Council for the Development of Social Science Research in Africa
Conseil pour le développement de la recherche en sciences sociales en Afrique
Conselho para o Desenvolvimento da Pesquisa em Ciências Sociais em África
مجلس تنمية البحوث الإجتماعية في أفريقيا
CODESRIA

Fourth CODESRIA Conference on Electronic Publishing

Quatrième conférence du CODESRIA sur l'édition électronique

The Open Access Movement and the Future of Africa's Knowledge Economy

Le mouvement pour le libre accès et l'avenir de l'économie africaine du savoir

Dakar, Senegal, March 30 - April 1, 2016

Dakar, Sénégal, 30 mars - 01 avril 2016

Open Research Data in Sub-Saharan Africa: A Bibliometric Study Using the Data Citation Index

Omwoyo Bosire Onyancha

CODESRIA
Avenue Cheikh Anta Diop X Canal IV
BP 3304, CP 18524, Dakar, Senegal
Phone: (+221) 33 825 98 22 / (+221) 33 825 98 23
Fax: (+221) 33 824 12 89
http://codesria.org/
https://www.facebook.com/CODESRIA-181817969495/
https://twitter.com/codesria

open.access@codesria.sn

Abstract
This paper underscores the need for sharing research data in the context of current world trends in open access scholarly publishing. It also explores the status of the publication of research data in sub-Saharan Africa (SSA) as a way of determining how research data is shared among researchers in the region and internationally. In order to explore the status of research data publication in SSA, relevant data was extracted from the Data Citation Index (DCI) hosted at Thomson Reuters' Web of Science (WoS). An advanced search, limited to the publication years between 2009 and 2014, was conducted using the names of the 50 countries that constitute SSA. Data was analysed using the DCI built-in "data analysis" and "citation analysis" features in order to obtain the number of data records by country, institution, subject category, year of publication, and document type as well as the number of citations. A Spearman's correlation analysis was also conducted to gauge the relationship between the data records and research articles. Preliminary findings indicate that only 20 countries produced at least one data record in the DCI, with South Africa leading the pack with 539 (61.39%) records followed by Kenya, Cameroon and Ghana. SSA contributes a mere 0.03% of the world's research data as compared to an average of 1.4% of the world's research articles. Research institutions and universities are the major contributors of research data, which largely focuses on Genetics and Heredity (61.3%), Biochemistry and Molecular Biology (61.3%), Agriculture (29.2%) and Forestry (27.3%). Citation-wise, the research data has attracted fewer average citations than the articles. A correlational analysis of the data reveals that although there is a significant correlation between the publication of data and research articles, the relationship is not strong. The implications of sharing research data for scholarly publishing in SSA are discussed in the paper and recommendations for further research are offered.

1. Introduction

Although open access in the context of scholarly publishing is a relatively new concept, the idea and practise of providing free online access to educational resources as well as journal articles is much older. The Budapest Open Access Initiative of February 2002, the Bethesda Statement on Open Access Publishing in June 2003, and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities in October 2003, as well as the Organization for Economic Co-Operation and Development (OECD) Principles and Guidelines for Access to Research Data from Public Funding in 2007, are credited with clarifying what the term "open access" entails. In its definition of open access, the Budapest Open Access Initiative states:

> By 'open access' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited (Budapest Open Access Initiative, 2002).

Since its inception, the concept of open access has attracted unprecedented attention from stakeholders, including the scholarly community. A search within Google Trends, for instance, shows that although the search volume for open access has dwindled over time, it still remains as high as 65% of all searches conducted in Google (as at September 2015). The number of articles published on open access, as indexed in the Elton B. Stephens Co (popularly known as EBSCO)-hosted Library and Information Science Abstracts (LISTA)

and Library and Information Science Source (LISS) grew from just less than 20 articles in 2003 to approximately 180 articles in 2011 (Chilimo, 2014). The Directory of Open Access Journals (DOAJ) consists of 10 499 journals that have published a total of 2 070 861 articles spanning 134 countries (Directory of Open Access Journals, 2015). The institutional repositories (IRs) have also increased tremendously, from just 105 in 2005 to 2 972 in 2015. The world total as at 11 September 2015 was 2 972 IRs spread across different continents as follows: Europe (44.1%), Asia (20.1%), North America (19.2%), South America (8.8%), Africa (4.3%), Australasia (2.2%), Central America (0.6%), and Caribbean (0.5%) (OpenDOAR, 2015). In line with the traditional means of disseminating research, most (if not all) of these IRs were initially intended for research publications such as journal articles, theses and dissertations, books, book chapters and sections of books, unpublished reports and working papers, conference and workshop papers, multimedia and audio-visual materials.

While the focus on open access has concentrated on the abovementioned outputs, the ground is gradually shifting to include data. Lortie (2014) has predicted the death of papers and the rise of in scholarly publishing. Similarly, in October 2014, **Martin Vetterli**, President of the National Research Council, Swiss National Science Foundation, observed and predicted thus:

> The open-data movement has already reached almost the whole of society. For example, digital content can be used freely (open content), computer programs perused and altered (open source), official data consulted (open government) and educational courses pursued free of charge (open education). Research, too, is affected. At present, the demand for free access to scientific literature is a major talking point. For scientists, the open-access movement is however only the beginning. The next big challenge will be free access to the data from work that has been published. This will bring in its wake complex questions regarding the storage and shared use of data, but it will also prove positive for the scientific community, since it will allow for a whole new culture of reproducibility of scientific experiments, which has become a matter of concern in recent years (Vetterli, 2014).

Martin Vetterli's prediction has indeed come to pass as the open data movement has reached all societies, including sub-Saharan Africa (SSA). In fact, some governments in SSA had already deposited their data in open repositories even before 2014, when Martin Vetterli made his prediction. Worldwide, there were 146 open data repositories registered in the Directory of Open Access Repositories (OpenDOAR) as at September 2015 (OpenDOAR, 2015), accounting for a mere 4% of the total number of repositories, as shown in figure 1.

