we set the three ranges of data sharing to emphasize the interviewees' actual behaviors and perceptions of data sharing.

Finally, we offered five options for the fourth aspect, "data type," because researchers handle various types of data in their research activities, and this variation affects perceptions about research data. The first five options are "analyzed data," "material data," "clinical data," "statistical data," and "raw data." "N/A" in the data type means that a statement either mentioned nothing about data type or included a mention of more than one data type.

Ultimately, we assigned four options (one for each aspect) for each item in the statements. For example, we chose "restricted," "my data," "open to the public," and "analyzed data" for the statement, "We paid 500,000 yen for the analysis, so others being able to see that data for free doesn't feel too great. We won't share the data with others if we don't need to do so" (28). We chose the option "restricted" because the statement said, "We won't share the data with others." Next, we considered that the researchers' paying money for the analysis themselves generated the idea that the data are mine, leading us to assign the option "my data" to the statement. The statement mentioned that allowing other researchers with no relationship to this research to use that data for free does not feel too great. Therefore, we chose the option "open to the public" for the range of data sharing. We chose "analyzed data" because it said, "We paid 500,000 yen for the analysis."

In another example, we assigned "interpretation," "specific opinion on data," "open to the public," "N/A" for the statement, "Production of data is not a science. It only becomes science after interpretation of the data, expressing some new idea, and then having someone else understand it" (382). This statement does not determine whether research data should be open or restricted; instead, it mentions the nature of research data. In other words, research data needs to be interpreted by researchers to exist as itself. Therefore, we assigned the option "interpretation" to the statement. Moreover, the statement "production of data is not a science" is this interviewee's distinctive opinion, so we assigned the option "specific opinion on data" to the statement. By contrast, the statement did not mention the specific ranges or types of data. Therefore, we assigned the option "open to the public" and "N/A" to the statement.

Finally, we conducted the SPSS TwoStep cluster analysis using the dataset obtained through these methods, which involved examining all 406 statements with each option from four aspects.

Results

We identified 14 clusters from the SPSS TwoStep cluster analysis for each option from the four aspects to best represent the characteristics of each statement. We used the TwoStep cluster analysis because it is effective for very large datasets, and it is an exploratory tool designed to reveal natural clusters within a dataset. The silhouette coefficient (SC) value shows cluster quality. If its score is between 0.2 and 0.5, the cluster quality is "fair" (cluster quality is "good" if >0.5, and "bad" if <0.2). The SPSS TwoStep cluster analysis identified 14 clusters that best represented the characteristics of each of the statements. The value of the SC in the analysis was 0.4, so the quality of our clusters was "fair."

We divided these 14 clusters into five groups to obtain a complete view as a matter of practical convenience. We based the division of groups, labeled "A," "B," and "C," around the options chosen by the researchers for the "stance on research data" aspect. We found that in the three clusters of Group "A," almost 100% of the statements were "open" options regarding the "stance on research data" aspect. The results of the two clusters of Group "B" indicated more than 90% of statements were assigned the "restricted" option, and 100% of statements in three clusters of Group "C" were assigned the "interpretation" option. We also included Groups "D" and "E," which have no features regarding the "stance on research data" aspect, but instead have characteristics regarding the "practices and perceptions of data use" aspect. We found that in the three clusters of Group "D," more than 80% of statements were assigned to the "processing" or "preservation" option. Group E had three clusters with statements that were predominately assigned to the "national or international policy," "specific opinion on data," or "issues on research data" option in the "practices and perceptions of data use" aspect.

Table 3 presents the labels for the 14 clusters as well as the percentages representing the distribution of the statements across the four aspects. The figure noted in parentheses under the cluster number indicates the number of statements included in each cluster. The following section describes the characteristics of the 14 clusters with reference to the specific statements.

Group A: Open Access for Data

In the three Group "A" clusters (Clusters 1, 2, and 3), almost 100% of responses were assigned the option "open" within the "stance on research data" aspect. For Clusters 1, 2, and 3, almost 100% of the statements were assigned the option "for science and society" within the "practices and perceptions of data use" aspect.

