This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

Data Journals: A Survey

Leonardo Candela, Donatella Castelli, Paolo Manghi, and Alice Tani
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Italian National Research Council, via G. Moruzzi, 1, 56124, Pisa, Italy. E-mail: {candela, castelli, manghi, tani}@isti.cnr.it

Data occupy a key role in our information society. However, although the amount of published data continues to grow and terms like "data deluge" and "big data" today characterize numerous (research) initiatives, a lot of work is still needed in the direction of publishing data in order to make them effectively discoverable, available, and reusable by others. Several barriers hinder data publishing, from lack of attribution and rewards, vague citation practices, quality issues, to a rather general lack of data sharing culture. Lately, data journals came forward as a solution to overcome some of these barriers. In this study of more than 100 currently existing data journals, we describe the approaches they promote for datasets description, availability, citation, quality and open access. We close by identifying ways to expand and strengthen the data journals approach as a means to actually promote datasets access and exploitation.

Introduction

Data – that serving "*Big Science*" as well as that serving "*Long-tail Science*" (Murray-Rust, 2008) – is emerging as a driving instrument in science. Benefitting from data availability researchers are envisaging a large variety of new research patterns that are revolutionizing how science is being conducted. The full realization of this paradigm shift, however, requires addressing many onerous and challenging issues (Bell, Hey, & Szalay, 2009; Hey, Tansley, & Tolle, 2009; Halevi & Moed, 2012).

Although there is an almost universal agreement on the benefits of "data sharing and re-use" as a means to accelerate science performance, there are a number of barriers hindering the realization of this objective in a systematic and effective way (Borgman, 2011; Tenopir et al., 2011; Pampel & Dallmeier-Tiessen, 2014). These barriers are methodological, legal, technical, and often related to the lack of incentives for researchers to share their data (Douglass, Allard, Tenopir, Wu, & Frame, 2014; Asher et al., 2013; Bourne et al., 2012; Bourne, 2010). The effects of these obstacles on science is deleterious, e.g., Vines et al. (2014) demonstrates how the availability of research data was strongly affected by article's age when no policy is in place. Thus, proper data sharing practices and policies must be introduced to foster the data availability. Moreover, mechanisms must be identified to make the scientific community aware of the available data sets, to facilitate their understanding and to foster their effective re-use.

In this changing landscape *Data Journals* have been proposed as first step solution to some of the above discussed barriers. They realize the "data publication" concept by mirroring the scientific publication model. They promote the publication of *data papers*, "scholarly publication of a searchable metadata document describing a particular on-line accessible dataset, or a group of datasets, published in accordance to the standard academic practices" (Chavan & Penev, 2011). Their final aim being to provide "information on the what, where, why, how and who of the data" (Callaghan et al., 2012). Thus data publication is a pre-requisite to enable data sharing and reuse.

Despite their potentiality, data journals are not the ultimate and complete solution for all the data sharing and reuse issues and, in some cases, they are considered to induce false expectations in the research community (Parsons & Fox, 2013).

In this survey we review current data journals to discuss the different approaches put in place to overcome the data sharing barriers. In particular, the rest of the survey is structured as follows. The rationale, motivations and initiatives leading to data journals are described. Then a survey of over 100 data journals is discussed by comparing their approaches to data papers concept implementation including how to properly describe a dataset, how to promote datasets availability, how to properly cite a dataset and guarantee rewards, how to guarantee dataset quality, and how to guarantee open access to dataset. The paper ends by giving some suggestions aiming at enhancing the role of data journals as a true data sharing means.

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

The beginnings

Data sharing, intended as the release of (research) data for use by others, is a discussion topic since the 1980's (Fienberg, Martin, & Straf, 1985; Borgman, 2011; Costas, Meijer, Zahedi, & Wouters, 2013). Data "publication" is an approach that creates and curates data as first class objects (Klump et al., 2006; B. Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011; Costello, Michener, Gahegan, Zhang, & Bourne, 2013).

The issues related to formally publishing and citing datasets are clearly enumerated and discussed by others, e.g., B. Lawrence et al. (2011); Callaghan et al. (2012). The aim is to have datasets as a "first class research output" that will be available, peer-reviewed, citable, easily discoverable and reusable. Proposed plans for data publication identified by Callaghan et al. (2012) "involve working with academic publishers to develop a new style of article: a data paper, which would describe the dataset, providing information on the what, where, why, how and who of the data. The data paper would contain a link back (a DOI) to the dataset in its repository, and the journal publishers would not actually host the data. This means that even in situations where the data paper might be restricted access, the dataset could still be open".

In fact, in the data publication approach there are three main actors with different perspectives: researchers, publishers and Data centres/Libraries. From an analysis conducted by Reilly, Schallier, Schrimpf, Smit, and Wilkinson (2011) it emerged that *(a)* researchers were fully aware of the benefits and value of sharing data, yet they need facilities for storing and maintaining data, for controlling access, for getting credits, and for someone that pays the costs for this; *(b)* publishers were willing to embrace the approach, yet they are struggling with the costs and alternatives for data management (e.g., data as supplementary files, data in external repositories, journals dedicated to "data papers" only); *(c)* Data centres and Libraries were fully aware of their mission of research output custodians, yet they need to reconsider their mission in modern scientific communication.

The early attempts of publishers to support data publication were: *(a)* data are an integral part of the paper, or *(b)* data resides in supplementary files attached to the paper. The first approach is the traditional one and it is affected by a number of drawbacks including the difficulties in separating the data from the rest and re-use them. The second approach goes beyond the motivation to share data. In fact, around 2009 the large majority of journals were accepting data (and other material) as supplementary files to be "published" in the online version of research articles only, often under heavy restrictions on volume and total number of supplementary items as well as heavy conditions on copyright (Reilly et al., 2011). The drawback of this publishing model is that it requires curation and preservation of such files and lacks allowing readers to find and link data independently of the main publication.

Because of these limitations, the need to establish a new data publishing paradigm based on the concept of "data paper" started to be generally recognized (Kunze et al., 2011), especially in the biodiversity community (Chavan & Penev, 2011). Publishers have readily accepted to deal with this emerging needs. At the 8th Research Data Management Forum, R. Lawrence (2012) affirmed that "there is now a general consensus that sharing and publishing data is good" and "each stakeholder group has made some steps forward". But the same does not happen for the idea of publishing journals exclusively dedicated to data papers. There are reasons why editors are so cautiously accepting this idea. Some of these reasons have been discussed in the same forum with contributions by speakers from Nature, Elsevier, Dryad, the International Union of Crystallography and Faculty of 1000. One central point was the economic risk of going to publish new journals before their acceptance is assured. Tempest (2012) affirmed that publishers recognize that scientists' investment in creating and interpreting data, and their intellectual and financial contributions need to be recognized and valued. However, he added that where publishers "add value and/or incur significant cost" then their "contributions also need to be recognized and valued". He also resumed the pros and cons in making data available and accessible concluding that publishers were investing in innovation, but there were still issues and still a lot of work to do. A slight different position of "small" publishers was brought by Wilson (2012) of the Nature Publishing Group. She terminated her presentation stating that "there needs partnership by institutions, repositories, publishers, researchers and funders, even though roles are to be well established and business models well determined". Similarly, an editorial of Nature Genetics solicited interested people to be careful to avoid potential problems ("It's not about the data", 2012).

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

However, a number of initiatives supporting data papers were arising. Callaghan, Hewer, Pepler, Hardaker, and Gadian (2009) investigated the idea of an *overlay journal*, i.e., a journal that "consists of a number of overlay documents, which are structure documents created to annotate another resource with information on the quality of the resource". Each overlay document was expected to contain "*(i)* metadata about the overlay document itself; *(ii)* information about and from the quality process for which the document was constructed; and *(iii)* basic metadata from the referenced resource to aid discovery and identification". Newman and Corke (2009) announced that The International Journal of Robotics Research started soliciting a new genre of paper, a "data paper". In their editorial, they underlined that their primary goal was "to facilitate and encourage the release of high quality peer-reviewed datasets to the robotics community" as well as to help authors "to publish and gain credit for their valuable data" since "data papers will be treated in the same fashion as regular papers". Pfeiffenberger and Carlson (2011) launched Earth System Science Data, a journal specifically conceived to data papers to "provide reward for data "authors" through fully qualified citation of research data, classically aligned with the certification of quality of a peer reviewed journal". Chavan and Penev (2011) promoted "biodiversity data papers" to incentivize data publication. Kennedy, Ascoli, and De Schutter (2011) promoted a "Data Original Article" in neuroscience research to "support the publication of high quality, richly reusable, fully described data". This realized the vision that "promotes the primacy of data in the scientific endeavor" (De Schutter, 2010). In April 2013, Nature Publishing Group announced the launch of Scientific Data (Scheer, 2013), an open-access journal for the publication of descriptions of scientifically valuable datasets.

The number of Data Journals is rapidly growing, thus the time is ripe for an analysis of the approaches and trends that publishers and journals are implementing for data publication.

The data journals landscape

A number of initiatives have been launched to realize data journals in various domains ranging from archeology to chemistry, ecology and oceanography. To identify the journals target of this investigation we conducted a Web-based inventory study. In particular, we identified an initial set of journals through Google searches and then supplemented by investigating the "related links" pages. The Web sites of these journals were analysed to identify the core characteristics. When necessary, the editorial teams were contacted to clarify doubts and acquire additional information from their valuable and informative feedback.

For this study, 116 data journals published by 15 different publishers have been identified. We have clustered journals into sets corresponding to publishers since it is quite common that publishers use shared approaches and policies for their journals. This is the case of *(i) BioMed Central Journals*, a set of 85 data journals published by BioMed Central, *(ii) Chemistry Central Journals*, a set of 3 data journals published by Chemistry Central, *(iii) Pensoft Journals*, a set of 7 data journals published by Pensoft Publishers, *(iv) SpringerOpen Journals*, a set of 8 data journals published by SpringerOpen, and *(v) Ubiquity Press Journals*, a set of 3 data journals published by Ubiquity Press.

Although software can be seen as data, e.g., Marcus and Menzies (2010), and this leads to "Software Journals" supporting "Software papers" such as *Chemistry Central Journal* and *SpringerOpen*, in this survey such a type of journals / papers is not included.

An overview of the studied journals is given in Table 1. The table reports for each journal or set of journals: the *nature*, i.e., whether the journal publishes only data papers ("pure") or any typology of paper including data papers ("mixed"); the *subject extent*, i.e., the number of subjects covered; the *exploitation*, i.e., an indicator of the amount of data papers currently published (up to December 2013); the *offering*, i.e., the number of journals supporting data papers (up to December 2013); the *index*, i.e., the number of journals indexed by professional services (e.g. we used Thomson Reuters Web of Science); the *length*, i.e., an average "size" for data papers ("n" stands for "normal" meaning that there is no limitation for data papers, "s" stands for "small" meaning that data papers are expected to be up to 4 pages); the *open access nature*, i.e., whether the journal is an open access one or not. The entire list of journals including a description of each journal and a reference to its website is given in a supplemental file associated with the paper.

Table 1. Generic characteristics of analysed journals. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}:

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

Ecology; J_{F1}: F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}: International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals. The Nature of journals is indicated as follows: "m" for mixed and "p" for pure. Length of data papers is indicated as follows: "n" for normal and "s" for short.

Characteristic	J_{BM}	J_{CC}	J_{DP}	J_{ES}	J_{Ec}	J_{F1}	J_{GD}	J_{GS}	J_{RR}	J_{Ni}	J_{Pe}	J_{PO}	J_{SD}	J_{SO}	J_{UP}
Nature	m	m	p	p	m	m	m	p	m	m	m	m	p	m	p
Subject extent	15	4	3	1	1	2	1	1	3	2	2	3	3	9	3
Exploitation	534	5	39	79	89	7	1	3	10	2	36	0	0	3	18
Offering	85	3	1	1	1	1	1	1	1	1	7	1	1	8	3
Indexed	72	2	0	0	1	0	0	0	1	1	2	0	0	2	0
Length	n	n	n	n	s	n	s	n	s	n	n	n	n	n	s
Open access	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓

In terms of nature, the large majority of journals are mixed (cf. Fig. 1). Only 4 out of the 15 considered publishers release journals exclusively dedicated to data papers. The overall number of pure data journals is 7 (6% of the sample).

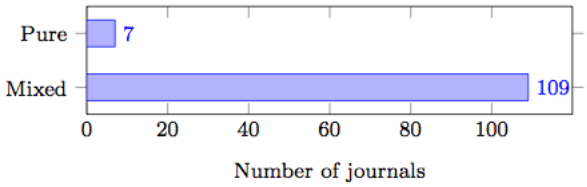

Figure 1. Nature of data journals.

In terms of subject extent, journals cover all the four Scopus subject clusters, i.e., Health Science, Life Sciences, Physical Sciences, and Social Sciences & Humanities (cf. Tab. 2). In particular, by analysing the Scopus journals classification (and complementing it for the missing titles) it emerges that the three most represented subjects (in terms of number of journals) are *Medicine* (52.67%), *Biochemistry*, *Genomics and Molecular Biology* (25.89%), and *Agricultural and Biological Sciences* (16.07%) (cf. Fig. 2). However, these figures are partially biased by the number of journals a publisher supports. In fact, the three most represented subjects (in terms of number of publishers) are *Medicine* (46.66%), *Biochemistry, Genomics and Molecular Biology* (33.33%), and four subjects covered by the 13.33% of the journals, namely *(i) Immunology and Microbiology*, *(ii) Mathematics*, *(iii) Pharmacology, Toxicology and Pharmaceutics*, and *(iv) Psychology*.

Table 2. Number of journals by subject (Scopus classification).

Subject	Number of journals
Health Sciences	66
Life Sciences	59
Physical Sciences	28
Social Sciences & Humanities	9
Multidisciplinary	1

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

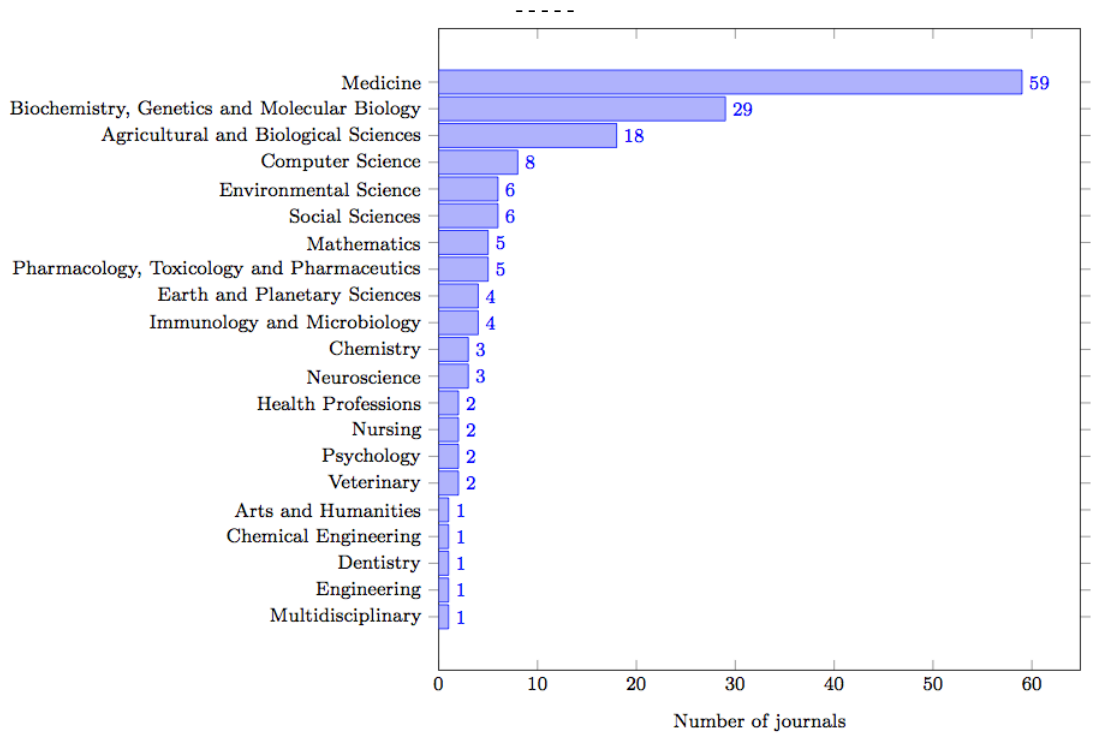

Figure 2. Number of published journals by topic.

In terms of exploitation, from 2000 up to 2013 a total of 826 data papers have been published. The number of data papers published by year is growing (cf. Fig. 3), in the last year the 23.5% of the total amount of currently existing data papers was published. This trend seems to continue, in fact in the first month of 2014 a total amount of 47 data papers have been already published.

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

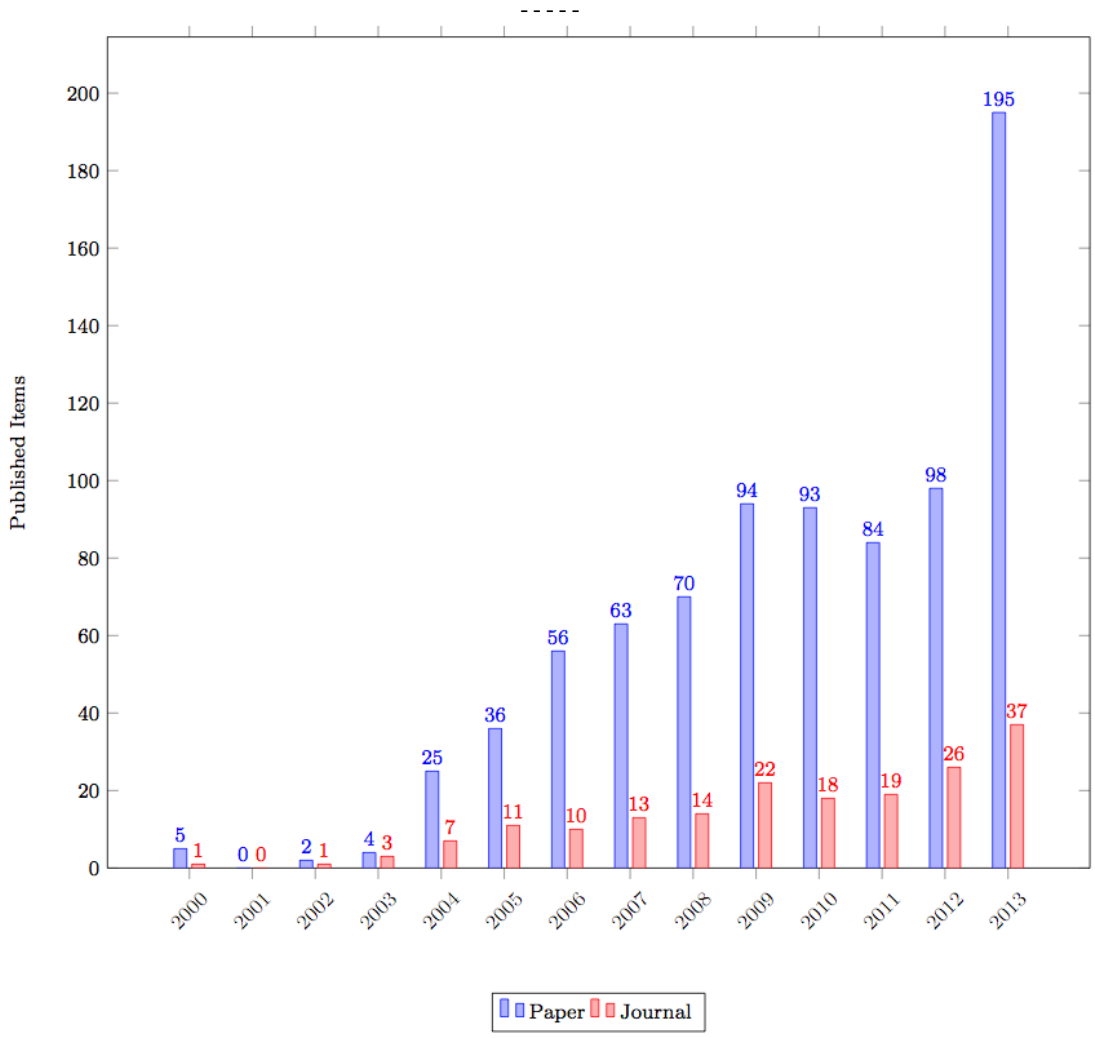

Figure 3. Number of published papers and operational data journals, by year.

In terms of offering, although we found that there are 116 Data Journals promoting data papers, only 60 of them have published at least one data paper in the period January 2000 – December 2013. However, the number of journals publishing at least a data paper is growing (cf. Fig. 3). In 2013, 37 diverse journals published 195 data papers while in the first month of 2014 9 diverse journals have published 47 data papers.

In terms of index, the 69.82% of the journals of the sample are indexed by Thomson Reuters (cf. Fig. 4). However, this figure is biased by the presence of "mixed" journals in the sample and it is very difficult to derive an indication for data papers. The only observation that can be safely reported is that no one of the "pure" journals is yet indexed either by this professional service, or by other services such as SCImago or Scopus.

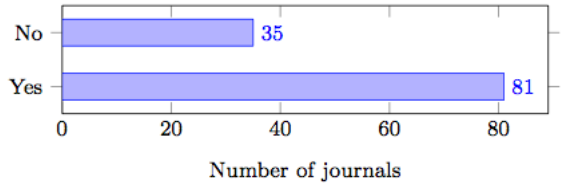

Figure 4. Number of data journals indexed by Thomson Reuters.

In terms of paper size, because of the fact that the majority of data journals are actually mixed journals there is no special arrangement for data papers including for the number of pages. A few journals only (6 out of the 116 analysed) – namely Ecology, Genomics Data, International Journal of Robotics Research and Ubiquity Press Journals – envisage data papers as artefacts consisting of few

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

pages. This somehow indicates that journals tend to provide authors of data papers with how many space as they need to properly describe their datasets and the process leading to them.

In terms of open access nature, almost all the analysed data journals are open access (cf. Fig. 5). Only three are considered to be non open access journals: *International Journal of Robotics Research* that is based on subscription fees but data papers are freely available, *Ecology* that is based on subscription fees, and *Neuroinformatics* journal that is offering the "open access choice" with a fee of €2,200. More on this aspect will be discussed later in the paper.

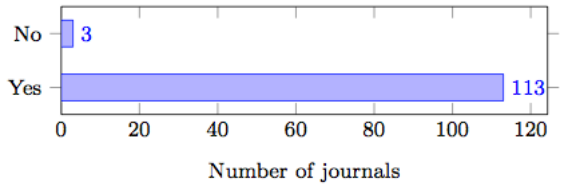

Figure 5. Number of open access data journals.

To analyse the effectiveness of data journals as means promoting data publication and reuse, a model describing major concepts and relationships is introduced in the next section along with a discussion on the different names used by diverse journals to refer to them. After that, the following problems affecting data publication and reuse are discussed in dedicated sections: how to properly describe a dataset, how to promote datasets availability, how to properly cite a dataset and guarantee rewards, how to guarantee dataset quality, and how to guarantee open access to dataset. Each of these sections end in a "observations" paragraph where a critical comment on the discussed approaches is given.

Data paper concepts and naming

The data paper landscape is highly varied and many are the different understandings and implementations of this concept. In order to better deal with this variety, we introduce a simple model for this concept (cf. Fig. 6) that will enable us to describe and analyse the different solutions in a more uniform way.

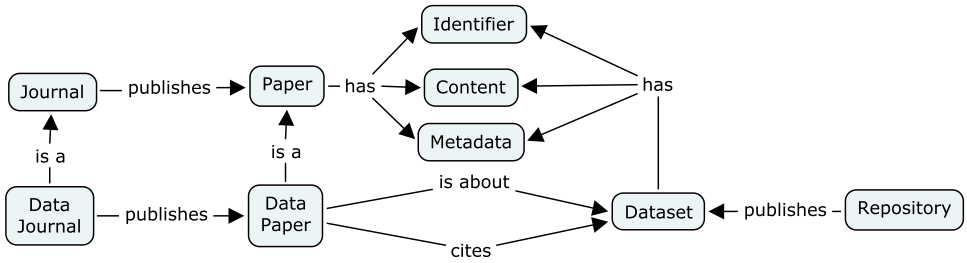

Figure 6. Data papers concept map.

The concept of data paper has at least two elements that have to be materialized into concrete and identifiable information objects in order to fully implement it: the *dataset*, i.e., the subject of the data paper, and the *data paper* itself, i.e., the artefact produced to describe the dataset. From this point on, the term data paper will be used to refer to the artefact only. This artefact is homologous to papers for traditional journals, it is expected to have an *Identifier* and a *Content* with title, authors, abstract, a number of sections and references.

Both the data paper and the dataset are associated with other information objects, their *metadata*. Metadata bring additional information that is useful for the management of the corresponding primary object. The metadata format and content is usually selected by the entity managing the two elements, i.e., the data journal editor for the data paper, and the data archive manager for the dataset (and often results in proprietary, ad-hoc-solutions).

Note that data paper metadata has not to be confused with data paper content. In some cases the confusion may be originated from the understanding of the content of a data paper as a sort of meta-data description of the dataset itself. The data paper is instead analogous to the traditional research papers. As such it can be processed by any of the traditional tools and services available for research papers, including those dedicated to indexing and citation analysis.

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

Different journals implement diverse approaches for the management of the data papers. The data journal publishes data papers, meaning that the data papers are primary objects of concerns for journals. Datasets are considered as secondary objects maintained in repositories that may be managed by the same journal editor or, most often, by third-party organizations specialized on data archiving. As a consequence of this state of affairs and of the lack of standards in this area, there is a large heterogeneity in the format of the metadata associated with the components, each reflecting the strategies of the corresponding managing entity, and in the technical solutions adopted.

Although "data paper" is the most commonly used name for the typology of artefacts that will be discussed in this paper, different journals call them differently often as a consequence of naming choices done when this concept was not as popular as it is now (cf. Tab. 3 and Fig. 7). The alternative names often also reflect specific purposes the data papers have been introduced for or specific understanding of them. Some examples are given below with reference to the journal exploiting them and the article intended purpose: "*Data article*" – in International Journal of Food Contamination is the main article type since the scope of the journal is to publish important data on prevalence and concentration of different food contaminants; "Dataset paper" – in Dataset Papers in Science is for papers describing any dataset; "Data descriptors" – in Scientific Data is for descriptions of scientifically valuable datasets; "Data in brief " – in Genomic Data is for detailed descriptions of genomic data including experimental methods and any quality control analysis; "Data note" – in BioMed Central Journals is for descriptions of biomedical data sets or databases, with the data being readily accessible and attributed to a source; "Data original article" – in Neuroinformatics is for documenting an original data release realizing a significant data contribution to the journal field; "Database article" – in BioMed Central Journals is for descriptions of novel biomedical databases likely to be of broad utility; "Database paper" – in Plos ONE is for descriptions of databases including details about how the data were curated, as well as plans for long-term database maintenance, growth, and stability; "Genome database" – in Human Genomics (Biomed Central) is for descriptions and/or evaluations of databases providing information regarding the human genome. In some cases the same journal have diverse types of data papers, e.g., SpringerPlus publishes both "Data notes" – describing a biomedical data set or database – and "Database articles" – describing a novel database likely to be of broad utility. In the case of Pensoft journals and the recently launched Biodiversity Data Journal, "Data papers" are for large data sets while "Taxonomic paper" and "Species inventory" are for domain specific "data", namely taxonomic or nomenclatural acts, systematic list of taxa with notes, species observations, and species inventory.

Table 3. Data paper names. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}: Ecology; J_{F1}: F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}: International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals.

Name	J_{BM}	J_{CC}	J_{DP}	J_{ES}	J_{Ec}	J_{F1}	J_{GD}	J_{GS}	J_{RR}	J_{Ni}	J_{Pe}	J_{PO}	J_{SD}	J_{SO}	J_{UP}
Data article														✓	
Data descriptor													✓		
Data in brief							✓								
Data note	✓					✓								✓	
Data original article										✓					
Data paper				✓	✓			✓	✓		✓				✓
Database article	✓	✓												✓	
Database paper												✓			
Dataset paper			✓												
Genome database	✓														

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

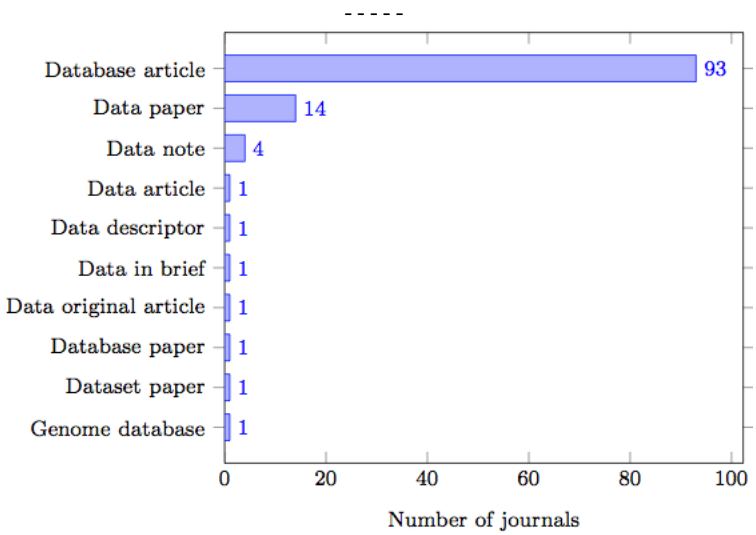

Figure 7. Number of data journals by data paper names.

Observations From this analysis it emerges that editors started assigning diverse names to the same conceptual entity, i.e., a "data paper". In some cases, names are intended to capture a specific typology of dataset the paper should be about, e.g., Genome Database. This proliferation of different names seems to have no real motivation and the risk is to confuse an average user and make some basic tasks more challenging, e.g., the discovery of "data papers" across journals.

How data journals describe datasets

As previously discussed, the purpose of a data paper is to describe a given dataset as a scientific paper describes a research outcome. The data paper is expected to promote the dataset exploitation and citation by giving details such as the methods and protocols used to create and process the dataset, the dataset structure and format, and the reuse potential. No scientific analysis made by using the data should be described, nor results nor conclusions drawn from it.

From our analysis, it emerges that there are no shared and fixed templates for data papers nor expected content for them. As for traditional research papers, every journal provides authors with its own set of instructions, guidelines and templates that indicate the typologies of papers to be accepted and how the data papers should be structured and formatted (see Table 4). Differently from the research papers, however, where over the time a number of common elements have emerged (e.g., each paper must have an abstract, an introduction, a related work section) for the data papers such de-facto common guidelines have not yet been identified. In some cases these guidelines are very detailed including information on how the manuscript should be structured, e.g., BMC Journals, in other cases the guidelines are generic leaving a certain degree of freedom on the content of the paper, e.g., Earth System Science Data. Some journals have developed their own templates to convey guidelines for manuscript production and expected content.

Table 4. Data papers production approaches promoted by data journals. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}: Ecology; J_{F1}: F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}: International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals.

	J_{BM}	J_{CC}	J_{DP}	J_{ES}	J_{Ec}	J_{F1}	J_{GD}	J_{GS}	J_{RR}	J_{Ni}	J_{Pe}	J_{PO}	J_{SD}	J_{SO}	J_{UP}
Detailed Guidelines	✓	✓	✓		✓		✓				✓		✓	✓	✓
Generic Guidelines				✓		✓		✓	✓	✓		✓			
Template	✓	✓		✓			✓	✓				✓	✓	✓	✓
Tool											✓				

The guidelines and templates for the content of the paper contain rules and advices on two classes of information: traditional scholarly communication related information and specific information on the dataset. Concerning the former, almost all the journals agree on including title, authors, abstract, keywords, and references. As far as the latter, the resulting picture is more

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

---

heterogeneous. By analysing existing templates and guidelines we have identified the following 10 classes of dataset information promoted by journals.

Availability to provide dataset access attributes, namely a DOI or a URI.

Competing interests to provide explicit declaration of any factor (including personal or financial relationship) that might influence the related dataset, including factors affecting the production or the presentation of the dataset.

Coverage to provide dataset "extent" attributes including spatial and temporal coverage.

Format to provide information oriented to promote the actual reuse of the dataset such as data format, encoding and language.

Licence to provide information oriented to dataset policies governing its use.

Microattribution to provide appropriate credits to each author of the paper by capturing in detail the contribution of each author.

Project to provide information on the initiative leading to the production of the dataset including goal and funding sources.

Provenance to provide information oriented to describe the methodology (including the tools) leading to the production of the dataset.

Quality to provide information oriented to assert qualitative aspects of the dataset including dataset limitations and anomalies.

Reuse to provide information oriented to promote potential exploitations of the dataset.

These classes of information are in some cases presented together, in other cases are split across diverse sections of the paper. A detailed picture of the classes of information used by each journal is given in Table 5. This table has been compiled by reporting the classes of information that are explicitly indicated via templates or guidelines. The lack of a tick does not mean that the journal is not promoting this information. Rather, the absence of the tick indicates that the editor decided to give their authors the freedom to describe their datasets, e.g. this is the case of Earth System Science Data journal that only prescribes the presence of availability-related information even if the Editors suggest to pay attention to the guidelines for reviewers to figure out what the potential content of a data paper might be.

Table 5. Dataset information explicitly promoted by journals via guidelines and templates. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}: Ecology; J_{F1}: F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}: International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals.

	J_{BM}	J_{CC}	J_{DP}	J_{ES}	J_{Ec}	J_{F1}	J_{GD}	J_{GS}	J_{RR}	J_{Ni}	J_{Pe}	J_{PO}	J_{SD}	J_{SO}	J_{UP}
Availability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Competing interests	✓	✓	✓			✓				✓			✓	✓	✓
Coverage					✓		✓				✓				✓
Format	✓	✓	✓		✓			✓			✓		✓	✓	✓
Licence					✓						✓		✓		✓
Microattribution	✓	✓				✓								✓	✓
Project					✓	✓		✓			✓				✓
Provenance	✓	✓	✓		✓	✓		✓			✓	✓	✓	✓	✓
Quality	✓	✓			✓	✓					✓		✓	✓	✓
Reuse	✓	✓						✓					✓	✓	✓

In addition to guidelines and templates, some data journals have also developed tools for supporting authors while producing their papers.

In particular, Pensoft developed its dedicated writing tool, the Pensoft Writing Tool (PWT) (Smith et al., 2013). This is an online tool that supports a collaborative production of the data paper. It is still based on a template approach (there are many templates including a data paper one) yet it guides the authors step by step in properly filling in the template sections by making it possible to automatically select author profiles, species classifications, and references from recognized information systems or controlled vocabularies.

Another tool supporting the authoring of data papers is the Integrated Publishing Toolkit (IPT) (Chavan & Penev, 2011) developed by the Global Biodiversity Information Facility (GBIF). Actually, this tool is specifically conceived to support the production of metadata for datasets of primary biodiversity data to be published through the GBIF network. However, the tool is also equipped with a

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

---

facility for automatically generating a data paper manuscript from the dataset metadata. The author is only requested to produce the Introduction section of the data paper. Any modification to the data paper resulting from the review process is actually performed by modifying the dataset metadata and regenerating the manuscript thus to actually improve simultaneously the data paper and the dataset description.

Observations From this analysis it emerges that the only information that the editors require to be necessarily specified in all the data papers is the datasets availability. The lack of a core and shared set of information to characterize datasets turns to be a strong limitation if data journals and data papers are expected to promote the real use of the datasets they are about. It would be fundamental to develop a shared, open, flexible and rich data characterization framework that should be used across disciplines, across the boundaries of the community a data journal is primarily oriented to. Such a framework should rely, as much as possible, on existing standards (that might be community specific or data type specific) and should accommodate the largest characterization of the datasets that is possible, no matter the primary domain such data are collected for. Moreover, it should be as open and flexible as possible to easily fit with diverse scenarios and domains. The availability of such a shared framework may enable the development of a number of tools, even by third party entities. The functionality offered by these tools can range from supporting the editing of data papers from a "syntactic" point of view, e.g., guaranteeing that the data paper contains all the sections envisaged by a given journal, to supporting the editing of data papers from a "content" point of view, e.g., automatically extracting information from the dataset for supporting the compilation of the data paper.

How data journals promote dataset availability

A data paper is always associated with a dataset. In the last years the solution of implementing this association by submitting supplementary files with the article is being progressively discontinued in favour of publishing the datasets in an appropriate repository and creating the association through a link among the two artefacts. Usually, journals propose a list of "recommended trusted data repositories", or "qualified data repositories", where datasets should be deposited. These are intended as repositories meeting certain criteria. From our research it has emerged that usually such repositories must fulfil the following basic requirements in order to be considered qualified: *(i)* they must be accredited, i.e., internationally or institutionally recognized; *(ii)* they must guarantee a long-term availability of the datasets and a permanent access; *(iii)* the datasets in the repositories must have a unique digital object identifier, e.g., DOI, that must be included in the related data paper; and *(iv)* the datasets in the repositories must be available free of charge and without any barriers, except for a possible registration to get a free login.

The recommended repositories may include Institutional repositories, e.g., UCL Discovery[1], national repositories, e.g., British Atmospheric Data Centre (BADC)[2], DANS-EASY[3], and international repositories, e.g., Dataverse repositories (King, 2007), Dryad (White, Carrier, Thompson, Greenberg, & Scherle, 2008), Figshare[4], PANGAEA[5], Zenodo[6]. Some of these repositories are discipline-specific, e.g., Worldwide Protein Data Bank[7], while others offer generic data hosting as-a-service, e.g., Dryad, Figshare and Zenodo.

The author guidelines provided by F1000 Research, for example, recommend that "For some datatypes, such as genetic sequences and protein structures, it is essential that the data are deposited in Genbank and Protein Data Bank, respectively". Similarly, the Pensoft data journals encourage authors to deposit their data underlying biological research articles in the Dryad data repository only in cases where no suitable more specialized public data repository exists, e.g., GBIF for species-by-occurrence data and taxon checklists, or GenBank for genome data.

Providing a 24/7 operational data repository service requires investments in specialized computing, software resources and skilled technical staff. Therefore, in the majority of the cases journal editors prefer to rely on third-party repository providers that are thus usually decoupled from the data journals. There are few exceptions to this rule like, for example, the Ecology journal that requires that the datasets are published in the Ecological Archives[8], i.e., a proprietary repository of the Ecological Society of America for publishing material associated with their journals. Another exception is the GigaScience journal that recommends dataset deposition in the GigaDB repository (Sneddon, Li, & Edmunds, 2012). This repository primarily hosts data and tools associated with articles in GigaScience, however it also includes datasets that are not associated with articles in this journal

(upon approval by the Editors of the journal). Yet another exception is represented by Dataset Papers in Science since it promotes a "hybrid" model, i.e., datasets may be published either as a zip file submitted during the data paper submission process, thus as supplemental material, or in a data repository. The zip file may include a combination of one or more tables, images or gene sequences.

In order to homogenize the data deposition strategy, journal editors have introduced a Joint Data Archiving Policy (JDAP)[9] which requires that supporting data for the papers they publish, be they traditional papers or data papers, must be publicly available via appropriate public archives.

In some cases, journals establish special arrangements with repositories that offer their service to authors via the payment of a fee. For example, BMC Bioinformatics authors can obtain a complimentary subscription to LabArchives[10] with an allotment of 100MB of storage. LabArchives is "an Electronic Laboratory Notebook which will enable scientists to share and publish data files in situ". So, data files linked to published articles are provided with DOIs and remain persistently available. Anyway, the journal guidelines remind that "Use of LabArchives or similar data publishing services does not replace pre-existing data deposition requirements, such as for nucleic acid sequences, protein sequences and atomic coordinates".

Observations The analysis conducted so far highlights that while editors are embracing the publication of new type of journal papers, they do not consider the publication of data as part of their own mission. On the contrary, there is currently a fast growth of data repository specialized providers who envisage in the scientific data publication and maintenance a new promising market. In parallel, datasets are progressively emerging as "first class entities", i.e., they have a value per-se and, consequently, their publication is no longer a side-effect of the publication of a scientific paper. In the future this new role played by datasets may completely change the panorama and also partially reduce the motivations that have pushed the introduction of the data papers. As a matter of fact, datasets objects can be cited and the number of citations can be used as a measurement of their relevance as research products and, consequently, as reward for their authors. Citations can also be used as an implicit measurement of the quality of the datasets testified by actual users. However, although innovative and enhanced data publication and citation practices will be put in place, it is improbable that this will lead to the end of data papers. Data papers have a role in scientific communication that can be neither played by dataset with their metadata nor by research papers, as they have an information payload that is unique (e.g. Callaghan, 2013).

How data journals support dataset citation

Data citation is intended for promoting a direct and unambiguous reference to a dataset used in a particular study (Mooney & Newton, 2012; Ball & Duke, 2012). However, data citation presents challenges not included when referencing research papers and literature (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013), e.g., how to specify a particular subset of a dataset in the absence of familiar conventions like page numbers or sections.

To some extent, the existence of data papers reconciles the data citation problem with the traditional reference strategy, i.e., authors of a research paper can cite data papers as they are used to for papers rather than citing the dataset per se. However, this does not solve all the data citation challenges, e.g., it does not offer any convention for specifying a subset of a dataset.

The approaches data journals promote for the citation of the datasets they are about are in Table 6. Journals tend to have a section specifically dedicated to report on data availability. In some case this section is the Abstract, e.g., in Earth System Science Data and Genomic Data. In some cases, see "Reference format" in the table, journals have developed the format authors should use to refer to data in such a dedicated section, e.g., BioMed Central journals. Very often, journals require that datasets should be included in the reference list of the paper, e.g., by relying on DCC recommendations (Ball & Duke, 2012) or on the DataCite recommendations (Starr & Gastl, 2011). Concerning the use of Persistent Identifiers, for obvious reasons almost all the journals promote this mechanism as "the" one to use for accessing to datasets in a univocal manner. In the majority of cases, journals recommend DOIs in accordance with data repositories that use to assign DOIs to deposited datasets. Very often the identifier is displayed as an URI.

Table 6. Dataset citation approaches promoted by journals for data papers. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}: Ecology; J_{F1}: F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}:

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals.

	J_{BM}	J_{CC}	J_{DP}	J_{ES}	J_{Ec}	J_{F1}	J_{GD}	J_{GS}	J_{RR}	J_{Ni}	J_{Pe}	J_{PO}	J_{SD}	J_{SO}	J_{UP}
Dedicated section	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓
In reference list	✓	✓				✓		✓			✓			✓	
Persisted ID	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓		✓
Reference format	✓	✓				✓		✓					✓	✓	

Observations From this it emerges that, in practice, data journals have not developed a strong and shared set of practices that data paper authors are requested to follow to promote dataset citation. If data papers should be the artefact to cite in research papers willing to cite the datasets they use, then it is fundamental that journals contribute to the development and diffusion of emerging attempts to systematize and standardize data citation practices, e.g., the CODATA data citation principles (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013) and the Force11 data citation principles (Force11 Data Citation Synthesis Group, 2014).

How data journals guarantee dataset(s) quality

Peer review is among the features characterizing traditional journals. In essence, it is a regulatory mechanism through which experts in a given domain appraise the quality of scientific work produced by others in their field. Because of this role it is highly discussed, e.g., Solomon (2007) examined the role of peer review for journals and the benefits resulting from new review modes, Lee, Sugimoto, Zhang, and Cronin (2013) highlighted the bias affecting the process.

The peer review concept when applied to the data journals domain may create false assumptions simply because there is no shared understanding of what peer-review and data quality mean in this context (Parsons & Fox, 2013). B. Lawrence et al. (2011) provided an extensive discussion on data peer-review. For the sake of this study, rather than discussing dataset peer-review, we analysed the practices data journals have in place for the review of data papers.

Almost all the data journals we have analysed perform peer review to some extent. The differences emerge when analysing *(a)* the review process approaches, i.e., whether they are "open", "semi-open", or "close", and *(b)* the features reviewers are requested to assess. A summary of the differences is given in Table 7.

Table 7. Data journals peer review approaches and criteria. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}: Ecology; J_{F1}: F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}: International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals.

	J_{BM}	J_{CC}	J_{DP}	J_{ES}	J_{Ec}	J_{F1}	J_{GD}	J_{GS}	J_{RR}	J_{Ni}	J_{Pe}	J_{PO}	J_{SD}	J_{SO}	J_{UP}
Open process				✓		✓									
Semi-open process	✓														
Close process	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Data Paper specific			✓	✓	✓			✓	n/a		✓		✓		✓
Manuscript quality	3	3	n/a	5	3	12	7	2	n/a	29	6	4	n/a	3	n/a
Consistency	1	1	n/a	3	3	1	2	3	n/a	9	5	1	1	1	1
Data quality	2	2	n/a	5	6	6	4	4	n/a	8	3	n/a	3	2	n/a
Data reusability	1	1	n/a	5	3	9	5	2	n/a	2	5	2	4	1	7
Utility	n/a	n/a	n/a	3	4	n/a	6	2	n/a	4	2	1	n/a	n/a	n/a

The majority of data journals adopt the conventional scheme of "*close*" peer-review, namely pre-publication, anonymous and private peer-review.

Few journals adopt an "*open*" peer-review in order to promote fairness and objectivity and to reduce the publication time, namely Earth System Science Data and F1000 Research. In the case of Earth System Science Data there is a two-stage approach. After submission, manuscripts are published on the Web as "discussion papers" as the result of a lightweight review performed by a dedicated editor which is requested to evaluate the paper except for its scientific content – e.g., if it fits with the journal scope – and suggest minimal technical corrections like typing errors. The paper remains at this stage for a 8 weeks period when the community can create review and discussion. Every discussion paper shall receive at least two referee comments. After the public discussion phase,

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

---

the authors are requested to publicly reply to comments and to produce a revised manuscript that, if approved by the editor, is finally published in the journal. In the case of F1000 Research, after passing some rapid initial checks performed by the in-house editorial team, manuscripts are published with the status "awaiting peer-review". Then, authors are asked to suggest five potential referees for their manuscripts, which will have to judge whether the work seems scientifically sound by giving it one of three different statuses: "Approved", "Approved with Reservations", and "Not Approved". All status and referees' comments are published along with the manuscript, actually a version of it. In fact, this process leads to the publication of multiple versions of the manuscript that is "accepted" only when receives either two "Approved" statuses or two "Approved with Reservations" and one "Approved" status. After the approval the paper becomes an "Indexed" paper that is discoverable via services like PubMed, Scopus and Google Scholar.

Some journals adopt a "semi-open" approach, i.e., they make available the pre-publication history of accepted papers that include the initial submission, reviews and revisions. This approach is used, for example, by some of the BioMed Central Journals such as BMC Anesthesiology and BMC Cancer.

In the case of Biodiversity Data Journal, in addition to the "community" peer review (a "close" peer-review), authors may opt to make their manuscript available for comment to all registered journal users ("public" peer-review) and reviewers may opt to stay anonymous or disclose their names.

Similarly to conventional journals, data journals have developed guidelines, criteria and instructions for reviewers. By analysing existing evaluation criteria we have identified the following 5 classes of criteria.

Quality of manuscript: the conventional criteria for assessing manuscript writing, clarity, organization, adherence to template;

Consistency between data paper and dataset(s): criteria for assessing the effectiveness of the data paper content as a means for accessing the dataset(s);

Data quality: criteria for assessing the methodologies leading to the production of the dataset(s);

Data reusability: criteria for assessing the actual reusability of the dataset(s);

Utility and contribution of data: criteria for assessing the potential of the dataset(s) for the community.

A detailed picture of the emphasis each journal associates with the key classes is given in Table 7 where the number of questions in the guidelines is reported by class. The majority of questions fall in the manuscript quality class. A significant amount of questions are dedicated to assess information on quality and reusability of the dataset(s) reported in the paper. For instance, Ubiquity Press journals includes among the top criteria methods and reuse sections.

Observations Our perspective on dataset(s) quality and peer-review is summarized by the following three key points *(a)* the "quality" of the data intimately depends on the application domain the data is expected to be used in, it is usually resulting from the "fitness-for-use" and there is no standard for data quality that is globally accepted; *(b)* in some cases datasets are simple facts collected by sensors and other automatic approaches, although there are well known approaches to remove "errors" there is no way to improve them in absolute terms; *(c)* asserting data "quality" is a challenging task that is fundamentally different from evaluating a scholarly paper, e.g., often it is not feasible for a human to analyse data as a reviewer is used to do for a paper, as data frequently are big in volume and complex in structure. From the analysis of the described data paper review processes it emerges that the majority of existing journals rely on a traditional review approach that remain almost hidden to the data paper audience, e.g. review reports are known to paper-related authors, reviewers and editors only. Moreover, only in few cases applied review strategies are specialized to deal with data papers while in the majority of the cases assessing the quality of these artefacts is conceived as equivalent to assessing traditional papers quality. Actually, data papers are particular kind of paper and dedicated quality assessment criteria should instead be defined. These criteria should be oriented to evaluate the capacity of the artefact to provide its readers with the minimal set of information needed to enable an effective reuse of the dataset(s) associated with the paper. Moreover, the "disclosure" of the review process should be encouraged as a mean for proving evidence of the data paper quality.

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

Data journals and open access

As already discussed, almost all the analysed data journals are open access. However, open access usually implies publication costs and it does not necessarily mean free access or license free. In this section we discuss three aspects related with open access practices: *(i)* the costs associated with the publication of the data paper; *(ii)* the licence associated with the data paper; and *(iii)* the licence associated with the dataset.

Journals publishing data papers tend to have article processing charges (Solomon & Björk, 2012; Van Noorden, 2013). Differently from previous analysis on open access journals publication fees conducted by Kozak and Hartley (2013), where only a small percentage of journals charge authors for publishing, in the case of data journals this is the standard approach. In the case of "mixed" journals charges may be specific for data papers, e.g., Ecology, or may be the same for all the types of papers accepted by the journal, e.g., BioMed Central journals. A detailed description of the costs is given in Fig. 8 and Tab. 8. In average, the article processing charge is approximately €1300 for the journals in our sample. However, this value is strongly influenced by the presence of a large amount of mixed journals. In fact, the average article processing charge is approximately €420 for "pure" journals while it is approximately €1360 for "mixed" journals. Journals tend to ask no charges during the launch phase, e.g., GigaScience (BioMed Central journal) is not levying charges after the publication of 2 volumes and 16 issues, and Journal of Systems Chemistry (Chemistry Central journal) is still paying on its own the publication costs after 4 volumes and 12 issues. In some cases, the publication of the dataset per se in a repository has a cost, thus the cost of publishing a data paper is actually the sum of the two costs. As already discussed, journals can establish agreement with data repositories to set up special arrangements for data publication.

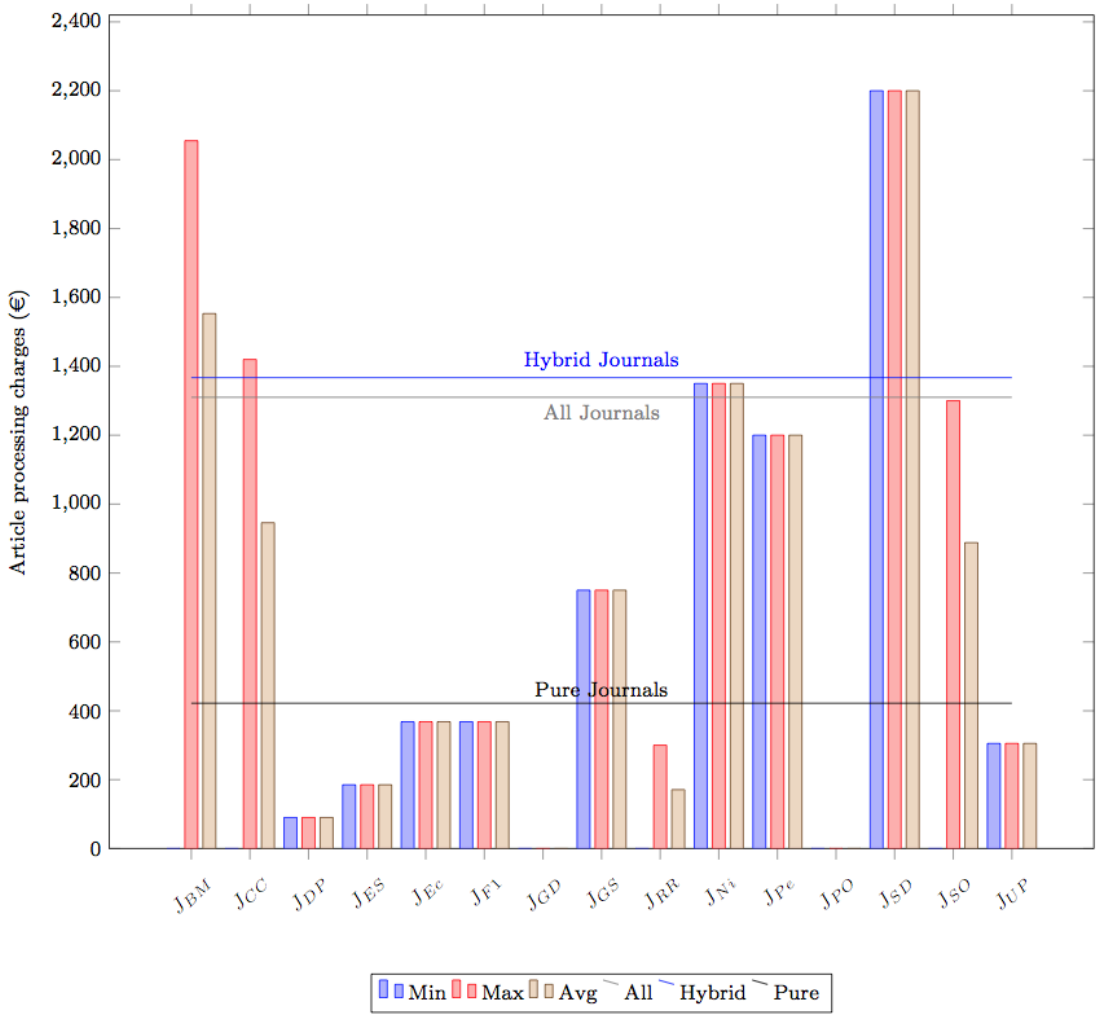

Figure 8. Article processing charges by journal. Journals are identified as follows. J_{BM}: BioMed Central Journals; J_{CC}: Chemistry Central Journals; J_{DP}: Dataset Papers in Science; J_{ES}: Earth System Science Data; J_{Ec}: Ecology; J_{F1}:

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

F1000 Research; J_{GD}: Genomics Data; J_{GS}: Geoscience Data Journal; J_{RR}: International Journal of Robotics Research; J_{Ni}: Neuroinformatics; J_{Pe}: Pensoft Journals; J_{PO}: PLOS ONE; J_{SD}: Scientific Data; J_{SO}: SpringerOpen Journals; J_{UP}: Ubiquity Press Journals.

Table 8. Data papers publication costs by journal.

BioMed Central journals have charges ranging from €840 for BMC Research Notes to €1,995 for Genome Medicine.
Chemistry Central journals have a charge of €1,315.
Dataset Papers in Science is currently not requiring any article processing charge.
Earth Science System Data has a charge of €3 for each page of a "discussion paper", i.e., an under review paper (cf. Sec. 3.5), while there is no charge for the publication of papers in their "final" version.
Ecology charges a one-time fee of $250 at the publication of the data paper that includes the possibility of store data up to 10MB (for data between 10MB and 1GB an extra charge of $250 applies).
F1000 Research data articles have an article processing charge of $500 which includes the possibility to publish up to 1 GB (for 1 to 5 GB of data an additional fee of $200 is requested while beyond 5 GB there is a negotiation for the cost).
Genomics Data has a publication fee of $500, excluding taxes (until December 31, 2013, there is an introductory offer of $100).
Geoscience Data Journal has a charge of €1,200 with a 20% discount for members of the Royal Meteorological Society.
Pensoft has a minimum fixed fee of €150/€200 for papers smaller than 10 printed pages. For papers exceeding this, there is a per page cost of €15/€20. In the case of ZooKeys, the fixed fee is €300 for papers smaller than 30 printed pages, while the cost of extra pages is €15.
PLOS ONE has a standard publication fee of $1,350 although there are special arrangements for "poor" countries that might be not charged at all or be charged a flat $500.
Scientific Data has an article processing charge that varies according to the licences authors are willing to use for their paper, i.e., €675 for CC BY-NC 3.0 / CC BY-NC-SA 3.0 and €750 for CC BY 3.0.
SpringerOpen journals have an article processing charge ranging from €840 for Health and Justice to €1,300 for International Journal of Food Contamination, while Botanical Studies has no charge since costs are covered by Academia Sinica.
Ubiquity Press journals have a publication fee of 25GBP.

For data paper copyright and licences, data papers are handled as scientific papers. For open access data papers, author(s) retain the copyright and grant *(a)* to the publishers the right to "publish" the paper and *(b)* to any third party the right to "use" the paper by giving credits to the original author(s). In the case of non open access data papers, the right to "use" the paper is granted to subscribers and authorized users only. In some cases, authors can decide which license they are willing to associate to their open access paper by selecting among a number of Creative Commons licences, e.g., Genomics Data and Scientific Data. However, there is no major difference among the selectable licences, all of them carry the obligation to properly attribute the paper when using it.

For dataset copyright and licences, the analysed journals agree on the need to make the dataset(s) described in the paper accessible free of charge for non-commercial uses both during the review phase and once the paper is accepted. Similarly to papers, any use of the dataset carries the obligation to give credit to the original author. Among the most common licences there are the CC0 and the Open Data Commons Attribution License. However, there is no guideline on whether the dataset user should give credits by using the data paper, the dataset per se or both.

Observations From this overview it emerges that data journals go in the direction to promote a free access to both the datasets and the papers describing them. However, the costs of this "openness" are charged to data owners when publishing their data and the accompanying papers. In fact, data owners are requested to pay the article processing fee to journal editors and to deposit the associated dataset(s) in a selected repository that may ask for another fee. These fees are part of the costs involved in data publication that is one of the major barriers affecting the whole data publication movement, e.g., Costas et al. (2013). The need to reduce these costs and make the whole process of data publication more efficient is a topic largely discussed and supported by the stakeholders involved in the data publication scenario, e.g. funding agencies are developing specific arrangements and policies for open access to research data in research programmes.

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

Conclusions and prospects

This article have surveyed existing data journals to analyse their approaches to data publication and to identify to what extent they contribute to facilitate sharing and re-use of data.

From our investigation, we found that data journals are now an established phenomenon in the scientific literature. In fact, the number of published data papers and data journals is rapidly growing. In fact, the 23.5% of the existing data papers has been published in 2013. The majority of data journals (69.82%) is indexed by well known professional services, namely Thomson Reuters Web of Science. Moreover, conferences have started promoting data papers, for example the International Symposium on Biomedical Imaging has just launched a call for data papers for the ISBI 2014 to be held in China, 29 April-2 May 2014[11].

We analysed current data papers practices and approaches implemented by journal editors from the perspective of five core aspects contributing to data sharing and reuse, i.e., how they promote dataset description, how they promote dataset availability, how they support dataset citation, how they guarantee dataset quality, and how they promote open access to datasets. From our observations it emerged that journal editors *(a)* have not yet a shared and consolidated strategy for promoting an effective dataset description favouring data reuse, e.g., the only dataset information that is promoted by all the journals is "availability", being other information such as "coverage", "quality" or "reuse" neglected in many cases; *(b)* usually rely on services offered by third-party data repositories and archives for making dataset available; *(c)* are overlooking many of the issues affecting dataset citation by assuming that citation approaches mirrored from the scientific publication model are sufficient; *(d)* have not yet deeply addressed the issues related with datasets quality. In essence they continue to rely on consolidated peer-review approaches focusing on the data paper only; *(e)* have embraced the open access model by generally relying on the "gold open access" model, i.e., authors are requested to pay an article processing fee, and asking to deposit the dataset(s) in selected repositories making them available free of charge. As observed by many, e.g., Vision (2010), journals are in a privileged position to promote data publication practices. However, a lot of work is needed to reach a common understanding on what these practices should be to contribute to data sharing and data re-use. Promoting data publication practices is not a sole responsibility of journals. Rather, journals are an integral part of the entire ecosystem underlying *scientific communication infrastructures* (Castelli, Manghi, & Thanos, 2013), thus they are called to develop innovative, comprehensive and effective data publication practices in such settings.

When designing data papers of the future it is fundamental to identify and highlight the added value a "data paper" has for data sharing and re-use with respect to rich, detailed and curated dataset(s) metadata that are managed by data repositories. Data papers are scientific communications, thus they have the benefits of this kind of communication with respect to metadata starting from the target audience to the intended mission and goal. However, scientific communication is potentially affected by the problems highlighted by Nosek and Bar-Anan (2012b), i.e., no, slow, incomplete, inaccurate, or unmodifiable communication. To design data papers of the future it is worth to discuss how each of these problems impact on data papers. *No communication* happens for a series of reasons including the tendency to publish positive results only. Fanelli (2012) demonstrated this tendency and highlighted how the absence of publication of negative results has impacts on scientific progress, e.g., cause a waste of resources replicating activities that has already failed. Data identification and collection – an activity preceding any data paper production – is an intellectual and time consuming activity, the capability to communicate negative results in producing them is important, e.g., it can prevent others to incur in similar issues, "buggy" datasets can be used to assess the robustness of certain algorithms when dealing with noisy in data. *Slow communication* is a problem not affecting data papers more than traditional papers. Data papers publication practices are already overcoming this problem and new one can be promoted, e.g., open peer review, to further reduce the communication time. *Incomplete or inaccurate communication* happens because authors do not report everything or because errors are difficult to detect. The lack of information underlying the production of the dataset a data paper describes is an issue that severely hinder dataset re-use, e.g., Thanos (2014). Authors should describe what is actually important rather then describing what they think is important. *Unmodifiable communication* happens because once published, articles are static entities. For data papers and datasets it is fundamental to guarantee that there is a powerful versioning system enabling researchers to identify the "right" research artefact.

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

When designing data papers of the future it is important to relate them to "enhanced publications" (Bardi & Manghi, 2014a, 2014b). In particular, the enhanced publication concept can be used to realize data papers of the future. This approach has the potentiality to enlarge the machinery the artefact author can use to convey its "content", namely the dataset description, e.g., at data paper access time the user can be provided with a number of graphs automatically produced out of the dataset the data paper is about, the current list of other research artefacts citing the data paper, the set of metrics highlighting the extent to which the artefact is recognized and "brought" by diverse communities, different "views" of the dataset tailored to serve different information needs.

Scientific communication is changing and data papers are part of this change. However, changes should not be driven by forces other than authors and scientists, e.g., P. A. Lawrence (2003); Nosek and Bar-Anan (2012a). It is a responsibility of scientists to assist the rest of scientific communication realm to remove the barriers affecting it, and this paper represents a contribution in this direction.

Acknowledgements

The work reported has been partially supported by the OpenAIREplus project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283595) and the iMarine project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644). The authors would like to thank M. B. Baldacci (ISTI-CNR) for her valuable support and the many helpful comments, and the anonymous reviewers for their helpful comments and suggestions.

Footnotes

[1] UCL Discovery http://eprints.ucl.ac.uk/
[2] British Atmospheric Data Centre (BADC) http://badc.nerc.ac.uk
[3] Data Archiving and Networked Services (DANS) EASY https://easy.dans.knaw.nl
[4] figshare http://figshare.com
[5] PANGAEA – Data Publisher for Earth & Environmental Science http://www.pangaea.de/
[6] Zenodo https://zenodo.org
[7] Worldwide Protein Data Bank http://www.wwpdb.org/
[8] ESA's Ecological Archives http://esapubs.org/archive
[9] Joint Data Archiving Policy (JDAP) http://datadryad.org/pages/jdap
[10] Electronic Laboratory Notebook – LabArchives http://www.labarchives.com/
[11] IEEE International Symposium on Biomedical Imaging http://biomedicalimaging.org/2014/

References

Asher, A., Deards, K., Esteva, M., Halbert, M., Jahnke, L., Jordan, C., Keralis, S. D. C., Kulasekaran, S. S., Moen, W. E., Stark, S., Urban, T., & Walling, D. (2013). Research data management: Principles, practices, and prospects. Washington: Council on Library and Information Resources.

Ball, A., & Duke, M. (2012). How to cite datasets and link to publications. Edinburg: Digital Curation Centre.

Bardi, A., & Manghi, P. (2014a). Enhanced publications: Data models and information systems. LIBER Quarterly, n.a., n.a. Retrieved from https://liber.library.uu.nl/index.php/lq/article/view/8445/9825

Bardi, A., & Manghi, P. (2014b). A rationale for enhanced publications. LIBER Quarterly, n.a., n.a. Retrieved from https://liber.library.uu.nl/index.php/lq/article/view/8445

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge. Science, 323 (5919), 1297-1298.

Borgman, C. (2011). The Conundrum of Sharing Research Data. Journal of the Association for Information Science and Technology , 63 (6), 1059-1078.

Bourne, P. E. (2010). What do i want from the publisher of the future? PLoS Computational Biology, 6 (5), e1000787. doi: 10.1371/journal.pcbi.1000787

Bourne, P. E., Clark, T., Dale, R., de Waard, A., Herman, I., Hovy, E. H., & Shotton, D. (2012). Improving the future of research communication and e-scholarship (Force11 White Paper). Force11. Retrieved April 4, 2014, from https://www.force11.org/white_paper

Callaghan, S. (2013) Data journals – as soon-to-be-obsolete stepping stone to something better? Blog post. Retrieved April 4, 2014, from http://citingbytes.blogspot.it/2013/01/data-journals-as-soon-to-be-obsolete.html

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERCs environmental data centres. International Journal of Digital Curation, 7 (1), 107-113. doi:10.2218/ijdc.v7i1.218

Callaghan, S., Hewer, F., Pepler, S., Hardaker, P., & Gadian, A. (2009). Overlay journals and data publishing in the meteorological sciences. Ariadne, Issue 60, –. Retrieved from http://www.ariadne.ac.uk/issue60/callaghan-et-al/

Castelli, D., Manghi, P., & Thanos, C. (2013). A vision towards scientific communication infrastructures. International Journal on Digital Libraries , 13 (3-4), 155-169. doi:10.1007/s00799-013-0106-7

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

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. Data Science Journal, 12, CIDCR1-CIDCR75. doi:10.2481/dsj.OSOM13-043

Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. (2013). The value of research data - metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. Leiden: Center for Science and Technology Studies. Retrieved from http://www.knowledge-exchange.info/datametrics

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution, 28 (8), 454-461. doi:10.1016/j.tree.2013.05.002

De Schutter, E. (2010). Data publishing and scientific journals: the future of the scientific paper in a world of shared data. Neuroinformatics, 8(3), 151–153. doi:10.1007/s12021-010-9084-8

Douglass, K., Allard, S., Tenopir, C., Wu, L., & Frame, M. (2014). Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. Journal of the Association for Information Science and Technology, 65(2), 251-262. doi:10.1002/asi.22988

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics, 90 (3), 891-904. doi:10.1007/s11192-011-0494-7

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.) (1985). Sharing research data. Washington: The National Academies Press.

Force11 Data Citation Synthesis Group. (2014). Joint Declaration of Data Citation Principles - FINAL. Retrieved April, 2014, from http://www.force11.org/datacitation

Halevi, G., & Moed, H. F. (2012). The evolution of big data as a research and scientific topic: Overview of the literature. Research Trends, 30, -.

Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). The fourth paradigm: Data-intensive scientific discovery. Microsoft Research.

It's not about the data. (2012). Nature Genetics, 44(2), 111. doi:10.1038/ng.1099

Kennedy, D. N., Ascoli, G. A., & De Schutter, E. (2011). Next steps in data publishing. Neuroinformatics , 9 (4), 317–320. doi: 10.1007/s12021-011-9131-0

King, G. (2007). An introduction to the Dataverse network as an infrastructure for data sharing. Sociological Methods & Research , 36 (2), 173- 199. doi:10.1177/0049124107306660

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., & Wächter, J. (2006). Data publication in the open access initiative. Data Science Journal, 5, 79-83.

Kozak, M., & Hartley, J. (2013). Publication fees for open access journals: Different disciplines – different methods. Journal of the Association for Information Science and Technology , 64(12), 2591-2594. doi:10.1002/asi.22972

Kunze, J., Cruse, T., Hu, R., Abrams, S., Hastings, K., Mitchell, C., & Schiff, L. (2011). Practices, trends, and recommendations in technical appendix usage for selected data-intensive disciplines (Tech. Rep.). California Digital Library.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. International Journal of Digital Curation , 6 (2), 4-37. doi:10.2218/ijdc.v6i2.205

Lawrence, P. A. (2003). The politics of publication. authors, reviewers and editors must act to protect the quality of research. Nature, 422, 259-261. doi:10.1038/422259a

Lawrence, R. (2012). Data publishing: Peer review, shared standards and collaboration. Presentation at 8th Research Data Management Forum, Southampton. Retrieved April, 2014, from http://www.dcc.ac.uk/webfm_send/798

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the Association for Information Science and Technology, 64(1), 2-17. doi:10.1002/asi.22784

Marcus, A., & Menzies, T. (2010). Software is data too. In G.-C. Roman & K. J. Sullivan (Eds.), Proceedings of the workshop on future of software engineering research (FoSER 2010) (pp. 229-232). ACM. doi:10.1145/1882362.1882410

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship and Scholarly Communication, 1(1), eP1035. doi:10.7710/2162-3309.1035

Murray-Rust, P. (2008, January). Big Science and Long-tail Science. Peter Murray-Rust Blog. Retrieved April, 2014, from http://blogs.ch.cam.ac.uk/pmr/2008/01/29/big-science-and-long-tail-science/

Newman, P., & Corke, P. (2009). Data paper – peer reviewed publication of high quality data. The International Journal of Robotics Research, 28(5), 587. doi:10.1177/0278364909104283

Nosek, B. A., & Bar-Anan, Y. (2012a). Scientific communication is changing and scientists should lead the way. Psychological Inquiry , 23 (3), 308-314. doi:10.1080/1047840X.2012.717907

Nosek, B. A., & Bar-Anan, Y. (2012b). Scientific utopia: I. opening scientific communication. Psychological Inquiry, 23 (3), 217-243. doi:10.1080/1047840X.2012.692215

Pampel, H., & Dallmeier-Tiessen, S. (2014). Open research data: From vision to practice. In S. Bartling & S. Friesike (Eds.), Opening science (p. 213-224). Springer International Publishing. doi:10.1007/978-3-319-00026-8_14

Parsons, M., & Fox, P. (2013). Is data publication the right metaphor? Data Science Journal, 12, WDS31-WDS46.

Pfeiffenberger, H., & Carlson, D. (2011). "Earth System Science Data" (ESSD) – A Peer Reviewed Journal for Publication of Data. D-Lib Magazine, 17 (1/2). doi:10.1045/january2011-pfeiffenberger

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). Report on integration of data and publications (Tech. Rep.). Opportunities for Data Exchange (ODE).

Scheer, R. (2013, April). NPG to launch Scientific Data to help scientists publish and reuse research data. Nature Publishing Group Press Release. Retrieved April, 2014, from http://www.nature.com/press_releases/scientificdata.html

Smith, V., Georgiev, T., Stoev, P., Biserkov, J., Miller, J., Livermore, L., Baker, E., Mietchen, D., Couvreur, T. L. P., Mueller, G., Dikow, T., Helgen, K. M., Frank, J., Agosti, D., Roberts, D., & Penev, L. (2013). Beyond dead trees: integrating the scientific process in the biodiversity data journals. Biodiversity Data Journal, 1, e995. doi:10.3897/BDJ.1.e995

Sneddon, T. P., Li, P., & Edmunds, S. C. (2012). GigaDB: announcing the GigaScience databases. GigaScience, 1(11), -. doi:10.1186/2047-217X-1-11

Solomon, D. J. (2007) The role of peer review for scholarly journals in the information age. Journal of Electronic Publishing, 10 (1). doi:10.3998/3336451.0010.107

Solomon, D. J., & Björk, B.-C. (2012). A study of open access journals using article processing charges. Journal of the Association for Information Science and Technology, 63(8), 1485-1495. doi:10.1002/asi.22673

This is a preprint of the article:
L. Candela, D. Castelli, P. Manghi, A. Tani Data Journals: A Survey. Journal of the Association for Information and Science Technology. Accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358

Starr, J., & Gastl, A. (2011). isCitedBy: A Metadata Scheme for DataCite. D-Lib Magazine, 17 (1/2). doi:10.1045/january2011-starr

Tempest, D. (2012, March). Journals and data publishing: Enhancing, linking and mining. Presentation at 8th Research Data Management Forum, Southampton. Retrieved April, 2014, from http://www.dcc.ac.uk/webfm_send/797

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. PLoS ONE , 6 (6), e21101. doi:10.1371/journal.pone.0021101

Thanos, C. (2014). Scientific data reusability: Conceptual foundations, impediments and enabling technologies (Tech. Rep.). Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR.

Van Noorden, R. (2013). Open access: the true cost of science publishing. Nature, 495, 426-429. doi:10.1038/495426a

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. Current Biology, 24(1), 94-97. doi:10.1038/495426a

Vision, T. J. (2010). Open data and the social contract of scientific publishing. BioScience, 60(5), 330-331. doi:10.1525/bio.2010.60.5.2

White, H. C., Carrier, S., Thompson, A., Greenberg, J., & Scherle, R. (2008). The dryad data repository: A singapore framework metadata architecture in a dspace environments. In Proceedings of international conference on dublin core and metadata applications (DCMI '08)(pp. 157-162).

Wilson, R. (2012, March). Integrating data and publications. Presentation at 8th Research Data Management Forum, Southampton. Retrieved April, 2014, from http://www.dcc.ac.uk/webfm_send/800