# The Value of Research Data – Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report (April 2013)

**Research** · September 2016

CITATIONS
22

READS
243

**4 authors:**

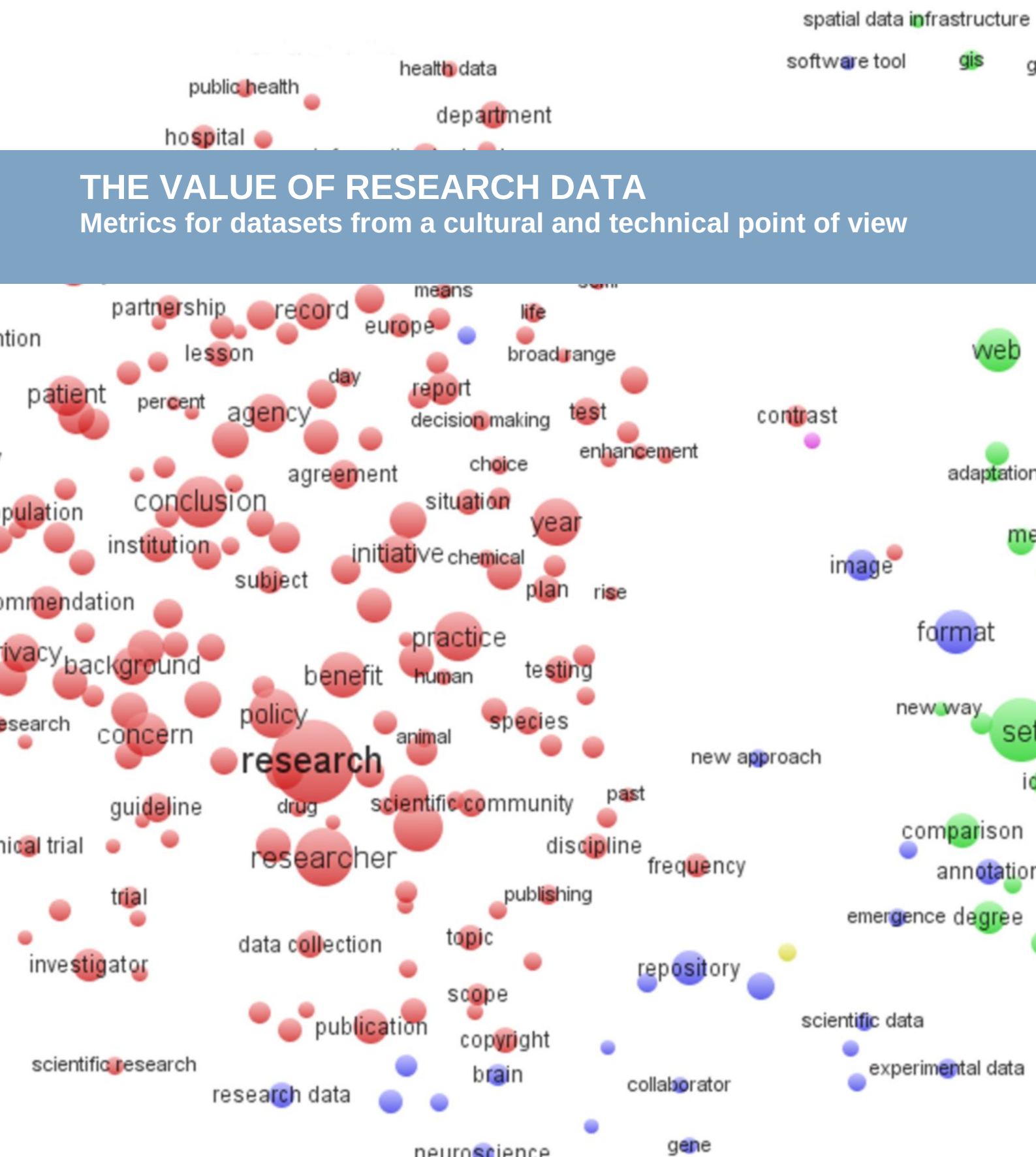
Ingeborg Meijer
Leiden University
**33** PUBLICATIONS **418** CITATIONS
SEE PROFILE

Rodrigo Costas
Leiden University
**115** PUBLICATIONS **2,369** CITATIONS
SEE PROFILE

Zohreh Zahedi
Leiden University
**29** PUBLICATIONS **727** CITATIONS
SEE PROFILE

Paul Wouters
Leiden University
**100** PUBLICATIONS **2,968** CITATIONS
SEE PROFILE

Knowledge Exchange




# THE VALUE OF RESEARCH DATA

## Metrics for datasets from a cultural and technical point of view

*Knowledge Exchange*
*Danish Agency for Culture*
*H.C. Andersens Boulevard 2*
*DK-1553 Copenhagen V*
*Denmark*

*Tel. +45 33 73 33 73*
*Fax +45 33 73 33 72*
*office@knowledge-exchange.info*
*www.knowledge-exchange.info*

**April 2013**

Authors:
Rodrigo Costas, Ingeborg Meijer, Zohreh Zahedi and Paul Wouters
Center for Science and Technology Studies (CWTS). Leiden University.
PO Box 905
2300 AX Leiden, the Netherlands
tel. +31 71 527 3909
e-mail: info@cwts.leidenuniv.nl



Please cite this document as: 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

**Knowledge Exchange is a co-operative effort that supports the use and development of Information and Communications Technologies (ICT) infrastructure for higher education and research. The Knowledge Exchange partners are:**

CSC – IT center for Science in Finland
Denmark's Electronic Research Library (DEFF) in Denmark
German Research Foundation (DFG) in Germany
Joint Information Systems Committee (JISC) in the United Kingdom
SURF in the Netherlands

# Executive summary

Scientific research revolves around the production, analysis, storage, management, and re-use of data. Data sharing offers important benefits for scientific progress and advancement of knowledge. However, several limitations and barriers in the general adoption of data sharing are still in place. Probably the most important challenge is that data sharing is not yet very common among scholars and is not yet seen as a regular activity among scientists, although important efforts are being invested in promoting data sharing. In addition, there is a relatively low commitment of scholars to cite data. The most important problems and challenges regarding data metrics are closely tied to the more general problems related to data sharing. The development of data metrics is dependent on the growth of data sharing practices, after all it is nothing more than the registration of researchers' behaviour. At the same time, the availability of proper metrics can help researchers to make their data work more visible. This may subsequently act as an incentive for more data sharing and in this way a virtuous circle may be set in motion.

This report seeks to further explore the possibilities of metrics for datasets (i.e. the creation of reliable data metrics) and an effective reward system that aligns the main interests of the main stakeholders involved in the process. The report reviews the current literature on data sharing and data metrics. It presents interviews with the main stakeholders on data sharing and data metrics. It also analyses the existing repositories and tools in the field of data sharing that have special relevance for the promotion and development of data metrics. On the basis of these three pillars, the report presents a number of solutions and necessary developments, as well as a set of recommendations regarding data metrics. The most important recommendations include the general adoption of data sharing and data publication among scholars; the development of a reward system for scientists that includes data metrics; reducing the costs of data publication; reducing existing negative cultural perceptions of researchers regarding data publication; developing standards for preservation, publication, identification and citation of datasets; more coordination of data repository initiatives; and further development of interoperability protocols across different actors.

# Table of contents

# 1. Introduction

Scientific fields differ in the nature of their data and in their methods and conventions about data use and re-use, but data are always at the core of empirically oriented science and scholarship (Halevi & Moed, 2012). These datasets can be very diverse: archaeological or biological material with personalized data attached to it, libraries of art, music or other digitized collections, digital observation tools in astronomy, extensive mathematical & statistical calculations or economic data pertaining to companies or countries as material. Whereas in the non-digital environment, research data could be classified as partly human effort and partly knowledge, in the digital environment automated collection and processing has made datasets more independent of specific researchers. Once collected, the same set of data can be used by a variety of researchers from different institutes, disciplines and nations for an unlimited period of time to produce novel science. Open accessibility and data sharing are now considered crucial for good science and scholarship. The increasing role of huge datasets in scientific research has important implications for the way research is conducted, for the way it should be organized and funded, and also for the way it should be evaluated and valued (Wouters & Schröder, 2003).

During the last few years, there has been an intense debate within the scientific community about the need of openly sharing the data that are a result of research (Torres-Salinas, Robinson-García, & Cabezas-Clavijo, 2012), particularly when this research is funded with public funds. The emerging awareness about data sharing throughout the scientific community (Schäfer et al., 2011) is reflected in the profusion of reports and scientific publications discussing the problems and challenges of data sharing (Van der Graaf & Waaijers, 2012). The promotion and general adoption of data sharing activities within the different scientific communities is regarded as an important strategic development in the pursuit of more Open Science, and good scientific practice in general.

Data sharing has been a relevant topic since the 1980's. In 1985 (Fienberg, Martin, & Straf, 1985) already pointed out some of the benefits, problems, controversies, and other challenges of sharing research data in an extensive report about the importance of data sharing for scientific development. Some of the conclusions of this report are still relevant, such as the need for developing guidelines on data sharing, the idea that multiple institutions (and stakeholders) should be involved in the process (e.g. scientific and professional associations, journals, foundations and research funds), and the need for government policies and standards for accessing, classifying, documenting and archiving data. More recently, other comprehensive documents dealing with the problems related to data sharing and data citation have been published. One of these reports is the summary of the *International Workshop on developing data attribution and data citation practices and standards* (Uhlir, 2012) in which a broad panel of experts, stakeholders and scientists discussed and debated the most important challenges and possibilities of data sharing, data publication and data citation. Another interesting collection of reports is the result of the Opportunities for Data Exchange (ODE) project (Dallmeier-Tiessen et al., 2012; Schäfer et al., 2011). The ODE project is a FP7 Project carried out by members of the Alliance for Permanent Access (APA) and its main aim has been to engage in dialogue with relevant stakeholders, in order to collect and document views and opinions on challenges and opportunities for data exchange. More recently also the idea of "Big Data"[1] has been highlighted as an emerging topic in the scientific landscape (Halevi & Moed, 2012). Big data has also been the crucial concept in the studies on e-science and e-research (GRDI2020, 2012; Hey, Tansley, & Tolle, 2009) The need to specify the relationship between data and the characteristics of the discipline involved, a topic which tends to be underemphasized in the literature on e-science, has also been raised (Arzberger et al., 2004; Borgman, 2007; Edwards, 2010; Wouters, Beaulieu, Scharnhorst, & Wyatt, 2013).

An important element that has been pointed out as a potential incentive for data sharing is the development of metrics for data (Piwowar, Becich, Bilofsky, & Crowley, 2008). These metrics could be incorporated in the framework of an appropriate professional and career reward structure and would take into account data sharing and data publication as important activities in the regular work of scholars (Arzberger et al., 2004). However, little has been studied and written about metrics for datasets, and actually "data metrics" or "dataset-level metrics" are quite new expressions that have not

[1] Understood as a wide range of datasets almost impossible to manage and processing using traditional data management tools due to their size or their complexity.

yet been broadly used in scientific publications (with Heather Piwowar probably as one of its main pioneers - see Piwowar, 2012). Thus, data metrics is still an underdeveloped concept.

*The Value of Research Data - Metrics for datasets from a cultural and technical point of view*

This report seeks to further explore the possibilities of metrics for datasets (i.e. the creation of reliable data metrics[2]) and an effective reward system that aligns the main interests of the main stakeholders involved in the process. Thus, this report presents a first landscape study on the possibilities of developing data metrics. These data metrics would be expected to play a role in research assessment and thus contribute to stimulate data sharing. As a result, this report will be of interest for the major stakeholders in science (i.e. governments, funders, data centres, universities, etc.). By providing them with more knowledge about tools to promote (and reward) data sharing and data publication within their scientific communities, they are able to choose among them for different purposes. Therefore, in this landscape study, all major stakeholders have been consulted about their main views, problems and challenges that need to be tackled in the development of metrics for datasets, and in the adoption and promotion of data sharing activities.

The report is organized as follows: chapter 2 presents the main objectives and the conceptual framework and methodological approach, discussing the main concepts and stakeholders in the area of data sharing and data metrics. Chapter 3 presents the results: the state of the art of data sharing, models for data metrics, and stakeholder perceptions. Chapter 4 presents the analysis of the existing repositories and tools in the field of data sharing that have special relevance for the promotion and development of data metrics. Chapter 5 summarizes the main problems and challenges existing in the development of data metrics, while chapter 6 outlines existing and possible solutions for these problems and challenges. Finally, chapter 7 introduces some recommendations and points at potential developments that can be regarded as necessary and strategic for developing robust and valid metrics for datasets.

[2] Other possibilities (e.g. “data altmetrics”) could be developed in the future as well (Piwowar, 2012).

# 2. Objectives and conceptual framework

The main objective of this research is to inform the scientific community, including information infrastructure providers, funding agencies and policy makers, about the state of the art in the area of data science metrics. This study aims to present an overview of the existing solutions, a critical assessment of possibilities for their use and suggestions for further actions. Central objectives in our study are the analysis of best practices that can encourage data metrics, under the assumption that data metrics could be an important asset in order to stimulate researchers to share research data.

## 2.1 Main concepts related with data sharing, data publication, data citation and data metrics

In this report the following concepts are used:

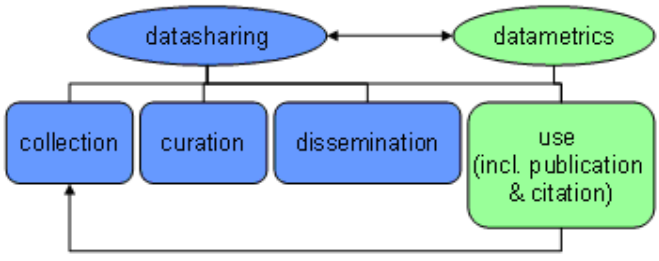
"**Data sharing"** has been defined as the "voluntary provision of information from one individual or institution to another for purposes of legitimate research" (Fienberg et al., 1985) or simply "the release of research data for use by others" (Borgman, 2012). This general concept is grounded in the assumption that data are a valuable long-term resource and that sharing them and making them publicly-available is essential if their potential value is to be realized (Swan & Brown, 2008). Data sharing requires the systematic *collection, curation* and *dissemination* of data.

**"Data citations"** have been defined as formal citations included in the reference list of published articles to data resources that led to a given research result (Mayernik, 2012). In this sense, the concept of data citation is tied to the idea that datasets should be published just as other kinds of scholarly products, being considered also as first class research outputs, both from social and funding policy perspectives (Lawrence, Jones, & Matthews, 2011).

**"Data publication"**: The idea of publication of datasets mirrors the scientific publication model, although some criticisms have been also raised (Mayernik, 2012) as this model does not fully fit all the idiosyncrasies related with the sharing and publication of datasets.

**"Data metrics"**: Data metrics are mainly related with data publication and data citation (but not exclusively, for example we could also potentially include 'altmetrics' on datasets here). Both data publication and data citation can be considered as signals of *use* of data. Use of data can generate new data, which may feed back into the collection phase (see Figure 1). Thus, for data metrics to build up, data sharing is a necessary prerequisite. Whether it will work the other way round (metrics leading to sharing) remains to be seen. In the rest of this report data sharing (i.e. collection, curation, dissemination) and data metrics (metrics on production and use) will be dealt with separately.