In Cluster 1, 93.8% of statements were assigned the option "for science and society" within the "practices and perceptions of data use" aspect, and all statements were assigned the option "open to the public" within the "range of

Table 3. Labels for the 14 Clusters and the Percentage of Statements Across the Four Aspects.

Group	Cluster number	Label	Stance on research data (%)		Practices and perceptions of data use (%)		Range of data sharing (%)		Data type (%)	
A	1	Data sharing in public for science and society	Open	100.0	For science and society	93.8	Public	100.0	Analyzed data	37.5
	(*n* = 32)				Specific opinion	6.2			Material data	31.2
									N/A (or multiple)	15.6
									Statistical data	12.5
									Clinical data	3.1
	2	A system for sharing raw data	Open	97.0	For science and society	100.0	Public	72.7	Raw data	100.0
	(*n* = 33)		N/A	3.0			Project	27.3		
	3	Data sharing within small-scale collaboration	Open	100.0	For science and society	73.1	Collaboration	80.8	Analyzed data	46.2
	(*n* = 26)				Processing	19.2	Project	19.2	Raw data	38.5
					Issues on research data	3.8			N/A (or multiple)	7.7
					Preservation	3.8			Statistical data	7.7
B	4	Restrictions in sharing and use of data	Restricted	100.0	My data	68.4	Collaboration	55.3	Raw data	47.4
	(*n* = 38)				Issues on research data	15.8	Project	44.7	Analyzed data	36.8
					Processing	5.3			Clinical data	13.2
					Preservation	5.3			Statistical data	2.6
					For science and society	2.6				
					National or institutional policy	2.6				
	5	Unwilling to share data due to my data	Restricted	92.6	My data	100.0	Public	100.0	Analyzed data	37.0
	(*n* = 27)		Interpretation	3.7					Raw data	29.6
			N/A	3.7					Clinical data	18.5
									Statistical data	14.8
C	6	Specific opinion on data	Interpretation	100.0	Specific opinion on data	100.0	Public	78.9	Analyzed data	57.9
	(*n* = 19)						Project	21.1	Raw data	21.1
									N/A (or multiple)	10.5
									Clinical data	10.5
	7	Standard for data interpretation and judgment	Interpretation	100.0	Issues on research data	100.0	Public	100.0	Analyzed data	100.0
	(*n* = 17)									
	8	Sole data cannot be interpreted	Interpretation	100.0	My data	50.0	Collaboration	55.6	Raw data	55.6
	(*n* = 18)				Issues on research data	44.4	Project	44.4	Analyzed data	38.9
					For science and society	5.6			Material data	5.6

(continued)

Table 3. (continued)

Group	Cluster number	Label	Stance on research data (%)		Practices and perceptions of data use (%)		Range of data sharing (%)		Data type (%)	
D	9	Practices of data processing	Open	50.0	Processing	100.0	Project	61.5	Raw data	61.5
	(*n* = 26)		Restricted	50.0			Public	38.5	Analyzed data	23.1
									N/A (or multiple)	11.5
									Material data	3.8
	10	Processing and preservation of analyzed data	N/A	100.0	Processing	60.5	Collaboration	55.3	Analyzed data	71.1
	(*n* = 38)				Preservation	23.7	Public	31.6	Clinical data	21.1
					Issues on research data	13.2	Project	13.2	Raw data	5.3
					Specific opinion on data	2.6			Statistical data	2.6
	11	Processing and preservation of raw data	N/A	100.0	Preservation	59.6	Collaboration	51.1	Raw data	100.0
	(*n* = 47)				Processing	25.5	Project	46.8		
					Issues on research data	8.5				
					National or institutional policy	4.3				
					My data	2.1				
E	12	Policy effects	Open	71.0	National or institutional policy	100.0	Public	93.5	Statistical data	54.8
	(*n* = 31)		Restricted	25.8			Project	6.5	Analyzed data	12.9
			N/A	3.2					Raw data	9.7
									N/A (or multiple)	9.7
									Material data	6.5
									Clinical data	6.5
	13	Data management and organization	Restricted	53.3	Specific opinion on data	40.0	Public	66.7	N/A (or multiple)	100.0
	(*n* = 15)		N/A	46.7	National or institutional policy	33.3	Project	26.7		
					My data	13.3	Collaboration	6.7		
					Issues on research data	6.7				
					Preservation	6.7				
	14	Standardization	Open	38.5	Issues on research data	100.0	Public	97.4	Raw data	30.8
	(*n* = 39)		Restricted	35.9			Project	2.6	Analyzed data	17.9
			Interpretation	20.5					Material data	17.9
			N/A	5.1					Clinical data	15.4
									Statistical data	10.3
									N/A (or multiple)	7.7

data sharing" aspect. Cluster 1 suggests that research data must be generally open to the public because data sharing can contribute to the advancement of the sciences and the general good of society. Specific statements included "General and widely used information . . . [such as] GenBank data . . . that is accessed by everyone should be left open" (325). Cluster 1 is labeled "Data sharing in public for science and society."

Although Cluster 2 is similar to Cluster 1 in that the option "for science and society" was assigned to all statements, Cluster 2 is unique in that all statements were assigned to the option "raw data" with regard to the "data type" aspect. Many of the interviewees' statements provided examples of how they implemented raw data-sharing systems into large research projects. For example, one statement mentioned that "with regard to a data-sharing platform . . . we are currently in the process of establishing a system using a GIS modified for outer space whereby various users can share data on the basis of its mapping information" (76). Due to the emphasis on raw data among its statements, Cluster 2 is labeled "A system for sharing raw data."

In contrast to Clusters 1 and 2, the statements in Cluster 3 describe a more primitive style of small-scale collaborations performed within individual university laboratories. In Cluster 3, 80% of statements were assigned the option "small-scale collaboration" with regard to the "range of data sharing" aspect. Example statements described the practice as, "We brought together in-progress data during analyses, and then we discussed the problem or further perspectives of research" (1), whereby research data are shared in conventional collaborative research as a part of an established, results-oriented process. Cluster 3 is labeled "Data sharing within small-scale collaboration."

Group B: Restricted Access for Data

In contrast to Group "A," the two clusters in Group "B" consisted of statements from interviewees who are either opposed to the idea of data sharing or who insist on some restrictions. Almost 100% of statements in both Clusters 4 and 5 were assigned the option "restricted" within the "stance on research data" aspect.

Statements in Cluster 4 typically described examples and situations in which data could not be shared within research projects—neither in small-scale collaborations nor in large-scale research projects. This is illustrated by the statement "In collaboration with private companies, we discussed and shared research results based on analyzed data, but they are unwilling to share raw data, even with members of the research team" (112). There were many cases in which collaborative research involved private companies that were reluctant to share data, although this stance varied among cases. We did not find a correlation between the stance on data sharing and the research field.

As with their small-scale counterparts, larger research projects experienced restrictions on data sharing. For example, in the A Toroidal LHC ApparatuS (ATLAS) project, which is one of the seven particle detector experiments conducted at the European Organization for Nuclear Research (CERN), only members conducting ATLAS experiments can access raw data from those experiments, and they are not able to share other experimental raw data that was obtained from the other six detectors. Therefore, Cluster 4 is labeled "Restrictions in sharing and use of data."

The most distinctive trait of Cluster 5 is that 100% of statements were assigned the option "my data" regarding "practices and perceptions of data use" aspect. The typical example in Cluster 5 is the following statement:

> Only the person who creates data can use the data. For example, in a clinical study, the attending physician derives a variety of data from his or her patients, which would not be shown to anybody else. So this in turn becomes "my data." (53)

This way of thinking, which considers data as belonging only to the researcher who gathered it, also leads to a lowered priority for sharing data with others, as well as feelings of being at a disadvantage in showing their data to others. Cluster 5 is labeled "Unwilling to share data due to my data."

