An examination of metadata practices for research data reuse: Characteristics and predictive probability of metadata elements

Min Sook Park* and Hyoungjoo Park
School of Information Studies,
University of Wisconsin Milwaukee, USA
e-mail: *minsook@uwm.edu (corresponding author); park32@uwm.edu

ABSTRACT

This study explores metadata practices in the relation to data reuse in biology. Metadata has long been viewed as a major constituent in research data management and reuse. However, the topic of whether metadata is used in a way that encourages data reuse has been understudied. The current study examined metadata elements used to describe datasets and the predictive probability of those metadata elements for data reuse under the assumption that citation frequency reflects the frequency of research data reuse. A total of 34,491 cited records from the biology category of the Clarivate Analytics Data Citation Index were analyzed using descriptive comparison and multiple regression analysis to compare usage patterns of metadata elements between data records cited more than twice and those cited only once. Of the five types of metadata elements identified and examined, metadata elements that provided descriptions about datasets and author-related information dominantly appeared across datasets, whereas DOI and ORCID identifier were scarce. Metadata related to author and funding resources were found to be positive influential factors in predicting data reuse, whereas data descriptions and identifiers appeared to have negative influences. This study contributed to a better understanding of metadata needs for data reuse.

Keywords: Metadata; Research data reuse; Data sharing; Research data; Scholarly communication.

INTRODUCTION

The exponential growth of research data has highlighted the importance of data sharing and reuse across disciplines (COSEPUP 2009; Schofield et al. 2009). The argument in favour of reusing published data is well established and demonstrates a variety of advantages of data reuse, such as increasing research reproducibility, intensifying the efficiency of funding use, and avoiding duplication of data collection (Piwowar and Vision 2013). A wide range of efforts have sought to boost research data reuse, including examination of data practices (Tenopir et al. 2015), implementation of polices (Tenopir et al. 2015), and metadata standard development and refinement (Faniel et al. 2013a).

Of the factors influencing research data reuse, metadata has been identified as a key component in both the storage and dissemination of large quantities of data for reuse (Gartner 2016; Leonelli 2016) as well as in substantiating persistent data reuse (Faniel and Zimmerman 2011; Leonelli 2016; Star and Gastl 2011) . Extensive efforts have been invested in developing metadata standards that more efficiently describe datasets for

https://doi.org/10.22452/mjlis.vol24no3.4

management, discovery, and reuse. Various metadata initiatives and ongoing refinement and modification of existing metadata standards reflect efforts to better describe datasets according to the local needs of different disciplines and ever-changing environments (e.g. digitalization and the growth of data-driven research) to make available datasets more reusable. In the field of biology, for example, data sharing models (e.g., the functional genomics experiment [FuGE]) and guidelines (e.g., findability, accessibility, interoperability, and reusability [FAIR] principles and FAIR-health) were developed with a specific emphasis on metadata to enhance the reusability of data.

Despite efforts to develop guidelines and principles, the extent to which these guidelines are carried out remains largely untested. The practices among researchers often determine the effectiveness of data- sharing and reuse (Kim and Burn 2016). Previous studies have largely focused on exploring a range of metadata elements in an attempt to better describe datasets and thus improve support for data reuse. Furthermore, extensive studies have examined many aspects of data reuse from data sharing perspectives. Few studies have considered whether metadata standards are being used in ways that may boost data reuse. Thus, this study reports on how datasets are described using metadata in relation to data reuse frequency. The field of biology is addressed in the current study because metadata practice varies across disciplines, reflecting local metadata needs.

LITERATURE REVIEW

Data Reuse and Metadata

An exponential rate of research data growth is occurring across a majority of disciplines. This phenomenon requires the development of rich datasets that are widely available for reuse by anyone anywhere (Borgman 2012; Li, Greenberg and Lin 2016) via information infrastructure such as digital libraries and data repositories (Qin and Li 2013,; Zimmerman 2007) . Data reuse refers to the use of one or more datasets by others to study new and different problems (Borgman 2012; Law 2005) . Data reuse can leverage available datasets with new methods and instrumentation, providing many benefits such as increased research reproducibility, avoidance of duplicated data collection efforts, and discovery of new knowledge (e.g., new drugs) (Borgman 2012; Cox and Pinfield 2014; Davis and Vickery 2007; Faniel and Zimmerman 2011; Fienberg, Martin and Straf 1985; Li, Greenberg and Lin 2016) . In other words, limited access to published datasets restricts research capacity. These benefits and concerns are well attested in the literature. In a study conducted by the Publishing Research Consortium (PRC) in 2010 with over 3000 researchers, respondents stated that access to datasets was highly important. However, only 38 percent of them felt that they had easy access to datasets, the lowest percentage among various types of information (e.g., patent information) (PRC 2010). This finding was confirmed in another survey where the respondents reported that the restricted accessibility of datasets generated by others limited their ability to answer their research questions (Tenopir et al. 2011).

Growing recognition of the importance of data reuse has made datasets a primary unit of information currency in research. This recognition has led to a shift in the focus of scholarly communication from journal articles to datasets, taking steps to link data with articles (Davis and Vickery 2007) . To those who share data, the frequency of citation acts as a currency of value that links publications and datasets and this metric is used in research funding, promotion decisions, and other benefits (Davis and Vickery 2007; Greenberg 2009; Piwowar, Day and Fridsma 2007) . In fact, a majority of Tenopir et al.’s (2011) survey respondents answered that they were willing to share and reuse others' data depending on certain conditions, such as getting credit through formal citations.