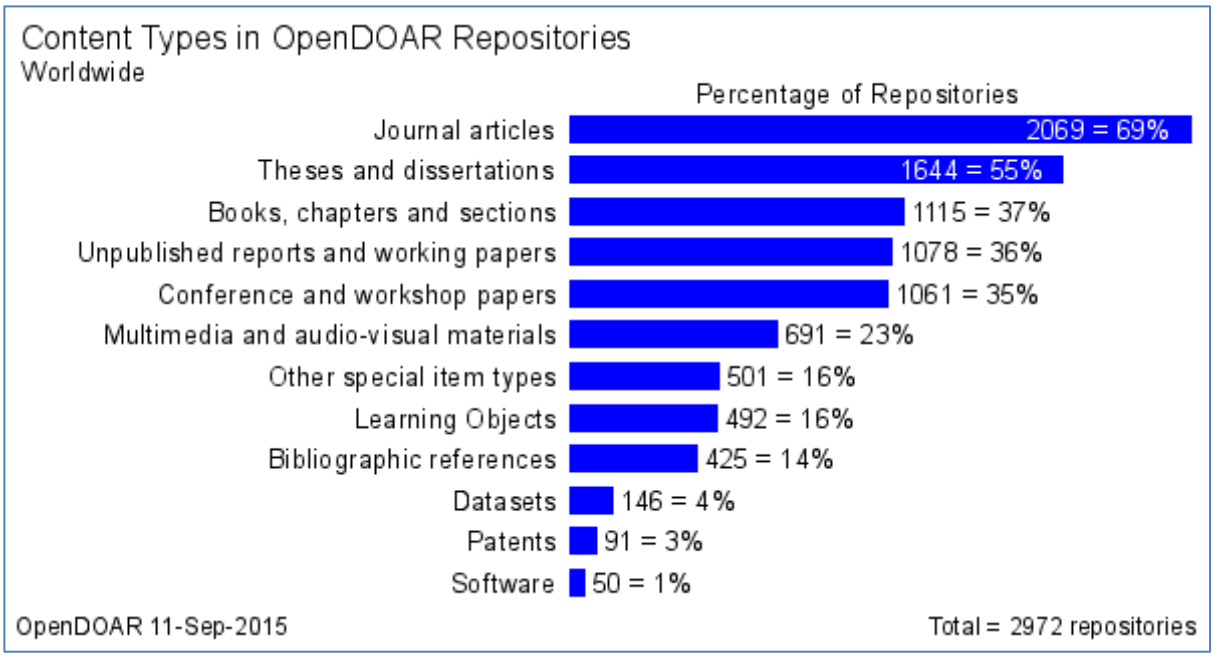

Figure 1: Distribution of IRs according to content type (Source: OpenDoar, 2015)

This does not however mean that there are only 146 open data repositories in the world, as it is possible that most open data repositories that are located in developing countries are not registered in the OpenDOAR. For instance, opendataforafrica.org provides links to 53 portals that can be used to access government open data of countries in Africa. Each of the 53 countries has its own link/portal. It follows therefore that only a fraction of data repositories are registered in the OpenDOAR.

A quick scan of the content of the portals to open data in Africa shows that the data spans several socio-economic indicators, e.g. gross domestic product (GDP), population, income and human development, food prices, energy statistics, country policies and development goals. The countries in Africa have shared various types of data in open platforms. The data covers different topics including GDP growth, population statistics, demographics, education and telecommunications. The specific subjects covered in the datasets include Africa, agriculture, business, commodities, mortality, poverty, education, electricity, environment, fertilisers, foreign trade, fragile states, gender, health, infrastructure, labour, living conditions, malaria, ratings and urbanisation, which yielded one dataset each. Similar patterns were witnessed with datasets of other countries in SSA.

2. Open research data: brief overview

Neuroth, Strathmann, Oswald & Ludwig (2013) opine that the term "research data" could refer to data from instruments such as a telescope or raw data from a mass spectrometer, and to digital maps or full-text documents such as those used in the creation of critical editions. The author however advises that research data must always be viewed in relation to a particular subject discipline. Research data can be either qualitative or quantitative (Krier & Strasser, 2014). Krier & Strasser (2014) further classify research data as follows:

a. observational data – data that has been gathered from observing a particular population or phenomenon
b. experimental data – derived from controlled, randomised experiments; and

c. computational data – output of a computer that has taken a large set of varied data and run it through a simulation

The OECD (OECD, 2007: 13) defines research data as follows:

> Research data are defined as actual records, numerical scores, textual records, images, and sounds used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.

A similar definition of research data has been offered by the US White House Office of Management and Budget (OMB, 2013: 78631) in its OMB Circular A-81. Both the OECD and the OMB have outlined what does not constitute research data. For instance, the OECD (2007: 14) advises that research data does not include "laboratory notebooks, preliminary analyses, and drafts of scientific papers, plans for future research, peer reviews, or personal communications with colleagues or physical objects (e.g. laboratory samples, strains of bacteria and test animals such as mice)". The OMB considers the same elements as non-research data.

What is open research data? Before we attempt to define it, we examine what the concept "open data" means. According to the Australian National Data Service (2013b), open data can be defined as data that is:

- freely available to download in a reusable form
- licensed with minimal restrictions to re-use
- well described with provenance and re-use information provided
- available in convenient, modifiable and open formats
- managed by the provider on an ongoing basis

On its part, the Open Data Handbook (2015) defines open data as data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirements to attribute and share alike. To expound on this definition, the Open Data Handbook (2015) summarises as follows:

- **Availability and accessibility:** the data must be available as a whole and at no cost than a reasonable reproduction cost, preferably by downloading over the Internet. The data must also be available in a convenient and modifiable form.
- **Reuse and redistribution:** the data must be provided under terms that permit re-use and redistribution including intermixing with other datasets.
- **Universal participation:** everyone must be able to use, re-use and redistribute – there should be no discrimination against fields of endeavour or against persons or groups.

It follows therefore that open research data is research data that combines the characteristics of open data and the types of research data spelt out in OMB (2013) and OECD (2007). The characteristics listed in the Australian National Data Service (2013b) and the Open Data Handbook (2015) are by no means exhaustive but are the ones most commonly associated with open data.