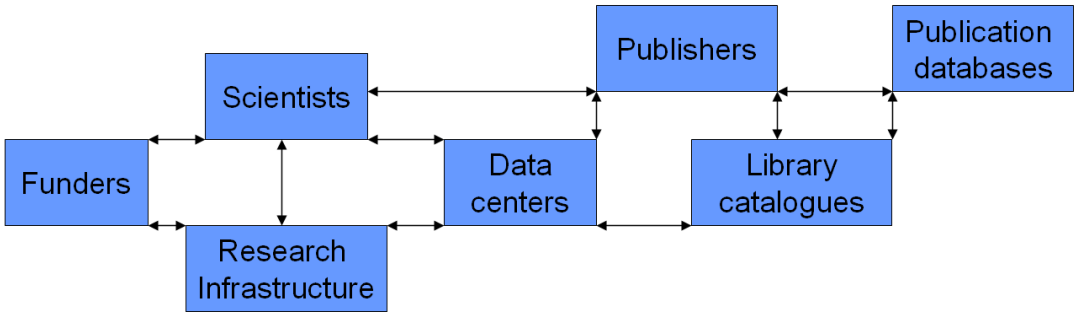
**Figure 1: Schematic view of the data sharing – data metrics model**




## 2.2 Main stakeholders in the development of data sharing and data metrics

A wide range of stakeholders have an important role in data sharing (Mayernik, 2012) and as such in the development of data metrics. In Figure 2 a schematic overview of the main stakeholder groups and their mutual relations and dependencies is presented.

**Figure 2: Schematic view of the main stakeholder groups**




In Table 1 we describe the main stakeholders, and their main interest regarding data sharing and data metrics.

**Table 1: Main stakeholders**

| Stakeholder | Main interest (based on interviews – see below) |
|---|---|
| Funders | - To fund data collection and make the most out of public research funding<br>- To encourage data use<br>- To promote recognition and rewards for data sharing and data use |
| Research Infrastructures | - From funding perspective: Long term added funding for curation and access (sharing);<br>- From providers perspective: To facilitate data use |
| Scientists | - Digital scholarship added to reward and esteem for tenure track<br>- To cite data in publications, and make data available |
| Data Centres | - To standardize storage and create metadata in a globally harmonized way (sharing).<br>- To track re-use and promote good scientific practice<br>- To provide data curation and recommend citations to datasets |
| Publishers | - To deal with the data that is provided with submitted publications.<br>- To enforce and facilitate data citation and data metrics policies and standards |
| Libraries | - To make data identifiable and accessible<br>- Coordination of scholars and data centres |
| Publication databases | - To link publications and data citations<br>- To enable data publication and citations counts and indicators |

## 2.3. Approach and methodology

We have applied different methodological approaches that are described below:

*1) Literature review.* This consists of a thorough study on the state of the art of the study of data sharing and data metrics (data publication and data citation). Desk research and a literature review have been performed to detect the main documents (more than 40 publications and reports have been examined) that discuss the most important issues regarding data sharing and data metrics. This literature review is supplemented with a term map, using the visualisation software VOSviewer (http://www.vosviewer.com) showing the overall main topics of research within the field of “data sharing”.

*2) Interviews with relevant stakeholders for data metrics.* This qualitative analysis has been performed through 9 telephone interviews with selected stakeholders in the area. The number of interviews was limited because of time constraints, but we obtained relevant information both on the technical and cultural aspects regarding data sharing and data metrics from a selected number of experts in the following three domains: repository caretakers, researchers promoting data sharing/citing and stakeholders from other audiences (e.g. research funders). The interviews were held in January/February 2013. The interviews were open-ended and provided a comprehensive view of the availability of solutions and tools, as well as on the critical possibilities they offer to the different users of data metrics. The results from the interviews are presented in an aggregated, anonymous way, so that no individual statements can be traced to any of the interviewees. The interviewees are listed in Appendix 6. The findings and perceptions are presented in chapter 3.3 and complement the findings of the literature review (3.1 and 3.2).

*3) Technical analysis of existing data repositories.* This consists of an extensive analysis of existing data repositories, paying special attention to elements that may play an important role in the development of data metrics. As an input list we have taken the list of more than 500 data repositories available in DataCite (http://datacite.org/repolist) which is itself based on DataBib (a searchable directory of research data repositories – http://databib.org). We have selected a sample of 35 repositories. A checklist has been developed and the sampled repositories were analysed according to the features in this checklist, focusing on issues such as data access, usage and validation of the repositories (see the full checklist in Appendix 1).

By using these three different methods, their triangulation strengthens the findings on data sharing, data metrics, and the role of stakeholders (see Table 2 below).

**Table 2: Overview of the three methodologies applied in this study**

| Focus/Method | Literature review | Interviews | Repositories/tools |
|---|---|---|---|
| Data sharing | X | X | X |
| Data metrics | | X | (X) |
| Stakeholders | X | X | |

# 3. Results

In this chapter we present the state of the art of the research regarding data sharing (chapter 3.1), models and the conceptualisation of data metrics (chapter 3.2), and reflections of stakeholders and global initiatives on data sharing and data metrics (chapter 3.3).

## 3.1. State of the art of the research on "data sharing"

*Term map analysis*

In this section, we present a summary of the main research related with data sharing and the potential development of data metrics. In the first place we present a term map of the literature covered in the Web of Science on data sharing. In Appendix 5 we present the most important methodological issues regarding this term map. We think that this first map can help to easily explore the "state of the art" of the current research in the field of data sharing and also to help to detect potential gaps that have not been tackled in the literature.

We have searched for the string "data sharing*" in the title, abstract or keywords of the publication in the Web of Science. A total of 1,460 documents have been identified and analysed with the standard options of the VOSviewer software (http://www.vosviewer.com). Figure 3 presents the results of this term map.

The map in Figure 3 presents three main clusters of terms that can be understood as follows. On the left side of the map (red colour) we find terms related to research, policies, science and scientific work in general. On the right side (green colour) there is a cluster related with technical and technological terms, including terms such as "application", "system", "user" or "paper", but also "computer", "architecture", "algorithm", "data integration", etc. Apparently, research on data sharing (or involving data sharing) comprises these two dimensions (the scientific/political dimension and the technological dimension).

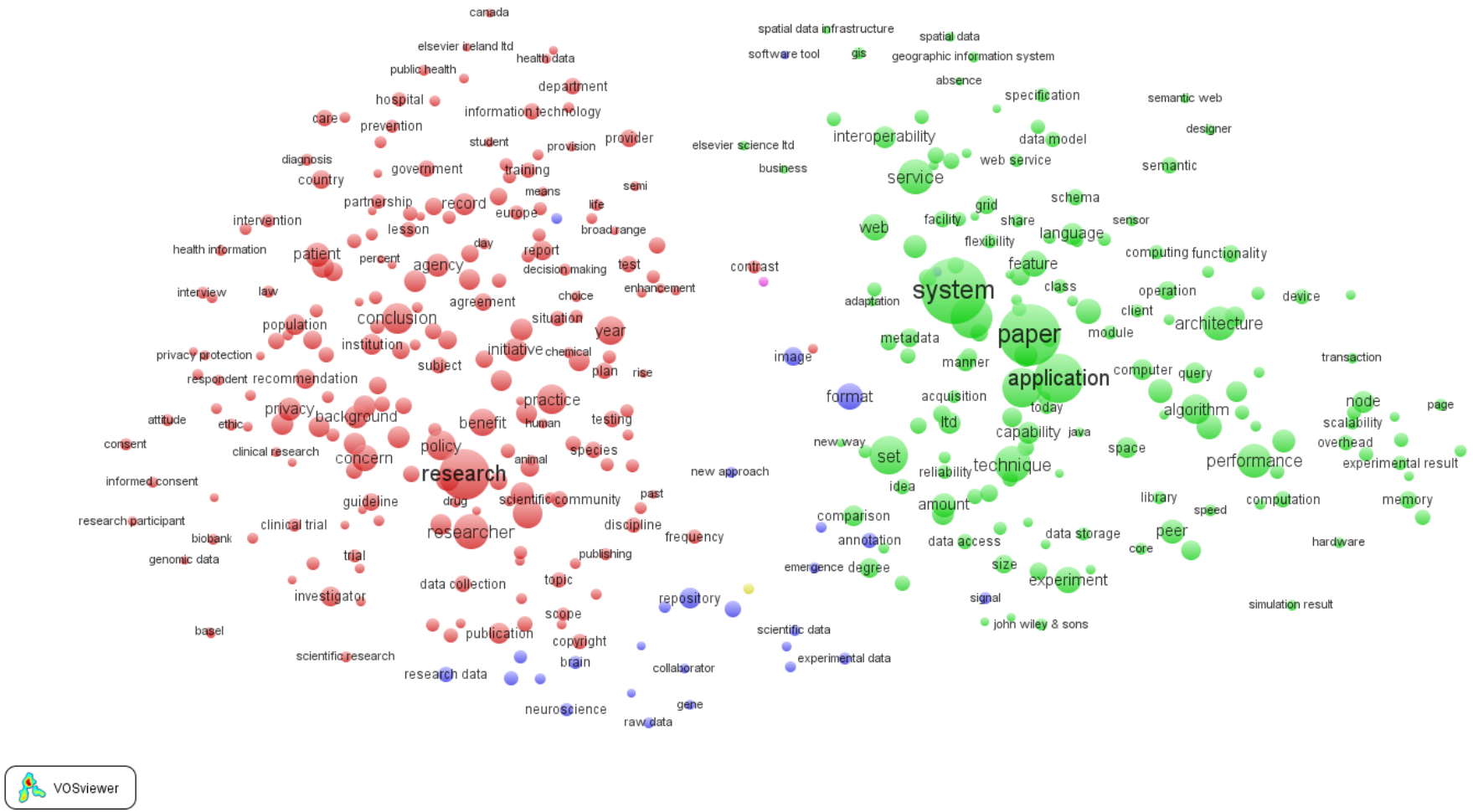
A remarkable element in the map is that these two main clusters seem to be joined by a third smaller cluster (bottom part of the map, in blue). This shows a focus on terms related to data themselves (i.e. "raw data", "scientific data", "experimental data" or "data mining"), "repositories", "software", "formats" and interestingly also with "motivation", "new technologies" and "new approaches" for data sharing. This third cluster can be understood as the practical conjunction of the other two dimensions, the scientific and political dimensions and the technological developments, through the organisation of repositories and consideration of the motivations of scientists to share their data through these new approaches and technologies.

The literature review that we present in the following paragraphs can best be positioned in this third cluster, due to its focus on current data initiatives (e.g. repositories, formats and approaches for data sharing) and on the motivations for, and cultural implications of, data sharing for scholars.

*Literature review*

Sharing data has always been regarded as an important activity in science, a point that is widely accepted by the scientific community (Fienberg et al., 1985). From the point of view of the development of data metrics, probably the most important benefit of data sharing activities is that if they can function as a potential source of scientific recognition, they can have an incentivizing effect in the promotion of data sharing among scholars. In this sense, proper curation and dissemination of datasets can be considered as another scientific activity subject to research assessment and accountability for hiring, promotion, allocation of funds, etc. (Uhlir, 2012). In addition, it has been suggested that sharing data can increase the citation rate of the publications of the authors who share their data (Piwowar, Day, & Fridsma, 2007) in a kind of open data citation advantage (Piwowar & Vision, 2012) and as a way of fostering responsible scholarship (Mooney, 2011). Some other relevant benefits described in the literature are the following:

**Figure 3: Term map of the WoS-covered literature about "Data sharing"**



- Stronger Open Science (Fienberg et al., 1985). Scientific findings must be made available to the entire scientific community and the availability of research data for scrutiny and reanalysis should be part of the presentation of scientific developments and results, and can be regarded as good science. Sharing data allows other researchers and perhaps also other individuals (e.g. clinicians, patients, etc.) to access and use data that otherwise would not be possible (Groves, 2009), in the same manner as publications contribute to the advancement of knowledge and the production of new publications. This openness in the sharing and publication of scientific data would also work as a protection against fraud and faulty data in scientific research and contributes to the improvement of data collection and management methods, and in general terms is regarded positively by researchers (Cragin, Palmer, Carlson, & Witt, 2010).
- More efficiency in the use of scientific resources (Piwowar, 2011). Data sharing allows not only verification or refuting of previous studies, but also the re-analysis of the data, their refinement and further use. Also it allows a wider use of empirical data (particularly when these data are difficult to obtain or collect) (Fienberg et al., 1985). In this line, data sharing enhances the utilisation of data and promotes competition of scientific ideas (Gardner et al., 2003) as well as promoting collaboration. For funding agencies data sharing promotes the wisest use of public resources by reducing repetitive collection of expensive or sensitive data (Kowalczyk & Shankar, 2010). It has been also argued that data sharing contributes to accelerating scientific progress, generating opportunities for additional publications through collaboration (Brase, Farquhar, Gastl, Gruttemeier, & Heijne, 2009; Piwowar & Chapman, 2008).
- Promotion of new research through existing data and encouragement of multiple perspectives (Fienberg et al., 1985). The availability of data can promote the development of new research lines, more exploration and study of the data as well as better theories and analytic techniques. In this regard we can also mention the importance of data sharing in the international context of global issues such as health, environmental change, and food production, with particular challenges for data and researchers in developing countries (Arzberger et al., 2004).
- Other possible uses of data can be made possible. Here we can particularly mention the provision of resources for training of new students and researchers (Piwowar, Day, & Fridsma, 2007, Tenopir et al., 2011). Increasingly, it is not just researchers who reuse data, but also educators, policymakers, and even the general public. Making data broadly available can promote public understanding of science, evidence based advocacy, educational uses or citizen-science initiatives (Kowalczyk & Shankar, 2010).

## 3.2 Models for data metrics and conceptualisation of potential data metrics

We have not found literature on the idea of “data metrics” and, as Ingwersen & Chavan (2011) already claimed, “no metrics exist for data usage” that “recognize all players involved in the life cycle of those data from collection to publication”. For this reason we focus here more on potential data metrics models. In this section we will provide a thorough discussion on the existing models that could be relevant for the development of data metrics. We also will provide a first conceptualisation of the types of metrics for datasets and discuss some of the first attempts already made in the literature (e.g. Ingwersen & Chavan, 2011; Moritz et al., 2011)

*Models*

One key element for the potential development of data metrics is the existence of an adequate model (or “metaphor”) that can represent and explain the process of how all the stakeholders involved in the process can perform, contribute and benefit from data sharing and data metrics. In this sense, it has been claimed that to promote data sharing, a citation and credit model must be developed (Gardner et al., 2003).