Group C: Data Interpretation

Group "C," which includes Clusters 6, 7, and 8, consists of statements regarding data or data sharing that were not categorized as either "open" or "restricted." Many of the individuals interviewed in this study found it difficult to explicitly state a stance on research data; some interviewees explained the reason for their difficulty in terms of the nature of data, the relationship between data and research practices, and the importance of data interpretation. During the reanalysis of statements from four aspects for cluster analysis, we assigned these statements to the option "interpretation" within the "stance on research data" aspect. All statements in these three clusters were assigned to this "interpretation" option, making this the unique characteristic of Group "C."

Cluster 6 is characterized by statements assigned to the option "specific opinion on data" within the "practices and perceptions of data use" aspect. For example, one interviewee stated his fundamental philosophy toward data and research as follows: "Production of data is not a science. It only becomes science after interpretation of the data, expressing some new idea, and then having someone else understand it" (382). This statement recognizes that data are not inherently valuable; rather, the value of the research is imparted to and derived from the process of data interpretation. Cluster 6 is labeled "Specific opinion on data."

Cluster 7 includes statements that indicate a level of self-confidence in interpreting and making value judgments about data. For example, one interviewee stated that researchers looking at graphs and charts would generally yield a similar interpretation of the data: "This is probably a very specialized example . . . and is an easy-to-understand event" (297). In this case, the subject suggested that although the authors did not understand the graph generated by someone else that the interviewee showed us during the interview, the interviewee would be able to explain the meaning of the graphical data at a glance. Another statement expressed suspicion toward data represented in academic articles in which "sometimes, data might appear too clean; we view them with a skeptical eye" (278). In other words, these interviewees have expertise that enables them to judge data produced by other researchers. Cluster 7 is labeled "Standard for data interpretation and judgment."

Cluster 8, which may appear to be juxtaposed with Cluster 7, consists of statements in which interviewees were unable to interpret research data that were not obtained from their own experiments. One statement categorized in Cluster 8 mentioned, "Experimental data are not sufficient for analyzing, just as data is. What method did they use, or which procedure did they conduct? Metadata are essential to properly analyze or understand" (181). Statements in this cluster stress a need for contextuality or background information to properly analyze data. Furthermore, some statements in this cluster suggested that an analysis can only be carried out by the researchers who participated in a given experiment; for example, "Ultimately, data have to be analyzed by people who conduct the experiments that generate them or who have the same level of expertise as the original researchers" (203). Cluster 8 is labeled "Sole data cannot be interpreted."

Group D: Data Processing and Preservation

Clusters 9, 10, and 11 are included in Group "D," and each cluster contains statements that describe the practices by which data are processed and preserved. Although data preservation is not synonymous with data sharing, data sharing cannot occur without data preservation; therefore, data preservation is understood to be a necessary precursor to data sharing.

In Cluster 9, 100% of the statements were assigned the option "processing" within the "practices and perceptions of data use" aspect. Interviewees' "stance on research data" was split between the "open" and "restricted" option, and in the "data type" aspect, both the "raw data" and "analyzed data" options were included. Many statements referred to instances in which the practice of data processing was employed to promote sharing, such as the development of software that could be used to standardize data for international-sharing protocols and convert it into widely used formats. A progressive opinion on data sharing is as follows: "It's not sharing in the physical sense as it was in the past. Instead of sending analyzed data, you could get answers whenever you make inquiries into the system" (159). In the system described in this statement, the data itself was retained by each researcher (or group), but researchers could obtain responses to inquiries for specific information. This process is a sophisticated mode of sharing in which the data itself is not distributed; instead, it is only the "information" obtained from the data that can be shared. Cluster 9 is labeled "Practices of data processing."

