



Analysis on open data as a foundation for data-driven research

Honami Numajiri¹ · Takayuki Hayashi¹

Received: 15 November 2023 / Accepted: 15 January 2024
© The Author(s) 2024

Abstract

Open Data, one of the key elements of Open Science, serves as a foundation for “data-driven research” and has been promoted in many countries. However, the current status of the use of publicly available data consisting of Open Data in new research styles and the impact of such use remains unclear. Following a comparative analysis in terms of the coverage with the OpenAIRE Graph, we analyzed the Data Citation Index, a comprehensive collection of research datasets and repositories with information of citation from articles. The results reveal that different countries and disciplines tend to show different trends in Open Data. In recent years, the number of data sets in repositories where researchers publish their data, regardless of the discipline, has increased dramatically, and researchers are publishing more data. Furthermore, there are some disciplines where data citation rates are not high, but the databases used are diverse.

Keywords Open science · Open data · Data citation index · DCI · OpenAIRE graph

Introduction: data-driven research by open data

Open Science (OS), a new research style in which data and findings from research activities are shared and utilized by the academic community, is being promoted in many countries. In the European Union, Open Science initiatives have been a fundamental policy since Horizon 2020 (European Commission, 2018). In 2022, the Council of the European Union adopted the “Research assessment and implementation of Open Science”, furthering the commitment to OS policies and proposing joint actions across the European Research Area. In addition, the UNESCO Recommendation on Open Science in 2021 has significantly strengthened the movement by providing a comprehensive framework at the international level, while advocating support for fair and equitable OS for all at individual, institutional, national, regional and international levels. The promotion of OS by major

✉ Honami Numajiri
doc22053@grips.ac.jp
Takayuki Hayashi
ta-hayashi@grips.ac.jp

¹ National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan

countries was also highlighted in the "G7 Science Ministers' Communiqué" (Joint Statement) adopted at the G7 Science and Technology Ministers' Meeting in Sendai in May 2023.

Notably, OS is defined as "an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community." (UNESCO, 2021). According to Fecher and Friesike (2014), four elements of Open Science can be identified as follows: Open Access, Open Data, Citizen Science, and Collaborative Research. Among these, Open Data is considered to be one of the central elements of OS and is expected to promote greater transparency and accessibility of research through the publication of data generated in the research process (Koznov et al., 2016).

Meanwhile, a new scientific research paradigm that centers on data as a driving source is emerging. Newman et al., (2003) defined "data-intensive science" as the fourth paradigm of science, which is an approach to scientific research in which new knowledge and insights are gained by collecting, organizing and analyzing large amounts of data. For example, in the discipline of biodiversity, the development of information and communications technology (ICT) has made data acquisition more efficient and less costly, thereby enabling research based on big data acquisition and analysis. This approach is now feasible even in traditionally less data-centric disciplines.

Open Data plays a crucial role in accelerating data-intensive science (Ramachandran et al., 2020). Making various types of data open and readily available to researchers, it is expected to stimulate research activities and improve cost-effectiveness (Kansa, 2012). The OECD describe the research activities conducted with big data that aggregate Open Data (including the sharing of research results and data) as "data-driven research" (OECD, 2015). Although still an evolving concept without a definitive definition, the OECD report suggests that data-driven research is part of the broader concept of "data-driven innovation" (DDI), which can be applied to research activities. The development of ICT in research activities enables the generation and collection of big data related to research activities, thereby stimulating such. Consequently, Open Data is increasingly being regarded as standard practice in some research and funding agencies (Maeda & Torres 2012).

With the promotion of Open Data in this way and the expected development of data-intensive science/data-driven research, it is necessary to evaluate the effectiveness of Open Data promotion in practice. To do this, it is necessary, in the first place, to understand the current status and diversity of data publication and use in each discipline, and to analyze the impact of Open Data is having. This study uses a database of research data to not only clarify the current status of data publication and citation but also determine how data publication affects citation and the interdisciplinary use of data.

Previous research and research questions

There has been research on data openness and use, although not much has been done. In research areas such as genetics and materials science, data sharing and use are considered to be active, and research has been conducted on the promotion of such Open Data, including the extent to which data are made publicly available and the motivations of researchers

when publishing and using data (Robinson-Garcia et al., 2015; Silvello, 2018; Suhr et al., 2020). However, these studies only focus few areas. On the one hand, analyses of specific data repositories have included research on the use of open data in the repository “Figshare”, among others (Quarati & affaghelli, 2020).

Since 2012, Clarivate has published the Data Citation Index (DCI), a database of data citations. Initial analyses of the overall picture of Open Data have been conducted using the DCI to determine which disciplines are publishing data and their citations (Robinson-Garcia et al., 2015), as well as analyses of published software (Park & Wolfram, 2019). However, these studies are from the early stages of the DCI release, and the current status needs to be clarified in more detail now that data openness has become a political priority. At another point in recent years, databases on data have also been developed in the open access database, such as the OpenAIRE Graph funded by the European Union. However, it is unclear whether these databases are sufficient for analyzing and using data publications.