One of the models that has been proposed consists of applying the publication and citation models to datasets (Brase et al., 2009; Lawrence et al., 2011; Newton, Mooney, & Witt, 2010; van der Graaf & Waaijers, 2012) in what is considered the “data publication approach” (Parsons & Fox, 2011). This approach is seemingly analogous to scholarly literature publication, and generally emerges from the culture of academic research and scholarly communication. Data publication seeks to define discrete, well-described datasets, ideally with a certain level of quality assurance or peer-review. The datasets often provide the basis for figures and tables in research articles and other publications. It is important to take into account that this model is the most mature of the “metaphors” in play (Parsons & Fox, 2011), although there is still incomplete agreement on the definitions and assumptions that arise from

the data publication model. In this regard, we can mention that two main publication models have been proposed, termed as "publishing" (with a small *p*) versus "Publishing" (with a capital *P*) (Callaghan et al., 2012), also discussed in (Uhlir, 2012). Thus,

- '*publishing*': this is a relatively simple model, consisting of the researchers publishing their data files on a website somewhere. This means that there are no guarantees that the data will be there after some time or that the files will not get corrupted. Furthermore, it is possible that a scientist who is not the data creator will not be able to understand or even open the data.
- '*Publishing*': in this model, for a dataset to "count" as a publication, it should follow a similar publication process as an article: be properly documented with metadata, be reviewed for quality (i.e. peer-review), be searchable and discoverable in catalogues (or databases), and be citable in articles (Brase et al., 2009). Two variants of this model have been suggested: the "stand alone publication" model and the "data journal publication" model (Lawrence et al., 2011; Schäfer et al., 2011), and both are expected to offer comprehensive review of the metadata and data itself and thus offer "true" data publication.
    - The "stand alone publication" model is when data are published as a stand alone dataset (which might include data sub-sets within the collection). In other words, the data are a publication in their own right, with no requirement for a co-existing standard journal article describing the data. The data archive provides systems which provide a data description document as the citable item, and the data is obtainable electronically.
    - The "data journal publication" model is based on "data journals" specialized in publishing "data papers". A data paper is a journal publication whose primary purpose is to describe data (providing information on the what, where, why, how and who of the data), rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments in support of those hypotheses based on data, as found in a conventional research article. Its purposes are threefold: to provide a citable journal publication that brings scholarly credit to data publishers; to describe the data in a structured human-readable form; and to bring the existence of the data to the attention of the scholarly community (V. Chavan & Penev, 2011). In appendix 3 we present some of the most important existing data journals (see Appendix 3).

In essence, it is expected that Data Publication (and also data citation) will ensure that data will potentially be considered as a first class research output. However, it has also been argued that the data publication model is not necessarily the most suitable for data (Mayernik, 2012; Parsons & Fox, 2011) and some limitations of the data publication can be pointed out:

- The exact meaning of data publication is not clear and can create misunderstandings among researchers. "Publication" carries many implicit assumptions that do not need to be true. Three main "frames" can exemplify this point (Parsons & Fox, 2011):
    - Peer review of data and articles are not parallel. An important element in data publication is how to ensure the quality of the data deposited in the repositories. In a number of scientific communities, there is no established data repository or data quality assessment protocol (Brase et al., 2009), and according to Dallmeier-Tiessen et al. (2012) there are two aspects to data quality: fitness for purpose and trustworthiness. These needs are particularly important for cross-disciplinary reuse, when the potential re-user might not have in-depth expertise and ability to evaluate the data being considered. In this sense, a peer-review quality control of datasets has been proposed (Lawrence et al., 2011). Thus, a dataset that has been peer-reviewed can be considered to have some guaranteed quality. However, traditional human refereeing is appropriate for certain datasets, but is too slow and will not scale to handle the current deluge of data.
    - Data citation is not necessarily a standardized and accepted concept. Some scientists can see it as a way of reducing citations to their papers, funding agencies sometimes question the idea of recognizing individuals as data authors, and from the bibliometric theory the value of "data citations" as a credit assess still needs to be established.
    - Copyright and restricted-access literature can have consequences for the data publication model, thus actually restricting the access to the data (contrary to the general idea of "open" data sharing").
- There are still important technical limitations restricting the development of data publications and data citations (and thus data metrics). These include problems related to incompatibilities in machine and software systems and data file structure, data storage, data management, data compatibility, etc. (Groves, 2010). To be shared effectively data must be stored, described and organized in such a way that others can find, access, understand, use and cite them (Dallmeier-

Tiessen et al., 2012). In other words, for data publication to be effective, the datasets must satisfy the following criteria: persistence, longevity, sustainability and quality (Kowalczyk & Shankar, 2010). However, all technical problems are essentially solvable. There are two technical problems that are of special interest regarding data citations and data metrics:

  - Dataset identification is a key element for allowing citations and long term integration of datasets. Scientists need to be convinced that preparing their data for online publication is a worthwhile effort (Brase et al., 2009). Brase et al. (2009) reviewed some of the most important persistent identifier that can be used for datasets, including Digital Object Identifiers (DOIs), Archival Resource Keys (ARK), Persistent Uniform Resource Locators (PURL) or Uniform Resource Numbers (URN). Persistent identifiers are also important elements in access and preservation of the data. Sometimes datasets are only published on the researcher's website, and if referenced at all, only referenced by the corresponding URL (Brase et al., 2009).
  - A further problem is the issue of granularity and versioning of datasets. Versioning of datasets is an important element in data citations (Kowalczyk & Shankar, 2010; Uhlir, 2012) and data metrics, because archived data can change substantively over time with new additions and changes. Granularity is also an important issue as there is a need to identify (and cite) smaller units or parts within the datasets (Ball & Duke, 2012).

- Costs involved in data publication. Probably the extra costs and effort that data sharing and data publication bring to the researchers have been pointed out in different surveys by researchers as an important barrier for researchers to share their data (Cragin et al., 2010; Dallmeier-Tiessen et al., 2012; Fienberg et al., 1985; van der Graaf & Waaijers, 2012). Essentially, data publication involves time and some extra costs for scholars (Stanley & Stanley, 1988). This is because datasets need to be properly documented (prepare all the metadata and the files, documentation, etc.) and stored, and sometimes tutorials need to be developed (Meyer, 2011).
- Organisational and legal problems. This limitation refers to aspects such as confidentiality (Savage & Vickers, 2009), privacy and ownership (Kowalczyk & Shankar, 2010) or legal national constrains regarding the publication of the data (Dallmeier-Tiessen et al., 2012). Some academic centres may view data sharing and publication as a threat to intellectual property, possibly impeding entrepreneurial spin-offs and technology transfers that bring revenue and act as incubators for future research (Piwowar et al., 2008).

Alternatively, other models could be also considered. Parsons & Fox (2011) have outlined some of the approaches or "metaphors" that could be an alternative to (or complement) the data publication model. These other approaches are the following:

- "Big Iron" approach: this approach comes from engineering culture and typically deals with massive volumes of data that are relatively homogeneous and well defined but highly dynamic and with high throughput. The Big Iron itself is a large, sophisticated, well-controlled, technical infrastructure potentially involving supercomputing centres and specialized interfaces. Big Iron systems rely on data and metadata standards and typically use relational and hierarchical data structures and organisational schemes.
- "Science support" is viewed as an embedded, operational support structure typically associated with a research station or lab. In environmental sciences, the focus is often on place-based research such as is conducted at long term research stations or sites. Data management is seen as a component or function of the broader "science support" infrastructure of the lab of the project. In this approach, data collectors at a field site may be either lead investigators on a given research project or lab technicians supporting many projects. In this context, data tend to be the research collections similar to those in the data publication metaphor but there is often a focus on creating community collections by characterizing important fundamental processes or particular representative conditions over time.
- "Map making" is seen as a central concept in so-called spatial data infrastructures and their associated geographic information systems (GIS). Map making could be seen as a subset of the data publication metaphor, but here the analogous publication is a map or an atlas rather than a journal article. The important metaphor here it is not the final product or the production process but rather the representation of the data and their associated science questions through a map. Data in this approach tend to be more fixed in time.
- "Linked Data" is based on computer science concepts of the "Web of data", relying on the underlying design principle behind the Semantic Web. The "data" in Linked Data are defined broadly and are envisioned as small, independent bits with specific names (URIs) interconnected through defined semantic relationships. The focus of this approach is more on interoperability and

capitalizing on the interconnected nature of the web, and less on preservation, curation or quality. In other words, the metaphorical emphasis is not on the product or process but on the data representations (e.g. as a network or graph).

Parsons & Fox (2011) argue that it is important to recognize these other existing metaphors and actively seek new metaphors that complement each other and help in conceiving all aspects of the e-science challenge. They also present two high-level metaphors (or models) that go beyond the data management enterprise, the "data infrastructure" and the "data ecosystem". The "data infrastructure" metaphor considers that an entire infrastructure helps us to recognize the scale of our endeavour (reaching across the entire scientific enterprise), but in many ways the concept of a data or information infrastructure is not yet defined. The "data ecosystem" metaphor considers the people and technologies collecting, handling, and using the data and the interactions between them, thus the focus is on interactions and relationships. However, this model currently presents some missing or unclear elements (e.g. what is the equivalent of publishing a dataset in a "data ecosystem"?).

### 3.2.1 Conceptualisation of data metrics

In this section we provide a first conceptualisation of metrics based on data sharing, data publication and data citation activities of scholars. We take the data publication model as the basic model, since this is the most developed one. It is important to highlight that, as mentioned before, little has been done about the concrete development of metrics for datasets. Actually, even infrastructures like DataCite indicate that for now they are "paving the way for new metrics and publication models that recognize and reward data sharing" (http://datacite.org/whatdowedo) but without actually developing any of these indicators.

For this reason, in order to conceptualize potential "data metrics" the best approach is to envision potential data metrics that could be applied to measure the "usage" of datasets, based on the "Data Publication" model(s) previously explained, and compare them with current metrics available for scientific publications. In this sense, we describe possible "metric scenarios" for datasets as well as how current available tools can be applied to the development of these metric scenarios.

In the first place, it is important to take into account, that several broad dimensions of metrics can be considered (Costas, Leeuwen, & Bordons, 2010; Waltman & Eck, 2009). These dimensions of indicators are described as follows:

- Size dependent indicators (total performance indicators): these are metrics that present the raw performance of a given unit of analysis (e.g. the total number of publications of a unit, the total number of citations, the h-index). In the framework of data metrics, potential metrics could be the number of data publications or datasets published by a researcher, a research group, a university, etc. as well as all the "data citations" received by these data publications. Thus, we could calculate the total number of data publications, the total number of data citations, even a data h-index, as potential indicators.

- Size independent indicators (average performance indicators): these are indicators that somehow measure the mean or median performance of a given unit (e.g. the average number of data citations per data publication, the median impact, etc.). We can distinguish two other groups of metrics[3], depending on what is the focus of the assessment (the direct performance of the unit or the sources of platforms where the unit is publishing):
  - *Direct average impact performance*: these are indicators that are based on the direct average data impact performance of the unit of analysis. For example, the average data impact of the data publications of the unit, the field normalized data impact of the unit, the share of 'top' data publications (e.g. datasets that are among the top 10% most cited in their fields), etc.
  - *Source-based performance*: these are indicators that measure the performance of the publication venues of a unit of analysis (i.e. the publication journals, the conferences, the repositories where the data are stored, etc.). In this dimension, a possible indicator would be a variant of a well-known indicator, the Journal Impact Factor (Fersht, 2009; Garfield, 1955), although the important criticisms to this indicator must be taken into account (Brembs & Munafó, 2013). This indicator is intended to be a measure of the performance of the journals

[3] Although it could also be argued to suggest size-dependent indicators on publication-venue performance, we include them here as they are mainly 'impact-factor"-like indicators.

in which a unit is publishing in. Thus, we could argue that (depending on the data publication model) it could be possible to calculate a kind of "data-venue impact factor" as well as other publication venue-based performance measures (e.g. MNJS in the new CWTS set of indicators – cf. Waltman, Van Eck, Van Leeuwen, Visser, & Van Raan, 2010) thus helping to assess the performance of for example, data repositories, data journals, etc.

Secondly, regarding our description of metric scenarios we also envision two main types of metrics that may play an important role in the development of metrics for datasets:

- Data publication & citation-based indicators. Considering the data publication model and assuming the existence and traceability of data citations it is not difficult to envision indicators that would be based on the numbers of data publications and data citations, following a similar model as in the scientific publication framework.

- Altmetrics-based indicators (Priem, Piwowar, & Hemminger, 2011). Altmetrics are being developed rapidly and they are meant to measure other types of impact apart from those that can be measured through citations. Examples are mentions in social media such as Facebook or Twitter, readers in Mendeley, comments in blogs, etc. Although they are not yet fully developed and important limitations must be taken into account when working with them (Wouters & Costas, 2012) it is clear that they could be also an important source of impact useful for the development of data metrics (or 'data altmetrics'). In this group we could also include the "Data Usage Index" (DUI) metrics suggested by Chavan & Ingwersen (2009) and Ingwersen & Chavan (2011) based on "search events and dataset download instances" obtained through usage logs. In fact these authors prefer this approach to data publications and citations because as they claim "no data citation mechanism now exists". Although these authors have made a quite interesting suggestion of 14 DUI-based indicators, they also acknowledge that for now these indicators are only possible in the GBIF website (http://www.gbif.org/) and based on the logs collected by its staff. So their possible extension and general adoption by the scientific community will depend very much on the availability of searching, viewing and downloading data provided by the different data repositories and data publishers.

**Table 3: Metrics scenarios for the data publishing models (and comparison with the current scientific publication model)**

| Types of metrics | Currently available tools with possibilities for "data metrics" | Metric dimensions | Models | | | |
|---|---|---|---|---|---|---|
| | | | Scientific publication | Data publication | Data Publication: Stand-alone data publications | Data Publication: Journal data publications |
| Data publication & citation-based metrics | - Data Citation Index (Web of Science)<br>- Google Scholar<br>- Scopus<br>- Microsoft Academic Search<br>- DataCite | Size-dependent | Yes | Difficult (1) | Yes (4) | Yes |
| | | Size-independent | | | | |
| | | - Direct average performance | Yes | No | Yes | Yes |
| | | - Source-based performance | Yes | No | Yes (3) | Yes |
| Altmetrics-based metrics | - ImpactStory<br>- Twitter, Facebook | Social media indicators | Yes | Yes | Yes | Yes |
| | - Mendeley<br>- CiteULike | Readership counts | Yes | No | Yes | Yes |
| | - Repositories<br>- Data Journals | Downloads & views counts (DUI metrics) | Yes | Difficult (2) | Yes | Yes |

(1) Careful and difficult data collection across repositories and data sources (e.g. acknowledgements, references, full texts, etc.) would be necessary.
(2) Depending on availability of the hosting websites (i.e. that downloading accounts are accessible to the public or analysts).
(3) Considering the publication venue (e.g. a repository) it may be more or less feasible depending on the type of publication venue (e.g. general and disciplinary repositories would be more useful than institutional or local repositories).
(4) Depending on the availability of meta-data for the datasets.

In Table 3, a general landscape of metric scenarios based on the dimensions and types of metrics previously mentioned is presented. This landscape scenario is meant to provide a first approximation of the potential development of metrics for datasets (particularly as compared to the already existing "scientific publication" model, and by no means is intended to be exhaustive in suggesting new indicators or metrics.
As shown in Table 3, the Data *P*ublication model(s) is the one most suitable for the extraction and development of metrics, particularly considering the current available tools for metrics. Altmetrics and particularly social media indicators seem very suitable for the collection of impact evidence for all models of data publication. Also download and view counts could be used, although they depend heavily on the availability of download and view metrics across the different websites and repositories.

We can argue that the Journal data publication model is the one most similar to the scientific publication model and therefore it can potentially profit from all existing indicators both based on publications/citations and altmetrics. The stand-alone data publication model, potentially, could have the same possibilities, but then we have to address issues such as the availability and type of repositories, where the datasets are published, or the type and amount of meta-data that is made available, in order to be able to obtain similar indicators.