Cluster 10 comprises statements about the processing and preservation of previously analyzed data. Many statements indicated an abundance of standardized software or tools for data analysis: "We're developing software that would allow the computer to analyze results from mass spectrometry analyses" (262) and "they've begun providing Linux-compatible versions of legitimate analysis software" (268). Although these statements are not directly discussing data sharing, the standardization of analysis methods enables the comparison of results derived using those methods, which thus may be read as an indication of the necessity of data sharing. Conversely, some statements in Cluster 10 indicated that the utility of other researchers' data would be limited, because "there probably wouldn't be what we could refer to as commonality, or universality [in the data]" (386). Furthermore, a number of interviewees suggested that analyzed data do not necessarily need to be preserved for long periods of time: "Increased computer performance would likely allow calculations to be performed in a split-second, so there is little point in the storage of analyzed research data" (347). Cluster 10 is labeled "Processing and preservation of analyzed data."

In Cluster 11, 100% of the statements were assigned to the option "raw data" within the "data type" aspect. For the aspect "practices and perceptions of data use," approximately 60% of the statements were assigned to the option "preservation," many of which refer to small-scale collaborations; approximately 25% were assigned to the option "processing" of massive amounts of data in large-scale research projects. Examples included statements about data preservation in university laboratories (small-scale collaborations), such as "The graduate students would report once every two weeks. We would have them create a summary and constantly include raw data in them. . . . The summaries were then all stored on a shared computer in the lab" (356). As previously mentioned, raw data, especially in massive amounts, cannot be analyzed by researchers in their natural form. The statements in this cluster thus provide specific examples of automatic processing and observations on how massive the data being processed can become. For example, "a few gigabytes of data were generated only by a one-time measurement" (158) and "trigger system selects promising events to store its data for further analysis" (169). Cluster 11 is labeled "Processing and preservation of raw data."

Group E: Data Infrastructure

Group "E" is similar to Group "C" in that both incorporate statements in which the stance of data was not categorized as either "open" or "restricted." However, the three clusters (Clusters 12, 13, and 14) in Group "E" contain statements that address the issue or factor supporting research data management and the proceeding data sharing. These clusters usually address behind-the-scenes roles or issues; meanwhile, Group "E" is referred to "Data infrastructure."

Cluster 12 was characterized by 100% of statements being assigned to the option "national or institutional policy" within the "practices and perceptions of data use" aspect. Although statements in Cluster 12 varied at the international, national, and institutional levels, they tended to highlight examples of the ways that policies promote open access to data as well as the problems caused by delays in establishing related regulations and how current laws can act as obstacles in efforts toward open access. One positive perspective on the relationship between policy and data sharing can be found in the statement, "in Finland, the national database was constructed to follow the prescriptions of every single citizen" (54). A negative example is, "In Japanese legislation, we cannot create a National Death Index (a database on causes of death)" (243). Because it includes statements that refer to the influence of policies on data sharing, Cluster 12 is labeled "Policy effects."

In Cluster 13, half of the statements expressed the "restricted" option in the "stance on research data" aspect, whereas the other half did not express any stance (these statements were assigned the option "N/A" in "stance on research data" aspect). Although all options regarding the "practices and perceptions of data use" aspect were assigned, many statements in this cluster referred to issues regarding data management and organization. The statements about data organization expressed that "it seems more convenient to have a unified data format with regards to standardization, but it's rare to have such rules established" (402) and alternatively that "data should be provided in an open format because data can be freely analyzed" (376). This statement (376) argued that the free format was indispensable. In other words, the researcher behind this statement is suspicious of the idea of formatting data for its own sake in case the original results are based on freewheeling thinking.

Although these two statements represent opposing sides of an argument, both indicate the importance of considering data formats when organizing data. Another statement, assigned to the option "specific opinion on data" within the "practices and perceptions of data use" aspect, highlighted the manner by which the construction of databases contributes to the improvement of the discipline of science. For example, one interviewee mentioned that "[for science], only understanding the data is insufficient; to overcome is essential. Data must be essentially organized [for science] to improve" (374). Cluster 13 is labeled "Data management and organization."

