Data Journals: A Survey

Leonardo Candela
Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo,” Italian National Research Council, via G. Moruzzi, 1, Pisa 56124, Italy. E-mail: candela@isti.cnr.it

Donatella Castelli
Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo,” Italian National Research Council, via G. Moruzzi, 1, Pisa 56124, Italy. E-mail: castelli@isti.cnr.it

Paolo Manghi
Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo,” Italian National Research Council, via G. Moruzzi, 1, Pisa 56124, Italy. E-mail: manghi@isti.cnr.it

Alice Tani
Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo,” Italian National Research Council, via G. Moruzzi, 1, Pisa 56124, Italy. E-mail: 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 such as data deluge and big data today characterize numerous (research) initiatives, much 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, and quality issues to a rather general lack of a data-sharing culture. Lately, data journals have overcome some of these barriers. In this study of more than 100 currently existing data journals, we describe the approaches they promote for data set description, availability, citation, quality, and open access. We close by identifying ways to expand and strengthen the data journals approach as a means to promote data set access and exploitation.

Introduction

Data—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; Halevi & Moed, 2012; Hey, Tansley, & Tolle, 2009).

Although there is an almost universal agreement on the benefits of “data sharing and reuse” 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; Pampel & Dallmeier-Tiessen, 2014; Tenopir et al., 2011). These barriers are methodological, legal, and technical and are often related to the lack of incentives for researchers to share their data (Asher et al., 2013; Bourne, 2010; Bourne et al., 2012; Douglass, Allard, Tenopir, Wu, & Frame, 2014). The effects of these obstacles on science is deleterious; for example, Vines et al. (2014) demonstrate how the availability of research data was strongly affected by an article’s age when no policy was in place. Thus, proper data-sharing practices and policies must be introduced to foster data availability. Moreover, mechanisms must be identified to make the scientific community aware of the available data sets, facilitate their understanding, and foster their effective reuse.

In this changing landscape, data journals have been proposed as a first-step solution to some of these barriers. These journals 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 online accessible data set, or a group of data sets, published in accordance to the standard academic practices” (Chavan & Penev, 2011),...
the final aim being to provide “information on the what, where, why, how, and who of the data” (Callaghan et al., 2012, p. 112). Thus data publication is a prerequisite for 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).

This survey reviews 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 more than 100 data journals is discussed by comparing their approaches to the data papers concept implementation, including how to describe a data set properly, how to promote data set availability, how to properly cite a data set and guarantee rewards, how to guarantee data set quality, and how to guarantee open access to data sets. The article ends by giving some suggestions for enhancing the role of data journals as a true data-sharing means.

The Beginnings

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

The issues related to formally publishing and citing data sets have been clearly enumerated and discussed by others (e.g., Callaghan et al., 2012; B. Lawrence et al., 2011). The aim is to have data sets 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, p. 112) “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 centers/libraries. From an analysis conducted by Reilly, Schallier, Schrimpf, Smit, and Wilkinson (2011) it emerged that (a) researchers are fully aware of the benefits and value of sharing data, yet they call for support (e.g., facilities for storing and maintaining data, facilities for controlling access, facilities for getting credits, someone that pays the costs underlying data sharing); (b) publishers are 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 centers and libraries are fully aware of their mission as 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 being an integral part of the article or (b) data residing in supplementary files attached to the article. The first approach is the traditional one, and it is affected by a number of drawbacks, including the difficulties of separating the data from the rest of the material and reusing them. The second approach goes beyond the motivation to share data. In fact, by about 2009, most 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 does not allow 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 a “data paper” started to be generally recognized (Kunze et al., 2011), especially in the biodiversity community (Chavan & Penev, 2011). Publishers have readily accepted dealing with this emerging need. At the 8th Research Data Management Forum, R. Lawrence (2012) affirmed that “there is now a general consensus that sharing and publishing data is [sic] 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 the Faculty of 1,000. One central point was the economic risk of publishing new journals before their acceptance is ensured. Tempest (2012) affirmed that publishers recognize that scientists’ investment in creating and interpreting data and their intellectual and financial contributions have to be recognized and valued. However, he added that, when publishers “add value and/or incur significant cost,” their “contributions also need to be recognized and valued.” He also restated the pros and cons in making data available and accessible, concluding that publishers were investing in innovation, but there were still issues and still much work to do. The slightly different position of “small” publishers was brought up by Wilson (2012) of the Nature Publishing Group. She terminated her presentation stating that “there needs [to be] partnership by institutions, repositories, publishers, researchers, and funders, even though roles are to be well established and business models well determined.” Similarly, an editorial in Nature Genetics urged interested people to be careful, to avoid potential problems (“It’s not about the data,” 2012).
However, several initiatives supporting data papers have arisen. Callaghan, Hewer, Pepler, Hardaker, and Gadian (2009) investigated the idea of an overlay journal, 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 “(a) metadata about the overlay document itself, (b) information about and from the quality process for which the document was constructed, and (c) 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.” Their editorial underscored 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” because “data papers will be treated in the same fashion as regular papers” (Newman & Corke, 2009, p. 587). Pfeiffenberger and Carlson (2011) launched Earth System Science Data, a journal devoted specifically 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, Chavan and Penev (2011) promoted “biodiversity data papers” to incentivize data publication. Kennedy, Ascoli, and De Schutter (2011, p. 318) 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” (Kennedy, Ascoli, & De Schutter, 2011, p. 318; 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 data sets. The number of data journals is rapidly growing, so 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

Several initiatives have been launched to establish data journals in various domains ranging from archeology to chemistry, ecology, and oceanography. To identify the target journals 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 this by investigating the “related links” pages. The web sites of these journals were analyzed to identify their core characteristics. When necessary, the editorial teams were contacted to clarify issues and acquire additional information from their valuable and informative feedback.

For this study, 116 data journals published by 15 different publishers were identified. We have clustered journals into sets corresponding to publishers because it is common for publishers to use shared approaches and policies for their journals. This is the case for (a) BioMed Central journals, a set of 85 data journals published by BioMed Central; (b) Chemistry Central journals, a set of three data journals published by Chemistry Central; (c) Pensoft journals, a set of seven data journals published by Pensoft Publishers; (d) SpringerOpen journals, a set of eight data journals published by SpringerOpen; and (e) Ubiquity Press journals, a set of three data journals published by Ubiquity Press.

Although software can be seen as data (see, e.g., Marcus and Menzies, 2010), and this leads to “software journals” supporting “software papers” such as Chemistry Central and SpringerOpen journals, in this survey such types of journals/articles are not included.

An overview of the journals studied is given in Table 1. The table reports for each journal or set of journals: the nature, that is, whether the journal publishes only data papers (“pure”) or any type of paper including data papers (“mixed”); the subject extent, that is, the number of subjects covered; the exploitation, an indicator of the amount of data papers currently published (to December 2013); the offering, that is, the number of journals supporting data papers (to December 2013); the index, that is, the number of journals indexed by professional services (we used Thomson Reuters Web of Science); the length, 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 four pages); and the open access nature, that is, whether the journal is open access. The entire list of journals including a description of

| Characteristic | JBM | JCC | JBP | JES | JEC | JF1 | JGD | JGS | JRR | JNi | JPe | JPO | JSD | JSO |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Nature        | m   | m   | p   | p   | m   | m   | m   | p   | m   | m   | m   | m   | m   | p   |
| Subject extent| 15  | 4   | 3   | 1   | 1   | 2   | 1   | 1   | 3   | 2   | 2   | 3   | 3   | 9   |
| Exploitation  | 534 | 5   | 39  | 79  | 89  | 7   | 1   | 3   | 10  | 2   | 36  | 0   | 0   | 3   |
| Offering      | 85  | 3   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 7   | 1   | 1   | 8   |
| Indexed       | 72  | 2   | 0   | 1   | 0   | 0   | 0   | 1   | 1   | 2   | 0   | 0   | 2   |
| Length        | n   | n   | n   | n   | s   | n   | s   | n   | n   | n   | n   | n   | s   |

Note. Journals are identified as follows. JBM = BioMed Central journals; JCC = Chemistry Central journals; JBP = Dataset Papers in Science; JES = Earth System Science Data; JBP = Ecology; JF1 = F1000 Research; JGD = Genomics Data; JGS = Geoscience Data Journal; JRR = International Journal of Robotics Research; JEC = Neuroinformatics; JPE = Pensoft journals; JPO = PLOS ONE; JSD = Scientific Data; JPO = SpringerOpen journals; JUP = 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.
each journal and a reference to its website is given in the Supporting Information.

In terms of nature, most journals are mixed (cf. Figure 1). Only four of the 15 publishers considered release journals exclusively dedicated to data papers. The overall number of pure data journals is seven (6% of the sample).

In terms of subject extent, journals cover all the four Scopus subject clusters, that is, health science, life sciences, physical sciences, and social sciences and humanities (cf. Table 2). In particular, in analyzing the Scopus journals classification (and supplementing it with 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. Figure 2). However, these figures are partially biased by the number of journals that 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 13.33% of the journals, namely, (a) immunology and microbiology; (b) mathematics; (c) pharmacology, toxicology, and pharmaceutics; and (d) psychology.

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

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 to December 2013. However, the number of journals publishing at least one data paper is growing (cf. Figure 3). In 2013, 37 diverse journals published 195 data papers, and, in the first month of 2014, nine diverse journals published 47 data papers.

In terms of index, 69.82% of the journals of the sample are indexed by Thomson Reuters (cf. Figure 4). However, this figure is biased by the presence of “mixed” journals in the sample, and it is difficult to derive an indication for data papers. The only observation that can be safely reported is that none of the “pure” journals is yet indexed either by this professional service or by other services such as SCImago or Scopus.

### 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                  |

FIG. 1. Nature of data journals. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

FIG. 2. Number of published journals by topic. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
In terms of paper size, because of the fact that most data journals are actually mixed journals, there is no special arrangement for data papers, including for the number of pages. Only a few journals (six of the 116 analyzed), namely, Ecology, Genomics Data, International Journal of Robotics Research, and Ubiquity Press Journals, envisage data papers as artefacts consisting of a few pages. This indicates that journals tend to provide authors of data papers with how much space they need to describe their data sets properly along with the process leading to them.

In terms of open access nature, almost all the analyzed data journals are open access (cf. Figure 5). Only three are considered not to be open access journals: International Journal of Robotics Research is based on subscription fees, but data papers are freely available; Ecology is based on subscription fees; and Neuroinformatics, a journal that is offering the “open access choice” with a fee of €2,200. This is discussed in greater detail later.

To analyze the effectiveness of data journals as a means to promote data publication and reuse, a model describing major concepts and relationships is introduced in the next section along with a discussion of 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 describe a data set properly, how to promote data set availability, how to cite a data set and guarantee rewards properly, how to guarantee data set quality, and how to guarantee open access to data sets. Each
of these sections ends in an “observations” paragraph in which critical comments are provided on the approaches discussed.

**Data Paper Concepts and Naming**

The data paper landscape is highly varied, and there are many different understandings and implementations of this concept. To deal with this variety, we introduce a simple model for this concept (cf. Figure 6) that allows us to describe and analyze the different solutions in a more uniform way.

The concept of a data paper has at least two elements that have to be materialized into concrete and identifiable information objects: the data set (the subject of the data paper) and the data paper itself (the artefact produced to describe the data set). From this point onward, the term data paper is used to refer to the artefact only. This artefact is homologous with articles for traditional journals; it is expected to have an identifier and a content with title, authors, abstract, number of sections, and references.

Both the data paper and the data set 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 are usually selected by the entity managing the two elements, that is, the data journal editor for the data paper and the data archive manager for the data set (and often results in proprietary, ad hoc solutions).

Note that data paper metadata should not be confused with data paper content. In some cases, the confusion may originate from the understanding of the content of a data paper as a sort of metadescription of the data set itself. The data paper is instead analogous to 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.

Different journals implement diverse approaches for the management of data papers. The data journal publishes data papers, meaning that the data papers are primary objects of concern for the journals. Data sets are considered as secondary objects maintained in repositories that may be managed by the same journal editor or, most often, by third-party organizations specializing in data archiving. As a consequence of this state of affairs and the lack of standards in this area, there is great 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 discussed here, different journals name them differently, often as a consequence of naming choices made when this concept was not as popular as it is now (cf. Table 3 and Figure 7). The alternative names often

![FIG. 6. Data papers concept map. [Color figure can be viewed in the online issue, which is available at wileyonlineibrary.com.]](image)

| Name                  | JBM | JCC | JDP | JES | JEc | JF1 | JGD | JGS | JRR | JNi | JPe | JPO | JSD | JSO | JUP |
|-----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Data article          | ✓   |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Data descriptor       | ✓   |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Data in brief         | ✓   |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Data note             |     | ✓   |     |     |     |     |     |     |     |     |     |     |     |     |
| Data original article |     | ✓   | ✓   |     |     |     |     |     |     |     |     |     |     |     |
| Data paper            |     | ✓   | ✓   | ✓   |     |     | ✓   | ✓   |     |     |     |     |     |     |
| Database article      | ✓   | ✓   | ✓   | ✓   | ✓   |     |     | ✓   |     |     |     |     |     |     |
| Database paper        |     | ✓   | ✓   | ✓   | ✓   | ✓   |     |     |     |     |     |     |     |     |
| Dataset paper         |     | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |     |     |     |     |     |     |     |
| Genome database       | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |     |     |     |     |     |     |

**Note.** Journals are identified as follows. JBM = BioMed Central journals; JCC = Chemistry Central journals; JDP = Dataset Papers in Science; JES = Earth System Science Data; JEc = Ecology; JF1 = F1000 Research; JGD = Genomics Data; JGS = Geoscience Data Journal; JRR = International Journal of Robotics Research; JNi = Neuroinformatics; JPe = Pensoft journals; JPO = PLOS ONE; JSD = Scientific Data; JSO = SpringerOpen journals; JUP = Ubiquity Press journals.
also reflect specific purposes for which the data papers have been introduced or a specific understanding of them. Some examples are given below with reference to the journal using them and the article’s intended purpose. Data article in the International Journal of Food Contamination is the main article type because 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 used for papers describing any data set; data descriptors in Scientific Data is used for descriptions of scientifically valuable data sets; data in brief in Genomic Data is used for detailed descriptions of genomic data, including experimental methods and any quality-control analysis; data note in BioMed Central journals is used 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 used for documenting an original data release realizing a significant data contribution to the journal field; database article in BioMed Central journals is used 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; and genome database in Human Genomics (BioMed Central) is used for descriptions and/or evaluations of databases providing information regarding the human genome. In some cases, the same journal has diverse types of data papers, for example, 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 and taxonomic paper and species inventory are for domain-specific data, namely, taxonomic or nomenclatural acts, systematic lists of taxa with notes, species observations, and species inventory.

**Observations.** From this analysis it emerges that editors started assigning diverse names to the same conceptual entity, that is, a data paper. In some cases, names are intended to capture a specific typology of data set that the paper should be about, for example, 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, such as the discovery of data papers across journals.

**How Data Journals Describe Data Sets**

As previously discussed, the purpose of a data paper is to describe a given data set, as a scientific paper describes a research outcome. The data paper is expected to promote exploitation and citation of the data set by giving details such as the methods and protocols used to create and process the data set, the data set structure and format, and the reuse potential. No scientific analysis made by using the data should be described, nor results or 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). Unlike research papers, however, for which over the time a number of common elements have emerged (e.g., each paper must have an abstract, an introduction, a related work section), data papers have no de facto common guidelines. In some cases these guidelines are very
detailed, including information on how the manuscript should be structured, for example, *BMC Journals*; in other cases the guidelines are generic, leaving a certain degree of freedom on the content of the paper, for example, *Earth System Science Data*. Some journals have developed their own templates to convey guidelines for manuscript production and expected content.

The guidelines and templates for the content of the paper contain rules and advice on two classes of information, traditional scholarly communication-related information and specific information on the data set. For the former, almost all the journals agree on including title, authors, abstract, key words, and references. As for the latter, the resulting picture is more heterogeneous. By analyzing existing templates and guidelines, we have identified the following 10 classes of data set information promoted by journals.

- Availability, to provide data set access attributes, namely, a DOI or a URI.
- Competing interests, to provide explicit declaration of any factor (including personal or financial relationships) that might influence the related data set, including factors affecting the production or the presentation of the data set.
- Coverage, to provide data set “extent” attributes, including spatial and temporal coverage.
- Format, to provide information oriented to promote the actual reuse of the data set, such as data format, encoding, and language.
- License, to provide information oriented to data set 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 data set, including goal and funding sources.
- Provenance, to provide information describing the methodology (including the tools) leading to the production of the data set.
- Quality, to provide information on qualitative aspects of the data set, including data set limitations and anomalies.
- Reuse, to provide information promoting potential uses of the data set.

These classes of information are in some cases presented together; in other cases they are presented across diverse sections of this article. A detailed picture of the classes of information used by each journal is given in Table 5, which 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 the authors the freedom to describe their data sets. This is the case for *Earth System Science Data Journal*, which prescribes only the presence of availability-related information even if the editors suggest paying attention to the guidelines for reviewers to determine the potential content of a data paper.

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

### Table 4. Data paper production approaches promoted by data journals.

| JBM | JCC | JDP | JES | JEC | JFI | JGD | JGS | JRR | JN1 | JPO | JSD | JSO | JUP |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |

**Note.** Journals are identified as follows: JBM = BioMed Central journals; JCC = Chemistry Central journals; JDP = Dataset Papers in Science; JES = Earth System Science Data; JEC = Ecology; JFI = F1000 Research; JGD = Genomics Data; JGS = