An introduction to the coverage of the Data Citation Index (Thomson-Reuters): disciplines, document types and repositories

Daniel Torres-Salinas[1], Alberto Martín-Martín[2], Enrique Fuente-Gutiérrez[3]

[1]EC3 Research Group & EC3metrics Spin-Off, Universidad de Navarra, Pamplona (Spain)
[2]EC3 Research Group & EC3metrics Spin-Off, Universidad de Granada, Granada (Spain)
[3]C3 Research Group & EC3metrics Spin-Off, Universidad de Granada, Granada (Spain)

ABSTRACT

In the past years, the movement of data sharing has been enjoying great popularity. Within this context, Thomson Reuters launched at the end of 2012 a new product inside the Web of Knowledge family: the Data Citation Index. The aim of this tool is to enable discovery and access, from a single place, to data from a variety of data repositories from different subject areas and from around the world. In this working paper we present some preliminary results from the analysis of the Data Citation Index. Specifically, we address the following issues: discipline coverage, data types present in the database, and repositories that were included at the time of the study.

KEYWORDS

Data; Research Data; Open Access; Data Sharing; Scientific Communication; Information Sources Repository; Databases; Citation Indexes; Web of Science; Thomson Reuters

Grupo de Investigación EC3
Evaluación de la Ciencia y de la
Comunicación Científica

Spin-Off
UNIVERSIDAD DE GRANADA

EC3metrics

EC3's Document Serie:
EC3 Working Paper, Nº 11

Document History
Versión 1.0, Published on 11 June 2013, Granada

Cited as
Torres-Salinas, D.; Martín-Martín, A.; Fuente-Gutierrez, E (2013). An introduction to the coverage of the Data Citation Index (Thomson Reuters): disciplines, document types and repositories. EC3 Working Papers, N 11, June 2013

Corresponding author
Daniel Torres-Salinas, torressalinas@gmail.com

1. Introduction

1.1. The Data Sharing Context

During the last decade, there has been a heated debate among the scientific community about the need of releasing research data, a movement commonly referred to as data sharing. Although the practice of sharing data has been present among researchers for a long time (Hrynaszkiewicz; Altman, 2009), the movement of data sharing is currently enjoying great popularity due to the convergence of a number of circumstances, two of the most important being the development of the information technologies, and researcher's ever more open attitude towards their findings (as exemplified by movements like Open Access).

Currently there are a large number of initiatives, commonly called data banks or data repositories, dedicated to store, describe and disseminate scientific data. Unlike pre-prints or post-prints repositories, which deal only with one bibliographic format for the items they contain, there is a great variety of data repositories and the solutions adopted are different in each case, and often this makes them difficult to use to people without knowledge of the data bank's subject area (Torres-Salinas, Robinson-García, Cabezas-Clavijo, 2012).

The benefits of data sharing have already been studied and identified (Arzberger et al., 2004; Vickers, 2006). In the first place, data sharing contributes to make the most of the funds invested in science because it helps prevent duplication of efforts and also because it makes possible the development of new studies that reuse these data. This is worth considering in the present situation of economic crisis, especially when research is government funded. Secondly, these data can be used as a tool to detect fraud, since they would enable other researchers to verify or disprove the results of an experiment through its replication (Renolls, 1997). Thirdly, there is evidence that published studies whose data are openly available receive more citations (Piwowar; Day; Fridsma, 2007). Lastly, it is possible that these practices open the way for the creation of data metrics that complement existing indicators for scientific evaluation (Wouters & Schröder, 2003; Costas, R., Meijer, I., Zahedi, Z. and Wouters, P., 2013).

1.2. The Data Citation Index – Thomson Reuters

Within the context described above, Thomson Reuters has added a new member to the Web of Knowledge family of databases: the Data Citation Index (DCI). The DCI, released in November 2012, is described as a tool to discover and access, from a single place, data from a variety of repositories from the three major subject areas (Science & Technology, Social Sciences, and Arts & Humanities) and from around the world. In order to be included in the DCI, a data repository must first undergo a process of evaluation in which a number of factors are considered, including the repository's basic publishing standards, its editorial content, the international diversity of its authorship, and the citation data associated with it (Thomson Reuters, 2012). At the same time, records in the DCI are linked to the publications they inform, thus providing citation information for the data sets, and opening the way to data citation analysis. However, even though the DCI is the first tool that allows us to quantify the impact and reutilization of research data, it is as of yet a young product that needs to be assessed in order to comprehend its strengths and limitations. This assessment will allow bibliometricians, librarians, and the rest of potential users of this tool to better understand for what purposes it may be used and how.