In Cluster 14, there are no characteristic tendencies on "stance on research data" aspect. Of the statements, 100% were assigned to the option "issues on research data" within the "practices and perceptions of data use" aspect. Many statements in this cluster addressed standardization, a standard data archive, and standard software across different disciplines. Some statements expressed the belief that data sharing has become widespread, pointing to the construction of data collection, such as Text REtrieval Conference (TREC), which serves as a benchmark for information retrieval experiments, or GenBank, which has become a fundamental data archive that almost all researchers in genetic research fields use to promote their research. Other interviewees presented opposing perspectives in comments such as, "Even NASA has not provided observational data in the standard way that would allow all researchers can use conveniently" (93). These contradictory statements expose the difficulty in providing a standard way to access research data. Each researcher has his or her individual research purpose, style, and methods for using data.

Certain statements in Cluster 14 highlight the logistical obstacles to comparing shared research data, such as the following: "It's essentially just photographing what naturally occurs, so it's easy to make comparisons. But if we have to actively intervene to measure humans' reaction, then the results differ greatly depending on how the intervention is made, so making comparisons through commonality becomes much more difficult" (135). The comparability of data is partially determined by the nature of the data itself and by the process through which the data are generated. Therefore, standardization is necessary for the comparison of data. Whether experimental tools and environments can be standardized depends on the research methods and object of the experiment (see "nature" and "human reaction" in Statement 135). This example highlights the difficulty involved in comparing research data and includes determining data contextuality in a broader sense. Cluster 14 is labeled "Standardization."

Discussion

This study focused on the position of data in research practices and the meaning of data use and data sharing for researchers. To understand the complex and diverse relationship between research data and actual research practices, we adopted the approach of combining qualitative and quantitative analyses of the data obtained from interviews of 23 natural science researchers. In the first step, we analyzed researchers' statements on research data from the bottom up (detailed content analysis). Many elements and combinations were extracted from more than 400 statements; thus, we recoded statements based on four aspects and then conducted a cluster analysis for the recoded data. We obtained 14 clusters automatically using this cluster analysis method. We based the labels and characteristic descriptions of these clusters on our interpretation of the statements included in each cluster. Our results revealed a representative, integrated set of categories of diverse researchers' practices and perceptions with regard to data.

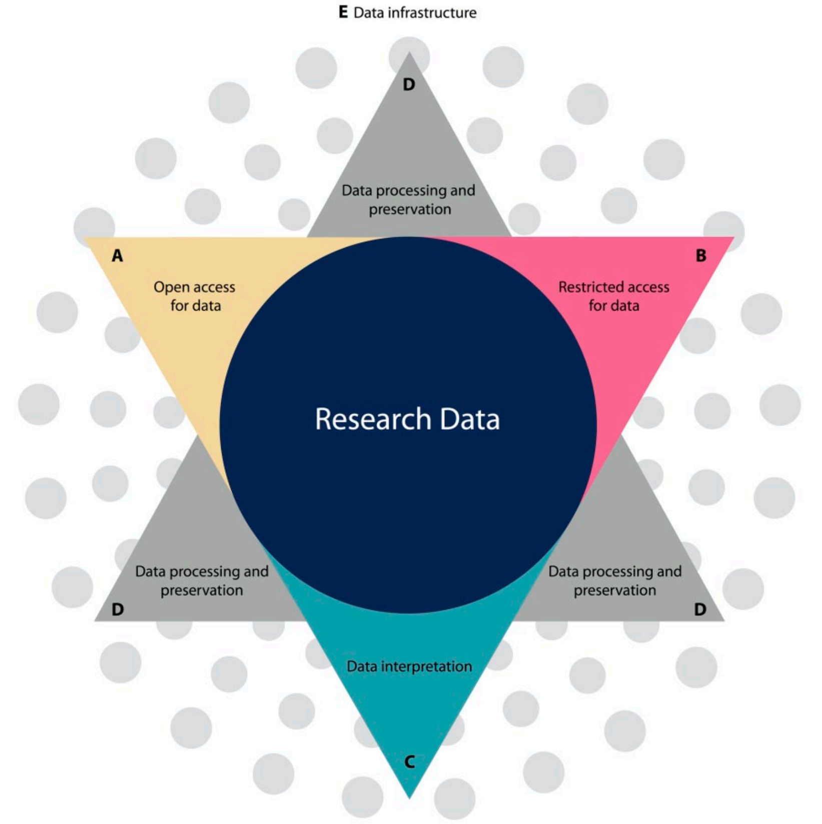
E Data infrastructure
D
Data processing and preservation
A
Open access for data
B
Restricted access for data
Research Data
Data processing and preservation
D
Data processing and preservation
D
Data interpretation
C

Figure 1. Concept image of the relationship among researchers' perceptions and practices regarding research data.