Research questions

Open data has been promoted internationally over the past few years, but its current situation of open data is not clear. In particular, it is not clear which countries have the most data openly available data, although it is assumed that the state of development varies according to national policies and previous initiatives. In addition, some research disciplines may place a high value on open data, while many others may not rely on open data at the moment. Overall, the disciplines that focus more on open data are not clear. From these points of view, the first research question of this study is stated as follows: (RQ1) Which countries and research disciplines are active in Open Data?

However, as we saw in the previous chapter, progress is expected to be made in a new research style termed “data-driven research”, where new research is developed based on data, especially big data. To understand this situation, it is important to not only to understand the number of data published, but also to ascertain whether research is progressing using the data. Therefore, the second research question is presented thus: (RQ2) Which areas of ‘data-driven research are stimulated by the open data? Specifically, to verify whether research using open data has become important in the discipline, this study examines whether papers that use open data are cited more frequently than those that do not. It also checks whether data from diverse disciplines are being used, as the availability of big data is expected to stimulate research using a variety of data.

However, as discussed in the next section, the development of databases on open data is still in its early stages and it is not clear what the current state of databases and their limitations are for analysis are not clear. Therefore, before the analysis, we compare the DCI with the OpenAIRE Graph, a non-commercial open access database, to clarify the relative characteristics of the DCI.

DCI and comparison with the openAIRE graph

Status of DCI registration data

The DCI was launched by Clarivate Analytics in 2012 as a part of the Web of Science (WoS). The DCI contains over 14 million datasets and other items from approximately 450 repositories as well as information on citations from the articles in the Web of Science.

For this analysis, we used the DCI bulk data (XML data) obtained in March 2022 from Clarivate (i.e., data until the end of 2021). In addition, we used DCI data from the web interface to obtain get the latest status as of November 2023 for comparison with the recent OpenAIRE Graph data.

The DCI bulk data at the end of 2021 included 13,882,271 records and 1,371,848 citation information from the Web of Science articles. As Table 1 shows, the DCI contains four data types, where of which 12,227,647 are “datasets” corresponding to open data, accounting for 88% of all records in the DCI. The total number of citations to these data sets is 1,012,547, with an average of 0.083 citations per data set. The table also shows the current number for November 2023 in the last row.

After reviewing the actual data recorded in the DCI, it is unclear what kind of data is included, and what data should be counted as “*one case*.” Therefore, to understand which datasets are recorded in the DCI, the top 10 repositories with the most data sets recorded in 2019 with the highest number of data recorded are listed in Table 2.

The top 10 repositories ranked by frequency show diversity. Among them, four repositories (No.1, 2, 8 and 9) are for individual researchers. These include numerous figures and extensive data accompanying research articles. This particular phenomenon was not previously observed in the DCI analysis conducted in 2015 (Robinson-Garcia et al., in 2015). In the 2015 analysis, the Crystallography Open Database topped the list with 47.4% of all datasets, followed by the Protein Data Bank. This indicates that the number of data repositories in which researchers register their data has expanded significantly over the past seven years. Currently, the current top repository Zenodo mainly serves as a platform for researchers to publish their datasets. However, it should be noted that the time series of Zenodo data shows approximately 510,000 cases in 2019 and approximately 100,000 cases in other years, with the value for 2019, the year analyzed in this study, being particularly large.

The remaining 6 repositories encompass a wide range of data types. They include individual DNA information from the life sciences and biotechnology disciplines, geographic data originating from the United States, and chemical structure data, which can contain tens to hundreds of thousands of items. These findings suggest that Open Data repositories comprise a heterogeneous mix of various data categories. This mix includes data accompanying research papers, substantial volumes of domain-specific data managed by specific research organizations, and survey data collected by governmental agencies.

Comparison with the openAIRE graph

According to its website, the DCI selects repositories based on the repository’s basic publishing standards, its editorial content, the international diversity of its authorship, and its associated citation data. As a result of its selection on the basis of these criteria, the DCI contains, of course, only a part of the world’s data and may be unbalanced in its composition. However, there is no way to know all the data in the world. Therefore, we compared it with another large database, the OpenAIRE Graph.

The OpenAIRE Graph is a database from the OpenAIRE project, a European Commission-supported project that began in 2010 (Artini et al., 2015; Manola et al., 2015). OpenAIRE aggregates metadata from content providers and entity registries that comply with the OpenAIRE acquisition policy and transforms these according to an internal metadata model to ensure quality and provide high-quality, accessible scientific materials relevant to research.