1.3. Objectives

For this reason, the EC3 Research Group (University of Granada) is launching a new line of research to study the DCI. In this Working Paper we will present some preliminary results where we address the following issues:

1. Discipline coverage in the DCI.
2. What kinds of data types are present in the DCI, and what is their statistical distribution?
3. Which data repositories contribute a larger share of records to the DCI?

We believe these results are interesting and innovative since they are the first empiric results obtained from an analysis of the DCI as a scientific information and evaluation tool.

2. Methodology

For the purpose of this analysis, all records from the Data Citation Index were downloaded in April-May 2013, using the DCI web interface. The resulting text files were processed and added to a relational database, using the Accession Number field (UT) as the primary key for the data records. The rest of the fields analyzed were: Document Type (DT), Publication Year (PY), and Web of Science Category (WC). Regarding the issue of discipline coverage, two classification systems have been used in order to assign categories to the records: one of them comprises four major subject areas (Science, Social Sciences, Humanities & Arts, and Engineering & Technology), and the other is the one proposed by Moed (2005), with thirteen disciplines. These systems were built by aggregating Web of Science categories, in the same way as we did in other studies analyzing products by Thomson Reuters (Torres-Salinas et al., 2013).

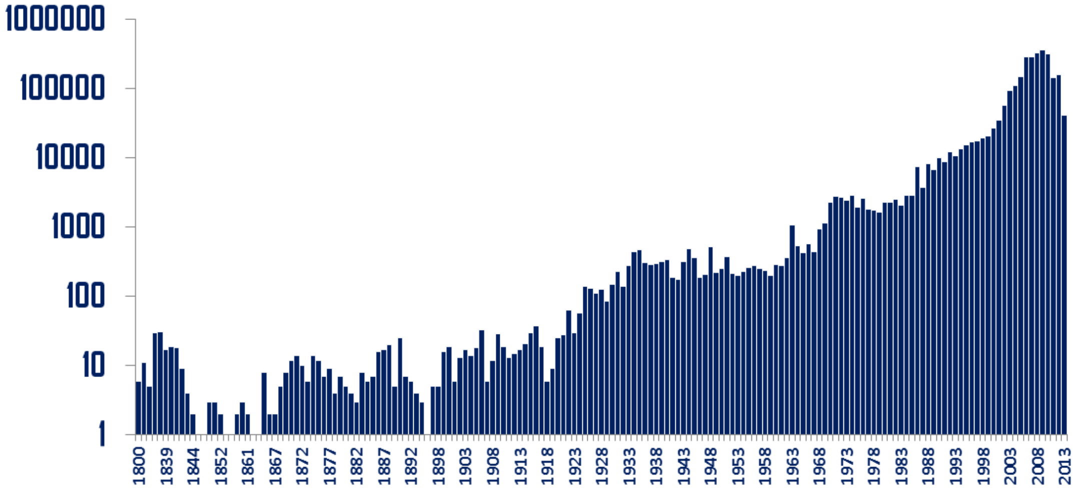
3. Results

3.1. General description & distribution per area and scientific field

At the time of the download, the Data Citation Index held a total of 2.623.528 records. The oldest of them can be traced back to the year 1800 (see Figure 1) but, as is natural, this database mainly deals with contemporary data, and 92% of records are dated between 2000 and 2013. The year where we can find more records is 2009, with a total of 365.381.

If we attend to their subject areas, it is clear that most of the records belong to the area of Science, with a crushing 80% (see Figure 2), well ahead of the Social Sciences with 18%, and Humanities & Arts with 2%. The presence of records in the area of Engineering & Technology is almost non-existent, with less than 0.1%. These results are consistent with the known issue of the under-representation of the Social Sciences and Arts & Humanities in other multidisciplinary databases of the WoK family, namely the Web of Science.

Figure 1
Record distribution by year of publication. 1800-2013.



If we consider the classification system proposed by Moed (see Figure 3), Clinical Medicine is the discipline that accounts for the largest share of the records (50.8%), closely followed by Molecular Biology and Biochemistry with 48%, and, at some distance, Geosciences with 20% (note that a record may be assigned to several disciplines).