Finally, we divided the 14 clusters into five groups based on the characteristics of each of the 14 clusters. Thus, we had five perspectives that provided an overview of the relationship between research data and research practices. Figure 1 shows the relationship among these five perspectives, which constitute three layers centering over "research data." The first layer is formed (framed) by three stances on data: "Open access for data (A)," "Restricted access for data (B)," and "Data interpretation (C)." This layer indicates a more surface perception of researchers regarding data. The second layer is "Data processing and preservation (D)," which is considered to embody actual research practices supporting the perception of data sharing. The third layer provides background for the awareness of data sharing and research practices. In other words, this layer functions or interferes with data sharing or data use openly.

Previous studies of data use and sharing have generally focused on the simple dichotomy between "open" and "restricted" stances regarding these issues. As described in the "Results" section, many researchers in our study expressed a feeling of strangeness or difficulty with this dichotomy. Many statements, as typified in the perspective "Data interpretation (C)," emphasized the interpretation of research data in a different context.

In addition, a number of previous studies focused on factors that promote data sharing (Fecher et al., 2015; Kim & Stanton, 2016). Some of the 14 clusters found in our studies are consistent with factors in previous studies. "Policy effects," "Data management and organization," "Standardization" in perspective, "Data infrastructure (E)" and "Data sharing in public for science and society" are common. However, a complicated awareness of data use or data sharing and actual practices regarding data processes or interpretation can be identified as our independent contribution. For example, "Unwilling to share data due to my data (Cluster 5)" expressed the researchers' opinion or feeling that data belong to the researchers who collected or produced them in their investigation. This awareness is not necessarily the same as the intention to avoid cost and efforts. It may be related to the contextuality of data in complex research processes, which we will examine below.

Borgman (2015) is one of a handful of researchers who addresses the complexities of data collection and data handling. She illustrated that the characteristics of data differ by research purpose or methods. Based on her point of view, we analyzed the relationship between position of data and research practice. In Figure 1 obtained from our analysis, the "Data processing and preservation (D)" perspective can be positioned as the second layer, which serves to connect the researchers' stance on data and data sharing (the first layer) with the data infrastructure (the third layer). However, we did not clarify the mechanism by which a specific research method or data type inevitably promotes data sharing. In specific research fields, such as genetic research, many researchers consider it to be the scientific community's norm to make genetic data open in GenBank. In contrast, there are cases in which researchers never provide genetic data even to research collaborators.

To clarify the relationship between researchers' stance on data use and data sharing and research practices, the concept of data contextuality is an important point. We examine this concept from two levels, explicit and implicit, which were mentioned briefly in the "Literature Review" section.

At a specific level, a significant number of statements mentioned the importance of the format and standardization of research data for data sharing, although researchers did not use the term "metadata." Some statements, however, expressed the researchers' fears or doubts about the effects of standardization. Regarding implicit data contextually, several statements referenced the belief that data cannot be fully analyzed or understood except by the researchers who conducted the experiments or investigation in which the data were obtained. These arguments may provide insight into the "my data" mentioned above. Although the perception of "my data" has been criticized by proponents of data sharing as being egotistic, it also reflects a fundamental suspicion about a researcher's ability to analyze data from other researchers. The doubt or fear that arises from using data in isolation from the context of the research that created it seems to be an important point when promoting data sharing.