Table 1 Basic Information on the DCI Recorded data (as of the end of 2021)

Document type	Data set	Data study	Software	Repository
Type overview	Single or consistent data set of collected data, research data, or software	Description stored in repository with research data or software	Software used in conducting the study	Databases containing research data and metadata
No. of data	12,227,647	1,400,409	253,772	443
Total citations	1,012,547	294,251	53,453	11,592
No. of data (as of Nov.2023)	14,246,953	1,628,668	430,864	447

Table 2 Information of top 10 repositories for 2019 data sets in the DCI

Repository name	No. of data entries	Organization	Start year	Discipline decided in DCI	Brief explanation
1 Zenodo	517,647	European Organization for Nuclear Research (CERN)	2013	Multidisciplinary Sciences	Open platform for researchers to freely share research data with anyone
2 figshare	221,661	Digital science	2011	Multidisciplinary Sciences	An online, open-access repository where researchers can store and share their research findings, including figures, data sets, images, and videos
3 BacDive	162,953	Leibniz Institute DSMZ—German Collection of Microorganisms and Cell Cultures GmbH	2012	Microbiology	Bacterial meta-database providing phylogeny-related information on bacterial and archaeal biodiversity
4 Barcode of Life Data System	102,910	Center for Biodiversity and Genomics (Canada)	2007	Genetics & Heredity; Biochemistry & Molecular Biology	Web platform dedicated to DNA barcodes
5 Cambridge Structural Database	59,717	Cambridge Crystallographic Research Center	1965	Crystallography	Repository of 3D structural data for a wide range of organic, organometallic, and organometallic molecules, including molecules
6 U.S. Census Bureau TIGER/Line Shapefiles	33,080	U.S. Census Bureau	2007	Geography	Spatial files extracted from the U.S. Census Bureau's MAF/TIGER database of roads, railroads, rivers, and other land features, as well as legal and statistical geographic areas
7 Mendeley Data	31,812	Elsevier, Inc	2015	Multidisciplinary Sciences	Public Repository of Research Data for Researchers

Table 2 (continued)

Repository name	No. of data entries	Organization	Start year	Discipline decided in DCI	Brief explanation
8 MassBank	16,079	Mass Spectrometry Society of Japan (MSSJ)	2006	Spectroscopy	Public repository designed to share mass spectrometry data for compound identification and structure elucidation with the scientific research community
9 Animal Quantitative Trait Loci(QTL)Database	12,591	The National Animal Genome Research Program (NAGRP), USA	2004	Genetics & Heredity	All publicly available trait mapping data, i.e., QTL (phenotype/expression, eQTL), candidate gene and association data (GWAS), and copy number variations (CNV) mapped to livestock animal genomes
10 Dryad	12,076	National Evolutionary Synthesis Center	2008	Evolutionary Biology; Ecology; Biodiversity Conservation	An international open-access repository of research data, especially data underlying scientific and medical publications (mainly of evolutionary, genetic, and ecology biology)

We used snapshot data of August 2023 from the OpenAIRE site. The OpenAIRE Graph contains 234,321,533 records including a large volume of publication data (168,043,128). In relation to the data recorded by the DCI, among the above, there are 58,576,399 “datasets” (Metadata records about research data) and 343,168 “software” (Metadata records about research software). Compared to the data sets in the DCI (14,246,953) simply by numbers, this is more than four times the number of data sets in the DCI.

Table 3 compares information on the top 10 repositories in the DCI and the top 10 ‘publishers’ in the OpenAIRE Graph in two tables. In the OpenAIRE Graph, the “publisher” item contains the equivalent information to the repository information in the DCI, but it may not always correspond to the repository, as discussed below. Our analysis focuses on the datasets published in 2019.

Only 4 of the top 10 repositories in the DCI overlap with the top 10 publishers in the Open AIRE Graph. Only 6 of the top 10 in the DCI are included in the Open AIRE Graph and only 5 of the top 10 issuers in the OpenAIRE Graph are included in the DCI. In other words, rather than one almost encompassing the other, a large part of both are not included in the other.

In detail, the OpenAIRE Graph shows that, similar to the DCI, Zenodo is the most used data repository in 2019. Close behind is the Global Biodiversity Information Facility (GBIF), a database that provides metadata on various organisms on a global scale and plays a crucial role in the discipline of biodiversity, which is not included in the DCI. The third source is the German Collection of Microorganisms and Cell Cultures (DSMZ) by the Leibniz Institute, which is a comprehensive resource focusing on biodiversity in prokaryotes and eukaryotes. There is a noticeable divergence in the number of entries on Figshare between the OpenAIRE Graph and the DCI, with the OpenAIRE Graph having fewer entries with only 52,239. The data entries in 2019 on Figshare’s official website are close to the DCI number. However, in OpenAIRE Graph, even if the “publihsrer” is the name of a different data provider, many of the data site addresses list Figshare, which adds up to 77,156 entries.

Although, as mentioned above, both databases contain data that the other does not have, the fact that the DCI does not contain big data for the GBIF and DSMZ entries, which are included in the OpenAIRE Graph, suggests that the DCI may not be fully representative of all research areas. Important biodiversity data may not be given sufficient attention in the DCI and the results of the DCI analysis should be discussed with this in mind.