Corti, Van den Eynden, Bishop & Woollard (2014: viii) summarise the precursors of research data management by stating that "research funders are increasingly mandating open access

for research data; governments internationally are demanding transparency in research; the economic climate is requiring much greater re-use of data; and fear of data loss calls for more robust information security practices". The fear of data loss and/or manipulation has also been reported (see Agorist, 2015). Agorist reports that a Centre for Communicable Diseases scientist acknowledged that they (the scientist and co-researchers) destroyed data that showed vaccines caused autism in children. Agorist quotes the scientist's regret thus:

> I regret that my co-authors and I omitted statistically significant information in our 2004 article published in the journal *Pediatrics*. The omitted data suggested that African American males who received the MMR vaccine before the age of 36 months were at increased risk for autism. Decisions were made regarding which findings to report after the data were collected, and I believe that the final study protocol was not followed (Agorist, 2015).

The other compelling force behind the sharing of research is the journal publishers. A number of journal publishing companies now require authors to deposit data that informed their conclusions in the published papers/research articles (Australian National Data Service. 2013a). These journals and/or journal publishers include the Dryad consortium of journals; *Journal of the Royal Society; Interface*; *Nature*; *PloS*; *PNAS*; and Science and National Academies Press (Victoria University, Melbourne 2015).

Furthermore, there is a push from researchers to have **all** research products recognised by interested parties. For instance, among the general recommendations made at the San Francisco Declaration on Research Assessment in December 2012 was one that stated that, for funding purposes, funding agencies and other institutions should "consider the value and impact of all research outputs (including datasets and software) in addition to research publications, and consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice" (San Francisco Declaration on Research Assessment (DORA), 2012). Similarly, Piwowar (2013: 159) reports that the US National Science Foundation required all grant applicants, starting in January 2013, to list their research products (including datasets and software and other non-traditional research products such as journal articles) in the biographical sketch section.

In South Africa, the National Research Foundation (NRF) has mandated authors of research papers generated from research either fully or partially funded by the NRF, when submitting and publishing in academic journals, to a) deposit their final peer-reviewed manuscripts that have been accepted by the journals in the administering IR and b) to deposit the data supporting the publication in an accredited Open Access repository, with the provision of a Digital Object Identifier for future citation and referencing (NRF, 2015). The agitation for the recognition of research data as a valuable research product is increasingly becoming popular among the proponents of altmetrics (alternative metrics) too (Piwowar, 2013). Altmetricians contend that one of the benefits of altmetrics is the diversity of products that they measure (Bornmann, 2014: 898). Bornmann (2014) observes that altmetrics can be used to measure the impact of scholarly products other than papers, which were the traditional formats that were largely considered in the measurement of scholarly impact.

But why has research data attracted so much attention? In their explanation giving the reasons for research data curation, Neuroth et al (2013) note that by ensuring that research data are accessible, shareable and re-usable over time, certain activities are possible. These activities, which Neuroth et al (2013) call steps, include:

- Research data are documented and could therefore be **validated**.

- Research data could be the **basis for other and new research questions**, since it could be an integral part of the (digital) research lifecycle from the beginning.
- Research data could be re-analysed by using new, innovative digital methods that were unknown at the moment of data acquisition; and
- Research data could be used by other disciplines, therefore encouraging interdisciplinary research.

In support of the opinions expressed by Neuroth et al (2013), Corti et al (2014: 2) observe that "access to research data means that scientific findings can be verified and scrutinized if needed". The authors further state that society demands access to data

- to enable businesses to employ new knowledge for the development of tools and applications
- to allow organisations to question governmental policies and decisions; and
- for thousands of citizens to engage in research processes, or "citizen science", to advance our collective scientific knowledge

The value of research data has also been expressed in relation to improving scientific transparency and accuracy (Bishop, 2014). Bishop (2014) argues that when one is forced to share data, one is obligated to ensure that it is accurate and properly documented. However, she hastens to say that errors in the data are inevitable and unavoidable, a situation that provides strong arguments for data sharing. She concludes that open research data is important for science but difficult for scientists, particularly those researchers who could be sloppy or inaccurate in their researches. She sums up the concerns of scientists as follows: "there are concerns that open data sharing could lead to scientists getting scooped, will take up too much time, and could be used to impose ever more draconian regulation on beleaguered scientists". Wicherts (2013) concurs with Bishop (2014) by opining that data sharing not only helps facilitate the process of psychology research but is also a reflection of rigour.

3. Purpose of the study

The purpose of this study was to explore the status of research data sharing in SSA with a view to comparing the findings with scholarly publishing of research articles. The specific objectives were:

- Determine the trend of research data publication in SSA.
- Identify the type of data records indexed in the DCI as shared by SSA.
- Find out the countries behind data sharing in SSA.
- Identify the institutions behind data sharing in SSA.
- Examine the subject coverage of the data records.
- Find out the relationship between data records and research articles.

4. Research methodology

In order to explore the availability and the status of the publication or sharing of research data in SSA, relevant data was extracted from Thomson Reuters' DCI. The DCI, which was launched in 2012, is hosted at the Web of Science (WoS) of Thomson Reuters. By August 2014, the DCI covered a total of 4 million records in the form of datasets, data repositories, microcitation and data studies. According to Robinson (2014), the DCI:

- enables the discovery of data repositories, data studies and data sets in the context of traditional literature
- links data to research publications
- helps researchers to find data sets and studies and track the full impact of their research output
- provides expanded measurement of researcher and institutional research output and assessment and
- facilitates more accurate and comprehensive bibliometric analyses

The DCI covers data repositories from around the world, based on several factors, which include the persistence and stability of the repository, peer review, links to the research literature, the age of the material, inclusion of funding statements and the language of the data and metadata (Swoger, 2012). The selection criteria imply that the DCI is therefore not exhaustive in its coverage of data repositories. Nevertheless, the DCI provides data that can be used to gauge the performance of researchers, institutions and countries in terms of their data-sharing practices.

Data was extracted from the DCI using the names of the 50 countries in SSA. The search was conducted using the country search tag, i.e. CU=Country Name (for example, CU=Kenya). An Advanced Search using the tag CU=X retrieved all records originating from country X. The search was limited to records published between 2009 and 2014. The choice of the start date of 2009 was Index-driven as the DCI's coverage only goes as far back as 2009. The data collection exercise was conducted between 7 and 11 September 2015.