From the previous results, we can conclude that the development of data metrics is feasible. Consequently, it could contribute to the acceptance of data sharing and data publication activities amongst researchers. Of course, for this to happen, stakeholders and scholars must understand and get committed to their development, application and fair use. If we argue that data metrics could have an incentivizing effect on data sharing, we also have to be aware of the attitudes of researchers towards them, the possible abuses of these metrics and unintended consequences of their adoption (cf. Brembs & Munafó, 2013; Weingart, 2005). Potential examples of these consequences, based on the current scientific publication model, could be: "data salami slicing" (i.e. the authors of data publishing their datasets in smaller pieces to increase the number of data publications), "data self-citations" (i.e. the creators of the datasets self-citing their own datasets in a disproportionate way), problems with authorship (e.g. "honorary data authorship", "ghost data authorship") among others. In a way, most of the unintended consequences that are found in the scientific publication model could be translated to the data metric dimension. Therefore, the implications of the adoption of a particular data metrics model will need further exploration and study.

## 3.3 Perceptions and views of stakeholders on data sharing and data metrics

When discussing data sharing and data metrics with stakeholders, important technical and cultural issues have been addressed at four different levels: at the level of collection, curation, dissemination and use of data. All have been considered as highly relevant from the perspective of good scientific practice. Sharing and using data should be at the core of scientific practice, but for this to happen it is critical that datasets can be found. And once they are found, it is critical to be able to access them. And if they are accessible, it is critical that they are interpretable. If these critical steps are taken, then it is in essence easy to use and reuse the data. During the discussions with the stakeholders, new tools that are being developed have been pointed out. These tools are presented in appendix 4, and discussed in more detail in chapter 4.

The main perceptions and views of the interviewees are summarized in table 4 (on collection and curation), and table 5 (on dissemination and use) and discussed in more detail below. These statements are aggregated from the interviewee's vision on necessary requirements and current hindrances for further development of data sharing and data metrics.

Summarizing the outcomes in table 4, from the interviews it is clear that regarding data sharing, the urgency to act comes from publishers and from data centres. For publishers it is urgent because they need to do something with the amount of data that they get when scientists submit papers with data. Publishers do not consider themselves the right place to store the data properly, so they have engaged with other stakeholders (data centres primarily, but also research infrastructures) in order to

**Table 4: Main perceptions of the interviewees on collection and curation (data sharing)**

| Stakeholder | Collection | Curation |
|---|---|---|
| Funder (only plans) | - Require data management plan for handling data that are collected as a result of funding<br>- Not allowed to fund infrastructural related costs<br>- Research project funding is short term<br>- Include data management in research evaluation.<br>- Recommendations, rather than rules/policies | - Require deposit in accordance with discipline-specific standards in subject-specific or institutional repositories<br>- Data and materials need to be prepared for unrestricted use of manual, automated and data mining tools |
| Scientist | - Thinks his/her data is too complicated for others to understand<br>- Big science vs individual science: scale of research | - Too much time and effort to curate data and provide metadata.<br>- Process need to be researcher-driven; top-down approaches will not work.<br>- Quality of the data depends on scientist |
| Research Infrastructure | - Data management plan required by NSF and in projects funded under Horizon 2020<br>- Includes withholding the last money until data are curated.<br>- Budget available for professionals<br>- Project funding not suitable for infrastructure: longevity 10-20 years instead of 3-5 years | - Provides international connection<br>- The petabytes are not the cost problem, it is the cost of staff for curation |
| Publisher | - 70% of all publications submitted with data, according to publishers<br>- Published as supplement: leads to fragmentation of data<br>- Don't want to pay for looking after the data | - Peer review of data for quality check<br>- Publishers cannot organize this by themselves, collaboration with data centres<br>- Collaboration with journals to demand proper data curation |
| Data Centre | - Type of data is irrelevant<br>- Protocol and consent relating to privacy; or anonymous.<br>- Humanities and social sciences have less centralized datasets except for big government data (OECD, WorldBank)<br>- Benefit of scalability is large; small datasets less interesting<br>- Charity funds have an interest in data collection and curation.<br>- Private organisations and business have an interest in data collection and curation as well. | - Need for Trusted digital repositories through certification<br>- Storage volume is critical.<br>- Need for search or viewing tools<br>- Rich metadata<br>- Interoperability<br>- Discipline specific, sometimes even topic specific repositories to define granularity, standards, audits & certification<br>- Some data that cannot be reproduced (climate/environment)<br>- Huge datasets like CERN use a scientific model<br>- Essentially, no technical problems |
| Libraries | - Indexing is necessary<br>- Synchronizing of data in data centres. | - Persistent Identifier systems (DOI, ISSN, or URN) are developing slowly<br>- Registries are crucial as backbone to link data to publications.<br>- DataCite discussion took 3 years, is now part of EU RI's |

address the commonly agreed on relevant issues in data sharing, as perceived from the interviews, such as:

- Standardisation of the metadata that accompany datasets;
- Organisation of persistent identifiers for datasets;

- Dealing with granularity and versioning of datasets;
- Organisation of the quality of datasets (trusted repositories, peer review);
- Dealing with scientific discipline specific aspects;
- Dealing with interoperability;
- Developing search and data mining tools.

These issues overlap and confirm the issues discussed in the literature. Essentially, all stakeholders agree that the technical problems cannot be the main issue, even though sometimes huge storage volumes will be needed. Currently, global efforts by the main players are ongoing to solve the issues above. The costs of data sharing efforts will increase substantially according to research infrastructures, especially in terms of staff involved in curation. This cost aspect is key in the

**Table 5: Main perceptions of interviewees on dissemination and use (of data metrics)**

| Stakeholder | Dissemination | Use (data metrics) |
|---|---|---|
| Funder (only plans) | - Open Access on internet within e.g. 2 years<br>- Return on Investment of funding is higher | - Unrestricted re-use of content with proper attribution (Creative Commons CC0 License)<br>- Project funding with OA money for data publishing<br>- Trendy topic, but scientists dominate proposal evaluation |
| Scientist | - The data are mine.<br>- Too much effort required to prepare for sharing<br>- Risk of replication/falsifying/ fraud.<br>- Copyright restrictions<br>- 60% of researchers does not want to share (half of them may be motivated by budget incentives)<br>- Embargo is important | - Little value compared to publication/citation (as it is not part of the reward system)<br>- A publication with data has more value<br>- Data citation is most straightforward (compared to data publication & co-authorship) |
| Research Infrastructure | - New research paradigm<br>- More collaboration; less competition | - E-scholarship and e-infrastructure |
| Publisher | - Standardisation of metadata<br>- Editors of journals are key people – these are scientists | - Data journals publishing metadata.<br>- Otherwise, data are always linked to publication (bi-directional)<br>- Publication provides context |
| Data centre | - Signalling of downloading is rare<br>- Data become actionable<br>- Ethical limitations in SSH, but if you do not see the benefits there always will be a reason not to share (like in the OA discussion)<br>- Three scenarios or basic models in which datasets develop were identified[4] | - Cost for re-use is high in publicly funded data centres (new management models)<br>- Use of datasets is growing faster than growth of datasets<br>- 80-20 rule: 20% of data is responsible for 80% of use<br>- New tools needed to properly reuse the data. |
| Libraries | - DOI will be dominant principle because scientist and publisher know it well.<br>- Quick open access, short embargo periods | - Need concept of how to track users – It is still manually provided by data centres<br>- New tools to search registries (full data, free text mining) |
| Publication databases | - No common database for data citations | -Transition period to full data citation model in publications.<br>- Proper citations of data in publications |

[4] The three scenarios are: 1) A dataset which is completed in a limited timeframe and will never be changed. This the easy scenario since one DOI will identify the dataset and citations will subsequently accrue over time. 2) A dataset that takes e.g. 20-30 years to complete by regularly adding on new data. Here you can create time series with assigned DOI's to each serie, and after 30 years close it off. 3) Building systematic and dynamic databases, such as the offices of national statistics, where every day the data change. In that case references can be made using one DOI and the day of access, although this does provide a complicated a citation model.

discussion with research funders. Likewise, data management plan requirements that are developed by the National Science Foundation in the USA (NSF) are a first step to acknowledge the importance of data sharing in science.

Not only are grantees expected to share with other researchers, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants[5], research proposals must also include a supplementary document labelled "Data Management Plan", describing how the proposal will conform to NSF policy on the dissemination and sharing of research results[6]. These requirements are in line with NSF's review of the grant proposal criteria. From this review, it is now recommended that meaningful assessment and evaluation of NSF funded projects should be based on appropriate metrics, keeping in mind the likely correlation between the effect of broader impacts and the resources provided to implement projects (National Science Board, 2011)[7].

Among research infrastructures and data centres there is anticipation that the EU Horizon 2020 research data management requirements will be as firm as those of the NSF, and some even hope for a stronger position of the EC towards data sharing. Altogether, interviewees confirm that the main hurdle in data sharing is the individual scientist who is reluctant to put effort into data sharing. This is mainly for cultural reasons: ownership of the data, workload to properly curate the data making them available for others, and lack of career-reward for making this type of effort. One explanation for this is that the research funding process is primarily geared at short term project funding, whereas data sharing is a long time effort. In any case, data sharing is a trendy topic, and the growing number of reports indicates that stakeholders have taken up their responsibility.

Summarizing the outcomes in table 5, from the interviews it is clear that examples of dissemination and data metrics in practice are limited and scarce. Conceptually, interviewees relate the different models as described in section 3.2.1 (stand alone data, citation in publication, data publication in data journals), but this is based on discipline specific initiatives rather than on rational choices. Dissemination and data metrics are related to open access practices in the sense that if data should be available for unrestricted use and re-use, the publication model should be open access as well. In terms of actual practice, some interviewees state that in some disciplines the use of datasets is growing faster than the growth of datasets. The results of use of datasets are visible in publications (article or data publication), which provide the context to the data; however the awareness to give credit to the dataset is often still lacking. This observation by interviewees confirms the literature that citing datasets is scarce. One interviewee estimates that two-third of the scientists lack awareness for data citation on a systematic basis. This is predominantly perceived as a cultural issue ('the data are mine') and a career issue (no reward is given for data publication or citation), rather than a technical issue. Interviewees however also point at emerging new activities where different stakeholders (e.g. data centres, publishers, libraries and scientific organisations) team up and try to link datasets with journal publications, or standards in mixed initiatives in different disciplines (see 4.1.1 for some examples). If these activities expand further, it is expected by the interviewees that, new tools will develop quickly to search registries, to track users and reuse of data. Interviewees wonder whether a data publication/citation database will provide a viable business model for publishers and publication databases, regardless of the fact that it will take a transition period of 5-10 years to build up a data citation model. Interviewees suggest that if funders, university human resources management and publishers include requirements for data sharing in their practice, the process will speed up and scientists will have to effectively comply and get committed From this data metrics may develop more easily.. The urgency for this commitment comes from the need for public trust in science and also to justify future research budgets (see also The Royal Society, 2012). Even though there are factors slowing this process down, the bottom line is that data have been used in scholarly practice for a long time already.

[5] See NSF's: Award & Administration Guide (AAG) Chapter VI.D.4 (http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)
[6] See NSF's Grant Proposal Guide (GPG) Chapter II.C.2.j (http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#IIC2j)
[7] http://www.nsf.gov/nsb/publications/2011/nsb1211.pdf

## 3.4. Global initiatives relevant for the development of data sharing and metrics

In this section we point at a few activities that stakeholders already have engaged in to increase data sharing and data metrics, especially with regard to offering rich metadata that allow discovery of data defining and exploring common standards of sustainability and establishing consistent citation of data. They serve as examples for inspiration and the list is by no means limitative.

*Standardisation/harmonisation*

Here we present some of the most important organisations taking up the challenge to discuss and negotiate international/global harmonisation and standardisation measures.

The **Research Data Alliance** (RDA) (http://rd-alliance.org) is being brought into existence by an initial three research funding organisations: The Australian Commonwealth Government through the Australian National Data Service; The European Commission through the iCordi project funded under the 7th Framework Program; and The United States of America through the RDA/US activity funded by the National Science Foundation.
The purpose of the Research Data Alliance is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonisation, and discoverability. This will be achieved through the development and adoption of infrastructure, policy, practice, standards, and other deliverables. The work of the Research Data Alliance will primarily be undertaken through its working groups.

The **International Council for Science** (ICSU) (http://www.icsu.org/) is a non-governmental organisation with a global membership of national scientific bodies. As virtually all international science depends on the production, use and integration of data and information, ICSU is keenly interested in all aspects of this issue. Today's environment raises new challenges related to standardizing the collection, analysis and dissemination of data, as well as to intellectual property rights and data access. Some of ICSU's Data and Information bodies are specific to a particular scientific domain; others are concerned with broad issues that affect the entire scientific community. Two that are particularly relevant for data metrics are:
*ICSU World Data System (WDS)* that aims at a transition from existing stand-alone services to a common globally interoperable distributed data system that incorporates emerging technologies and new scientific data activities, striving to become a worldwide '*community of excellence*' for scientific data. This data system should offer searchable common data directories and catalogues, which ensures the long-term stewardship and provision of quality-assessed data and data services to the international science community. The new system builds on the potential offered by advanced interconnections between data management components to foster disciplinary and multidisciplinary applications for the benefit of the international scientific community and other stakeholders. Applications for the new WDS are already being investigated, including data publication schemes and a WDS online portal. WDS will strive for a broader disciplinary and geographic base than its predecessor bodies and will encourage the establishment of nodes in emerging countries.
*ICSU Committee on Data for Science and Technology* (CODATA), its mission is to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use. CODATA is concerned with all types of quantitative data resulting from experimental measurements or observations in the physical, biological, geological and astronomical sciences. Particular emphasis is given to data management problems common to different scientific disciplines and to data used outside the field in which they were generated. The general objectives are the improvement of the quality and accessibility of data, as well as the methods by which data are acquired, managed and analysed; the facilitation of international cooperation among those collecting, organizing and using data; and the promotion of an increased awareness in the scientific and technical community of the importance of these activities.

*Searching*

The Australian National Data Service (ANDS) has as its vision 'more researchers reusing more data more often'. In order to make this possible, ANDS is building the *Australian Research Data Commons (ARDC)*. The ARDC is a combination of the shareable Australian research collections, the descriptions of those collections including the information required to support their re-use, the relationships between the various elements involved (the data, the researchers who produced them, the instruments that collected them and the institutions where they work), and the infrastructure needed to enable, populate and support the commons. This combined information can then be used to help

people discover data in context[8]. ANDS will build a set of interlinked web pages and make them available for harvesting by web search engines at Research Data Australia (which is part of the Research Data Alliance described above).

*Analysis*

Integrated Earth Data Applications (IEDA) (http://www.iedadata.org/). IEDA is a community-based facility that serves to support, sustain, and advance the geosciences by providing data services for observational Geoscience data from the Ocean, Earth, and Polar Sciences. IEDA is funded by the US National Science Foundation. Apart from the IEDA data collections, they also provide tool for searching and visualisation. IEDA develops and supports tools such as GeoMapApp, Virtual Ocean, and the EarthChem Portal that facilitate data access and visualisation, and that provide analytical capabilities to promote the use of Earth, Ocean, and Polar data within the scientific community and for educational purposes. IEDA applications are designed to enable a diverse community of researchers to dynamically interact not only with the continually expanding data collections, but also with complementary data held in other repositories. Both desktop and web-based applications are available for example to create maps, visualize and interrogate data, integrate data from disparate sources and create customized data compilations.