There are also pros and cons regarding to the way disciplines are handled: the DCI, on the one hand, has the advantage of having consistent disciplines assigned by Clarivate. However, disciplines are assigned to repositories rather than individual data. In the OpenAIRE Graph, on the other hand, although disciplines are assigned to individual data, the determinants of the disciplines are either freely assigned by the data provider or are “keywords” automatically generated by the OpenAIRE Graph, making consistent aggregation extremely difficult.

Citation practices and self-citations

Table 3 also lists the numbers of the data-related citations. However, the aforementioned study highlightsthat most of the DCI citations are self-citations. Therefore, we checked the WoS papers for several samples of the data sets cited in 2019. We found that, especially in the case of data publication by individual researchers, data and figures accompanying such papers and stored in repositories were also treated as “citations.” (note

Table 3 Comparison of data and citation counts between the DCI and the OpenAIRE Graph published in 2019

DCI (As of Nov. 2023)		OpenAIRE graph									
Rank	Repository	No. data entry	No. data cited >= 1	No. data cited >= 2	% data cited >= 2	Rank	Repository	No. data entry	No. data cited* >= 1	No. data cited* >= 2	% data cited >= 2
1	Zenodo	525,625	706	160	0.03%	1	Zenodo	502,104	2642	437	0.09%
2	Figshare	224,032	452	48	0.02%	8	Figshare	52,239	818	99	0.19%
3	BacDive	162,953	0	0	0%	-	-	-	-	-	-
4	Barcode of Life Data System	102,910	2	0	0%	4	Barcode of Life Data Systems	102,296	7	0	0%
5	Cambridge Structural Database	59,983	49,541	663	1.11%	7	Cambridge Crystallographic Data Centre	59,770	33	1	0.00%
6	U.S. Census Bureau TIGER/Line Shapefiles	33,080	0	0	0%	-	-	-	-	-	-
7	Mendeley Data	31,747	333	61	0.19%	-	Mendeley Data	11,902	320	77	0.65%
8	MassBank	16,079	712	0	0%	-	-	-	-	-	-
9	Animal Quantitative Trait Loci (QTL) Database	12,619	12,497	0	0%	-	-	-	-	-	-
10	Dryad	12,191	3,980	400	3.28%	2	Dryad	13,084	1615	235	1.80%
-	-	-	-	-	-	2	The Global Biodiversity Information Facility	165,169	2340	52	0.03%
-	-	-	-	-	-	3	DSMZ	118,206	0	0	0%

Table 3 (continued)

DCI (As of Nov. 2023)		OpenAIRE graph							
Rank	Repository	No. data entry	No. data cited >= 1	No. data cited >= 2	% data cited >= 2	No. data entry	No. data cited* >= 1	No. data cited* >= 2	% data cited >= 2
	Harvard Data-verse	4	0	0	0%	102,296	228	45	0.04%
-	-	-	-	-	-	60,807	683	128	0.21%
-	-	-	-	-	-	36,920	0	0	0%
-	-	-	-	-	-	34,169	0	0	0%
ALL		1,297,231	99,735	3196	0.25%	5,303,367	14,695	2,395	0.05%

*For OpenAIRE Graph, cited times are counted for the citation from the publication whose sub-type is categorized as "Article"

that information that can systematically confirm whether the data are self-citations is not included in the DCI). Although 99,735 datasets were cited at least once in the DCI, it is possible that self-citations might have accounted for a large proportion. Therefore, Table 3 presents the number of datasets that have received more than one citation as a proxy for the number of the cited datasets without self-citation. Therefore, in terms of data citations from other papers, data-driven research may not yet be sufficiently advanced in the whole picture. However, the number of self-citations is not a meaningless number. Considering that the number of self-citations reflects the status of the publication of the data accompanying a paper, it can be seen as an indicator of the progress of data-driven research through the publication of data.

A similar calculation was made for the OpenAIRE Graph: In the “Relationship” object, relation of type “IsCitedBy” was counted, focusing only on citations from the sub-type “Article” within the type “Publication”. Although “Publication” includes various types (e.g., Book, Master thesis, Doctoral thesis, Preprint, etc.) other than Article, only citations from “Article” are counted to make a comparison with the DCI data citations.

When comparing the citation data of both in Table 4, it is not clear which is better, even when the values are compared for the same repositories. However, in the OpenAIRE Graph, although the number of data citations appeared to be high (e.g. Zenodo), there were some cases where citations to the papers that generated the data were measured as citations to the data. This is because of the current situation, in which the manner of data citations is not clearly defined. In this respect, based on the explanation on the DCI website, the DCI is working with some repositories to develop a data citation format that is expected to provide citation relationships with a certain degree of accuracy.

In summary, there are no complete databases for analyzing datasets, including citations, and the data they contain vary from one database to another. Despite these limitations, the analysis below uses the DCI (bulk data as of the end of 2021), as it is easier to relate to WoS articles and is considered to have a certain level of quality control over citation relationships. The repositories in the DCI are also selected based on clear criteria. However, it should be noted that the results may be unbalanced due to the composition of the repositories included.

Analysis on data release and its use by country and research disciplines

As for RQ1, to ascertain which countries and disciplines are actively publishing data, we tabulated the number of databases and time-series changes in 10 major disciplines (Figs. 1 and 2). It should be noted again that research disciplines are assigned to the repository containing the dataset. For Fig. 2, a three-year moving average I as taken.

The most common disciplines in the DCI-recorded data set are Genetics & Heredity (31.5%), followed by Multidisciplinary Sciences (23.3%), Biochemistry & Molecular Biology (19.6%), and Crystallography (10.4%). A review of the chronological changes by disciplines reveals that all disciplines have been increasing since 2005; however, Multidisciplinary Sciences have increased rapidly since 2015, while all other disciplines have been decreasing. The data categorized as Multidisciplinary Sciences are those published in repositories that allow researchers to publish data accompanying their papers, regardless of the disciplines, which also appear at the top of Table 2. In other words, the number of individual researchers releasing data to non-discipline-specific repositories has been increasing rapidly since around 2015, and it is possible that some of the datasets that have

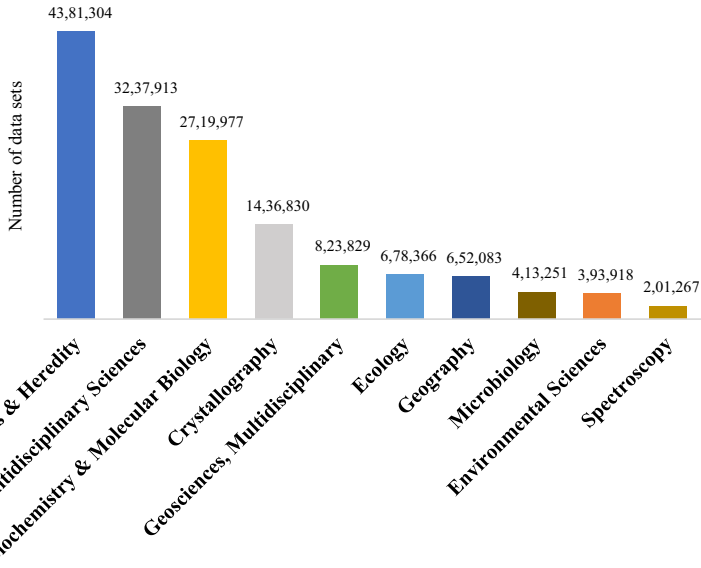


Fig. 1 Number of datasets in 10 major disciplines (-2021)

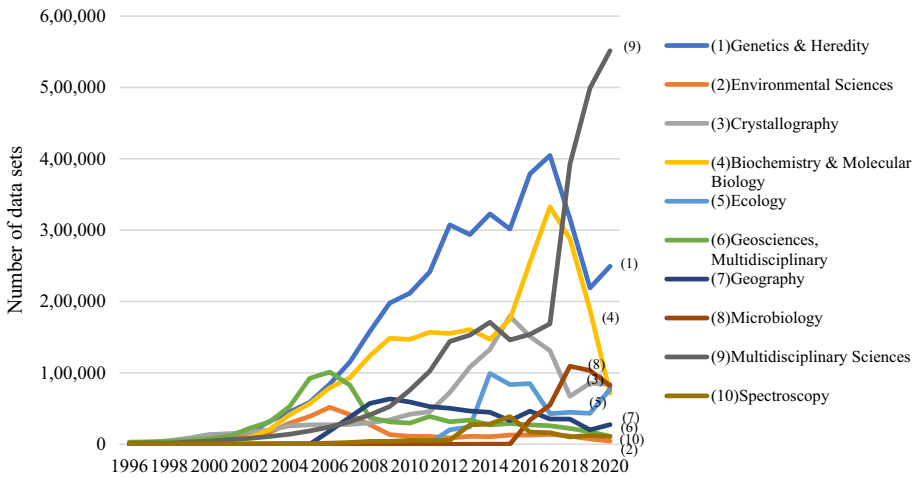


Fig. 2 Time-series changes in the number of datasets in 10 major disciplines

Table 4 Number of data sets and share (%) of the 10 main disciplines for the 5 major countries (-2021)

	USA		China		Japan		Germany		Norway		Other countries or N/A		Total
	#Data sets	%	#Data sets	%	#Data sets	%	#Data sets	%	#Data sets	%	#Data sets	%	
Genetics & Heredity	1,068,211	24	141,983	3	70,931	2	101,772	2	17,940	0	2,980,435	68	4,381,304
Multidisciplinary Sciences	38,943	1	17,330	0	5,675	0	12,714	0	87,065	3	3,076,184	95	3,237,913
Biochemistry & Molecular Biology	927,307	34	48,520	2	61,967	2	81,463	3	8,653	0	1,595,026	59	2,722,977
Crystallography	442	0	10	0	4	0	4	0	-	-	1,436,370	100	1,436,830
Geosciences, Multidisciplinary	26,109	3	78	0	207	0	1,027	0	59	0	796,346	97	823,829
Ecology	59,383	9	8	0	55	0	16	0	19	0	618,876	91	678,366
Geography	26,109	3	78	0	207	0	1,027	0	59	0	799,377	97	826,860
Microbiology	-	-	-	-	-	-	-	-	-	-	413,251	100	413,251
Environmental Sciences	64,168	6	28	0	82	0	125	0	48	0	1,004,335	94	1,068,792
Spectroscopy	7,416	4	310	0	33,776	17	9,782	5	4	0	149,954	75	201,267

been previously published in discipline-specific repositories are now also being published in these repositories.

In the Table 4, we present the number of data sets for each of the 10 major disciplines of each major country. Note that the country of the creator of the data set is assigned to 2,058,624 data sets in total, which means that only 16.8% of the total data sets are assigned the name of the country. Therefore, the overall percentage of countries is not large.

The U.S. has a 20–30% share in the Biochemistry & Molecular Biology and Genetics & Heredity disciplines. China has almost no data releases outside of the top three disciplines. For Spectroscopy, Japan has released the top 17% of the data sets. According to Table 2, the repository of Spectroscopy is operated by the Mass Spectrometry Society of Japan (MSSJ), which is one of the disciplines where Open Data has advanced in Japan. In Norway, there is a particularly large amount of data in the disciplines of Multidisciplinary Sciences, which is because a national repository called “DataverseNO” has been in operation since 2017. For major disciplines, while the U.S. has a large share of data released by its government, other countries also differ in the percentage of datasets released, and it is possible that there are differences in the disciplines in which each country has historically focused on data release.

Analysis on the development of data-driven research by research disciplines

For RQ2, we quantitatively and qualitatively analyzed the disciplines in which data-driven research through data publication was active. The first step is to calculate the percentage of papers in each discipline with data citations in 2019, the most recent year with the highest number of data sets loaded in the DCI bulk data. A total of nine disciplines are targeted, those in which data sets registered in the DCI are cited more than 2,500 times in WoS papers in total for all years. The second part is to measure the diversity of the disciplines in the DCI data set cited by the paper for each discipline of papers. Here, we use the Rao-Stirling diversity Index, which measures diversity and considers the potential distance between disciplines:

$$\text{Rao – Stirling diversity index } D = \sum_{ij(i \neq j)} d_{ij} p_i p_j, \tag{1}$$

where p_i is the proportion of discipline i among all cited datasets, and d_{ij} is calculated by 1-similarity, which is calculated by dividing the number of data sets to which both disciplines i and j are assigned by the sum set of disciplines i and j .

In general, the larger the population, the higher the diversity index; therefore, so the effect of differences in population size was mitigated by resampling 1000 times with a sample size of 2500, the smallest sample size among the disciplines.

First, the percentage of data citations (percentage of papers that cite at least one data item on the DCI) in each discipline is presented in Table 5 for all papers and for the top 10% of the most cited papers. However, as previously mentioned, as “self-citations” are included here, this indicator should be interpreted as the progress of data-driven research, including the trend of data publication, rather than pure citation. The results reveal that Crystallography is more advanced in data-driven research than in other disciplines, accounting for more than 30% of all papers. Ecology is next with 12%. In six of the nine disciplines, the percentage of papers in the top 10% most cited tends to be statistically

Table 5 Percentage of papers with data citations in each discipline

	Entire paper			Top 10% papers only		
	Papers with any data citation	All papers	Ratio	Papers with any data citation	All papers	Ratio
Biochemistry & Molecular Biology	1801	56,419	3.19%	319	6395	4.99% ** <i>p</i> < 0.01
Biodiversity Conservation	399	6399	6.24%	57	679	8.39% * <i>p</i> < 0.05
Genetics & Heredity	602	21,281	2.83%	129	1999	6.45% ** <i>p</i> < 0.01
Crystallography	1960	6274	31.24%	196	634	30.91%
Geosciences, Multidisciplinary	592	26,965	2.20%	104	3003	3.46% ** <i>p</i> < 0.01
Meteorology & Atmospheric Sciences	441	14,536	3.03%	71	1480	4.80% ** <i>p</i> < 0.01
Multidisciplinary Sciences	1612	61,635	2.62%	474	6068	7.81% ** <i>p</i> < 0.01
Social Sciences, Interdisciplinary	59	6684	0.88%	6	789	0.76% ** <i>p</i> < 0.01
Ecology	1635	20,369	8.03%	248	1981	12.52% ** <i>p</i> < 0.01

significantly higher than that of data citations for the entire paper. The results indicate that papers with high citation counts tend to have data citations (including self-citations). We reviewed the submission policies of the 10 journals to which the top 10% of WOS papers in crystallography belong and found that all journals require data sharing.

Next, the Rao-Stirling diversity index of the disciplines of data cited by the nine WoS paper disciplines is illustrated in Fig. 3. The median values of resampling 1,000 times range from 0.343 to 0.470 depending on the discipline, indicating the diversity of disciplines. The overall results indicate that there are differences in the use of the disciplines of the data cited by each discipline. Crystallography is more concentrated than the other disciplines, ranging from 0.20 to 0.25. In contrast, Multidisciplinary Sciences has the highest value, indicating that the types of data cited in this discipline are very diverse, ensuing because this research discipline category includes various research. This is not surprising, especially given the definition of Multidisciplinary Science. Biochemistry & Molecular Biology, Ecology, and Social Sciences, Interdisciplinary have values slightly above the average, indicating that these disciplines exceed the overall average in terms of the diversity of cited data.

Figure 4 illustrates the characteristics of the development of data-driven research in each discipline by showing the relationship between the total number and percentage of papers citing data and the Rao-Stirling diversity index shown in Fig. 3. Crystallography has a high data citation ratio in the discipline; however, but the diversity of the cited data is more limited than that in other disciplines. Ecology not only has the second highest data citation rate after Crystallography, but also suggests a high diversity of data to cite. Social Sciences, Interdisciplinary does not have a high citation rate for data, but their diversity is relatively high. Multidisciplinary Sciences has the highest median position in terms of diversity, indicating that the types of data cited are more diverse than in any other discipline. This situation varies considerably across disciplines.

Discussion

After the comparison with the OpenAIRE Graph, the DCI data were analyzed from various perspectives to gain a cross-disciplinary understanding of the current status of the promotion of data-driven research. The results revealed that the countries and disciplines in which Open Data are thrives tend to differ. The U.S., which promotes Open Data on a

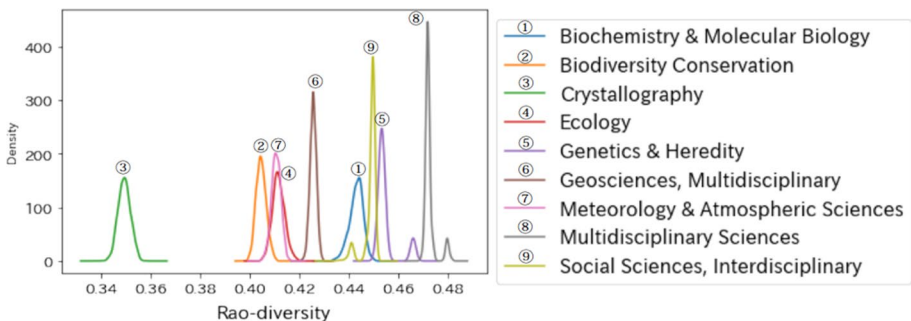


Fig. 3 Rao-Stirling diversity index of data cited by papers in each discipline

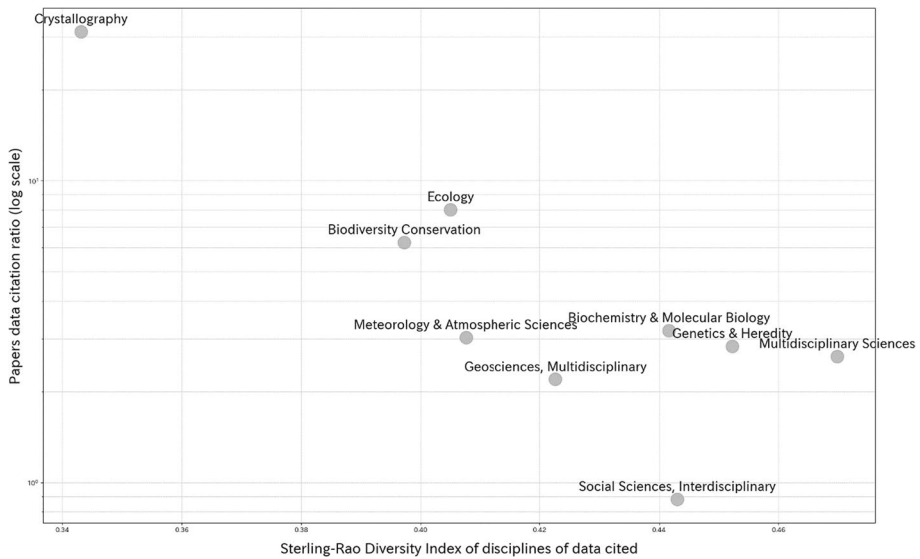


Fig. 4 Relationship between citation data ratio and Rao-stirling diversity index by discipline

national scale, tends to release numerous datasets in multiple disciplines. Japan has been particularly active in releasing spectroscopic data. However, the number of data sets in repositories (classified as Multidisciplinary science disciplines), where researchers can load data regardless of the discipline, has increased rapidly in recent years. In terms of data citation rates, it is particularly high in the discipline of Crystallography, but the disciplines of cited data sets are concentrated. Meanwhile, there are disciplines in which data citation rates are low, but the disciplines whereby databases are used are diverse. We find that the foundations of data-driven research are characterized by diverse types, each with a distinct citation behaviour and disciplinary focus.