Figure 2
Record distribution per scienfific field in the Data Citation Index

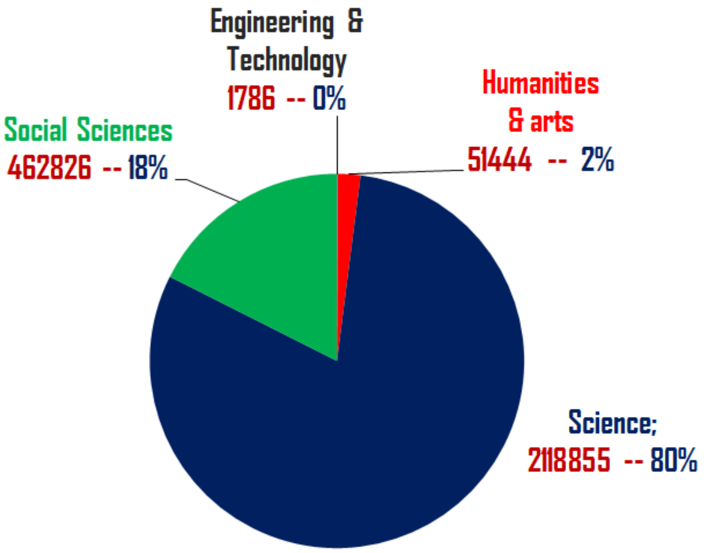
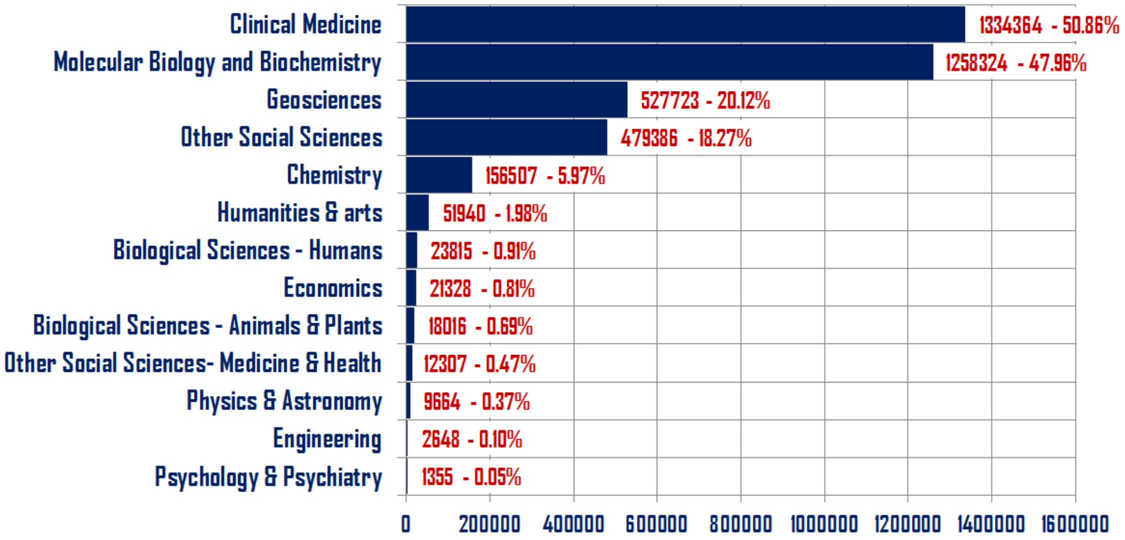


Figure 3
Record distribution per discipline in the Data Citation Index



3.2. Distribution per type of document: data repository, data study and data set

The Data Citation Index contains at the moment three different document types: data repositories, data studies, and data sets. The definitions that Thomson Reuters gives to each one of these document types can be seen in Table 1. Data sets are the basic unit of information and are usually, but not necessarily, part of a data study. Data studies are, in turn, part of a data repository. The distribution of all records among each of these document types is presented in Table 2, broken down by subject areas. There are a total of 2.475.534 records in the data set category, which makes it the most common document type in the database by far, with 94% of the total number of records. Only 159.280 are classified as a Data Study (6%) and 97 as a data repository.

Table 1
Document types in the Data Citation Index, according to Thomson Reuters.

Data repository

A database or collection comprising data studies, and data sets which stores and provides access to the raw data. Constituent data studies, and sometimes individual data sets, are marked up with metadata providing a context for the available raw data.

Data study

Description of studies or experiments held in repositories with the associated data which have been used in the data study. (Includes serial or longitudinal studies over time). Data studies can be a citable object in the literature and may have cited references attached in their metadata, together with information on such aspects as the principal investigators, funding information, subject terms, geographic coverage etc. The level of metadata provided varies between repositories.

Data set

A single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment. Data sets may exist in a number of file formats and media types: they may be number based files such as spreadsheets, images, video, audio, databases etc. Data sets can be a citable object in the literature and may include cited references attached in their metadata, but more commonly they inherit the metadata of the overall study in which they are used.