In an effort to encourage and increase the ability to access and reuse datasets, extensive studies on data reuse have identified a wide variety of underlying factors, including policies such as pressure from funding agencies (Davis and Vickery 2007; Piwowar, Day and Fridsma 2007; Piwowar and Chapman 2010), journal policies (Hanson, Sugden and Alberts 2011) and archiving policies (Dryad 2018; Whitlock et al. 2010), psychological and behavioural factors such as motivations and incentives (Curty 2015; Kim and Stanton 2016), ethical reasons (Fienberg, Martin and Straf 1985) , perceived data quality or relevance (Faniel et al. 2013b; 2013, Niu 2009); technological dimensions (Hilgartner 1995; Hine 2006; Zimmerman 2007) ; and disciplinary traditions (Brown 2003; Faniel and Jacobsen 2010; Glasner 2002). These findings demonstrate the intricate nature of data reuse.

Among a wide variety of influencing factors, metadata is recognized as the foundation of data preservation and discovery to support data reuse (Christel 2009; Gartner 2016; Qin, Ball and Greenberg 2012; Star and Gastl 2011) as well as is central to the data citation framework (Li, Greenberg and Lin 2016). Well-formed and right-sized metadata is recognized as a critical element of a working infrastructure capable of meeting the needs of data access, identification, and reuse (Star and Gastl 2011) . Collecting and providing access to datasets is no longer a trivial matter for researchers because they can discover, access, and repurpose datasets using data that are at exponential rates. In fact, approximately 2.8 million research outputs were indexed in Scopus in 2014 and the global research output doubles every nine years (Stang 2016). The proliferation of research datasets drives research and efforts that seek to organize and share datasets and increase the expectations for libraries' archived research datasets (Davis and Vickery 2007; Stang 2016) . Much effort has also been dedicated to the development of metadata standards that describe datasets (Sterling 1988) . The National Science Foundation (NSF) requires a data management plan for all proposals so that data are "deposited in a well-documented form" (i.e., metadata) "and are regularly and easily consulted and analysed by specialists and non-specialists alike" (i.e., data reuse) (NSF 2007). DataCite Consortium (DataCite Metadata Working Group 2016), the adoption of the Digital Object Identifier (Paskin 2005), Force 11 (Data Citation Synthesis Group 2014), and the Content Standard for Digital Geospatial Metadata (CSDGM) (Federal Geographic Data Committee 1998) are some examples of collaborative efforts seeking to develop metadata schema specific to digital data. Ball's (2009) survey reported on many efforts in metadata initiatives related to different types of research data.

Elements in metadata standards for research data convey essential information needed to describe datasets, such as information on creators, storage format, and instrument (Qin and Li 2013) . These elements serve different functions in supporting data discovery, selection, and location for data users and are required to fulfill the goals of preservation, interoperability, and reuse of datasets (Ailamaki, Kantere and Dash 2010, Michener 2006, Qin and Li 2013) . These requirements for research data have shifted to meet changing needs. The development of a metadata infrastructure (e.g., digital object identifier [DOI], uniform research identifier [URI]) and technical standards (e.g., resource description framework [RDF]) demonstrate how data-driven science, new technologies, and digitalization have grown to meet these needs (Qin and Li 2013). The continuous modification and extension of existing metadata standards and variations of metadata application across different disciplines are additional aspects of metadata that reflect specific disciplinary searching and browsing needs for describing distinctive research data formats, types, processing, and requirement (Qin, Ball and Greenberg 2012; Riall, Marincioni and Lightsom 2004) . In this sense, the variation of research data across different disciplines reflects researchers' discipline-specific strategies in seeking and using data as well as the requirements of metadata for data reuse support.

Metadata and Data Reuse in Biology

Culture and the amount of data sharing vary among disciplines. Biology is a discipline, in which data grow fast, and researchers are willing to share their data with others. In Tenopir et al.'s (2011) survey of 1,329 scientists, 85 percent of biologists reported that they share their data with others, a percentage that is relatively higher than in other disciplines such as medicine (65%) and social science (58%). The situation in life science is growing increasingly complex as data are being used for more advanced modeling, which creates new datasets (Tenopir et al. 2011). To illustrate, in computational biology, data reuse occurs in combination with the use of other data such as comparison studies (Berger, Daniels and Yu 2016).

Effective tools (e.g., guidelines, models, and metadata standards) play a major role in boosting data management, sharing, and access to datasets for reuse (Tenopir et al. 2011). The field has developed its own centralized repositories and common data formats to share, integrate, and standardize data. Several standard bodies have begun to make use of common data sharing models and minimum guidelines such as the FuGE (functional genomics experiment), systems biology markup language, and systems biology ontology (Linkert et al. 2010; Tenopir et al. 2011). Of these programmes and guidelines in biology, some consider metadata a core activity of centralization for data management (Linkert et al. 2010; Tenopir et al. 2011). For example, the National Biological Information Infrastructure (NBII) made the metadata programme a core of its activities when it advanced the ability of aggregating and disseminating biological databases (Gil et al. 2010). The NBII metadata programme makes biological data discoverable for both scientists and the public by unifying diverse biological databases (Gil et al. 2010). The FAIR principles (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson et al. 2016) and FAIR-health (Holub et al. 2018) are another examples of guidelines that put specific emphasis on metadata to enhance the reusability of data holdings, including biological data. The metadata describing repositories and stored samples are components of these principles (Holub et al. 2018). As such, metadata often serves as the key component for integration of scaled bio data such as DNA sequences, microarray experimental data, and mutational data (Li et al. 2005). Due to these efforts, there are several metadata standards in biology. Abundant metadata specifications in biology are designed to capture precise descriptions of research data in biology, which is crucial for data reuse. Examples of metadata standards in biology include access to biological collection data, Darwin Core, ecological metadata language, genome metadata, Open Microscopy Environment eXtensible Markup Language, the protein data bank exchange dictionary and micromolecular crystallographic information framework, protocol data element definitions, and repository-developed metadata schemas. These examples demonstrate efforts that have largely focused on the development of metadata standards within the field of biology.