The analysis of data was conducted using the DCI's built-in analysis tool. The tool provides the searcher with options on how to analyse the results. One can analyse the results in terms of the following items: authors, countries/territories, document types, editors, group authors, institutions, languages, source titles, subject areas, WoS categories and years published. As this study sought to examine the trend of data sharing/publication in SSA, it focused only on the countries/territories of origin of the data; document types; institutions of author affiliations; subject areas; WoS subject categories; and years of publication. A correlation analysis between the data records and research publications was conducted to gauge the performance of researchers in the region in terms of data sharing/publication and research articles.

5. Results

This section presents and discusses the findings under the following sub-headings:

- Trend of research data publication in SSA
- Distribution of the data records by document type
- Distribution of data records by countries of author affiliation
- Distribution of data records by institutions of author affiliation
- Subject coverage of data records
- Correlation analysis of data records and research articles

5.1 Trends in research data publication in SSA

Table 1 shows the trend of publication of data records in the DCI from 2009 to 2014. The number of data records from SSA countries in DCI grew from just 178 in 2009 to 208 in 2010. Thereafter the number of data records fell by 98 to stand at 110 in 2011 but grew again slightly in 2012, to 196. The years that followed recorded the lowest number of records. This pattern seems to be in line with that in the rest of the world, whereby 2009 recorded a higher number of data records (i.e. 461 238) than 2010 and 2011, which recorded a total of 439 426 and 351 659 records respectively. The following years nevertheless witnessed an upward trend in the number of records, which peaked at 812 535 in 2013. The decline witnessed in 2014 in both cases (i.e. SSA and the world) could be attributed to indexing time lag. It often takes some time before a published article can be indexed in an indexing service such as bibliographic databases (Diodato, 1994: 157). The indexing time lag for research data could be longer, especially in cases where research data is not published together with the article referring to the data.

Table 1: Coverage of datasets and data publications in the Web of Science, 2009–2014

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**TOTAL**
Data							
SSA	178	208	110	196	96	58	846
World	461 238	439 426	351 659	666 231	812 535	370 625	3 101 714
% of world	0.04	0.05	0.03	0.03	0.01	0.02	0.03
Publications (articles)							
SSA	14 542	15 981	17 645	18 214	19 609	21 831	107 822
World	1 138 238	1 185 762	1 261 038	1 320 619	1 390 714	1 423 236	7 719 607
% of world	1.28	1.35	1.40	1.38	1.41	1.53	1.40

Table 1 further reveals that whereas SSA countries' contribution to the world's total number of research data records was 0.04% in 2009 and 0.05% in 2010, the countries performed worse in the next four years; their share of the world's total fell from 0.03% in 2012 to 0.01% in 2013 and 0.02% in 2014. On average, SSA contributes a mere 0.03% of the world's total number of data records. When we examined SSA performance in terms of the number of articles published in the same period, i.e. 2009 to 2014, we found that the countries in the region had performed much better than they did in terms of data records. The countries witnessed a continuous growth of the number of articles from just 14 542 in 2009 to 21 831 in 2014, a growth rate of approximately 50%. This pattern is also reflected in the countries' percentage share of the world total number of articles shown in table 1. SSA contributed 1.28% to the world's total number of articles in 2009, 1.35% in 2010, 1.40% in 2011, 1.38% in 2012, 1.41% in 2013 and peaked at 1.53 in 2014. On average, SSA yielded 1.4% of the world's total number of articles between 2009 and 2014.

5.2 Distribution of the data records by document type

As mentioned in the methodology section, the DCI indexes three different document types, namely datasets, data studies and repositories. Swoger (2012) defines a dataset as "a single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment" and a repository as "a database or collection comprising data studies, and data sets which stores and provides access to the raw data". A data study is defined as a "description of studies or experiments held in repositories with the associated data which have been used in the data study" (Swoger, 2012: 110).

An analysis of the data records produced in SSA and indexed in the DCI revealed that datasets and data studies are the only document types that have been covered. Table 2 shows that whereas SSA countries contributed 0.09% of the world's data studies indexed in the DCI, the region performed worse in terms of its contribution to the world's datasets. The DCI contained more datasets than it did in terms of data studies and repositories. This observation is in line with world trends wherein more datasets are shared than any other document type (Salinas, Martin-Martin & Fuente-Gutierrez, 2014).

Table 2: Document types in the DCI, 2009–2014

	Dataset	**Data study**	**Repository**
World	2 480 200	266 077	61
SSA	612	234	0
% of world	0.02	0.09	0.00

Salinas et al (2014) note that datasets were the most common document types in the DCI, accounting for 94% of all records indexed therein. This seems to have changed slightly since, as the current study found that the 2 480 200 datasets accounted for 90.3% of the total data records indexed in the DCI between 2009 and 2014. It may be, therefore, that data studies are gaining popularity among researchers. In the case of SSA, the datasets constituted 72.3% while data studies accounted for 27.7% of SSA countries' data records in the DCI. SSA did not contribute to the world's total of 61 repositories.

5.3 Distribution of data records by countries of author affiliation

The distribution of data records by the contributing countries in SSA is provided in table 3. At the top of the table is South Africa, which contributed a total of 539 data records, accounting for 63.7% of SSA's 846 total records indexed in the DCI. In the second position is Kenya with 121 (14.3%) records, followed by Cameroon with 94 (11.1%) and Ghana, which yielded 23 (2.7%) data records. Out of 20 countries that produced at least one data record in SSA, seven were from East and West Africa each while Southern African region consisted of five countries. Southern African countries yielded a combined total of 566 records while West Africa and East Africa posted a total of 163 and 148 records respectively. The analysis of the data further revealed that a number of foreign countries (outside Africa) participated in the research that generated the data indexed in the DCI. These countries include the US, which featured in 55 (6.5%) data records, and the UK which collaborated in 14 (1.7%) research studies. Italy, France and Sweden co-authored/published eight (0.9%), two (0.2%) and one (0.1%), respectively, together with SSA countries. Whereas the presence of the UK and France can be attributed to their colonial legacies in Africa (Narváez-Berthelemot, Russell, Arvanitis, Waast & Gaillard (2002)), the role of Italy, the US and Sweden might be related to funding some research conducted in the region, hence their collaboration with countries in SSA. These countries, among others, have been regarded as key international collaborators in research on various topics in SSA (see Onyancha & Ocholla, 2007; Sooryamoorthy, 2009). Sooryamoorthy (2009: 425), for instance, observes South African scientists largely collaborate with their counterparts from the US, the UK, Germany, Australia, Canada, France, the Netherlands, Italy, Belgium, Israel, Scotland, Switzerland, Japan, Sweden and Spain.

Table 3: Number of data records by country (N=846)

No.	Country	Data records	%	No.	Country	Data records	%
1	South Africa	539	63.7	11	Malawi	6	0.7
2	Kenya	121	14.3	12	Rwanda	5	0.6
3	Cameroon	94	11.1	13	Senegal	3	0.4
4	Ghana	23	2.7	14	Cape Verde	2	0.2
5	Mali	22	2.6	15	Lesotho	2	0.2
6	Benin	16	1.9	16	Sierra Leone	2	0.2
7	Angola	12	1.4	17	Sudan	2	0.2
8	Somalia	10	1.2	18	Zambia	2	0.2
9	Ethiopia	8	0.9	19	Burundi	1	0.1
10	Mozambique	7	0.8	20	Tanzania	1	0.1

5.4 Distribution of data records by institutions of author affiliation

The analysis of the data records according to authors' institutional affiliation indicates that a total of 52 institutions were involved in the research conducted in SSA between 2009 and 2014 as indexed in the DCI. Table 4 provides the top 20 institutions, which were led by South Africa's Council for Scientific and Industrial Research (CSIR), which yielded 162 (191.1%) data records in the DCI. Second-placed in table 4 was the World Agroforestry Centre (ICRAF) with 160 (18.9%) records, followed by the University of Cape Town with 97 (11.4%), CSIR Biosciences with 76 (9.0%) and the University of Pretoria with 75 (8.9%) data records.

Among the international institutions that featured in the top 20 institutions are: World Agroforestry Centre (ICRAF), Population Services International, NIH NIAID, University of Arizona, Michigan State University and the Medical Research Council in the USA. The visibility of international institutions implies collaborative ventures among researchers in SSA with their international counterparts. It was encouraging to note that universities, which are major participants in research in any country, are involved in sharing research data. Among the local universities (i.e. universities in SSA) that appear in the top 20 in table 4 are: University of Cape Town, University of Pretoria, University of Stellenbosch, University of the Free State, and the University of the Witwatersrand. It is worth noting that all the aforementioned universities are located in South Africa, which is the leading country in SSA in terms of research output (see Narváez-Berthelemot et al, 2002; Onyancha, 2008).

Table 4: Institutions behind the publication of the data records in SSA

No	Institutions	No of records	%
1	Council for Scientific and Industrial Research (CSIR)	162	19.1
2	World Agroforestry Centre (ICRAF)	160	18.9
3	University of Cape Town	97	11.4
4	CSIR Biosciences	76	9.0
5	University of Pretoria	75	8.9
6	University of Stellenbosch	53	6.3
7	Population Services International	22	2.6
8	University of the Free State	19	2.2

9	NIH NIAID	19	2.2
10	University of Arizona	18	2.1
11	Michigan State University	18	2.1
12	Center of Democratic Development	18	2.1
13	Medical Research Council of South Africa (MRC)	11	1.3
14	Medical Research Council	11	1.3
15	PSI Somaliland	10	1.2
16	University of the Witwatersrand	9	1.1
17	Africa Rice Center Africarice	9	1.1
18	Università degli Studi di Verona	8	1.0
19	PSI Mozambique	7	0.8
20	PSI Angola	7	0.8

5.5 Subject coverage of data records

Out of the 24 research areas that were covered in the data records indexed in the DCI, 519 (61.3%) were Genetics and Heredity and Biochemistry and Molecular Biology. Agriculture, Forestry and Business Economics took positions three, four and five with 247 (29.2%), 231 (27.3%) and 230 (27.2%) respectively. This pattern is a reflection of worldwide research wherein, in the same period of study (i.e. 2009 to 2014), Genetics and Heredity led the pack with 1 455 301 (46.9%) out of the total 3 101 698 records in the DCI. In the second position are Biochemistry and Molecular Biology, which posted 997 545 (32.1%) records between 2009 and 2014, followed by Crystallography with 555 315 (17.9%), Multidisciplinary Sciences with 502 229 (16.2%) and Geography with 318 346 (10.3%), just to name the research areas that posted more than 100 000 records each. Apparently, research data sharing is most common in the disciplines or research fields that, to some extent, dominate scholarly publishing, in terms of research outputs. Explaining the dominance of Clinical Medicine and Agriculture (including Biology) in scientific research in Africa, Narváez-Berthelemot et al (2002: 239) opine:

> The contributions made predominantly in the fields of Clinical Medicine and in Biology by African science according to our SCI data have possible explanations. Agricultural sciences are considered an important research area for many of the African countries, particularly for the Sub-Sahara countries (with the exception of South Africa). In the classification scheme used in the present study agricultural journals are included in the field of Biology thus giving this field added weight. It is tempting to speculate that the predominance of papers in Clinical Medicine is related to the AIDS epidemic devastating the African continent but we would need to carry out a detailed content analysis of published documents to sustain this hypothesis.

Table 5: Research areas sharing research data in SSA, 2009–2014

No	**Research Area**	**No. of records**	**%**
1	Genetics and heredity	519	61.3
2	Biochemistry/Molecular Biology	519	61.3
3	Agriculture	247	29.2
4	Forestry	231	27.3
5	Business Economics	230	27.2
6	Health Care Sciences/Services	54	6.4
7	Ethnic Studies	21	2.5

8	Communication	21	2.5
9	Women's' Studies	9	1.1
10	Sociology	5	0.6
11	Education/Educational Research	4	0.5
12	Water Resources	3	0.4
13	Social Issues	3	0.4
14	Urban Studies	2	0.2
15	Social Work	2	0.2
16	Nutrition/Dietetics	2	0.2
17	Government/Law	2	0.2
18	Family Studies	2	0.2
19	Meteorology/Atmospheric Sciences	1	0.1
20	Geography	1	0.1
21	Film/Radio/Television	1	0.1
22	Demography	1	0.1
23	Behavioural Sciences	1	0.1
24	Area Studies	1	0.1

A report commissioned by the World Bank and Elsevier and published in 2016 reveals that the most researched subject or discipline in SSA is Health Sciences (Blom, Lan & Adil, 2016). As Blom et al (2016: 17) observe, research on Health Sciences comprised 45.2% of SSA's total research outputs between 2003 and 2012. The dominance of Health Sciences and Agriculture in SSA's research outputs dates as far back as the 1990s (Narváez-Berthelemot et al, 2002). It is not therefore surprising to note that the most highly ranked subject categories in which data is largely shared are Health Sciences and Agriculture. Subjects or research areas in the social sciences as well as those in arts and humanities have continued to perform worse than natural and applied sciences.

5.6 Relationship between data and articles: a Spearman's correlation analysis

Table 6 provides the raw data on research data and articles produced in SSA as indexed in the DCI and the Science Citation Index (SCI), Social Science Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI), respectively, while table 7 provides the Spearman's correlation coefficients based on the data in table 6.

Table 6: Sub-Saharan African research data and articles as well as citation impact, 2009–2014

	Data		**Articles**			
	No. of records	Citations	No. of Articles	Citations	H-index	Average citations per article
South Africa	539	7	55 267	370 367	160	6.70
Kenya	121	40	7 443	73 748	85	9.91
Cameroon	94	0	3 699	21 686	47	5.86
Ghana	23	1	3 308	35 458	52	7.7
Mali	22	8	843	9 354	40	11.1
Benin	16	0	1 376	7 899	31	5.74
Angola	12	0	279	1 320	18	5.87

Somalia	10	0	28	148	6	5.29
Ethiopia	8	0	4 363	24 244	46	5.56
Mozambique	7	0	859	12 403	39	14.44
Malawi	6	2	1 805	19 473	49	10.79
Rwanda	5	0	648	5 152	32	7.95
Senegal	3	3	2 026	13 356	40	6.59
Cape Verde	2	0	88	575	12	6.53
Lesotho	2	0	150	840	15	5.6
Sierra Leone	2	0	198	1 610	17	8.13
Sudan	2	0	1 718	12 453	33	7.25
Zambia	2	12	1 314	13 037	46	9.92
Burundi	1	0	118	604	12	5.12
Tanzania	1	2	3 991	37 786	63	9.47

Table 6 shows that the number of citations associated or generated by research data (otherwise known as data citations) have remained low when compared with citations associated with articles (otherwise called article citations). For instance, whereas South Africa's data citations were seven, the country's articles citations totalled 55 267 between 2009 and 2014. On average, the number of data citations per data document was 0.09 while the average number of citations per article was 7.4. It may be speculated that data sharing in the region is still a new phenomenon and therefore even the researchers' citation behaviour is still not clear. Nevertheless, it has recently been revealed that data may attract more citations than articles do. For instance, in his study that compared the impact of a few openly accessible data sets and journal articles, Belter (2014) observes that "the production, archival, and sharing of data may actually be a more effective way to contribute to the advancement of scientific knowledge". In support of the aforementioned observation, Belter (2014) says:

> My results suggest that all three data sets are more highly cited than most journal articles. Each data set has probably been cited more often than 99% of the journal articles in oceanography that were published during the same years as the data sets. One data set in particular, the World Ocean Atlas and World Ocean Database, has been cited or referenced in over 8,500 journal articles since it was first released in 1982. To put that into perspective, this data set has a citation count over six times higher than any single journal article in oceanography from 1982 to the present.

Notable also in table 6 is that the countries would take different positions if they were ranked using each indicator of research performance.

With regard to the correlation between data and articles on the one hand, and their citations on the other, the Spearman's correlation test was conducted and the results in table 7 reveal that there is significant correlation between research articles and data, both in terms of output and citation impact. It was however noted that the data citations and article citations correlated significantly, thereby implying that there is a strong relationship between the two as far as the countries in table 6 are concerned. For instance, the analysis of data citations and article citations returned a correlation coefficient of 0.628 while the relationship between the data citations and the H-Index returned a correlation coefficient of 0.713.

Table 7: Spearman correlation between SSA's data and articles in DCI

	Articles

		Publications	Citations	H Index	Average Cites
Data Documents	Correlation Coefficient	.461*	.421	.452*	.050
	Sig. (2-tailed)	.041	.065	.046	.834
	N	20	20	20	20
Data Citations	Correlation Coefficient	.533*	.628**	.713**	.577**
	Sig. (2-tailed)	.015	.003	.000	.008
	N	20	20	20	20

6. Conclusions and recommendations

Data sharing in SSA can be said to be at its "initial formation stage" in which, as Crane (1972) (cited in Jacobs, 2004: 211) opines, the absolute number of publications is small and the growth rate shows signs of increasing. Indeed, the absolute number of data records shared by SSA is small but the trend has shown signs of growth, albeit slow, of the number of data records. Whereas SSA produced, on average, approximately 1.4% of the world's total number of articles per year between 2009 and 2014, the number of data records (i.e. datasets and data studies) has remained well below 0.05% throughout the study period. This trend is not peculiar to SSA as, Bryn Nelson had noted in 2009 (Nelson 2009) that the uptake of data repositories by researchers was minimal, a situation that had resulted in empty repositories. Nelson (2009) provides possible reasons why researchers, despite their acceptance that open access to data is the scientific ideal, choose not to share their data. Some of the reasons outlined in Nelson's article include: data loss; researchers' lack of skills to use the repositories; inadequate time on the part of researchers to deposit data in the repositories; researchers' fears that the data will be scooped, poached, or misused; lack of funding data sharing activities; researchers' lack of understanding on how much they have to relinquish when they share their data, and so on. Whyte (2014) stresses that funding for research data management is a great concern to many researchers. These issues, among others, may be impeding research data sharing the continent. It is encouraging however to note that a number of countries and institutions are involved in research data sharing. As is the case with the worldwide trend, datasets are the most common types of record shared by SSA, when compared with data studies and repositories. South Africa has continued to dominate research activities in the region, both in research output and data sharing. The country has a number of initiatives and incentives that are geared towards enhancing research productivity of her scientists (see Pouris, 2005; Pouris & Richter, 2000). Furthermore, the universities in South Africa, which are the major research-producing institutions in the country, are among the most prestigious universities in SSA, as exhibited in their performance in international ranking systems. It was not therefore surprising to note that institutions in South Africa topped the list of institutions behind the data that is indexed in the DCI. It was, however, sad to observe that only a fraction of the data generated from the research conducted in these institutions is shared.

Data sharing in SSA is largely concentrated in the natural and applied sciences as opposed to the arts and humanities and social sciences. Agriculture, which is the mainstream activity of a large population in SSA, takes centre stage as far as the sharing of the data in the region is concerned. Health Sciences and Forestry as well as Business Economics have shared their data equally. These research areas may be an indication of the problematic socio-political and

economic areas that the SSA governments might be most concerned about. Further research is however required to ascertain the reasons behind the high performance of these research areas, both in terms of research articles and data sharing.

There was no significant difference between data citations and article citations, implying that, despite some studies indicating that data may attract higher scientific impact, the difference in selected SSA countries is negligible. However it is noted that data sharing in the region is a recent practice and therefore the findings of this study might not reflect the worldwide pattern. We recommend that similar research is conducted in the future to examine the correlation between data citations and article citations of research conducted in the region.

What is the implication of open research data for SSA research and researchers?

We believe that open research data or research data sharing in open platforms will result in most (if not all) benefits associated with open scholarship. These include those listed in section 2. Sharing data in SSA may mean the following, among other things:

- Higher research output: a single data set may be used by more than one researcher/author, thereby generating more than one research article as different authors may offer different perspectives on a single data set.
- Improvement of scientific transparency and accuracy: authors will be compelled to handle the data carefully for fear of being embarrassed if its accuracy is compromised.
- Higher research/scientific impact: it has been shown that data attracts higher scientific impact than the articles that refer to the data.
- Accelerated socio-economic development in SSA: it has been shown that relevant research may lead to development of communities, on one hand, and the countries in which the research is conducted (Cohen, Manion & Morrison, 2013: 39; Guston, 2010: 354) on the other. Guston's (2010: 354) argument that "research leads to development of new industries or products, which leads to economic and/or social benefits" may apply to research data.
- There will be increased research collaboration among researchers with similar research interests.
- There will be increased accessibility and availability of research findings, which, when indexed in international key bibliographic databases, may pose serious accessibility and availability challenges for researchers in SSA.

For purposes of conducting further research, we believe that an investigation into the motivations for sharing research data and the challenges that researchers in SSA (and any other developing regions) face will shed more light on open research data in developing countries. Furthermore, a survey of the institutions that are sharing their research data through open platforms will help reveal more realistic statistics on data sharing in SSA. An investigation into governments' role in ensuring that data generated through publicly funded research is shared is another area proposed for further research in SSA.

Which way forward in research data sharing in SSA?

It is no secret that international databases and other online tools used to share knowledge are inclined to cover research publications emanating from developed nations and more specifically Europe and North America (see Luwel, 1999; Harzing, 2010; Nwagwu, 2010; Onyancha, Ngoepe & Maluleka, 2015). It is perhaps this bias in the coverage of research publications that has led some regions to develop autonomous databases similar to the Web

of Science. Examples of such databases include the Chinese Citation Index (in China), the Scientific Electronic Library Online (SCiELO) (originally in Brazil), and the Indian Citation Index (in India). Whereas the infrastructure for publications (for example, articles, conference papers, reviews, books, abstracts, letters, notes, etc) has seen put in place in some countries such as the aforementioned, data citation indexes are rare in the world. So far there is only one data citation index, namely the Thomson Reuters' Data Citation Index. As is the case with the WoS citation indexes and as has been mentioned in the methodology section, the DCI has strict selection criteria, which might lock out some data repositories, especially those located in the developing countries, thereby perpetuating Western and European hegemony as far as the visibility and impact of research is concerned. Despite African posting a sizable number of institutional repositories, the number of publications published by Africa remains less than 1.5% of the world output in the international databases. This study has further revealed that the situation is worse when it comes to the indexing of data records in the DCI. In view of these circumstances, there is need to not only explore the possibility of developing an African Citation Index (see Nwagwu 2010) for publications, but also to consider a system that can integrate data into the African Citation Index or one that can separately index data generated in Africa.

References

Agorist, M. 2015. CDC scientist admits they destroyed data that showed vaccines caused autism in children. http://thefreethoughtproject.com/cdc-scientist-admits-destroyed-data-showed-vaccines-caused-autism-children/#t5WXp5MdO39yU67h.99. (Accessed 06 January 2015).

Australian National Data Service. 2013a. Data journals. http://ands.org.au/guides/data-journals-awareness.pdf (Accessed 1 June 2015).

Australian National Data Service. 2013b. Open data. http://ands.org.au/discovery/opendata.html (Accessed 11 September 2015).

Belter, C. 2014. Global-level data sets may be more highly cited than most journal articles. http://blogs.lse.ac.uk/impactofsocialsciences/2014/05/15/global-level-data-sets-highly-cited/ (Accessed 9 June 2015).

Bishop, D. 2014. Data sharing may lead to some embarrassment but will ultimately improve scientific transparency and accuracy. http://blogs.lse.ac.uk/impactofsocialsciences/2014/05/29/data-sharing-exciting-but-scary/ (Accessed 9 June 2015).

Blom, A, Lan, G & Adil, M. 2016. *Sub-Saharan African science, technology, engineering, and mathematics research: A decade of development*. Washington, DC: International Bank for Reconstruction and Development/The World Bank.

Bornmann, L. 2014. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Infometrics*, 8 (4): 895-903

Budapest Open Access Initiative. 2002. Read the Budapest Open Access Initiative. http://www.budapestopenaccessinitiative.org/read (Accessed 10 September 2015).

Chilimo, W. 2014. Bibliometric analysis of open access research publications. Paper presented at the 15th Department of Information Studies Conference, University of Zululand, South Africa, 3rd – 5th September 2014. http://www.lis.uzulu.ac.za/conference/docs/pp/day1/session%203/13_Mapping%20Open%20Access%20Research%20Trends%20A%20Bibliometric%20Analysis_W_Chilimo%20.pdf (Accessed 12 January 2016)

Cohen, L, Manion, L & Morrison, K. 2013. *Research methods in education*. New York, NY: Routledge.

Corti, L, Van den Eynden, V, Bishop, L & Woollard, M. 2014. *Managing and sharing research data: A guide to good practice*. Los Angeles, CA: Sage.
Crane, D. 1972. *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago, IL: University of Chicago Press.
Diodato, VP. 1994. *Dictionary of Bibliometrics*. New York, NY: Haworth.
Directory of Open Access Journals. 2015. https://doaj.org/ (Accessed 11 September 2015)
Guston, DH. 2010. Encyclopedia of nanoscience and society. Los Angeles, CA: Sage.
Harzing, A.W. 2010. *Citation analysis across disciplines: the impact of different data sources and citation metrics*. http://www.harzing.com/data_metrics_comparison.htm (Accessed 18 June 2015).
Jacobs, D. 2004. Growth and development of knowledge management research: A bibliometric study. In Bothma, TJD & Kaniki, A. 2004. ProLISSA 2004. *Proceedings of the 3rd biennial DISSAnet Conference*, Pretoria, 28–29 October 2004. Pretoria: Infuse: 211–220.
Krier, L & Strasser, CA. 2014. *Data management for libraries: A LITA guide*. Chicago. IL: American Library Association.
Lortie, C. 2014. The citation revolution will not be televised: The end of papers and the rise of data. http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/05/citation-revolution-end-of-papers-rise-of-data/ (Accessed 9 June 2015).
Luwel, M. 1999. Is the Science Citation Index US-biased? *Scientometrics*, 46(3): 549-562
Narváez-Berthelemot, N, Russell, JM, Arvanitis, R, Waast, R & Gaillard, J. 2002. Science in Africa: An overview of mainstream scientific output. *Scientometrics*, 54(2): 229–241.
National Research Foundation (NRF). 2015. Statement on open access to research publications from National Research Foundation (NRF)-funded research. http://www.nrf.ac.za/media-room/news/statement-open-access-research-publications-national-research-foundation-nrf-funded (Accessed 16 October 2015)
Nelson, B. 2009. Data sharing: Empty archives. *Nature*, 461: 160–163.
Neuroth, H., Strathmann, S., Oswald, A. & Ludwig J. (eds). *Digital curation of research data: experiences of a baseline study in Germany*. Gottingen: Verlag Werner Hulsbusch.
Nwagwu, W.E. 2010. Cyberneting the academe: centralized scholarly ranking and visibility of scholars in the developing world. *Journal of Information Science*, 36(2): 228-241.
Office of Management and Budget (OMB). 2013. Circular A-110: Uniform administrative requirements for grants and agreements with institutions of higher education, hospitals, and other non-profit organizations," 2 C.F.R. 215.
Onyancha, OB. 2008. Authorship patterns of the literature on HIV/AIDS in Eastern and Southern Africa: An exposition of the responsible authors, institutions and countries, 1980–2005. *South African Journal of Libraries and Information Science*, 74(1): 9–22.
Onyancha, OB., Ngoepe, M & Maluleka, JR. Trends, patterns, challenges and types of archival research in sub-Saharan Africa. *African Journal of Archives, Libraries and Information Science*, 25(2): 145-159.
Onyancha, OB & Ocholla, DN. 2007. Country-wise collaborations in HIV/AIDS research in Kenya and South Africa, 1980–2005. *LIBRI*, 57(4): 239–254.
Open Data Handbook. 2015. http://opendatahandbook.org/ (Accessed 16 December 2015).
OpenDOAR. 2015. Proportion of repositories by continent – worldwide. http://www.opendoar.org/onechart.php?cID=&ctID=&rtID=&clID=&lID=&potID=&rSoftWareName=&search=&groupby=c.cContinent&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Proportion%20of%20Repositories%20by%20Continent%20-%20Worldwide (Accessed 11 September 2015).

Organisation for Economic Co-Operation Development. 2007. OECD principles and guidelines for access to research data from public funding. Massachusetts: OECD. http://www.oecd.org/sti/sci-tech/38500813.pdf (Accessed 16 July 2015).

Piwowar, H. 2013. Value all research products. *Nature*, 493: 159.

Pouris, A. 2005. An assessment of the impact and visibility of South African journals. *Scientometrics*, 62(2): 213–222.

Pouris, A & Richter, L. 2000. Investigation into state-funded research journals in South Africa. *South African Journal of Science*, 96: 98–104.

Robinson, N. 2014. The data citation index and datacite. http://datacite.inist.fr/IMG/pptx/datacite_robinson.pptx (Accessed 10 June 2015).

San Francisco Declaration on Research Assessment (DORA). 2012. Contestation of Impact Factor as a measure of journal quality. http://am.ascb.org/dora/ (Accessed 23 July 2015).

Sooryamoorthy, R. 2009. Collaboration and publication: How collaborative are scientists in South Africa? *Scientometrics*, 80(2): 419–439.

Suber, P. 2009. Timeline of the opne access movement. http://legacy.earlham.edu/~peters/fos/timeline.htm (Accessed 25 September 2015).

Swoger, B. 2012. Thomson Reuters Data Citation Index. *Library Journal*, 137(20): 110

Thomson Reuters. 2012. Repository evaluation, selection, and coverage policies for the Data Citation Index within Thomson Reuters Web of Knowledge http://wokinfo.com/media/pdf/DCI_selection_essay.pdf (Accessed 22 December 2015).

Torres-Salinas, D, Martin-Martin, A & Fuente-Gutierrez, E. 2014. Analysis of the coverage of the Data Citation Index – Thomson Reuters: Disciplines, document types, and repositories. *Revista Española de Documentación Científica*, 37(1): 1–6. http://arxiv.org/pdf/1306.6584.pdf (Accessed 22 December 2015).

Vetterli, M. 2014. Open access, open data, open science. http://library.epfl.ch/page-114954-en.html (Accessed 12 December 2015)

Victoria University, Melbourne. 2015. Research data management: data deposit requirements of selected science journals. http://guides.library.vu.edu.au/content.php?pid=489543&sid=4015042 (Accessed 7 December 2015).

Whyte, A. 2014. Opportunities for 'data intensive' social research are growing but funding for data management remains a challenge. http://blogs.lse.ac.uk/impactofsocialsciences/2014/03/25/research-data-management-strategy-funding-whyte/ (Accessed 9 June 2015).

Wicherts, J. 2013. Data sharing not only helps facilitate the process of psychology research, it is also a reflection of rigour. http://blogs.lse.ac.uk/impactofsocialsciences/2013/10/24/journal-of-open-psychology-data/ (Accessed 9 June 2015).