There is a high frequency of data citations in the disciplines of Crystallography; however, a significant amount of research has focused on crystallographic data. This trend suggests that data sharing and citation are becoming increasingly integrated into the research process, indicating that a data-centric approach within the discipline is thriving in this discipline. Conversely, disciplines such as interdisciplinary science and genetics have modest levels of data citation, but reflect a greater diversity in the types of data used and cited in research. This pattern indicates a diverse landscape of data-driven research, collecting multiple types of data from different disciplines, and suggests an integrated approach that draws from a wide range of data sources to support scientific inquiry.

Furthermore, the DCI does not include data from the GBIF repository related to biology and environmental studies, which has the second highest number of data entries in the OpenAIRE Graph, and the number of datasets in the discipline of ecology and environmental sciences ranks below sixth in the DCI. Therefore, it is difficult to conclude that this DCI analysis reflects the actual situation.

This analysis also revealed limitations of the DCI database. The range of data covered by the DCI might be biased, as shown in the comparison with the OpenAIRE Graph. Further research is needed to determine how this affects the results of the analysis. For databases of research articles, coverage has also been investigated, and it has been shown

that the conclusions of the bibliometric analysis depend on the data source and the indicators used (Harzing and Alakangas 2016; Visser et al., 2021). The same can be said for the analysis of databases of data. Comparisons need to be made not only with the OpenAIRE Graph but also with various databases.

In addition to this coverage issue, the lack of established data citation conventions, such as the lack of uniformity in how data are described in papers and other documents when they are cited, and the maintenance of citation information in repositories, have also created problems (Silvello, 2018). Furthermore, the relationship between articles and the accompanying publicly available data also requires a conceptual examination. One view regards this as “self-citation” and removes it from the analysis. However, another view that places the data at the center implies that whoever the users are, they the ones who used the data. In the latter case, even if it is self-cited, the data have value and need not be removed. It is necessary to re-examine the indicators by examining the concept of data citation itself. Future research should investigate how Open Data affects research activities from various perspectives.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Artini, M., Atzori, C., Bardi, A., La Bruzzo, S., Manghi, P., & Mannocci, A. (2015). The open AIRE literature broker service for institutional repositories. *D-Lib Magazine*. <https://doi.org/10.1045/november2015-artini>
- Fecher, B., & Friesike, S. (2014). Open science: One term, five schools of thought. In S. Bartling & S. Friesike (Eds.), *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing* (pp. 17–47). Springer.
- Harzing, A.-W., & Alakangas, S. (2016). Google scholar, scopus and the web of science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*, 787–804. <https://doi.org/10.1007/s11192-015-1798-9>
- Harzing, A. W., & Alakangas, S. (2016). Google scholar, scopus and the web of science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*, 787–804. <https://doi.org/10.1007/s11192-015-1798-9>
- Kansa, E. (2012). Openness and archaeology’s information ecosystem. *World Archaeology*, *44*, 498–520. <https://doi.org/10.1080/00438243.2012.737575>
- Koznov, D., Andreeva, O., Nikula, U., Maglyas, A., Muromtsev, D., & Radchenko, I. (2016). A survey of open government data in Russian federation. In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016)—KMIS*. <https://doi.org/10.5220/0006049201730180>
- Maeda, E., & Torres, J. (2012). Open environmental data in developing countries: Who benefits? *Ambio*, *41*, 410–412. <https://doi.org/10.1007/s13280-012-0283-4>
- Manola, N., Rettberg, N., & Manghi, P. (2015). Open AIRE plus project executive report. *Zenodo*. <https://doi.org/10.5281/zenodo.15464>
- OECD. (2015). *Data-driven innovation: Big data for growth and well-being*. OECD.
- Park, H., & Wolfram, D. (2019). Research software citation in the data citation index: Current practices and implications for research software sharing and reuse. *Journal of Informetrics*, *13*(2), 574–582.

- Quarati, A., & Raffaghelli, J. (2020). Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the figshare case. *Journal of Information Science*, 48, 423–448. <https://doi.org/10.1177/0165551520961048>
- Ramachandran, R., Bugbee, K., & Murphy, K. (2020). From open data to open science. *Earth and Space Science*. <https://doi.org/10.1029/2020EA001562>
- Robinson-Garcia, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, 67(12), 2964–2975. <https://doi.org/10.1002/asi.23529>
- Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1), 6–20. <https://doi.org/10.1002/asi.23917>
- Suhr, B., Dungal, J., & Stocker, A. (2020). Search, reuse and sharing of research data in materials science and engineering—a qualitative interview study. *PLoS ONE*, 15(9), e0239216. <https://doi.org/10.1371/journal.pone.0239216>
- UNESCO (2021). *UNESCO Recommendation on Open Science*. UNESCO, <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>
- Visser, M., Jan, N., van Eck, L., & Waltman, . (2021). Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic. *Quantitative Science Studies*, 2(1), 20–41. https://doi.org/10.1162/qss_a_00112

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.