Despite these models and guidelines, few policies specifically address metadata, and researchers also have reported a lack of tools to sufficiently describe datasets such as metadata (Tenopir et al. 2011). Only a quarter of the over 1,300 respondents (26%) in Tenopir et al.'s 2011 survey answered that they were satisfied with the available tools for preparing metadata. Another barrier to effective data sharing and preservation are often deeply rooted in the practices and culture of the research process as well as the researchers themselves (Kim and Burn 2016; Tenopir et al. 2011). Data sharing is not always perceived as a virtue, and perspectives often vary by field of study. For example, Blumenthal et al. (1997; 2006) found that geneticists were more likely to refuse to share their data with non-geneticists. A total of 44 percent of geneticists and 32 percent of other life scientists reported withholding their data for various reasons, including further publishing opportunities, avoidance of required administrative work (Kim and Burns 2016), and lack of time and funding (Tenopir et al. 2011). Although molecular biology has a strong culture of data sharing (Nelson 2009), biology is also a discipline where higher percentages of "no access" to datasets for data accessibility have been observed (Zhao, Yan and Li 2018). In sum, repository practices and metadata standard development have focused on supporting data collection, deposit, and access rather than data reuse. It is less obvious that metadata is being used in a manner to boost citation rates, which is a benefit for those who have shared the datasets with others. In exploring ways to fill this gap, this study addresses the relationship between research data citation rates and how metadata is used to describe research datasets. In this study, the usage patterns discerned from metadata practices are in turn translated into how researchers search and browse databases for available datasets to reuse. Using biology data records, this study addresses the following research questions:

(a) What metadata is used to describe datasets?
(b) What are the likelihoods of identified metadata types being cited?
(c) Are there differences in the metadata distributed in data records that is cited more than twice and that is cited fewer times?

Since the primary purpose of this study was to explore the latent relation between metadata practices and possible impacts on data reuse, only data records that searchers can utilize from a database system were used in the analysis. The assumptions made in this study were: searchers who attempt to retrieve published data records for their studies would examine the given records (e.g., author name, data type, and data titles) when selecting a certain data record over others; and the citation frequency reflects their choice.

MATERIALS AND METHOD

Data records were collected from the Clarivate Analytics' Data Citation Index (DCI). The DCI compiles research data from over 300 repositories worldwide and hosts over 7.4 million records. A total of 34,491 records in the DCI's biology subject category that had been formally cited at least once were selected and downloaded. No time-based selection parameters were employed. The metadata usage process was collapsed into a dichotomous outcome i.e., 1 for using the metadata element and 0 for not using the metadata element.

Descriptive comparison and multiple regression analysis were conducted to examine the relationship between different metadata and data citation frequency. Dependent variable of this study is the total times cited count (TCC), which is a sum of the number of cited counts of the Web of Science (WoS), Core Collection, BIOSIS Citation Index (BCI), Chinese Science Citation Database (CSCD), DCI, Russian Science Citation Index (RSCI), and SciELO Citation Index (SciELO CI). This study assumed that data description, identifiers, funding information, and author information are potential features for data reuse, which are a subset of the 124 attributes from Piwowar (2011). These features include document type, provision of abstract, source title, open researcher and contributor ID (ORCID), digital object identifier (DOI), source URL, taxonomy, data type, author name, and author e-mail addresses.

RESULTS

Metadata elements describing data records (e.g., abstract, author keywords, and version) were the most diverse and predominantly employed, followed by metadata elements about authors, identifiers, and other additional information. Of the nine metadata elements describing data records, four metadata elements (i.e. *d*ocument type, source title, documents title, and subject area) were the most prevalent across all data records. However, "documents types" categorized datasets only in three document types: data set, data study, and repository. Of those document types, slightly over half were data set (62.3%, n=21,495 out of 34,491), followed by data study (37.6.0%, n=12,976) and repository (0.06%, n=20). Despite its small size, the repositories appeared to be cited the most frequently. Eighteen out of 20 repositories had been cited more than twice, and this category included the record with the highest number of citations (n=53).

Of the five metadata elements pertaining to author information, "author names" was most present and other author related metadata element, such as "author e-mail addresses", appeared in less than half records. Of the four different identifiers, "accession number" and "source URL", appeared in the majority of records while other identifiers, including "DOI"*,* were comparatively scarce. Of the other additional metadata elements, "language" was dominant, followed by "miscellaneous" and "method" (see the total column in Table 1). "Miscellaneous" includes domain specific information such as '*cancer,*' '*acid stress*,' and '*enzyme*.' "Method" encompasses diverse biomedical data generation methods such as '*sugar analysis,*' '*chemical synthesis*,' and '*chemical method*.' Despite pressure from funding agencies, funding related metadata appeared to be rarely used (See the total field in Table 1).

In seeking to discern whether the identified five different types of metadata elements contributed to the prediction of citation, the multiple regression model with all five types of metadata together was found to account for 6 percent of data reuse variance, producing R^2 = .06 (p<.01). This result indicates that, five features together have a positive influence on data reuse, wielding significant regression weights. When examined the potential influence of different metadata types, "author", "identifiers" and "funding resources" were identified as positive influential factors, while data description and others appeared to exact negative influences on the data reuse. The results indicate that metadata elements describing data records and additional metadata elements, such as "language" and "method", may be rather problematic for the maintenance of research data and scholarly credits to data sharers (see Table 2).

When comparing data records used more than twice with those utilized only once, only 3.6 percent (n=1,238) of records were cited more than twice, while the majority of data records (96.4%) were used only once. The citation frequency ranged from 53 to 1. Both those records cited twice and those records cited only once show similar patterns (see Figures 1 through 4). To illustrate, particular metadata, such as "document type", "source title", "documents title", "subject area", "author name", "accession number", and "language" were appeared the most often across both of groups of data records, while "editors", "group authors", and "ORCID" were scarce. Among extra metadata that appeared to reflect disciplinary needs of biology, *gene name* and *geospatial* metadata were rarely used in both data record groups. Slight differences were observed as well. Data records cited more than twice appeared to include "source URL", "author address", and "author e-mail" metadata elements more frequently than the other data record group. Of the metadata elements describing data records, records cited only once appeared to more frequently provide “author keywords”, “data types” and “abstract information” (see Figure 1). Regarding author related metadata elements, “author name” and “author address” and “author e-mail” were slightly more frequently used in the group of data records cited twice or more (see Figure 2). Those observation results align with the multiple regression results that found metadata elements describing data records are negatively associated with the possibility of data reuse, whereas author related metadata were found positively associate. Of those identifiers, data records cited over twice appeared to have records of “source URL” more frequently, while the data records cited only once provide DOI information more frequently (Figure 3). Among other additional metadata elements use of the “method” element significantly differed between the data records cited twice or more (Figure 4). This finding may be the result of specific biology needs, as understanding of biology data is crucial to data reuse decisions.

Table 1: Frequencies and Percentage of Metadata Elements Use among Data Records.

		Cited Over Twice (n=1,238)		Cited Once (n=33,253)		Total (%)
		Freq.	%	Freq.	%	
Metadata Elements Describing Data Records	Document Type	1,238	100.0	33,253	100.0	34,491 (100.0)
	Source Title	1,238	100.0	33,193	99.8	34,431 (99.8)
	Document Title	1,238	100.0	33,193	99.8	34,431 (99.8)
	Subject Area	1,238	100.0	33,193	99.8	34,431 (99.8)
	Taxonomical Data	1,115	90.1	31,726	95.4	32,841 (95.2)
	Abstract	642	51.9	31,333	94.2	31,975 (92.7)
	Data Type	531	42.9	21,351	64.2	21,882 (63.4)
	Author Keywords	302	24.4	14,635	44.0	14,937 (43.3)
	Version	7	0.6	943	2.8	950 (2.8)
Metadata about Authors.	Author Name	1,232	99.5	33,174	99.8	34,406(99.8)
	Author Address	561	45.3	11,267	33.9	11,828 (34.3)
	Author E-mail	498	40.2	9,143	27.5	9,641 (2.8)
	Group Authors	17	1.4	556	1.7	573 (1.7)
	Editors	5	0.4	0	0	5 (0.01)
Identifiers	Accession Number	1,238	100.0	33,193	99.8	34,431 (99.8)
	Source URL	1,235	99.8	20,541	61.8	21,776 (63.1)
	ORCID	9	0.7	204	0.6	213 (0.6)
	DOI	3	0.2	12,652	38.0	12,655 (36.7)
Metadata about Funding	Funding Info.	10	0.8	7	0.02	17 (0.05)
	Funding Text	6	0.5	11	0.03	17 (0.05)
Other Metadata Elements	Language	1,238	100.0	33,193	99.8	34,431 (99.8)
	Miscellaneous	456	36.8	15,426	46.4	15,882 (46.0)
	Method	164	13.2	27,096	81.5	27,260 (79.0)
	Geospatial	42	3.4	207	0.6	249 (0.7)
	Gene Name	0	0	31	0.1	31 (0.1)

Table 2: Multiple Regression Analysis Results that Predict Impacts on the Total Number of Citation by Metadata Types

Variables	Frequency	*b*	*β*	*sig*
Metadata Elements Describing Data Records	205,878	-0.06^{*}	-13.08	0.00
Identifiers	69,075	0.02***	0.57	0.57
Metadata about Authors.	56,453	0.004**	1.11	0.27
Other Metadata Elements	77,853	-0.05*	-8.06	0.00
Metadata about Funding	34	3.52*	38.08	0.00
		R^2 =.06*		Intercept = 6.98

*p <.01, **p<.05, ***p<.1

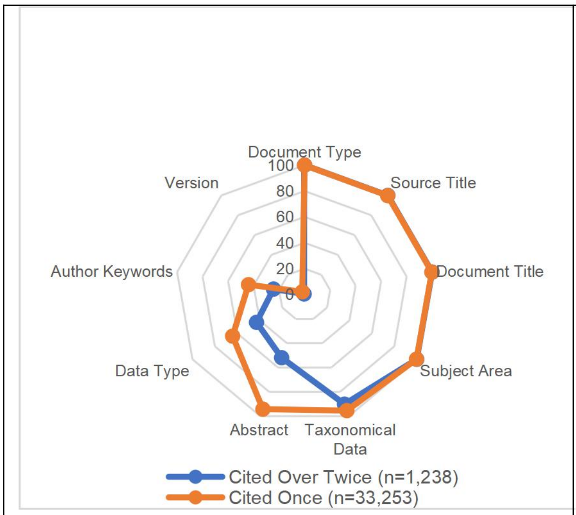

Figure 1: Metadata Describing Data Records.

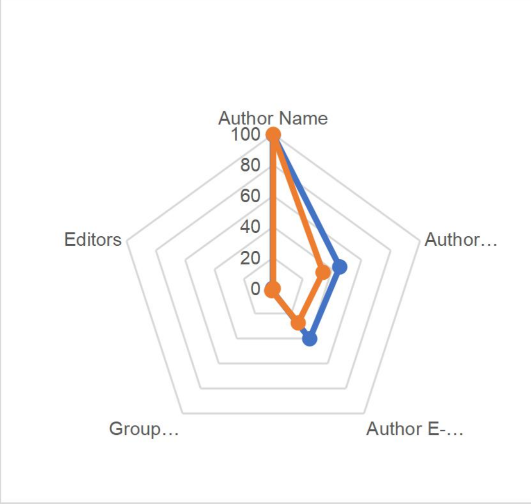

Figure 2: Author related Metadata.

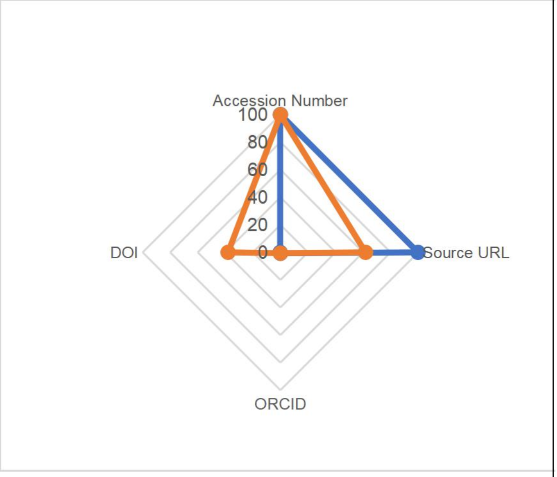

Figure 3: Identifier Metadata.

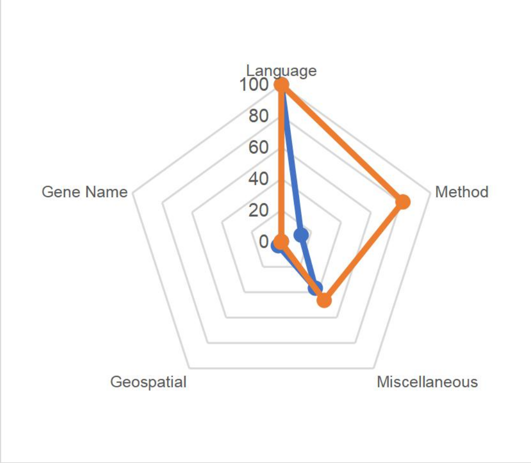

Figure 4: Other Metadata Elements

DISCUSSION

The current study sought to better understand the relationship between metadata practices and research data reuse based on formal data citation frequency. The investigation was performed on (a) the characteristics of metadata, (b) the predictive probability of metadata types with respect to data reuse, and (c) differences in metadata features of frequently and less frequently reused datasets. Understanding current practices and anticipating future metadata reuse is a requirement for facilitating research data reuse with sufficient context and clarity. Effective data management and use often rely on effective tools (Tenopir et al. 2011). In this sense, standard bodies have developed tools, such as data models, standards, and repositories to support data management and access to published datasets for reuse. In the repositories, data need to be stored and organized in a way that researchers easily access, share, and analyze the data (Tenopir et al. 2011). In this sense, metadata is considered as the key foundation of data preservation and discovery that should support data reuse (Christel 2009; Gartner 2016; Qin, Ball and Greenberg 2012; Star and Gastl 2011) . Yet, as Kim and Burn (2016) and Tenopir et al. (2011) assert, the effectiveness of data sharing is often grounded in the practices among researchers. This study revealed that data sharers more actively employ descriptive metadata. Addressing provenance metadata and methods metadata may be needed to automatically track and index research data.

Metadata elements usage patterns appeared to be skewed to particular types of metadata elements: descriptive metadata (e.g. “document type” and “author names”) appeared most often across data records, whereas other types of metadata elements (e.g. “funding information” and “identifiers”) occurred less frequently. DOI and ORCID identifiers were surprisingly scarce across all data records. The scarce use of ORCID may prohibit author identification and ambiguation. Considering the finding that data sharers more frequently use URLs than DOIs, automatic indexing by major scholarly databases (e.g., Web of Science or Scopus) can prove problematic because of the lack of permanent identifiers. Furthermore, despite pressure from funding agencies and policies, the use of metadata elements pertaining to funding appears to be insufficient in describing when datasets are shared. This result could be interpreted in two ways: a lack of funds to support data management or lack of awareness about particular metadata elements such as funding information. The low use of metadata fields describing funding information may be a source of concern for funders such as the National Institutes of Health, which propel data sharing requirements. In Tenopir et al.’s (2011) survey, more than half of the respondents (59%) indicated a lack of funds and training on best practices for data management from their organization or the projects. Considering that data managers, researchers, and publishers overwhelmingly agreed that public funding is the most important reason for data preservation (PARSE Insight 2009; Tenopir et al. 2011), funds for data management may help in sharing funding information along with datasets. On the other hand, funding information could have been simply missed by data sharers because of indifference or unfamiliarity with the information. In this case, there are several implications: funding agencies’ pressure on researchers to share did not significantly influence data sharing as results of prior studies’ regarding data sharing indicated (Piwowar and Chapman 2010; Kim and Burns 2016; Kim and Stanton 2015). Another possible implication is that data sharers were simply unsure about what metadata means. In Tenopir et al.’s (2011) survey, nearly half (46%) of the respondents answered that they use no metadata to describe their data. Given the finding of this study that “funding resources” metadata element positively influence data reuse in biology, there could be room for education on metadata concepts as a component of data management as Tenopir and colleagues discussed.

Among the five different types of metadata element groups, "author", "identifiers", and "funding resources" were identified as positive influential factors. In contrast, metadata elements describing data records (e.g. "author keywords") and elements providing additional information about data records (e.g. "language") appeared to negatively influence the data reuse in biology. This finding was confirmed through the study's comparison of data records cited twice or more and only once. Data records cited only once more frequently included "author keywords", "abstract", and "data type" elements than did records cited twice ore more. Considering detailed description of research data increased bibliographic citations by 69 percent (Piwowar, Day and Fridsma 2007) , the negative impact of data description on data reuse may merit concern when attributing scholarly credit to data sharers within the field of biology. Slightly more "source URL" and "author addresses" were found in the more frequently cited databases. This result aligns with previous findings that author reputation serves as a determining factor for data reuse (Zimmerman 2008). Wider use of URL over DOI as the preferred identifier challenges the sustainable and permanent identification of research data.

Quality metadata is essential for ensuring findability and reusability of the published datasets, because robust metadata can result in research data being lost or faced with indexing difficulties. Adherence to formal metadata standards is crucial to retrieval effectiveness, which may affect accessibility to published datasets. Low accessibility to data generated by others is reported to hinder researchers' abilities to answer scientific questions (PARSE Insight 2009; Tenopir et al. 2011). Besides, metadata standards influence norm of data sharing (Kim and Burns 2016), which leads to data reuse, future guidelines, and metadata standards in biology needs to incorporate metadata practices among researchers.

The work presented in this article is based on the assumption that the usage patterns of metadata used to describe data records in databases impacts researchers' searches of published datasets and their decisions to reuse data as reflected in the citation frequency. The impacts of self-citation in data reuse, domain specific features, and other potential social-cultural factors that may impact data reuse were beyond the scope of this work. This limitation was inevitable because the study aimed to understand the relation between metadata practices and possible impacts on data reuse, and because there was a lack of cited references or cited reference information limited the investigation of self-citation or the presence of a Matthew effect. Nevertheless, the findings of the current study suggest a direction for future research. Future research could observe and interview researchers to understand what metadata they use when searching and selecting a particular research data for data reuse. Socio-cultural and other factors that may intervene in data reuse decisions can be also investigated.

CONCLUSION

This study's focus on general-purpose metadata may provide implications for metadata findability for a single search interface. Specifically, this study examined metadata elements currently used by data sharers. DCI's metadata is general-purpose metadata that reflected, at least to some degree, many data repositories as required fields (e.g., author name) rather than optional metadata elements (e.g., geographic information). This simple metadata element allows data sharers to achieve long-term and greater interoperability across multiple data repositories. The DCI's general-purpose metadata are basic and understandable to research data sharers. Basic and simple metadata is easier to manipulate for metadata quality control, but simple metadata can elicit challenges for data re-users. Metadata is used as a resource discovery to transfer contextual information to data re-users (Niu and Hedstrom 2008). For instance, data re-users need to fully understand a prior experiment through sufficient data description because data are context-dependent. Each context demands sufficient data description to enable the development of new scientific inquiries by data re-users. It is also important to be aware of discipline-specific metadata with general-purpose metadata due to disciplinary variation. Discipline-specific metadata needs to include rich description with accurate and relevant attributes and detailed provenance metadata that meets discipline-relevant community standards (Wilkinson et al. 2016). Discipline-specific metadata allows more granular metadata discovery and help researchers identify the most appropriate research data for their contextual needs. The major challenges facing data sharing and reuse is the creation and standardization of metadata that can provide appropriate and adequate information to enable data re-users to develop to develop new scientific inquiries. With the emphasis on metadata standards and interoperability, research data can facilitate interdisciplinary research across disciplines via use of multiples data types.

ACKNOWLEDGEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors. We thank the reviewers for their helpful suggestions to improve the article.

REFERENCES

Ailamaki, A., Kantere, V and Dash, D. 2010. Managing scientific data. *Communications of the ACM,* Vol. 53, no.6: 68-78. Available at: doi: 10.1145/1743546.1743568.

Ball, A. 2009. Scientific data application profile scoping study report. Available at: http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf.

Berger B., Daniels, N.M. and Yu, Y.W. 2016. Computational biology in the 21st century: Scaling with compressive algorithms. *Communications of the ACM*, Vol.59, no.8: 72. doi: 10.1145/2957324. Available at: doi: 10.1145/2957324.

Blumenthal, D., Campbell, E.G., Anderson, M.S., Causino, N. and Louis, K.S. 1997. Withholding research results in academic life science: evidence from a national survey of faculty. *Journal of the American Medical Association*, Vol.277, no.15: 1224-1228. Available at: doi:10.1001/jama.1997.03540390054035.

Blumenthal, D., Campbell, E.G., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S. and Holtzman, N.A. 2006. Data withholding in genetics and the other life sciences: prevalences and predictors. *Academic Medicine*, Vol.81, no.2: 137-145. Available at: doi: 10.1097/00001888-200602000-00008.

Borgman, L.C. 2012. The conundrum of research data sharing. *Journal of the American Society for Information Science and Technology,* Vol.63, no.6: 1059-1078. Available at: doi:10.1002/asi.22634.

Brown, C. 2003. The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology,* Vol.54, no.10: 926-938. Available at: doi: 10.1002/asi.10289.

Christel, G.M. 2009. Automated metadata in multimedia information systems: creation, refinement, use in surrogates, and evaluation. In: *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Vol.1. no.1: 1-74. Available at: doi:10.2200/S00167ED1V01Y200812ICR002.

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age (COSEPUP). 2009. Ensuring the integrity, accessibility, and stewardship of research data in the digital age. Washington, DC: The National Academices Press. Available at: http://doi.org/10.17226/12615.

Cox, A.M. and Pinfield, S. 2014. Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*, Vol.46, no.4: 299-316. Available at:doi: 10.1177/0961000613492542.

Curty, G.R. 2015. Beyond "data thrifting": An investigation of factors influencing research data reuse in the social sciences. Syracuse: Syracuse University, 2015.

Data Citation Synthesis Group. 2014. Data Citation Synthesis Group: Joint declaration of data citation principles. Edited by M. Matone. San Diego, CA: FORCE11.

DataCite Metadata Working Group. 2016. DataCite metadata schema documentation for the publication and citation of research data. Available at: https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf.

Davis, M.H. and Vickery, J.N. 2007. Datasets, a shift in the currency of scholarly communication: Implications for library collections and acquisitions. *Serials Review*, Vol.33, no.1: 26-32. Available at: doi: 10.1016/j.serrev.2006.11.004.

Dryad. 2018. Joint Data Archiving Policy (JDAP). Available at: https://fairsharing.org/bsg-p000082.

Faniel, I.M. and Jacobsen, T.E. 2010. Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Journal of Computer Supported Cooperative Work*, Vol.19, no.3-4: 355-375. Available at: doi: 10.1007/s10606-010-9117-8.

Faniel, I.M., Kansa, E., Kansa, S.W., Barrera-Gomez, J. and Yakel, E. 2013a. The challenges of digging data: a study of context in archaeological data reuse. Paper presented at *the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, July 2013, at Indianapolis, Indiana. Available at: doi: 10.1145/2467696.2467712.

Faniel, I.M., Yakel, E., Kriesberg, A. and Daniels, M. 2013b. Can quantitative social scientists get data reuse satisfaction. Paper presented at the *Research Data Access and Preservation Summit*, April 2013, at Baltimore, MD.

Faniel, I.M. and Zimmerman, A. 2011. Beyond the data reuse deluge: A research agenda for large-scale data sharing and reuse. *International Journal of Digital Curation,* Vol.6, no.1: 58-69. Available at: doi: 10.2218/ijdc.v6i1.172.

Federal Geographic Data Committee. 1998. Content standard for digital geospatial metadata. Available at: https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata.

Fienberg, S.E., Martin, M.E. and Straf, M.L. 1985. Sharing research data. Washington, D.C.: National Academy Press.

Gartner, R. 2016. Metadata: Shaping knowledge from antiquity to the semantic web. London, UK: Springer International.

Gil, I.S., Hutchison, V., Frame, M., and Palanisamy, G. 2010. Metadata activities in biology. *Journal of Library Metadata*, Vol.10, no.2-3: 99-118. Available at: doi: 10.1080/19386389.2010.506389.

Glasner, P. 2002. Beyond the genome: Reconstituting the new genetics. *New Genetics and Society*, Vol.21, no.3: 267-277. Available at: doi: 10.1007/978-981-10-8441-6.

Greenberg, J. 2009. Metadata research supporting the Dryad data repository. Available at: https://ecommons.cornell.edu/bitstream/handle/1813/12247/DryadCornell.pdf?sequence=1&isAllowed=y.

Hanson, B., Sugden, A. and Alberts, B. 2011. Making data maximally available. *Science*, Vol.331, no.6018: 649. Available at: doi: 10.1126/science.1203354.

Hilgartner, S. 1995. Biomolecular databases: New communication regimes for biology? *Science Communication,* Vol.17, no.2: 240-263. Available at: doi: 10.1177/1075547095017002009.

Hine, C. 2006. Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science*, Vol.36, no.2: 2269-298. Available at: doi: 10.1177/0306312706054047.

Holub, P., Kohlmayer, F., Prasser, F., Mayrhofer, M.T., Schlünder, I., Martin, G.M., Casati, S. et al. 2018. Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-health. *Biopreservation and Biobanking*, Vol.16, no.2: 97-105. Available at: doi: 10.1089/bio.2017.0110.

Kim, Y. and 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*, Vol.67, no.4: 776-799. Available at: doi: 10.1002/asi.23424.

Kim, Y. and Burns, C.S. 2016. Norms of data sharing in biological sciences: The roles of metadata, data repository, and journal and funding requirements*. Journal of Information Science*, Vol.42, no.2: 230-245. Available at: doi: 10.1177/0165551515592098.

Law, M. 2005. Reduce, reuse, recylce: Issues in the secondary use of research data. *IASSIST Quarterly*, Vol.29, no.1: 5-10. Available at:doi: 10.29173/iq599.

Leonelli, S. 2016. Data-centric biology: A philosophical study. Chicago and London: University of Chicago Press.

Li, K., Greenberg, J. and Lin, X. 2016. Software citation, reuse and metadata considerations: An exploratory study examining LAMMPS. Paper presented at *the 79th Association for Information Science and Technology*, October 2016, at Copenhagen, Denmark. Available at: doi: 10.1002/pra2.2016.14505301072.

Li, L., Singh, R.G., Zheng, G., Vandenberg, A., Vaishnavi, V. and Navathe, S. 2005. A methodology for semantic integration of metadata in bioinformatics data sources. Paper presented at *the 43rd annual Southeast Regional Conference ACM*. Vol.1:131-136. ACM. Available at: doi: 10.1145/1167350.1167393.

Linkert, M., Rueden, C.T., Allan, C., Burel, J., Moore, W., Patterson, A., Loranger, B. et al. 2010. Metadata matters: access to image data in the real world. *The Journal of Cell Biology*, Vol.189, no.5: 777-782. Available at: doi: 10.1083/jcb.201004104.

Michener, K.W. 2006. Meta-information concepts for ecological data management. *Ecological Informatics*, Vol.1, no.1: 3-7. Available at:doi: 10.1016/j.ecoinf.2005.08.004.

National Science Foundation (NSF). 2007. NSF 07-28, Cyberinfrastructure Vision for 21st Century Discovery. National Science Foundation, Available at: http://www. nsf.gov/pubs/2007/nsf0728/index.jsp. Accessed 2010 Sep 30.

Nelson, B. 2009. Data sharing: Empty archives. *Nature News*, Vol.461, no.7261: 160-163. Available at: doi: 10.1038/461160a.

Niu, J. 2009. Perceived documentation quality of social science data. Ann Arbor, MI: University of Michigan - Ann Arbor

Niu, J. and Hedstrom, M. 2008. Documentation evaluation model for social science data. Paper presented at *the American Society for Information Science and Technology,* Vol.45, no.1: 11-11. Available at: doi: 10.1002/meet.2008.1450450223.

PARSE Insight. 2009. First insights into digital preservaation of research output in Europe. Available at: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf.

Paskin, N. 2005. Digital object identifiers for scientific data. *Data Science Journal*, Vol.4, no.28: 12-20. Available at: doi: 10.2481/dsj.4.12.

Piwowar, H. 2011. Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS One*, Vol.6, no.7: e18657. Available at: doi: 10.1371/journal.pone.0018657.

Piwowar, H., Day, R.S. and Fridsma, D.B. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS One,* Vol.2, no.3: e308. Available at: doi:10.1371/journal.pone.0000308.

Piwowar, H. and Chapman, W.W. 2010. Public sharing of research datasets: A pilot stuy of associations. *Journal of Informetrics*, Vol.4, no.2: 148-156. Available at: doi: 10.1016/j.joi.2009.11.010.

Piwowar, H. and Vision, T.J. 2013. Data reuse and the open data citation advantage. *PeerJ*, Vol.1: e175. Available at: doi: 10.7717/peerj.175.

Publishing Research Consortium (PRC). 2010. Access vs. Importance. A global study assessing the importance of and ease of access to professional and academic information (Phase I Results). Available at: http://publishingresearchconsortium.com/ index.php/prc-projects.

Qin, J., Ball, A. and Greenberg, J. 2012. Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. Paper presented at *the International Conference on Dublin Core and Metadata Applications*, September 2012, Kuching, Sarawak, Malaysia.

Qin, J. and Li, K. 2013. How portable are the metadata standards for scientific data? a proposal for a metadata infrastructure. Paper presented at *the International Conference on Dublin Core and Metadata Applications*, September 2013, at Lisbon, Portugal.

Riall, R., Marincioni, F. and Lightsom, F.L. 2004. Content metadata for marine science: A case study. US Department of the Interior, US Geological Survey. Available at: https://doi.org/10.3133/ofr20041002.

Schofield P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., Einhorn, D., Tocchini-Valentini, G., de Angelis, M.H. and Rosenthal, N. 2009. Post-publication sharing of data and tools. *Nature*, Vol.461, no.7261:171. Available at: doi:10.1038/461171a.

Stang, T. 2016. How growth in research data is spurring a shift in the librarian's role. Available at: https://www.elsevier.com/connect/librarians-the-new-research-data-management-experts

Star, J. and Gastl, A. 2011. isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, Vol.17, no.1. Available at: doi: 10.1045/january2011-starr.

Sterling, T. D. 1988. Sharing scientific data. *The ANNALS of the American Academy of Political and Social Science*, Vol.33, no.8: 49-60.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. and Frame, M. 2011. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, Vol.6, no.6: e21101. Available at: doi: 10.1371/journal.pone.002110.

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. and Dorsett, K. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE.* Vol.10, no.8:e0134826. Available at: doi:10.1371/journal.pone.0134826.

Whitlock, M.C., McPeek, M.A., Rausher, L.R. and Moore, A.J. 2010. Data archiving. *The American Naturalist*, Vol.175, no.2: 145-146. Available at: doi: doi.org/10.1086/650340.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, Vol.3. Available at: doi: 10.1038/sdata.

Zhao, M, Yan, E. and Li, K. 2018. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, Vol.69, no.1: 32-46. Available at:doi: 10.1002/asi.23919.

Zimmerman, S.A. 2007. Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, Vol.7, no. 1-2: 5-16. Available at: doi: 10.1007/s00799-007-0015-8.

Zimmerman, S.A. 2008. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science. *Technology & Human Values*, Vol.33, no.5: 631-652. Available at: doi: 10.1177/0162243907306704.