Source: Repository Evaluation, Selection and Coverage Policies for the *Data Citation Index* (Thomson Reuters, 2012)

Table 2
Document type distribution by subject areas in the Data Citation Index.

	Data set	Data study	Repository	Total
Engineering & Technology	1.545	240	1	**1.786**
Humanities & Arts	44.588	6.847	9	**51.444**
Science	2.004.449	114.338	68	**2118.855**
Social Sciences	424.952	37.855	19	**462.826**
Total	**2.475.534**	**159.280**	**97**	**263.4911**

	Data set	Data study	Repository	Total
Engineering & Technology	86,51%	13,44%	0,06%	**100%**
Humanities & Arts	86,67%	13,31%	0,02%	**100%**
Science	94,60%	5,40%	0,00%	**100%**
Social Sciences	91,82%	8,18%	0,00%	**100%**
Total	**93,95%**	**6,04%**	**0,00%**	**100%**

	Data set	Data study	Repository	Total
Engineering & Technology	0,06%	0,16%	1,11%	**0,07%**
Humanities & Arts	1,81%	4,43%	10,00%	**1,96%**
Science	81,19%	73,92%	75,56%	**80,76%**
Social Sciences	17,21%	24,47%	21,11%	**17,64%**
Total	**100%**	**100%**	**100%**	**100%**

As shown in Table 2, Science accumulates 81% of all the data sets, 73, 92% of data studies, and 75.56% of data repositories. Data sets are also the predominant typology in every major subject area. It is also worth noticing that there seems to be a larger presence of data studies in the areas of Engineering & Technology, and Humanities & Arts (around 13% of the total of records in both areas) which doubles the average percentage for that document type if we consider the entire database (6%).

3.3. Main repositories and distribution

Lastly, in Table 3 we present the names and record count of the main repositories that are indexed in the DCI. We only consider those repositories which contain at least 100 records, regardless of the document type. Only 64 repositories met this requirement. As can be seen, there is a very high concentration of records in a set of four repositories, which account for 75% of records in the *DCI: Gene Expression Omnibus*, *UniProt Knowledgebase*, *PANGAEA* and *U.S. Census Bureau TIGER/Line Shapefiles*. The first two repositories belong to Biochemistry & Molecular Biology, and Genetics & Heredity, while the other two fall within the scope of Geosciences, Social Sciences, and Geography. The best represented disciplines in the DCI in terms of number of repositories are Genetics & Heredity (24), Biochemistry & Molecular Biology (16), Social Sciences, Interdisciplinary (13), Astronomy & Astrophysics (9) and Geosciences, Multidisciplinary (9).

Table 3.
Main repositories in the Data Citation Index sorted by number of records

Gene Expression Omnibus	654917	24,96%	Mouse Phenome Database	2723	0,10%
UniProt Knowledgebase	496803	18,94%	Biological Magnetic Resonance Data Bank	2597	0,10%
PANGAEA	447137	17,04%	Electron Microscopy Data Bank	2525	0,10%
U.S. Census Bureau TIGER/Line Shapefiles	358957	13,68%	Human Metabolome Database	2433	0,09%
Crystallography Open Database	150917	5,75%	Australian Data Archive	2107	0,08%
ArrayExpress Archive	91846	3,50%	Australian Antarctic Data Centre	1765	0,07%
Protein Data Bank	76563	2,92%	Midbody, Centrosome and Kinetochore	1490	0,06%
Inter-university Consortium for Political and Social Research	72637	2,77%	Cancer GEnome Mine	935	0,04%
Roper Center for Public Opinion Research	25384	0,97%	Oak Ridge National Laboratory Distributed ...	905	0,03%
U.S. National Oceanographic Data Center	25370	0,97%	cancer Nanotechnology Laboratory	861	0,03%
EMAGE Gene Expression Database	23566	0,90%	British Oceanographic Data Centre	856	0,03%
miRBase	18222	0,69%	Finnish Social Science Data Archive	825	0,03%
Animal QTL Database	16636	0,63%	REFOLD	714	0,03%
NOAA National Geophysical Data Center	16500	0,63%	Database of Protein Disorder	650	0,02%
Institute for Quantitative Social Science	16196	0,62%	U.S. National Archives and Records Administration Dataverse	584	0,02%
Odum Institute Data Archive	10516	0,40%	eCrystals - Southampton	537	0,02%
IEDA: Marine Geoscience Data System	9110	0,35%	Dataweb	407	0,02%
nmrshiftdb2	8962	0,34%	Archaeology Data Service	405	0,02%
Chemical Effects in Biological Systems	8939	0,34%	Cell Centered Database	374	0,01%
The Cell: An Image Library	8789	0,34%	PseudoBase	360	0,01%
Dryad	6639	0,25%	British Geological Survey	333	0,01%
NOAA Paleoclimatology	6522	0,25%	1.2 Meter CO Survey Dataverse	306	0,01%
Cancer Models Database	5935	0,23%	COordinated Molecular Probe Line Extinction Thermal ..	302	0,01%
Nucleic Acid Database	5596	0,21%	British Atmospheric Data Centre	211	0,01%
The Association of Religion Data Archives	5405	0,21%	ShareGeo Open	204	0,01%
Eurostat	5366	0,20%	RESID Database of Protein Modifications	179	0,01%
UK Data Archive	4965	0,19%	caArray	173	0,01%
DrugBank	4743	0,18%	NASA Socioeconomic Data and Applications Center	169	0,01%
International Food Policy Research Institute	4351	0,17%	British Antarctic Survey	163	0,01%
Compendium of Protein Lysine Acetylation	3312	0,13%	Michigan Corpus of Academic Spoken English	151	0,01%
TreeBASE	3057	0,12%	QTL Archive	141	0,01%
GWAS Central	2763	0,11%	South African Data Archive	108	0,00%