Finally, we indicated that the scholarly information in research frontiers is not a "public good" but rather a "club good" or a "contribution good" (Kealey & Ricketts, 2014) in an economic sense. Generally, information provided by governments is considered a public good because any use (consumption) by individuals does not reduce the amount of information and does not exclude any other use. Although scholarly knowledge has also been regarded as a "common" resource, sometimes only researchers are able to truly understand it, allowing them to taking advantage and contributing to its accumulation. Current movements promoting open research data are considered to be disrupting the exclusive characteristics that research naturally has.

From any standpoint—the definition of data, the contextuality of data, and the exclusive property of research—the attempt to promote data sharing will force the research community to remodel or reconstruct the conventional system or norms of research practice and researchers' perception.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:: This article was prepared with support from JSPS KAKENHI grant numbers JP23300089 and JP26280121.

References

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. In M. M. Tremaine & C. Simone (Eds.), *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 339-348).Sanibel Island, FL, New York: ACM Press.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059-1078.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world* (1st ed.). Cambridge, MA: MIT Press.

Collins, H. M. (1983). The sociology of scientific knowledge: Studies of contemporary science. *Annual Review of Sociology*, *9*(1), 265-285.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023-4038.

Davies, T., & Edwards, D. (2012). Emerging implications of open and linked data for knowledge sharing in development. *IDS Bulletin*, 43(5), 117-127.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing: Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, *11*, 25-33.

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS ONE*, *10*(2), e0118053.

Fischer, B. A., & Zigmond, M. J. (2010). The Essential nature of sharing in science. *Publishing Research Quarterly*, *16*(4), 783-799.

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. London, England: Sage.

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data Access, ownership, and control: Toward empirical studies of access practices. *Knowledge: Creation, Diffusion, Utilization*, *15*(4), 355-372.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. *Conservation Letters*, *5*(5), 399-406.

Jones, S. (2012). Research data policies: Principles, requirements and trends. In G. Pryor (Ed.), *Managing research data* (1st ed., pp. 47-66). London, England: Facet Publishing.

Kealey, T., & Ricketts, M. (2014). Modelling science as a contribution good. *Research Policy*, *43*(6), 1014-1024.

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776-799.

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library & Information Science Research*, *37*(3), 189-200.

Kowalczyk, S., & Shankar, K. (2013). Data sharing in the sciences. *Annual Review of Information Science and Technology*, *45*(1), 247-294.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts* (2nd ed.). Princeton, NJ: Princeton University Press.

Longo, D. L., & Drazen, J. M. (2016). Data sharing. *New England Journal of Medicine*, *374*(3), 276-277.

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century: Report of the National Science Board*. Arlington, VA: National Science Foundation.

Nelson, B. (2009). Data sharing: Empty archives. *Nature*, *461*(7261), 160-163.

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE*, *9*(8), e104798.

Pryor, G. (2009). Multi-scale data sharing in the life sciences: Some lessons for policy makers. *International Journal of Digital Curation*, *4*(3), 71-82.

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*(1), S19-S31.

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE*, *8*(6), e67111.

Stvilia, B., Hinnant, C. C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., . . . Marty, P. F. (2014). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology*, *66*(2), 246-263.

Swanson, J., & Rinehart, A. K. (2016). Data in context: Using case studies to generate a common understanding of data in academic libraries. *The Journal of Academic Librarianship*, *42*(1), 97-101.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), e21101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., . . . Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*, *10*(8), e0134826.

Wallis, J. C., Rolando, E., & Borgman, C. L. (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), e67332.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6*(11), e26828.

Author Biographies

Keiko Kurata is a professor of the school of Library and Information Science, Faculty of Letters at Keio University, Japan. Her academic interest is the scholarly communication process, especially open access and open science.

Mamiko Matsubayashi is an assistant professor of the Faculty of Library, Information and Media Science, University of Tsukuba, Japan. Her academic interest is researchers; information practices.

Shinji Mine is an associate professor of Library and Information Science at Faculty of Humanities, Law and Economics, Mie University, Japan. His research lies in scholarly communication, particularly Open Access movement.