*Use and impact analysis*

STAR METRICS (https://www.starmetrics.nih.gov/) is the acronym for Science and Technology for America's Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science. It is a multi-agency venture led by the National Institutes of Health, the National Science Foundation (NSF) and the White House Office of Science and Technology Policy (OSTP). The project is a partnership between science agencies and research institutions, and aims to document the outcomes of science investments to the public. The benefits are that a common empirical infrastructure will be available, and it is perceived critical that this effort takes a bottom up approach that is domain specific, generalizable and replicable.
At present there is no data infrastructure that systematically couples science funding with its outcomes and there are also no mechanisms to engage the public with scientific funding. The aim of STAR METRICS is to create a repository of data and tools that will be useful to assess the impact of federal R&D investments. They will set up uniform standardized measures of the impact of science by including metrics such as publications and citations, but also social outcomes, workforce outcomes, and economic growth. In such a scheme, also data metrics could also get a suitable place.

[8] see http://ands.org.au/guides/discovery-ardc.pdf; http://services.ands.org.au/home/orca/rda/

# 4. Repositories and current tools relevant for the development of data sharing and metrics

## 4.1 Repositories and their features

In this section we analyse a sample of existing repositories from a technical point of view in order to determine their main features regarding the potential development of data sharing and data metrics. A total of 35 different repositories have been checked. A table showing the main features of the analysed repositories, particularly focusing on those elements that have relevance for data citations and for potential metrics for datasets, is attached in appendix 2 of this report.

Based on this analysis, several important aspects can be highlighted as relevant for data citations and the development of data metrics. These are basically the characteristics that belong to the category of "Access, usage, validation and metrics" in the checklist.

- All the repositories analysed offer "open" access to their data, however for 12 of them (34%) registration or some restrictions are involved (e.g. only for research and academic purposes, only for academic users, etc.).
- The repositories analysed are not necessarily very young and almost half of them (48%) were started before 2000.
- Almost all of the repositories analysed offer some possibilities of searching and browsing the datasets and records stored in them. This is an important element because the users of the repositories can easily "discover" datasets that can be of their interest.
- Most of the repositories are of a medium to large size, with numbers of datasets ranging from ~5,000 to 11 million records. Not all repositories mention the content that they offer.
- The type and format of data varies strongly and is clearly discipline and community related (this is also suggested by Kowalczyk & Shankar, 2010). The data repositories studied include formats such as photographs, maps, genomes, surveys, polls, proteins and nucleotides, nuclear structure properties, etc.
- The information in the repositories about validation or whether any reviewing takes place is not always clear. Twenty repositories (57%) show some level of validation of the data deposited. This is an important element regarding trust on future data metrics, as this is an element expected to ensure (and increase) the validity and usability of the data by future potential users.
- Another important element regarding the potential application of data citations (and future data metrics) is the presence of resource identifiers (i.e. unique identifiers of the datasets contained in the repositories, e.g. DOIs, URIs, ARKs, etc.) for the datasets deposited in the repositories. The majority of repositories (74%) show some kind of identifiers. At the same time, they present a broad diversity in types of identifiers (frequently internal identifiers are provided by the repositories). It is remarkable that only 8 repositories (23%) present DOIs for their datasets. This indicates that standardisation of identifiers for datasets may be important for the extension of data citations.
- Regarding the presence of metrics across the different repositories, for 18 of them (51%) we found some kind of metrics. In most of the cases (11 repositories) they do not offer metrics at the dataset level (which would be the best situation – cfr. Wouters & Costas, 2012) but only general, top-level or aggregated metrics at the repository level. The presence of metrics as such is an indication that the repository has (at least) the technology (and willingness) to collect them. All in all, given the fact that only half of the repositories offer some metrics, we can argue that the potential development of metrics based on usage (i.e. downloads, views, etc.) still needs to be realized by many of the current data repositories.
- Finally, the presence of guidelines for citation standards guidelines, recommendations and formats available across the data repositories can be considered as a good way to encourage the citation of data. The search for the presence of citation guidelines/recommendations to cite the datasets contained across the different data repositories showed that 20 of them (57%) present such information. This indicates that for a substantial amount of data repositories, albeit not all of them, data citations are seen as a way to acknowledge the use of the datasets contained in them.

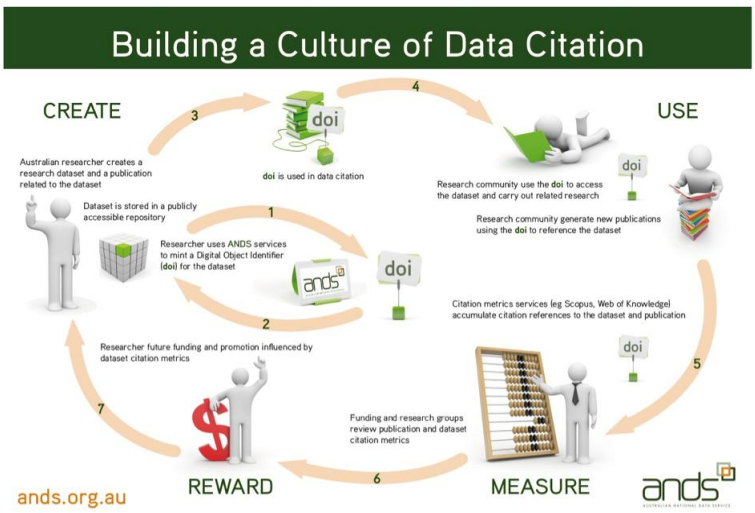
### 4.1.1 Stakeholders working together in combined initiatives regarding data sharing

Different data centres with different types of repositories (some of them in our sample above) are currently exploring models to link their data to other resources or to classic publications. This requires a collaborative effort of different stakeholders. Examples of some of the most important practices and initiatives are:

- GigaScience (http://gigadb.org/): Data repository and journal.
  GigaScience is a new i**ntegrated database and online open-access open-data journal** co-published in collaboration between BGI Shenzhen (the largest genomic organisation in the world) and BioMed Central, to meet the needs of a new generation of biological and biomedical research as it enters the era of "big-data". It aims to revolutionize data dissemination, organisation, understanding, and use from the entire spectrum of life and biomedical sciences. The journal has a novel publication format: one that links standard manuscript publication with an extensive database (providing DOI assignment to every dataset) that hosts all associated data and provides data analysis tools and cloud-computing resources. It includes not just 'omic[9]' type data and the fields of high-throughput biology currently serviced by large public repositories, but also the growing range of more difficult-to-access data, such as imaging, neuroscience, ecology, cohort data, systems biology and other new types of large-scale sharable data.

- The Inter-university Consortium for Political and Social Research (ICPSR) (http://www.icpsr.umich.edu/icpsrweb/landing.jsp) and the Data Preservation Alliance for the Social Sciences (DATA-PASS) alliance (http://www.data-pass.org/): professional organisation, data repository standards, and journals.
  **Professional associations in the social sciences** in the USA are increasingly recognizing the importance of properly citing data in their publications to encourage the replication of scientific results, to improve research standards, and to give proper credit to data producers. E.g. ICPSR has been working with DataPASS (a voluntary partnership of 6 professional organisations) to archive, catalog and preserve data used for social sciences research. They are promoting standards and improving practices for the citation of data. The American Sociological Review has already adopted a set of standards for citing data after an appeal from the Data-PASS partners. As other **peer-reviewed journals and data stakeholders** follow suit, consistently applied data citation standards will ensure that research data can be: discovered; reused; replicated for verification; credited for recognition; and tracked to measure usage and impact. ICPSR is also an associate member of DataCite, another key player in promoting data citation. They build on a model that was developed by Australian National Data Service (Figure 4).

**Figure 4: Model for a culture of data citation**
(http://www.ands.org.au/guides/data_citation_poster.pdf)




[9] The English-language neologism **omics** informally refers to a field of study in biology ending in *-omics*, such as genomics, proteomics or metabolomics. The related suffix **-ome** is used to address the objects of study of such fields, such as the genome, proteome or metabolome respectively (http://en.wikipedia.org/wiki/Omics ).

- Pangaea (http://www.pangaea.de/) data repository library and publisher Elsevier (http://www.elsevier.com/).
  PANGAEA is an information system operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research, linking primary data related to articles in earth and environmental science journals. PANGAEA is open to any field of earth system research, enabling a bibliographic citation of datasets with identification via DOI and a widespread distribution through portals, library catalogues and search engines, which is currently unique on the Internet. PANGAEA started a collaboration with Elsevier, **a publisher**, to interconnect the diverse elements of scientific research. Elsevier articles at ScienceDirect are now enriched with graphical information linking to associated research datasets that are deposited at PANGAEA. This enrichment functionality offers a blueprint of how Elsevier would like to work with dataset repositories all over the world. In the first phase, more than 1,000 articles from various earth science journals were linked. This includes 'reciprocal linking' – automatically linking research datasets deposited at PANGAEA to corresponding articles in Elsevier journals on its electronic platform ScienceDirect and vice versa.

- Dryad (http://datadryad.org/)
  Dryad is both an international **repository of data** underlying peer-reviewed articles in the basic and applied biosciences, and a membership organisation, governed by **journals, publishers, scientific societies, and other stakeholders**. Dryad welcomes data submissions related to published, or accepted, scholarly publications, in particular for tables, spreadsheets, and all other kinds of data that do not have another discipline-specific repository. Dryad also welcomes the involvement of **journals, editors, publishers, authors** and others who support **data archiving**. Authors may submit data files associated with their publications. Editors and journals can facilitate their authors' data archiving by setting up automatic notifications to Dryad of accepted manuscripts, streamlining the authors' process for depositing data. Dryad is being developed by the National Evolutionary Synthesis Center and the University of North Carolina Metadata Research Center, in coordination with a large group of Journals and Societies. The National Evolutionary Synthesis Center is a joint effort of Duke University, the University of North Carolina, and North Carolina State University.

## 4.2 Tools for data metrics

In this section we describe the most relevant infrastructures that are paving the way for new data metrics and publication models, helping to recognize and reward data sharing. This in turn would allow researchers and statistical or bibliometric experts to calculate adequate metrics. They are briefly described below (see Appendix 4 for overview).

**DataCite** (http://datacite.org/) is an international not-for-profit organisation formed in London on 1 December 2009. The aims of DataCite are to establish easier access to research data on the Internet, to increase acceptance of research data as legitimate, citable contributions to the scholarly record, and to support data archiving that will permit results to be verified and re-purposed for future study. DataCite seeks to support researchers by helping them to find, identify, and cite research datasets with confidence (i.e. discoverability of datasets), to support data centres by providing persistent identifiers for datasets, workflows and standards for data publication (i.e. helping to solve to problem of identification and traceability of datasets); and to support journal publishers by enabling research articles to be linked to the underlying data. DataCite also contributes to assign persistent identifiers to datasets, by developing an infrastructure that supports simple and effective methods of data citation, discovery, and access. DataCite is leveraging the Digital Object Identifier (DOI) infrastructure, which is well-established and already widely used for identifying research articles (although they also keep an open approach by considering also other identifier systems). In this regard, all DataCite DOIs resolve to a public landing page that contains information about the associated dataset and a direct link to the dataset itself. DataCite is a membership organisation, and the globally active partners all use the same standards (DataCite, 2011).

**CrossRef** (http://www.crossref.org/). CrossRef's goal is to be a trusted collaborative organisation of the world's leading scholarly publishers focusing on libraries and scientists. Its specific mandate is to

be the citation linking backbone for all scholarly information in electronic form. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content, but rather effects linkages through CrossRef Digital Object Identifiers (CrossRef DOI), which are tagged to article metadata supplied by the participating publishers. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article.

**ORCID** (http://about.orcid.org/) ORCID aims to solve the name ambiguity problem in research and scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current researcher ID schemes. These identifiers, and the relationships among them, can be linked to the researcher's output to enhance the scientific discovery process and to improve the efficiency of research funding and collaboration within the research community. ORCID is governed by an elected Board of Directors, majority non-profit, comprised of fourteen members of the global scholarly research community. The Board is responsible for establishing general policies for the governance of ORCID, based on a set of core principles, among them openness and transparency. In addition to the Board of Directors, ORCID has several Working Groups, open to the research community.

**Data Citation Index**. Thomson Reuters has recently launched a new product called “Data Citation Index”. This product is the result of the collaboration with some of the most important research libraries and digital repositories (e.g. California Digital Library, Protein Data Bank, PANGAEA, UK Data Archive, etc.) in order to design a single source of data discovery for the sciences, social sciences, and arts and humanities. The Data Citation Index claims to fully index a significant number of the world’s leading data repositories of critical interest to the scientific community, including over two million data studies and datasets. The records for the datasets, which include authors, institutions, keywords, citations and other metadata, are then connected to related peer-reviewed literature indexed in the Web of Knowledge. Thus, it is possible to track and count the citations that an individual dataset has received in the scientific literature. The Data Citation Index intends to solve three of the major issues that frustrate and discourage researchers from submitting their data to repositories: Discovery: as a database the Data Citation Index allows the user to search by different parameters and thus be able to retrieve and discover datasets that could be of their interest. Attribution: each result page contains a *how to cite this resource* link with a recommended citation format. This could help to establish citation conventions that could also help to get better data metrics. Thomson Reuters is partnering with researchers to recommend and standardize how citation should be collected and cited for datasets. Measurement: as a result of the linkages from data to literature it is possible to calculate the number of citations that a dataset has received and thus being able to “assess” the use of the dataset by other researchers. Unfortunately, testing the Data Citation Index falls beyond the scope and the possibilities of this report. It is a new service, which requires thorough testing before its value can be fully established (Thomson Reuters, 2012). Regardless of the value of such a service, the commercialisation of indices could have a negative impact on ranking and the system of research as a whole (Brembs & Munafó, 2013).

# 5. Challenges for the development of data metrics

As has been discussed in this report, data sharing offers important benefits for scientific progress and advancement of knowledge. However, several limitations and barriers in the general adoption of data sharing are still in place which limit the development of data metrics. Probably the most important challenge is that data sharing is not yet very common among scholars (Borgman, 2012) and is not yet seen as a regular activity among scientists, although important efforts are invested in promoting data sharing. As a result the most important problems and challenges regarding data metrics are closely tied to the more general problems related to data sharing. In addition, there is relatively low commitment of scholars to cite data. For example, many journals routinely require authors to share their data with other investigators, either by depositing the data in a public repository or making it freely available upon request (Groves, 2010; Kowalczyk & Shankar, 2010; Savage & Vickers, 2009). However, Savage & Vickers (2009) tested how well authors comply with such policies, found a very low rate of response of authors actually providing the data of their studies, and concluded that explicit data sharing policies of journals do not lead authors to share data.

This low involvement of researchers in data sharing has mostly to do with their perceptions and with cultural issues related to these activities. Tenopir and colleagues (Tenopir et al., 2011) performed a survey exploring the perceptions of researchers regarding data sharing and data publication. They found that important barriers perceived by the scholars are lack of time and lack of funding. Similar observations were made by (Nicholson & Bennett, 2011). These are the same concerns as were expressed by most of the stakeholders interviewed in this project. Other important perceptual problems among researchers are the following:

- There is a general complaint by scholars that data publication and citation is not an element considered for promotion and research assessment (Schäfer et al., 2011). Citation of data is not currently standard behaviour in scholarly writing. Actually it is a “rare” activity, the majority of scientific publications still fails to provide adequate data citation (Mooney & Newton, 2012). This seriously jeopardizes the development of a reward system. If scholars do not publish and cite datasets in a systematic and standardized manner, the development of data metrics will be difficult and probably not reliable (and trustworthy). This actually creates an interesting circular paradox: scholars do not share their data because they feel that they are not rewarded for this, although the development of data metrics could improve this situation; however, this does not happen because the volume of data publications and data citations (and the tools available) are still poor and unreliable, and this situation limits the development of a reward system that would incentivize more data sharing
- Stanley & Stanley (1988) argued that one important drawback for data sharing is the idea of the “loss of control” over the data of the creators of the data. Scholars may fear that the data can be misused by the users, thus the creators of the data would still have some responsibility in order to ensure that the data are used in an ethical manner. In other words, the problem of data misuse is still strongly perceived by scholars (Cragin et al., 2010). Misuse can also be considered as the lack of acknowledgement to the creators of the data.
- Stanley & Stanley (1988) also identify as a perceived drawback the potential decrease in the quality of science. This is based on the idea that more researchers would be more inclined to simply reuse the datasets of others instead of collecting new ones, ending up with many fewer original datasets examining the same or similar research questions.
- Restriction to access and use the datasets and embargos for researchers to exploit the data and to maximize its benefits could also have an effect in the development of data metrics. This problem has to do with data ownership and permission for data release (Groves, 2010). This is not necessarily a limitation, actually embargo periods have been suggested in the literature (Savage & Vickers, 2009) and they are sometimes considered as an important element to protect data (and also junior or postdoctoral researchers) and its incidence and time depends on the disciplines (Uhlir, 2012). Public-private research collaborations in which partners have different motivations for producing data pose unique challenges (Arzberger et al., 2004). Despite embargos, there is a strong cultural perception among scientists that the data are ‘theirs’, which is not fully supportable from a perspective of accountability of public funding.
- Researchers can also fear that potential errors can be exposed which would make them more vulnerable (Fienberg et al., 1985). However, the prevention of fraud and other misuse actually

could be one of the benefits of data sharing. In fact, the Dutch fraud case in psychology where Diederik Stapel, a highly regarded scientist at Tilburg University, who fabricated fraudulent datasets, has increased the call for more transparency (Callaway, 2011). The investigating committee revealed that Stapel often refused to share his research data with colleagues, even with co-authors. This is less uncommon than it may seem: In 2006 Wicherts et al. (2006) showed that that almost three-quarters of researchers who had published a paper in a high-impact psychology journal had not shared their data (Wicherts, Borsboom, Kats, & Molenaar, 2006), confidentiality being the most common excuse that psychologists offer for not sharing data. In practice, they simply fail to document their data in a way that allows others to quickly and easily check their work. Misconduct flourishes in such secrecy, and this calls for improving checks and balances to avoid a repeat. A suggested possibility is that the obligatory archiving of raw data in online appendices to journal articles or in repositories should be a precondition for publication.

- Scientific disciplines differ in their needs for data reuse (Arzberger et al., 2004) and (probably) also in data citation behaviours. In a way, data practices are an integral part of scientific communication and, therefore, subject to the same social and organisational constrains that shape disciplinary differences in the development and adoption of other communication practices and systems (Cragin et al., 2010). Disciplinary differences have been observed and resistance of researchers to share their data is not uncommon across fields of science and even in the disciplines with more sharing traditions (Milia et al., 2012).

Other important concerns, challenges and topics that have been raised by the interviewed stakeholders, can be summarized as follows:

- Will the old publishing model disappear? And what will be the business case for data publication? The open access debate is not the same as making data citable. And now that the Data Citation Index is developing, do we want data to be barrier based and subscription based? Is there a proper model for data citation? The current efforts are still limited.
- Should we consider data collection, curation, and dissemination as similar to publication/citation, or should we consider it as a different species in its own right? Some argue that data publication and data journals are not favourable because plain publication doesn't say anything about the scientific quality of the data. Others claim that a data Publication is a useful contribution (because of the peer review) and may show preliminary use. There is a clear distinction between the datasets themselves and dataset articles. A paper about a dataset is an incentive to collect at least some formal citations
- DataCite was a demand driven process, and the solution was there before the question came. But, despite the global registry, data citation 'bibliometrics' still has to develop. And will this always be linked to publication/citations?
- In order for data citation to be a first class object that stands the comparison with publication citations, one could consider including citation flavours and context. Citation flavours, according to Piwowar (2012) refer to datasets making different types of impact. Some are useful for example for training, or calibration, or for testing methods or exploring hypotheses. From a plain citation count this type of impact 'flavour' is not visible, but this may be derived from a combination of impact indicators and context (Priem, Piwowar, & Hemminger, 2012).
- When developing a data citation model and the consecutive data metrics, one has to take into account that datasets develop in different ways with different timelines and specificities, and also show differences by disciplines and in their purposes.
- A data citation model could be used for different purposes, depending on the objectives of the stakeholder when applying them.

# 6. Solutions and necessary developments

In spite of the limitations and challenges previously pointed out, step by step, scholars, research funders, data centres, libraries, and other data providers are intensifying their activities in the field of research data management over the last few years, and publishers are also beginning to develop strategies to support the sharing of research data (Schäfer et al., 2011). Most of the difficulties and limitations for data sharing can be overcome if the scientific community and the different funding agencies are able to commit substantial resources to data sharing (Fienberg et al., 1985). As Schäfer et al. (2011) pointed out, despite the high level and general enthusiasm for data sharing, its successful implementation will require detailed understanding of a complex landscape of intertwined issues, which are related to data sharing. In this section the most important potential solutions and developments needed regarding data sharing and particularly for data metrics are discussed.

Need for a reward system for scientists that considers data metrics. Appropriate professional and career reward structures are necessary in order to incentivize data sharing and data publication (Arzberger et al., 2004). The lack of recognition of data publication and other forms of sharing incentives is regarded as an obstacle to establishing a data metrics culture. This is one of the most important challenges for the acceptance and standardisation of data publication. In the current scientific paradigm, academic recognition is mainly achieved through scientific publications (e.g. articles, conference proceedings, books, book chapters, etc.), where sharing datasets is a time consuming task not adequately compensated (Brase et al., 2009). As a response to this situation, there is a general claim in the scientific literature that reward structures must be in place to encourage data sharing and data publication, and that data citation should be the appropriate tool for scholarly acknowledgment (Mooney & Newton, 2012). However, this challenge of a lack of a reward system needs to be approached from different perspectives:

- From a citation metric point of view, the importance of measuring data citations to provide an indication of impact on the scientific community and a driver of academic recognition has been pointed out (Brase et al., 2009). Instead of formally citing datasets, the users typically acknowledge data use in the text of the document or in the acknowledgements section. This needs to be changed by promoting data publication and data citation among researchers, and particularly developing a publication model where they can see the general advantages (both from a general scientific point of view, but also from an individual point of view) that data sharing can bring to their scientific careers and the development of their work.
- From an institutional perspective, it is important that research institutions and research funding organisations develop and track metrics for data sharing contributions as part of their academic research environments (Piwowar, Becich, Bilofsky, & Crowley, 2008). In the same line of argument, the consideration of data sharing activities during hiring, tenure, and promotion decisions, for example by providing a bonus to a publication's impact if the authors have shared the raw research data could be a challenging change (Piwowar et al., 2008). The development of a data sharing citation index, as a concrete tool with metrics for tracking the reuse and citation of datasets is a necessary solution (Piwowar et al., 2008). The scientific community should realize that part of the research budgets will have to be attributed to data curation and data dissemination.

Development of standards for data citation (Gardner et al., 2003). No universal standard exists for citing datasets and dataset identification and cross-referencing shall be accomplished at a global level (Brase et al., 2009) in order to be able to provide valid metrics crediting the proper datasets and data creators. However, for datasets there is no generic standardized format and among the most general citation styles there are no guidelines for the description of datasets (Newton et al., 2010). Data citation standards should be developed in order to be able to identify subsets of the data as well as the whole dataset (Ball & Duke, 2012). They must provide the reader with enough information to access the dataset and it should provide a mechanism for accessing the dataset through the Web infrastructure. Dataset identification is a key element for allowing citation and long term integration of datasets into texts as well as supporting a variety of data management activities (Brase et al., 2009). They must be usable not only by humans but also by software tools, so that additional services may be built using these citations. In particular services must be developed in order to use data metrics to support the academic reward system, as well as services that can generate complete citations (Ball & Duke, 2012).

Institutional commitment is an important challenge in the development of data sharing (Altman & King, 2006) and a reward system based on data metrics that would promote data sharing is essential. Research funding agencies need to recognize that preservation and access to digital data are central to their mission and support these tasks accordingly (Campbell, 2009) by promoting data publication and data metrics. The persistence of the connection between data citation and the actual data ultimately must also depend on some form of institutional commitment and the widespread adoption of a data sharing culture needs leaders (Piwowar, Becich, Bilofsky, & Crowley, 2008). Leading institutions must fund and maintain infrastructures of data sharing (Piwowar et al., 2008). Although there have been concerns about the potential problems that data sharing becomes a mandatory activity (Stanley & Stanley, 1988) the general trend now is that research funding organisations such as the National Institutes of Health (NIH) are requesting for plans of data sharing and management for the project they are funding (Gardner et al., 2003; Piwowar et al., 2008; Torres-Salinas et al., 2012). However, the results are not yet clear (Schäfer et al., 2011) and more research is necessary in this line. Digital equivalents of libraries are also necessary, through institutions that can take the responsibility for preserving digital data and making them accessible in the long term (Campbell, 2009). Data curation services will need to accommodate a wide range of subdisciplinary data characteristics and sharing practices (Cragin et al., 2010). As part of a larger set of strategies emerging across academic institutions, institutional repositories will contribute to the stewardship and mobilisation of scientific research data (Cragin et al., 2010).

# 7. Recommendations & evolution in the coming years

In this section we point out recommendations regarding the most important challenges that would need a response by the scientific community and main stakeholders in data metrics during the coming years. These recommendations are also targeted at the most important stakeholders involved in the promotion and generation of data sharing and potential data metrics activities in the next years.

In general, it is clear that there is growing awareness of the importance of data sharing amongst the most important scientific stakeholders. In this sense the benefits of data sharing and data publication are broadly accepted, although as discussed in this report important challenges are still in place. The advancement of data metrics (lifting it from a second-class to a first-class scholarly record) requires integration of data collections, curation and dissemination and the subsequent tracking into the different databases (e.g. the Web of Knowledge, Google Scholar, Scopus, Microsoft Academic Search, etc.). However, this cannot be achieved without the concurrence and collaboration of all stakeholders to build community engagement on data metrics, accompanied by open sharing and open access. In particular, it is necessary to reinforce access and visibility, to strengthen recognition and attribution, to elevate the status of research data, and to broaden resources. Recommendations for the potential solution are presented below, with indication of the most important stakeholders that should play a role in their solution:

- *General adoption of data sharing and data publication among scholars*. Data sharing is still a marginal activity in the daily activities of researchers, with important differences across disciplines and even sub-disciplines. Data citation is also a rare activity among scholars who mostly still do not publish data citations to the sources they use, although they acknowledge them by other forms (e.g. acknowledgements or mentions in the body of the texts).
- *Development of a reward system that includes data metrics.* Clearly, some form of a reward system that stimulates data sharing will be in demand in the near future. There are several possibilities for this reward system: one would be the creation of a completely separate reward system only focused on data publications and data citations; a second option would be to incorporate data metrics in the current reward system as a complement to the current evaluation systems. It is not clear which one of the options would be best, but both of them will need to be studied and discussed in the near future. Moreover, it needs to be considered whether a new reward system should be aimed at the individual scientist level or rather at the research group.
- *Reduce costs and make the whole process of data publication more efficient.* The most important barriers perceived by the researchers in order to share their data are the costs involved in data sharing and data publication, mostly in terms of time and money. Research infrastructures have an important influence in this point by providing services and resources that can reduce the costs involved in data publication.
- *Reduce the negative cultural perceptions of researchers regarding data publication.* The perceptions of losing control over their data, the possibility of missing research opportunities, the possibility of receiving criticism as the creators of the datasets or not perceiving that data sharing is a rewarded activity, as well as cultural differences among disciplines are important limitations that are slowing down the acceptance and development of a data sharing culture among scholars. Policies in order to inform and create awareness among scholars on the importance of data sharing, data publication and data citation are needed. Also policies in order to mitigate the potential damage or loss of reputation with respect to mistakes in datasets, could help reducing the resistance of scholars.
- Solution to the most important technical problems & lack of standards for preservation, publication, identification and citation of datasets. This includes addressing problems such as versioning and granularity. Data centres and research infrastructures must create systems for the clear identification of datasets and their versions, together with standards for data publication and citation. Different guidelines and solutions would probably be necessary regarding the different disciplines; therefore the collaboration with libraries and scientists would also be necessary at this point.
- *Solution to organisational problems*, this includes legal issues, confidentiality, embargo periods, and particularly a system of validation of the deposited datasets. These are a barrier that needs to

be tackled in the near future. All stakeholders must be included in the solution of the limitations regarding organisational aspects.

- *Reduce the dispersion of data repositories and coordinate initiatives.* There is a significant profusion of data repositories (e.g. more than 500 detected by DataBib), leading to a strong dispersion of sources for data discovery and data metrics. This emphasizes the important effort that an organisation such as DataCite is placing on concentrating and standardizing the different data repositories all around the world and their datasets, as well as attributing DOIs to them that will facilitate their traceability, citation and measurement. This type of coordination is necessary in order to reduce the dispersion of initiatives in data sharing and thus being able to develop more centralized and robust data metrics.
- *Develop standards and interoperability protocols across the different actors*. There is an important lack of homogenisation among the different repositories. The majority part is disciplinary (only a few are multidisciplinary) and most of them lack standards, DOIs or permanent identifiers, suggestions or recommendations for citing, etc. The role of two emerging players in the future development of data metrics is notable. DataCite brings together many different data repositories, contributing to their standardisation and centralisation and also the feasible development of data metrics. The Thomson Reuters Data Citation Index as a commercial tool would merit further testing and risk assessment in the standardisation of data metrics. Although we have identified these two players (because they focus more specifically on data metrics), these developments should be seen in the context of research data infrastructures as a whole. In particular, the European infrastructural initiatives are important, since they may provide incentives and tools for data sharing and data metrics. The Research Data Alliance in particular can be expected to develop relevant frameworks in the years to come [10]..

Based on the previous recommendations, in table 6 we summarise the most important recommendations for the different stakeholders in data sharing.

*Other more general recommendations and lines for future research*

This study still leaves open many questions, debates and challenges that will require a more careful analysis in the near future, particularly regarding the development of data metrics with validity for research assessment. Some of these future lines and their potential development are outlined below.

- In the next few years, a good conceptual model for the development of data sharing and data metrics will need to be developed. The publication/citation model seems to be relatively well accepted by the proponents of data sharing, but other models could also be applied and would need to be considered as well.
- In line with the lack of a model for data sharing, and regarding the future development of data metrics, we still need to disentangle what would be the 'value' of data citations in research assessment. Initially a good solution could be to consider them to have the same value as regular article citations. This would encourage scholars to consider data publication and data citation as (equally) important activities as their scientific publication activities. However, it could be argued that a citation to a dataset is even more important because without the dataset the new publication (probably) would not even exist; on the other hand, a counter argument could be that the development of a dataset does not (always) require the same scientific/intellectual effort that a publication with new scientific theories and approaches. These questions require more research as well as discussion in the scientific community.
- Incentives for data sharing should be connected to two key moments in the academic life cycle: 1) the moment of receiving a grant, which means that research funders or university administrators can demand an effort of the researchers in sharing their data and later consider it to decide on tenure tracks. 2) the moment of getting a publication accepted, which means that publishers can demand the curation of the underlying data. In practice, it is helpful if people are confronted with having to tick a box on data when publishing, this would contribute to making them aware of the importance of data sharing (although the literature also shows that even with these requirements scholars are still reticent to share their data, even before agreeing with the data sharing policies of the journals). Alternatively, the responsibility for data sharing practices could also be placed at the research group level, and not at the individual researcher level. This would be part of a paradigm

[10] See the RDA Launch Press Release (http://static.squarespace.com/static/50ad9169e4b00ca12a884beb/t/513f8affe4b0314c7e48ef2f/1363118847181/RDAEvent_pressrelease.pdf).

**Table 6: Recommendations for the different stakeholders**

| Stakeholders | Recommendations |
|---|---|
| Funders | * Demand and reward data sharing activities<br>* Consider data metrics in assessments<br>* Inform policy about the importance and benefits of data sharing<br>* Promote open access of data |
| Research infrastructures | * Promote policies of data sharing<br>* Promote arguments and incentives in favour of data sharing<br>* Provide options and alternatives to the different types of data sharing activities<br>* Professionalize staff and standardize data sharing activities (collection, curation, dissemination) |
| Scientists | * Include data sharing as good scientific and scholarly practice<br>* Promote data citation as the formal way of acknowledging data sharing<br>* Perform more research on benefits and possibilities of data sharing<br>* Define codes of conducts for disciplines considering appropriate regulations, i.e. embargo periods, anonymisation etc. |
| Data centres | * Inform the scientific community about data activities and services<br>* Contribute to reduce the dispersion of data repositories<br>* Develop robust solutions for the preservation and standardisation of the data storage and citations<br>* Develop tools for tracking the users of the repositories |
| Publishers | * Promote data sharing in their publications and journals<br>* Inform authors about other data sharing stakeholders (e.g. repositories, data centres)<br>* Support open access to data |
| Libraries | * Promote data publications and data citations<br>* Coach scholars and research managers in their data publication and citation activities<br>* Inform authors about other data sharing stakeholders (e.g. funders, repositories, data centres)<br>* Develop tools to find data repositories<br>* Develop and test appropriate metrics |
| Publication databases | * Collect and measure data publications and data citations<br>* Facilitate the analysis and metrics of data publications and data citations |

shift in science in which collaboration becomes more predominant. European policy strongly supports building research infrastructures that include 'big data', as well as further integration of the ERA (European Research Area). Both policy developments require collaboration at the organisational level, which is opposite to competition at the individual scientist level.

- Regarding the costs involved in data sharing and data publication (e.g. creation of metadata, curation, etc.), following Borgman (2012), it can be questioned if all data and forms of data are worthwhile to be shared and published. If so, mechanisms and criteria for selection would be necessary. More research would be necessary in this sense. However, we can hypothesize that the existence of data metrics could also play a role, by allowing the detection of datasets with a high demand among scholars, or to detect the "life cycle" or the "durability" (Costas, Van Leeuwen, & Van Raan, 2009) of datasets, allowing the development of indicators that would help in establishing more efficient preservation policies.

*Acknowledgements*

The authors want to show their gratitude to the Knowledge Exchange working group on Research Data, Magchiel Bijsterbosch, Niels Jørgen Blaabjerg, Joy Davidson, Ingrid Dillo, Rob Grim, Simon Hodson, Hans Pfeiffenberger, Laurents Sesink, Stefan Winkler-Nees, and particularly Angela Holzer and Keith Russell for their help and insights in the development of the research and the writing of this report. The authors also want to acknowledge the generous participation of all the interviewed stakeholders for their kind collaboration.

# 8. References


Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., et al. (2004). Promoting Access to Public Research Data for Scientific , Economic , and Social Development. *Science*, *3*(November), 1777–1778.

Ball, A., & Duke, M. (2012). The Scientific Article in the Age of Digitization. *JISC*, *11*. doi:10.1007/1-4020-5340-1

Borgman, C. (2007). *Scholarship in the Digital Age*. Cambridge: MIT Press.

Borgman, C. L. (2012). The Conundrum of Sharing Research Data, *63*(6), 1059–1078. doi:10.1002/asi

Brase, J., Farquhar, A., Gastl, A., Gruttemeier, H., & Heijne, M. (2009). Approach for a joint global registration agency for research data. *Information Services and Use*, *29*(1), 13–27. doi:10.3233/ISU-2009-0595

Brembs, B., & Munafó, M. (2013). Deep Impact : Unintended consequences of journal rank. *Arxiv*, 1–38.

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., et al. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7*(1), 107–113. doi:10.2218/ijdc.v7i1.218

Callaway, E. (2011). Report finds massive fraud at Dutch universities. *Nature*, *479*(7371), 15. doi:10.1038/479015a

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12 Suppl 1*(Suppl 15), S2. doi:10.1186/1471-2105-12-S15-S2

Chavan, V. S., & Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC bioinformatics*, *10 Suppl 1*(Lmmc), S2. doi:10.1186/1471-2105-10-S14-S2

Costas, R., Leeuwen, T. N. Van, & Bordons, M. (2010). A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level : The Effects of Age on Productivity and Impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564–1581. doi:10.1002/asi.21348

Costas, R., Van Leeuwen, T. N., & Van Raan, A. F. J. (2009). Is scientific literature subject to a sell-by-date? A general methodology to analyze the durability of scientific documents. *Journal of the American Society for Information Science*, *61*(2), 26. Retrieved from http://arxiv.org/abs/0907.1455

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosofical Transactions of the Royal Society A*, *368*(1926), 4023–38. doi:10.1098/rsta.2010.0165

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Suhonen, J., Wilson, M., Science, H., et al. (2012). *Compilation of results on drivers and barriers and new opportunities*.

DataCite. (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data*. doi:10.5438/0005

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (p. 552). The MIT Press. Retrieved from http://www.amazon.com/dp/0262013924

Fersht, A. (2009). The most influential journals: Impact Factor and Eigenfactor. *Proceedings of the National Academy of Sciences of the United States of America*.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (1985). *Sharing Research Data*.

Gardner, D., Toga, A. W., Ascoli, G. a, Beatty, J. T., Brinkley, J. F., Dale, A. M., Fox, P. T., et al. (2003). Towards effective and rewarding data sharing. *Neuroinformatics*, *1*(3), 289–95. doi:10.1385/NI:1:3:289

Garfield, E. (1955). Citation Indexes for Science through Association of Ideas. *Science*, *122*, 108–111.

GRDI2020. (2012). *Global Research Data Infrastructures: The Big Data Challenges. GRDI2020 Final Roadmap Report*. *Framework*.

Groves, T. (2009). Managing research data for future use. *Bmj*, *338*(mar24 2), b358–b358. doi:10.1136/bmj.b358

Groves, T. (2010). The wider concept of data sharing : view from the BMJ. *Biostatistics*, *11*(3), 391–392. doi:10.1136/bmj.b3928.G

Halevi, G., & Moed, H. F. (2012). The evolution of big data as a research and scientific topic: overview of the literature. *Research Trends*, (30). Retrieved from http://www.researchtrends.com/issue-30-

september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/
Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. (Tony Hey, S. Tansley, & K. Tolle, Eds.). Microsoft Research. Retrieved from http://www.amazon.com/dp/0982544200
Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC bioinformatics*, *12 Suppl 1*(Suppl 15), S3. doi:10.1186/1471-2105-12-S15-S3
Kowalczyk, S., & Shankar, K. (2010). Data Sharing in the Sciences. *Annual Review of Information Science and Technology*, *45*, 247–294.
Lawrence, B., Jones, C., & Matthews, B. (2011). The International Journal of Digital Curation Citation and Peer Review of Data : Moving Towards Formal Data Publication, *6*(2), 4–37.
Mayernik, M. S. (2012). *Bridging Data Lifecycles : Tracking Data Use via Data Citations Workshop Report*.
Meyer, E. T. (2011). *Splashes and Ripples Synthesizing the Evidence on the Impacts of Digital Resources*. London: JISC. Retrieved from http://ssrn.com/abstract=1846535
Milia, N., Congiu, A., Anagnostou, P., Montinaro, F., Capocasa, M., Sanna, E., & Destro Bisol, G. (2012). Mine, yours, ours? Sharing data on human genetic variation. *PloS one*, *7*(6), e37552. doi:10.1371/journal.pone.0037552
Mooney, H. (2011). Citing data sources in the social sciences: do authors do it? *Learned Publishing*, *24*(2), 99–108.
Mooney, H., & Newton, M. P. (2012). The Anatomy of a Data Citation: Discovery , Reuse , and Credit, *1*(1).
Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., Cockerill, M., et al. (2011). Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC bioinformatics*, *12 Suppl 1*(Suppl 15), S1. doi:10.1186/1471-2105-12-S15-S1
National Science Board. (2011). *National Science Foundation's Merit Review Criteria: Review and Revisions*.
Newton, M. P., Mooney, H., & Witt, M. (2010). A Description of Data Citation Instructions in Style Guides.
Nicholson, S. W., & Bennett, T. B. (2011). Data Sharing : Academic libraries and the scholarly enterprise, *11*(1), 505–516. doi:10.1353/pla.2011.0003
Parsons, M. A., & Fox, P. A. (2011). Is data publication the right metaphor? Kyoto (Japan): ICSU; Science Council of Japan.
Piwowar, H. (2012). Making data Count: tracking data impact through the scholarly literature and beyond. *Github*. Retrieved from https://raw.github.com/hpiwowar/datacitationstudy/master/citationpractices_knit_.md
Piwowar, H. a. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS one*, *6*(7), e18657. doi:10.1371/journal.pone.0018657
Piwowar, H. a, Becich, M. J., Bilofsky, H., & Crowley, R. S. (2008). Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS medicine*, *5*(9), e183. doi:10.1371/journal.pmed.0050183
Piwowar, H. a, Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one*, *2*(3), e308. doi:10.1371/journal.pone.0000308
Piwowar, H. A., & Chapman, W. W. (2008). Identifying Data Sharing in Biomedical Literature. *AMIA Annual Symposium Proceedings Archive*, 596–600.
Piwowar, H. A., & Vision, T. J. (2012). Data reuse and the open citation advantage. *Github*.
Priem, J., Piwowar, H. A., & Hemminger, B. H. (2011). Altmetrics in the wild : An exploratory study of impact metrics based on social media. *Metrics: Symposium on Informetric and Scientometric Research; 2011*. New Orleans, LA, USA. Retrieved from http://jasonpriem.org/self-archived/PLoS-altmetrics-sigmetrics11-abstract.pdf
Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv12034745v1 csDL 20 Mar 2012*, *1203.4745*, 1–23. Retrieved from http://arxiv.org/abs/1203.4745
Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*, *4*(9), e7078. doi:10.1371/journal.pone.0007078
Schäfer, A., Pampel, H., Pfeiffenberger, H., Dallmeier-Tiessen, S., Tissari, S., Darby, R., Giaretta, K., et al. (2011). *Baseline Report on Drivers and Barriers in Data Sharing* (pp. 1–75).
Stanley, B., & Stanley, M. (1988). Data Sharing. *Law and Human Behavior*, *12*(2), 173–180.

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network*.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., et al. (2011). Data sharing by scientists: practices and perceptions. *PloS one*, *6*(6), e21101. doi:10.1371/journal.pone.0021101
The Royal Society. (2012). *Science as an open enterprise*. The Royal Society Science Policy Centre.
Thomson Reuters. (2012). Collaborative Science: solving the issues of discovery, attribution and measurement in data sharing.
Torres-Salinas, D., Robinson-García, N., & Cabezas-Clavijo, Á. (2012). Compartir los datos de investigación en ciencia: introducción al *data sharing*. *El Profesional de la Informacion*, *21*(2), 173–184. doi:10.3145/epi.2012.mar.08
Uhlir, P. F. (2012). *For Attribution — Developing Data Attribution and Citation Practices and Standards Summary of an International Workshop*.
Van der Graaf, M., & Waaijers, L. (2012). *A surfboard for riding the wave: towards a four country action program on research data* (p. 39).
Van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer, 1–5.
Van Eck, N. J., Waltman, L., Dekker, R., & Van Den Berg, J. (2010). A Comparison of Two Techniques for Bibliometric Mapping: Multidimensional Scaling and VOS. *Journal of the American Society for Information Science*, *61*(12), 2405–2416. doi:10.1002/asi
Waltman, L., & Eck, N. J. Van. (2009). A Taxonomy of Bibliometric Performance Indicators Based on the Property of Consistenc. *ERIM Report Series Research in Management*.
Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. doi:10.1016/j.joi.2010.07.002
Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2010). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47. Retrieved from http://arxiv.org/abs/1003.2167
Weingart, P. (2005). Impact of bibliometrics upon the science system: inadvertent consequences? *Scientometrics*, *62*(1), 117–131.
Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728.
Wouters, P., Beaulieu, A., Scharnhorst, A., & Wyatt, S. (2013). *Virtual Knowledge*. The MIT Press.
Wouters, P., & Costas, R. (2012). *Users , narcissism and control – tracking the impact of scholarly publications in the 21 st century*. (M. Van Berchum & K. Russell, Eds.)*Image Rochester NY* (p. 50 pages). SURFfoundation. Retrieved from http://www.surffoundation.nl/en/publicaties/Pages/Users_narcissism_control.aspx
Wouters, P., & Schröder, P. (Eds.). (2003). *Promise and practice in data sharing: the public domain of digital research data*. The Netherlands: NIWI-KNAW.

# Appendices

## Appendix 1. Checklist for the analysis of the different repositories

| | | |
|---|---|---|
| **General Features (1)** | **Title** | General title of the repository |
| | **URL** | URL of the repository |
| | **Authority** | Indicates the organisation which maintains or holds the repository |
| | **Subject** | Main subjects covered by the repository |
| | **Discipline** | General subject area/field of the repository |
| | **Description** | General information about the repository |
| | **Access & download** | Possibility of accessing and downloading the data (e.g. whether it is free or by subscription/public or restricted/whether needs registration or not) |
| | **Start Date** | Start year of the repository |
| | **Location** | Geographical location of the repository |
| | **Reuse** | Possibility of reuse of data in the repository (e.g. open or closed, free or copyright protected) |
| | **Deposit** | Possibility of depositing the data to the repository (by public or affiliated personnel) |
| | **Type** | Organisation type of the repository |
| **Access, usage, validation & metrics (2)** | **Possibility of search/browse** | Searching or browsing the repository |
| | **Data availability** | Amount of data available |
| | **Data type/format** | Type/format of the data available |
| | **Data validation** | Review or quality control process of the datasets in the repository before publishing on the repository |
| | **Presence of Resource identifier** | Any identifier assigned to the dataset |
| | **Metrics Available** | Providing statistics/graphs/charts about the view/download and usage of the content of the repository |
| | **Refers to standard citation format** | Citation guidelines/recommendations to cite the content of the repository |

(1) General information about the sampled repositories retrieved from Databib.
(2) Information regarding the possibilities of use, download and citing the content of the sampled repositories.

## Appendix 2. Data repositories analysed – “Access, usage, validation & metrics” features only

| Title | National Digital Archive of Datasets | Council of European Social Sciences Data Archives (CESSDA) | DataFirst | 3TU.Datacentrum | Open Data Pilot Project | DataStaR | Association of Religion Data Archives (ARDA) | Chemical Database Service, The | National Nuclear Data Center |
|---|---|---|---|---|---|---|---|---|---|
| URL | http://www.nationalarchives.gov.uk/documentsonline/datasets.asp | http://www.cessda.org/ | http://www.datafirst.uct.ac.za/ | http://datacentrum.3tu.nl | http://www.data.gc.ca/default.asp?lang=En | http://datastar.mannlib.cornell.edu/ | http://www.thearda.com/ | http://cds.dl.ac.uk/ | http://www.nndc.bnl.gov/ |
| Possibility of search/ browse | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes (by subscription) | not found |
| Data availability & Data type/ format | Over 11 million historical government and public records (paper and parchment, electronic records and websites, photographs, posters, maps, drawings and paintings from Domesday Book to modern government papers and digital files) | Provides access to 25,000 data collections, delivering over 70,000 data collections per annum and acquiring a further 1,000 data collections each year, including sociological surveys, election studies, longitudinal studies, opinion polls, and census data. | Over 100 datasets are available publicly, many from Statistics South Africa.African census and Survey datasets includes census as well as books and journals on survey research, and statistical publications. | Over 5000 datasets | 273054 datasets are available | | Collection of surveys, polls, and other data submitted by researchers and made available online by the ARDA | | Nuclear structure properties of nuclides and the results of various experiments to derive those properties and made up of a collection of datasets |
| Data validation | not found | yes | yes | yes | not found | not found | not found | not found | not found | yes |
| Presence of Resource identifier | Reference code/hierarchy | not found | Reference ID (e.g.:eth-aau-aais-1993-v1) | At least one identifier (usually starting with *uuid:* and also DOIs) used as part of the URL. | Department Generated Identifier (e.g. AAFC-AIMIS-RP-1763 ) | Identifiers at the level of datasets and DC definition. | not found | not found | not found | Reference code and mass numbers |
| Metrics Available | not found | not found | not found | yes (1) | yes (1) | yes | yes (1) | not found | not found | yes (1) |
| Standard citation format | yes | not found | yes | not found | not found | not found | not found | yes | not found | yes |

| Title | National Archive of Criminal Justice Data | Speech and Language Data Repository | Advanced Cooperative Arctic Data and Information Service (ACADIS) | Protein Data Bank (PDB) | Biological Magnetic Resonance Data Bank | GenBank | Universal Protein (UniProt) | ArrayExpress [EBI] | Dryad | PRoteomics IDEntifications database |
|---|---|---|---|---|---|---|---|---|---|---|
| **URL** | http://www.icpsr.umich.edu/icpsrweb/NACJD/index.jsp | http://www.sldr.org | http://www.aoncadis.org | http://www.rcsb.org | http://www.bmrb.wisc.edu/ | http://www.ncbi.nlm.nih.gov/genbank | http://www.ebi.ac.uk/uniprot/index.html | http://www.ebi.ac.uk/arrayexpress/ | http://datadryad.org/ | http://www.ebi.ac.uk/pride/ |
| **Possibility of search/ browse** | Yes | Yes (with SLDR) | Yes | Yes | Yes | Yes | not found | Yes | Yes | Yes |
| **Data availability & Data type/ format** | Data on community studies, corrections, courts, drugs and alcohol, and police. | Tools for linguistic research; sound/video/image/ text corpora and any language-related signal; annotations of corpora, lexicons, reference databases, systems of representation, grammars etc. | Long-term observational timeseries, local-, regional-, and system-scale research, datasets and projects from many diverse domains | The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome | NMR results from peptides, proteins, and nucleic acids | Over 135,000,000 sequence records and is updated every two months. There are approximately 126,551,501,141 bases in 135,440,924 sequence records. | Curated protein information, including function, classification, and cross-reference. | Genomes | Over 2615 data packages and 7172 data files, associated with articles in 189 journals | 27020 Experiments; 11750365 Identified Proteins; 65288118 Identified Peptides; 5941667 Unique Peptides; 342784192 Spectra |
| **Data validation** | not found | yes | not found | yes | yes | yes | not found | not found | yes | yes |
| **Presence of Resource identifier** | study no | Sldr ids, Persistent Identifiers (PIDs), and Handles (e.g. hdl:11041/sldr000027 points to item sldr000027) | Project award no. and DOI for datasets | DOI and PDB IDs | Accession number and Deposition code | Sequence identifiers and annotation/accession numbers | Database identifier | Accession number | DRYAD DOI | Accession numbers |
| **Metrics Available** | yes (1) | not found | not found | yes (1) | yes | yes (1) | yes | yes | yes | not found |
| **Standard citation format** | not found | not found | not found | yes | yes | yes | not found | not found | yes | yes |

| Title | clinicaltrials | Energy Information Administration | TalkBank | History Data Service (HDS) | VectorBase | Data Bank, The | World Data Center | SeaDataNet | Cultural Policy and the Arts National Data Archive | PANGAEA - Data Publisher |
|---|---|---|---|---|---|---|---|---|---|---|
| URL | http://clinicaltrials.gov/ | http://www.eia.gov/ | http://talkbank.org/ | http://hds.essex.ac.uk/ | http://www.vectorbase.org/ | http://www.npdb-hipdb.hrsa.gov/ | http://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=wdc | http://www.seadatanet.org | http://www.cpanda.org/cpanda/ | http://www.pangaea.de/ |
| Possibility of search/ browse | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Data availability & Data type/ format | 137,949 studies contains study and often result in a standard, tabular formatting | Reports, web products, press releases, databases, and maps. | Transcripts, audio and video of communicative interactions for research in human and animal communication. | Over 650 separate studies transcribed, scanned or compiled from historical sources, in a range of different formats (Tabular data, Textual data, Images). | Sequence data, gene expression data, images, population data, and insecticide resistance data for arthropod vectors on Invertebrate Vectors of Human Pathogens | Data for statistical analysis in two formats: Tab Delimited and SPSS Version | Over 30,000 of these metadata descriptions of datasets, data services and ancillary descriptions from numerous government agencies, research institutions, archives and universities worldwide. | Meta-data and products deriving from in situ and remote observation of the seas and oceans. | Datasets, bibliographic materials, original analyses, and links to other web resources | Various scientific data types are archived with detailed description in a relational database. The description of each dataset is always visible and includes the principle investigator (PI) who may be asked for access. |
| Data validation | yes | yes | not found | yes | not found | yes | not found | yes | not found | yes |
| Presence of Resource identifier | Every clinical study associated with identification numbers, including identifiers from other registries and NIH grant numbers. | MSN (Comprehensive Data Series' Names and Descriptions ) and state codes (Two-character State Codes) | not found | not found | Internal identifiers. | Data Bank Control Number (DCN) is used to identify each query and report. | Metadata uuid and entry ID for datasets | not found | not found | DOI |
| Metrics Available | yes (1) | yes (1) | not found | not found | yes | yes (1) | not found | not found | not found | not found |
| Standard citation format | not found | yes | yes | not found | yes | not found | yes | not found | yes | yes |

| Title | GigaScience | UK Data Archive (UKDA) | World Bank Data Catalog | Figshare | National Center for Educational Statistics | VegBank |
|---|---|---|---|---|---|---|
| URL | http://gigadb.org/ | http://www.data-archive.ac.uk/ | http://data.worldbank.org/data-catalog/ | http://figshare.com | http://nces.ed.gov/ | http://vegbank.org/vegbank/index.jsp |
| Possibility of search/ browse | Yes | Yes | Yes | Yes | Yes | Yes |
| Data availability & Data type/ format | Biological data producers, integrates manuscript publication with a database contains both genomic and non-genomic datasets. | Research resources including key national and international survey data, qualitative data and a number of data services. | Access to World Bank data listing of available World Bank datasets, including databases, pre-formatted tables, reports, and other resources. | Figures, media (including video), papers (including pre-prints), posters and filesets (groups of files). | Publications and datasets including early releases, issue briefs, statistical reports, directories, and handbooks of standard terminology. | Plot records, vegetation types, and plant taxa. |
| Data validation | yes | yes | yes | not found | not found | yes |
| Presence of Resource identifier | DOI | DOI | Reference ID | DOI | not found | Accession Code |
| Metrics Available | not found | yes | not found | yes (1) | not found | not found |
| Standard citation format | yes | yes | yes | yes | not found | yes |

(1) Top counts or aggregated statistics on datasets downloads or usage.

## Appendix 3. Data Journals

Below is a non-exhaustive list of data journals in no particular order:

- Geoscience Data Journal http://www.geosciencedata.com
- Earth System Science Data http://earth-system-science-data.net/
- Ecological Archives - Data Papers http://esapubs.org/archive/archive_D.htm
- Hindawi publishing: http://www.datasets.com/
    - Dataset Papers in Agriculture
    - Dataset Papers in Biology
    - Dataset Papers in Chemistry
    - Dataset Papers in Ecology
    - Dataset Papers in Geosciences
    - Dataset Papers in Materials Science
    - Dataset Papers in Medicine
    - Dataset Papers in Nanotechnology
    - Dataset Papers in Neuroscience
    - Dataset Papers in Pharmacology
    - Dataset Papers in Physics
- Journal of Chemical and Engineering Data http://pubs.acs.org/journal/jceaax
- GigaScience http://www.gigasciencejournal.com/
- Journal of Physical and Chemical ResearchData http://jpcrd.aip.org/resource/1/jpcrbu
- Biodiversity Data Journal http://www.pensoft.net/journals/bdj/
- F1000 Research http://f1000research.com
- International Journal of Robotics Research http://ijr.sagepub.com/
- CODATA's Data Science Journal http://www.codata.org/dsj/index.html
- BMC Research Notes http://www.biomedcentral.com/bmcresnotes/
- Geoscientific Model Development (GMD) http://www.geoscientific-model-development.net/
- Journal of Open Archaeology Data: http://openarchaeologydata.metajnl.com/
- Journal of Open Public Health Data: http://openpublichealthdata.metajnl.com/
- Journal of Open Psychology Data: http://openpsychologydata.metajnl.com/
- Journal of Open Research Software: http://openresearchsoftware.metajnl.com/

## Appendix 4. Tools with relevance for data collection, curation, dissemination and citation

| | |
|---|---|
| Collection | DMP Tool (https://dmp.cdlib.org/)<br>DATA UP (http://dataup.cdlib.org/)<br>Cloud (web archiving) services |
| Curation | UC3 EZID (http://www.cdlib.org/services/uc3/ezid/)<br>e-scholarship (http://www.escholarship.org/)<br>UC3 Merritt (https://merritt.cdlib.org/) |
| Dissemination | ONE Share<br>DataONE (http://www.dataone.org/) |
| Use | DataCite (http://datacite.org/)<br>CrossRef (http://www.crossref.org/)<br>ORCID (http://about.orcid.org/)<br>Data Citation Index<br>(http://wokinfo.com/products_tools/multidisciplinary/dci/) |

## Appendix 5. Bibliometric mapping

The mapping methods presented in the term map using the VOSviewer are based on methods with a stronger mathematical and statistical foundations. Some of these characteristics are described below (together with the relevant related literature):

- VOS mapping technique (Van Eck, Waltman, Dekker, & Van Den Berg, 2010). Given a set of objects and a matrix that indicates for each pair of objects the strength of their relation (e.g., the number of co-occurrences), the VOS mapping technique locates the objects in a two-dimensional space in such a way that strongly related objects are located close to each other while less strongly related objects are located further away from each other. Because only two dimensions are available, distances usually do not give a perfect representation of the relatedness of objects, but the VOS mapping technique aims to provide a representation that is as accurate as possible. The VOS mapping technique provides an alternative to multidimensional scaling, which is a well-known technique from the statistical literature.
- VOS clustering technique (Waltman, Van Eck, & Noyons, 2010). Given a set of objects and a matrix that indicates for each pair of objects the strength of their relation (e.g., the number of co-occurrences), the VOS clustering technique assigns the objects to clusters. A so-called resolution parameter determines the level of detail of the clustering, where a higher level of detail means that there are more clusters. The VOS clustering technique is based on the same underlying mathematical principle as the VOS mapping technique, and therefore these two techniques together provide a unified framework for mapping and clustering. The use of such a unified framework is quite uncommon. Researchers often combine mapping and clustering techniques in a single analysis, but usually these techniques are based on very different principles and assumptions.
- Term identification techniques. We have developed our own techniques for identifying terms in texts, for instance in titles and abstracts of scientific publications. We first use natural language processing techniques for part-of-speech tagging, or in other words, for identifying nouns, adjectives, verbs, etc. When then convert plural nouns into singular ones. Next, we identify sequences of nouns and adjectives ending with a noun. We regard these sequences as terms. Finally, we have a technique that aims to determine which of the identified terms are most relevant in a given context (e.g., ‘result’ and ‘conclusion’ are usually not very relevant terms in a scientific context; they can be found in almost every scientific publication). A short description of our term identification techniques is available in (Van Eck & Waltman, 2011).

## Appendix 6. Interviewed stakeholders

| Name interviewee | Affiliation |
|---|---|
| Falk Reckling | FWF, Austria |
| John Wood | former chair ESFRI, EU and Secretary General of the Association of Commonwealth Universities; |
| Tom Heath | Open Data Institute, UK |
| Heather Piwowar | National Evolutionary Synthesis Center (NESCent), Durham, North Carolina, USA. Also co-founder of *ImpactStory* |
| Michael Diepenbroek | Pangaea & WDS |
| Sarah Callaghan | NCAS British Atmospheric Data Centre, UK |
| Peter Doorn | DANS, NL |
| Eefke Smit | STM, Dryad and WDS |
| Irina Sens | TIB (German Library) & DataCite |
| *Nigel Robinson (invited but no reply)* | *Data Citation Index, Thomson Reuters* |

This report presents a first landscape study of creating and promoting data metrics and the assessment of the use of datasets in scientific work as a tool to stimulate data sharing. This report will be of interest to all stakeholders in science and research, such as individual scientists, funders, research institutions, universities and data centres. It provides them with more knowledge about tools to promote and reward data sharing and data publication within their scientific communities. In this landscape study, all major stakeholders have been considered in order to summarise the main views, problems and challenges that need to be tackled in the development of metrics for datasets, and in the generalisation and promotion of data sharing activities.

Published April 2013

Download:
www.knowledge-exchange.info/datametrics