4. Final Remarks

In this working paper we have presented some preliminary results based on the analysis of the Data Citation Index. We have shown discipline coverage, the data repositories and document types that can be found in this new database. The main conclusions and findings can be summarized as follows:

1) It is heavily oriented towards the hard sciences (Science accounts for 80% of the records in the database). Within this area, the best represented disciplines are Clinical Medicine, Genetics & Heredity, and Biochemistry & Molecular Biology.

2) The DCI uses three document types. There are 96 data repositories, and the predominant typology is the data set, with 2.475.534 records, which is 94% of the entire database.

3) Even though there are a total of 64 repositories that contain at least 100 records, there are four repositories that contain 75% of all the records in the database: *Gene Expression Omnibus*, *UniProt Knowledgebase*, *PANGAEA*, and *U.S. Census Bureau TIGER/Line Shapefiles*.

5. References

Arzberger, Peter; Schroeder, Peter; Beaulieu, Anne; Bowker, Geof; Casey, Kathleen; Laaksonen, Leif; Moorman, David; Uhlir, Paul; Wouters, Paul (2004), "An international framework to promote access to data". Science, v. 303, n. 5665, pp. 1777-1778. doi:10.1126/science.1095958

Costas, R., Meijer, I., Zahedi, Z. and Wouters, P. (2013), The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. Available from: www.knowledge-exchange.info/datametrics

Hrynaszkiewicz, Iain; Altman, Douglas G. (2009), "Towards an agreement on best practice for publishing raw clinical trial data". Trials, v. 10, n. 17. doi:10.1186/1745-6215-10-17

Moed, H.F. (2005), Citation Analysis in Research Evaluation, Springer, Dordrecht, Netherlands

Piwowar, Heather A.; Day, Roger S.; Fridsma, Douglas B. (2007), "Sharing detailed research data is associated with increased citation rate". Plos one, v. 2, n. 3, p. e308. doi:10.1371/journal.pone.0000308

Rennolls, Keith (1997), "Science demands data sharing". British medical journal, v. 315, n. 7106, pp. 486.

Thomson Reuters (2012), Repository Evaluation, Selection, and Coverage Policies for the *Data Citation Index*^{SM} within Thomson Reuters Web of Knowledge. Available from: http://wokinfo.com/media/pdf/DCI_selection_essay.pdf

Torres-Salinas, Daniel; Robinson-García, Nicolás; Cabezas-Clavijo, Álvaro (2012), "Compartir los datos de investigación: introducción al data sharing". El profesional de la información, marzo-abril, v. 21, n. 2, pp. 173-184. doi:10.3145/epi.2012.mar.08

Torres-Salinas, D.; Robinson-García, N.; Campanario, J.M. & Delgado López-Cózar, E. (2013), Coverage, field specialization and impact of scientific publishers indexed in the 'Book Citation Index'. Online Information Review [In Press]

Vickers, Andrew J. (2006), "Whose data set is it anyway? Sharing raw data from randomized trials". Trials, v. 7, p. 15. doi:10.1186/1745-6215-7-15

Wouters, P., & Schröder, P. (Eds.) (2003), Promise and practice in data sharing: the public domain of digital research data. The Netherlands: NIWI-KNAW.