The types of research data receiving scholarly credit within and across the science, engineering, and mathematics fields

Hyoungjoo Park
Department of Library and Information Science,
Chungnam National University
99 Daehak-ro, Yuseong-gu, Daejeon 34134, KOREA
e-mail: hyoungjoo.park@cnu.ac.kr
ORCID ID: H.Park: 0000-0003-4271-1196

ABSTRACT
This study examined the types of data that receive formal scholarly credit within and across the science, engineering, and mathematics (SEM) fields. The topics of whether data types are used in a way that encourages data reuse has not been actively studied. This study applied an exploratory method because formal data citation is a relatively new area. The Data Citation Index (DCI) of the Web of Science (WoS) was selected because the DCI provides a single access point to 400 data repositories worldwide across multiple disciplines. Nearly all citations were of quantitative data. The types that received the most credit were, in descending order, ribonucleic acid (RNA), crystal structure, protein sequence data, crystallographic data, Sequence Read Archive (SRA), genomic, images, nucleotide sequencing information, molecular structure, and crystallographic information, though citation was diverse across the various disciplines within these fields. In particular, qualitative data received no scholarly credit. This study contributes to better understanding of data types for data reuse.

Keywords: Data citation; Data type; Data sharing; Data reuse; Research data.

INTRODUCTION

The term “research data” refers to any form of data obtained by researchers that is accepted or retained in scholarly communication and used in the production of original research outcomes or validation of research findings. The types of these data may include such information as details about research techniques and materials (Blumenthal et al. 2006) and may be raw or analyzed, observational, experimental, simulation, derived or compiled, and reference or canonical. Access to various types of research data across disciplines enables researchers to ask and answer synthetic, integrated, and broad-scale research questions in ways that allow for reproducibility. Thus, for example, combining geographic information systems with data relating to land use, climate, topography, and so on at the national level can assist in answering large-scale research questions about climate change. Precise descriptions are needed of the types of research data that allow for standardized and automated data registry across the data repositories (Lannom et al. 2015) where datasets are maintained for future reuse. Providing for the interoperability of interdisciplinary data across disciplines and discipline-specific cyberinfrastructures in modern science—that is, examining multiple disciplines both individually and in relation to one another—is, then, an

https://doi.org/10.22452/mjlis.vol28no1.1

essential aspect of data-intensive interdisciplinary research. Some work in this regard has already been done; thus, for instance, a multi-scale geospatial temporal ecology database has been formed from heterogeneous research data for the purpose of data reuse (Soranno et al. 2015). More detailed study of disciplinary differences with respect to data types and the awarding of scholarly credit across disciplines remains to be conducted, though. In order to help fill this gap in the literature on the impact of data types and data citation within and across diverse disciplines, two research questions were formulated, and the present study was conducted in order to answer these questions:

(a) Which data types are cited relatively frequently in the science, engineering, and mathematics fields?
(b) Which data types are cited relatively frequently across the science, engineering, and mathematics fields?

LITERATURE REVIEW

Data type can be defined as a "set of values, characterized by properties of those values and by operations on those values" (International Standard Organization for Standardization, 2013, p. 227). Data may be analyzed, compiled, derived, experimental, observational, simulated, or raw. Examples of data types include specimens, electronic microscopy results, physical collections, and software. The data type is often treated at the syntactic level, such as "integer," "string," and "float," but the syntactic definition of data types does not correspond well with the domain-specific meanings that are assigned to these types (Ma et al. 2016).

Previous studies found that certain data types were more likely to be reused or capable of reuse (Carlson and Anderson 2007; Moore 2006; Niu 2009) and documented disciplinary differences in the use of second-hand data (Dale et al. 2004; Peters, 2010). In particular, qualitative data tend to be shared only rarely (Faniel and Jacobsen 2010; Wallis et al. 2013), possibly owing to such characteristics as the tendency to include sensitive personal data. Rather, experimental data have more been often reused (He and Nahar 2016; Zhao, Yan and Li 2018); thus, for example, Korean social sciences researchers were found to prefer quantitative and survey data for reuse (Kim, Yoon and Chung 2020) and Chinese chemistry researchers to use experimental and observation data widely (Chen and Wu 2017). Biologists share their research data depending on the type (Kim 2022). Appropriate data sharing involves various forms of data and is contextual (National Institutes of Health 2020). Reusers' preferences for data types appear to be context-specific. For instance, surveys and aggregated and sequence data are more often cited (Belter 2014; Zhao, Yan and Li 2018). Regarding identifiers, surveys, aggregated, and clinical data have been cited more frequently than other data types using a Digital Object Identifier (DOI), while sequence, numerical, and individual-level data have been more frequently cited using a Uniform Resource Locator (URL; Peters et al. 2016).

Data citation is the practice of providing references to data recognizing them as primary research results. A reference to bibliographies to an author's own data is also regarded as a data citation. There have been efforts to encourage the authors of research articles to share their data in trusted data repositories so as to enable the bi-directional linkage of datasets and related publications by means of persistent and unique identifiers. Thus DataCite and the International Association of Scientific, Technical, and Medical Publishers (2012) released a joint statement outlining best practices for citing technical datasets in journals, and the Data Citation Synthesis Group (2014) released a Joint Declaration of Data Citation Principles

regarding the purposes, attributes, and processability (by both machines and humans) of data citation as well as the need for it. The practices of data citation in scholarly communication can be understood by examining a scholarly database that tracks and indexes data citation. For instance, the Data Citation Index (DCI) of the Web of Science (WoS) indexes over 13 million records of research data, 1.5 million data studies, and 3 billion records of software from over 440 repositories worldwide (Clarivate Analytics 2022). Thus, understanding the DCI helps to reveal the practices of formal data citation in scholarly communication.

Data citation provides references to research data and, thus, formal scholarly credit to data sharers. Nearly 1,600 Springer Nature journals have adopted standardized research data policies and encourage data citation (Editorial 2019). Data citation is important for data sharing and reuse, as evidenced by the finding that the citation rate of published research increased by more than two-thirds when a detailed description of the data was provided (Piwowar, Day and Fridsma 2007). This effect has been documented "independently of journal impact factor, date of publication, and author country of origin using linear regression" (Piwowar 2010, p.14). Likewise, Tenopir et al. (2011) reported that an overwhelming majority (91.7%) of researchers whom they surveyed agreed that data citation was at least somewhat important when their data were reused and nearly all (95%) agreed that it is "fair to use other people's data if there is formal citation of the data providers and/or funding agencies in all disseminated work making use of the data" (p. 10).

MATERIALS AND METHOD

Clarivate Analytics' DCI which tracks and indexes comprehensive datasets from all over the world, as discussed, served to collect the cited research data records for this study. The focus of the study was on the science, engineering, and mathematics (SEM) fields in the context of which research data are shared through data repositories worldwide. A comparison of the major National Science Foundation (NSF) discipline codes, the Research Areas of the WoS All Collections (Clarivate Analytics 2012), and the research areas of the DCI (Clarivate Analytics 2016) served to identify the disciplines for study within the SEM fields. In order to impose some level of quality control on the term "data-type," this study used the classification scheme of the WoS and the DCI in reference to other classification schemes because the classification of the ISI system is widely used and is based on expert judgment (though it is not without critics, e.g., Boyak, Klavans and Börner 2005). This study did not include the field of technology because the NSF did not include it among its major disciplines.

Table 1 presents the research areas sampled; the disciplines were astronomy/physics, biological sciences, chemistry, computing, earth sciences, engineering, and mathematical sciences. This study merged astronomy and physics (as astronomy/physics) though the NSF major discipline codes distinguish these fields because, as noted elsewhere, many universities house them within the same department. Interdisciplinary areas were excluded because of the difficulty inherent in assigning such areas to any one of the identified disciplines. Within the DCI, approximately 150 WoS Research Areas (i.e., the higher-level categories) were used rather than the 250 approximately Subject Categories (i.e., the lower-level categories) because the former contained more datasets than the latter.

Table 1: Comparison of NSF Major Disciplines the with Research Areas of the WoS (WoS All Collections and the DCI)

NSF major discipline	Research Areas of the WoS (WoS All Collections and the DCI)
Astronomy Physics	Astronomy & Astrophysics, Physics, Spectroscopy
Biological sciences	Genetics and Heredity, Biochemistry & Molecular Biology, Biotechnology & Applied Microbiology, Cell Biology, Developmental Biology, Evolutionary Biology, Marine & Freshwater Biology, Mathematical & Computational Biology, Microbiology, Plant Sciences, Reproductive Biology, Environmental Sciences & Ecology, Biodiversity & Conservation, Research & Experimental Medicine
Chemistry	Chemistry, Crystallography
Computing	Computer Science
Earth sciences	Geology, Oceanography, Geochemistry & Geophysics, Meteorology & Atmospheric Sciences, Water Resources
Engineering	Engineering
Mathematical sciences	Mathematics

The sample included the 30 most productive authors of published documents in each research area. Thirty authors were used as samples in each of the SEM fields because this number is considered sufficient, if small, for conducting quantitative statistical analyses. Productive authors were identified as the first authors of the most highly cited datasets in the DCI. The first author was assumed to be the one who made the most significant contribution and the last author to be the senior researcher with the most prestigious reputation; however, as Wren et al. (2007) noted, this is not always the case, so the first author was selected. If more than one highly cited dataset was attributed to the same first author, the next dataset on the list was selected. All records of the 30 most influential authors from the DCI in each SEM field were downloaded in tabular form. The total times cited count consisted of all cited counts, such as the WoS Core Collection, Biosis Citation Index (BCI), Chinese Science Citation Database (CSCD), DCI, Russian Science Citation Index (RSCI), and Scientific Electronic Library Online Citation Index (SciELO CI). Data citation served as an indication of data reuse, though additional reuse that is documented in the form of citations could occur. Descriptive analysis served to assess the types of SEM research data that were most often cited data in the DCI.

RESULTS

Table 2 displays the data types that were most cited in SEM fields in the DCI (RQ1). As the table shows, “ribonucleic acid (RNA)” was the type of SEM research data most often cited, followed by crystal structure, protein sequence data, crystallographic data, Sequence Read Archive (SRA), genomic, images, nucleotide sequencing information, molecular structure, and crystallographic information. Consistent with the previous findings discussed above, quantitative data were more often cited and shared than qualitative data in the SEM fields, presumably because these fields emphasize quantitative data. The sharing of such transcripts might require reading through many pages of text from multiple participants in a study to determine whether they mentioned names or key dates, such as of hospital admission, that would constitute privileged information unsuitable for dissemination. This concern to protect private information by removing direct and indirect identifiers is

motivated by the fact that nearly any American can be identified in datasets, even incomplete ones, based on only 15 demographic attributes (Rocher, Hendrickx and de Montjoye 2019). The International Committee of Medical Journal Editors (ICMJE) proposed requiring provision of the de-identified individual patient data as a condition of publication (Taichman et al. 2016).

Table 2*:* The 10 Most Cited Data Types in the DCI for SEM Fields

Data type	Total citations
ribonucleic acid (RNA)	931,673
crystal structure	754,913
protein sequence data	528,776
crystallographic data	490,252
Sequence Read Archive (SRA)	277,920
genomic	163,349
images	113,107
nucleotide sequencing information	109,135
molecular structure	91,870
crystallographic information	84,687
Total	**3,545,682**

Tables 3-10 display the disciplinary differences regarding the data types that were most cited in each SEM discipline (RQ2) in detail. Such a detailed examination of disciplinary differences was necessary because data sharing tends to be discipline-specific (Park and Wolfram 2017; Tedersoo et al. 2021). That is, the examination of individual disciplines allows for precise estimates of the types of data cited.

Table 3 displays the top 10 most cited data types in the DCI for the discipline of astronomy/physics. The fact that the total number of data types used was relatively small does not necessarily mean that research data in this discipline were not being reused, for such reuse may simply have gone unreported. Nevertheless, while data sharing is the norm at least in astronomy, as is preservation (Ivezić 2012), surprisingly little citation was evident in astronomy/physics in the DCI. Also noteworthy in this regard is the finding reported in another study that the least data sharing in journal articles occurred in physics (Keynon et al. 2016). In any case, the distribution of data types in astronomy/physics was quite skewed, with mass spectral data in particular accounting for more than 90% of the total, followed by Nuclear Magnetic Resonance (NMR) results and spectral data. The widespread use of mass spectral data in astronomy/physics is to be expected given the massive amounts of such data used to analyze large sky surveys. Research in astronomy is likely to involve observations, celestial coordinates, and actual astronomical objects (e.g., stars), and the integration of astronomical data relies on established constants and physical laws for the purposes of calibrating instruments and setting measurement standards. The use of “image files” and “final output pictures” may be attributable to the reliance on photographic observations in astronomy/physics. Interestingly, “data/dataset” was categorized as a type of data in the DCI though this term is not itself a type of research data but rather defines research data as a single, coherent set.

Table 3: Astronomy/Physics: The 10 Most Cited Data Types in the DCI

Data type	Total citations	Percentage
mass spectral data	31,072	70.80%
Nuclear Magnetic Resonance (NMR) results	6,157	14.04%
spectral data	3,723	8.48%
software	1,396	3.26%
image file	234	0.53%
Flexible Image Transport System (FITS) file	192	0.43%
data/dataset	170	0.39%
final output picture	163	0.37%
HRCROP	60	0.14%
TEX APPB	50	0.11%
Total	**43,280**	**98.55%**

Table 4 displays the most-cited data types in the DCI for the biological sciences. The relatively high rate of citation in the biological sciences may be due to the relatively high frequency of data sharing among biologists (Rung and Brazma 2013) as well as to the fact that biology was among the first SEM disciplines to adopt data sharing requirements. Thus, for instance, the National Institutes of Health (NIH) mandated data sharing in 2003 and the National Science Foundation (NSF) did so in 2011 (with subsequent revisions to its guidelines in 2013). The top biology journals have tended to make frequent and widespread use of official external data repositories for data sharing (Womack 2015). The actual data sharing among biologists increased significantly because of the journal publishers' data sharing policies (Kim & Burns 2016). In the biological sciences, data sharing and reuse to create new knowledge are more common than in other STEM fields (Yoon and Kim 2020). Biologists who anticipate scholarly citations are more likely to share data (Byrd et al. 2020). In some subdisciplines within the biological sciences, data sharing and reuse are common (Editorial 2018). The three most-cited data types (i.e., RNA, protein sequence data, and SRA) accounted for over half of the total citations for the disciplines in the biological sciences. The biological sciences in the DCI received the most citations from other fields in the WoS (Park 2022). The situation is clearly dynamic; thus, for instance, the number of sequences in the SRA—which stores raw sequencing data and alignment information from high-throughput sequencing platforms and makes biological sequence data available (National Center for Biotechnology Information, n.d.)—has doubled every 6-8 months on average over the past decade (Stephens et al. 2015). Returning to an earlier point, the diverse data types used in the biological sciences in general, and the life sciences in particular, are associated with a wide range of security and sensitivity considerations; thus, for instance, data relating to genetic mutations in humans may be highly sensitive while observations of diversity in non-human populations generally is not. Legal agreements prevent researchers working on human participants from revealing sensitive data (Lewandowsky and Bishop 2016).

As Table 5 shows, in chemistry, the two most common data types in the DCI accounted for 86.61% of citations. Specifically, the crystal structure consisted of 52.51% and crystallographic data consisted of 34.1% of the data in the DCI. The widespread use of crystallographic data corresponds with the frequent sharing of the chemical structure of newly synthesized compounds among chemists. Such derived data—that is, data that have been processed or reduced in some way—are in general widely used in the discipline. Crystallographic data are discipline-specific and shared in the form of a Crystallographic Information File (CIF), a standard machine-readable text file format was developed by the

International Union of Crystallography specifically for the exchange and curation such information.

Table 4: Biological Sciences: The 10 Most Cited Data Types in the DCI

Data type	Total citations	Percentage
ribonucleic acid (RNA)	931,673	29.67%
protein sequence data	525,973	16.75%
Sequence Read Archive (SRA)	277,920	8.85%
genomic	163,349	5.20%
images	113,107	3.60%
nucleotide sequencing information	109,135	3.48%
molecular structure	75,899	2.42%
FGEM	72,717	2.32%
processed data	72,717	2.32%
plant transcription factors and their annotation	65,536	2.09%
Total	**2,408,026**	**76.69%**

Table 5: Chemistry: The 10 Most Cited Data Types in the DCI

Data type	Total citations	Percentage
crystal structure	754,913	52.51%
crystallographic data	490,252	34.10%
molecular structure	91,870	6.39%
crystallographic information	84,687	5.89%
bacterial carbohydrate structure	4,298	0.30%
spectral data	3,720	0.26%
crystallographic structure	3,008	0.21%
dataset	2,410	0.17%
molecular data	954	0.07%
molecule	647	0.05%
Total	**1,436,759**	**99.95%**

Table 6 displays the most-cited data types in the DCI for the discipline of computing, of which software was by far the most common (accounting for 91.41% of the total). The low rate of citation in computing is quite counterintuitive because it is, of course, the discipline most concerned with software code, which is, again, a type of data. This result may be attributable to the frequent use of proprietary software while the focus of the present study was on scholarly communication. The low rate of citation in the DCI is consistent with the finding that the research data linked to articles in computing journals in the WoS have received very few citations therein (Park 2022).

Table 7 displays the most-cited data types in the DCI for the earth sciences. In this discipline, geospatial datasets associated with the Global Positioning System (GPS) were cited especially often, possibly owing to the frequency with which published research has been based on image datasets sampling three-dimensional spatial, temporal, and spectral characteristics of detected signals in, for instance, the measurement of cellular, tissue, and organizational processes and structures, as Williams et al. (2017) discussed. Examples of geospatial data

include Global Positioning System (GPS) tracks, Global Information System (GIS) map layers, and satellite observation data. GPS datasets represent a type that can be reused to visualize spatiotemporal analyses based on the computational use of source code. Generally speaking, the length of an earth observation time series corresponds positively with the accuracy and effectiveness of an assessment on which the series is based, such as predicting when a disaster may strike (e.g., pinpointing earthquakes and ground motion). In the earth sciences, then, it is especially important for researchers to be able to decode observation data from diverse sources, including those from the more or less distant past, and interpret them accurately.

Table 6: Computing: The 10 Most Cited Data Types in the DCI

Data type	Total citations	Percentage
software	18,246	91.41%
code	1,278	6.40%
model	416	2.08%
dataset	3	0.02%
database	2	0.01%
other	2	0.01%
raw experimental data	2	0.01%
chemistry data	1	0%
dataset used in the paper	1	0%
diagrams	1	0%
Total	**19,952**	**99.95%**

Table 7: Earth Sciences: The Most Cited Data Types in the DCI

Data type	Total citations	Percentage
dataset	32,975	30.64%
interactive resource	22,264	20.69%
GPS dataset	13,080	12.15%
geoscientific information	9,108	8.46%
GPS collection	5,741	5.33%
text	4,033	3.75%
navigation primary	3,691	3.43%
protein sequence data	2,803	2.60%
digital	2,699	2.51%
Total	**96,394**	**89.56%**

Table 8 displays the seven data types that together accounted for all of the citations in the DCI for the discipline of engineering. In engineering, 99.71% of citations were of test data; most of the rest were of datasets (0.13%). Data citation was less frequent in engineering than in other SEM disciplines, possibly because industrial and commercial contracts tend to be a more typical form of communication than scholarly papers (e.g., commissioned contracts in aerospace-, construction-, and defense-engineering contexts; again, the focus in the present study was on scholarly communication). That is, engineering is a discipline in which commercial enterprises, rather than public service, are often the primary concern and as such plays a role in a wide range of industries, each with its own working practices. Proprietary data from commercially sponsored research may have significance for future

patents and therefore be subject to an embargo period with respect to sharing. Researchers involved with industries are more reluctant to engage in data sharing owing to concerns about the reuse of the data for commercial purposes (Vogeli et al. 2006). Test data are important in engineering—and hence well-represented in Table 8—because the bulk of the research data are collected during the design and redesign phases. In product research, for instance, the performance of a series of designs and redesigns is usually monitored, recorded, and preserved for use throughout the whole product life cycle.

Table 8: Engineering: The Data Types Cited in the DCI

Data type	Total citations	Percentage
test data	3,749	99.71%
dataset	5	0.13%
GIS vector data	2	0.05%
Quartz Crystal Microbalance (QCM) data	1	0.03%
microscopy images	1	0.03%
fluorescence intensity data	1	0.03%
Microsoft Excel spreadsheet	1	0.03%
Total	**3,760**	**100%**

Table 9 displays the data types cited in the DCI for the mathematical sciences, of which, interestingly, there were only five. The software and matrix data types together accounted for nearly all of the citations (99.63%). The type "GEOID ondulation [*sic*] given on a grid" was counted with "GEOID undulation given on a grid" because "ondulation" is a typographical error.

Table 9: Mathematical Sciences: Data Types Cited in the DCI

Data type	Total Citations	Percentage
software	8,155	82.94%
matrix	1,640	16.69%
GEOID undulation given on a grid	35	0.35%
dataset	1	0.01%
academic test score data	1	0.01%
Total	**9,832**	**100%**

DISCUSSION

The accurate classification of data types is crucial to facilitate data reuse and citation and thereby promote scientific reproducibility. Such classification is time-consuming, though, and can require prior knowledge and additional work on the part of data sharers. Sharing would accordingly be facilitated with the implementation of procedures for identifying machine-actionable data types automatically and in a standardized way across disciplines, but this is no easy task, in particular because the machine-actionable classification of data types depends on machine-readable definitions. The task can be accomplished through the creation of data-type registries that feature a common and open interface and accurate descriptions of data; further, the establishment of a standardized set of elements describing data types would make possible automated typing of data (Lannom and Broeder 2014). By

means of a web interface, even researchers without advanced technical skills could easily manage, query, and annotate their data. Similarly, for instance, authorized users can currently create specific data types by submitting a JavaScript Object Notation (JSON) schema through a web interface (Izzo et al. 2014). In order to retrieve and use data from a repository quickly and accurately, a researcher needs to know their format, structure, and meaning as well as how to use the various tools and services available for data processing. However, as Read et al. (2015) discussed, a single dataset could be in the form of a discrete data type from a specific diagnostic device or consist of all of the data collected or analyzed during a research project or pre- or post-intervention or of every individual measurement reported in a bibliography. Shared data could also be in the form of tables or graphs that display trends in the sizes of structures in, for instance, treated or untreated cells over time by bringing the difficulties in ensuring accurate classification of data types. The Research Data Alliance Data Type Registries Working Group confirmed that a precise and detailed description of the data type is the essential consideration for data sharing and reuse so as to accommodate the requirements of each discipline (Lannom, Broeder and Manepalli 2015).

The wide use of multiple data types from various researchers across diverse disciplines can also impede data citation. In the environmental sciences, for instance, physically based distributed hydrologic models require geospatial and time-series data in order to be processed into model inputs, a task that involves considerable time and effort on the part of researchers (Gichamo et al. 2020). Data types affect the level of cyberinfrastructure needed for any given data. Considering the broad range of data types in use across the SEM fields found in the present study, the diversity of data management can and should correspond to the range of access conditions applied. Data types need to be standardized and made discoverable through one or more piece of interoperable cyberinfrastructure to facilitate the identification and distribution of those that are useful. Access to some data is limited by embargoes, while other data may be released immediately or deemed too sensitive to be released at all, so the license type and authorship need to be clarified. Data repositories likewise must be able to accommodate a broad range of data types, even if doing so requires large-scale investment in cyberinfrastructure. While some biomedical repositories, such as Gene Expression Omnibus (Brandt and Uden 2003) and ArrayExpress (Kolesnikov et al. 2015), feature combined interfaces to provide browsing and search features based on the filter options (e.g., experiment type or organism), such interfaces remain insufficient to address fully the demands for data reuse across the wide variety of data types currently used in the SEM fields.

CONCLUSIONS

The findings presented here indicate that various data types are frequently cited across SEM fields and among disciplines. All citations were of quantitative data. Overall, the most-cited types in SEM fields in the DCI were ribonucleic acid (RNA), crystal structure, protein sequence data, crystallographic data, Sequence Read Archive (SRA), genomic, images, nucleotide sequencing information, molecular structure, and crystallographic information. By discipline, the three most commonly cited types were, for astronomy/physics, mass spectral data, NMR results, and spectral data; for the biological sciences, RNA, protein sequence data, and SRA; for chemistry, crystal structure, crystallographic data, and molecular structure; for computing, software, code, and models; for the earth sciences, datasets, interactive resources, and GPS data; for engineering, test data, datasets, and GIS vector data; and, for the mathematical sciences, software, Matrix, and GEOID undulation on a grid. This study has clear implications for the implementation of information systems for research data regarding

the implementation of widely accepted standards for data identification and attribution. It is hoped that the findings presented here will contribute to efforts to make data citation more practical and advantageous for researchers and data hosting facilities. A future study will include a broader selection of disciplines to assess citation practices relating to the types of research data types and the awarding of scholarly credit in various fields. Despite the care taken in conducting this study, its limitations include a lack of detail regarding the contexts for the data types and citations. Future studies could overcome this limitation by employing a qualitative or mixed-methods approach that may better explain the underlying circumstances by taking into account more specific types of data and citations.

ACKNOWLEDGEMENT

No grant from any public, commercial, or non-profit funding agency was offered for the conduct of this research.

AUTHOR DECLARATION

The author has no conflict of interest to declare.

REFERENCES

Belter, W. C. 2014. Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE*, Vol.9, no.3: e92590.

Blumenthal, D., Campbell, E. G., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S. and Holzman, N.A. 2006. Data withholdings in genetics and the other life sciences: Prevalences and predictors. *Academic Medicine*, Vol.81, no. 2: 137-145.

Boyak, K. W., Klavans, R., and Börner, K. 2005. Mapping the backbone of science, *Scientometrics*, Vol.64, no.3: 351-374.

Brandt, D. S. and Uden, L. 2003. Insight into mental models of novice internet searchers. *Communications of the ACM*, Vol.46, no.7: 133-136.

Brueman, P. 2006. How to cite curated databases and how to make them citable. Paper presented at the *18th Scientific Database Management Conference*, 2006, at Los Alamitos, U.S.A.

Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., and Greene, C.S. 2020. Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, Vol.21: 615–629.

Carlson, S. and Anderson, B. 2007. What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, Vol.12, no.2: 635-651.

Chen, X. and Wu, M. 2017. Survey on the needs for chemistry research data management and sharing. *The Journal of Academic Librarianship*, Vol.43, no.4: 346-353.

Clarivate Analytics. 2012. Research areas. Available at: https://images.webofknowledge.com/WOKRS57B4/help/WOS/hp_research_areas_easca.html

Clarivate Analytics. 2016. Data Citation Index - Research area. Available at: http://images.webofknowledge.com/WOKRS523_2R2/help/DRCI/hp_research_areas_easca.html

Clarivate Analytics. 2022. Data Citation Index: connecting data to the research it informs.

Available at: https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/

Dale, A., Wathan, J. and Higgins, V. 2004. Secondary analysis of quantitative data. *The SAGE encyclopedia of social science research methods* (Thousand Oaks, CA: Sage), 1007-1008.

Data Citation Synthesis Group. 2014. *Joint declaration of data citation principles*. San Diego: FORCE11.

DataCite and the International Association of Scientific, Technical and Medical Publishers. 2012. STM-DataCite Joint Statement. Available at: http://www.stm-assoc.org/2012_06_14_STM_DataCite_Joint_Statement.pdf

Editorial. 2018. Data sharing and the future of science. *Nature Communications*, Vol.9, no.1: 2817.

Editorial. 2019. Data citation needed. *Scientific Data*, Vol. 6, no.27.

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.

Gichamo, T. Z., Sazib, N. S., Tarboton, D. G. and Dash, P. 2020. HydroDS: Data services in support of physically based, distributed hydrological models. *Environmental Modeling and Software*, Vol.125: 104623.

He, L. and Nahar, V. 2016. Reuse of scientific data in academic publications: An investigation of Dryad digital repository. *Aslib Journal of Information Management*, Vol.68, no.4: 478-494.

Howison, J. and Bullard, J. 2016. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology,* Vol.67, no.9: 2137-2155.

International Standard Organization for Standardization. 2013. ISO/IEC 11179-3:2013(E) Information technology - metadata registries (MDR)-Part 3: Registry metamodel and basic attributes. Available at: https://www.iso.org/standard/50340.html

Ivezić, Z. 2012. Data sharing in astronomy. In: K. B. Mathae and Uhlir P. F. (eds). *Committee on the case of international sharing of scientific data: A focus on developing countries* (pp. 41-45). Washington, D.C.: National Academic Press.

Izzo, M., Mortola, F., Arnulfo, G., Fato, M. M. and Varesio, L. 2014. A digital repository with an extensible data model for biobanking and genomic analysis management. *BMC Genomics*, Vol.15, no.Suppl 3: S3.

Keynon J., Sprague, N. and Flathers, E. 2016. The journal article as a means to share data: a content analysis of supplementary materials from two disciplines. *Journal of Librarianship and Scholarly Communication*, Vol.4: Ep2112.

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: 230-245.

Kim, N., Yoon, J. and Chung, E. 2020. What data characteristics are needed for data reuse in the domain of social sciences in Korea? Paper presented at the *iConference*, March 2020 at Borås, Sweden.

Kim, Y. 2022. Data sharing by biologists: A comparative study of genome sequence data and lab experiment data. *Library & Information Science Research*, Vol.44, no.1: 101139.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Megy, K., Pilicheva, E., Rustici, G., Tikhonov, A., Parkinson, H., Petryszak, R., Sarkans, U. and Brazma, A. 2015. ArrayExpress update-simplifying data submissions. *Nucleic Acids Research*, Vol.43: D1113-D1116.

Lannom, L., Broeder, D. and Manepalli, G. 2015. RDA data type registries working group output. Available at: https://zenodo.org/record/1406127#.Y6o1f3ZBy3A

Lannom, L. and Broeder, D. 2014. Data type registries: A research data alliance working group.

D-Lib Magazine, Vol.20, no.1/2.

Lewandowsky S. and Bishop, D. 2016. Research integrity: Don't let transparency damage science. *Nature,* Vol. 529: 459-461.

Ma, X., Erickson, J. S., Zednik, W., West, P and Fox, P. 2016. Semantic specification of data types for a world of open data. *ISPRS International Journal of Geo-Information*, Vol.53, no.3: 38.

Moore, N. 2006. The contexts of context: Broadening perspectives in the (re)use of qualitative data. *Methodological Innovations*, Vol.1, no.2: 21-32.

National Center for Biotechnology Information. [n.d.]. SRA: Now available on the cloud. Available at: https://www.ncbi.nlm.nih.gov/sra

National Institutes of Health. 2020. Final NIH policy for data management and sharing. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

National Science Foundation. 2010. Instructions and codes for completing project data form (Form 1295). Available at: https://nsf.gov/pubs/2010/nsf10034/nsf10034.docx

Niu, J. 2009. Overcoming inadequate documentation. Paper presented at the *72nd American Society for Information Science and Technology Conference,* November 2009, at Vancouver, Canada.

Park, H. 2022. The interdisciplinary of research data: How widely is shared research data reused in the STEM fields? *Journal of Academic Librarianship*, Vol. 48: 102535.

Park, H. and Wolfram, D. 2017. An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics*, Vol.111, no.1: 443-461.

Peters, P. C. D. 2010. Accessible ecology: Synthesis of the long, deep, and broad. *Trends in Ecology & Evolution*, Vol.25, no.10: 592-601.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. and Gorraiz, J. 2016. Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, Vol.107, no.2: 723-724.

Piwowar, A. H. 2010. *Foundational studies for measuring the impact, prevalence, and patterns of publicly sharing biomedical research data*. Pittsburgh: University of Pittsburgh.

Piwowar, H. A., Day, R. and Fridsma, D. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, Vol. 2, no.3: e308.

Read, K. B., Sheehan, J. R., Huerta, M. F., Knecht, L. S., Mork, J. G., Humphreys, B. L. and NIH Big Data Annotator Group. 2015. Sizing the problem of improving discovery and access to NIH-funded data: A preliminary study. *PLoS ONE*, Vol.10, no.7: e0132735.

Rocher, L., Hendrickx, J. M. and de Montjoye, Y.-A. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, Vol.10, no.1: 1-9.

Rung, J. and Brazma, A. 2013. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, Vol.14, no.2: 89-99.

Soranno, P. A., Bissell, E. G., Cheruvelil, K. S., Christel, S. T., Collins, S. M., Fergus, C. M. and Webster, K.E. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. *GigaScience*, Vol.4, no.28.

Stephens, D. Z., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J. and Robinson, G.E. 2015. Big data: Astronomical or genomical? *PLoS Biology*, Vol.13, no.7.

Taichman, D.B., Backus, J., Baethge C, Bauchner, H., Leeuw, P.W., Drazen J.M., Baethge, C. and Wu, S. 2016. Sharing clinical trial data. *BMJ*. 2016. Vol. 532: i255.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K. and Sepp, T. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, Vol.8, no.192. Available at: https://doi.org/10.1038/s41597-021-00981-0.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E. and Frame, M. 2011. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, Vol.6, no.6: e21101.

Vogeli, C., Yucel, R., Bendavid, E., Jones, L. M., Anderson, M. S., Louis, K. S. and Campbell, E.G. 2006. Data withholding and the next generation of scientists: Results of a national survey. *Academic Medicine*, Vol.81, no. 2: 128–136.

Wallis, J. C., Rolando, E. and 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*, Vol 8, no.7: e67332.

Williams, E., Moore, J., Li, S. W., Rustici, G., Tarkowska, A., Chessel, A. and Swedlow, J.R. 2017. Image data resource: A bioimage data integration and publication platform. *Nature Methods*, Vol.14, no.8: 775-781.

Womack, P. R. 2015. Research data in core journals in biology, chemistry, mathematics, and physics. *PLoS ONE*, Vol.10, no.12: e0143460.

Wren, J. D., Kozak, K. Z., Johnso, K., Deakynes, S. J., Schilling, L. M. and Dellavalle, R.P. 2007. The write position: A survey of perceived contributions to papers based on byline position and number of authors. *EMBO Reports*, Vol.8, no.11: 988-991.

Yoon, A. and Kim, Y. 2020. The role of data re-use experience in biological scientists' data sharing: an empirical analysis. *Electronic Library*, Vol. 38, no.1: 186-208.

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.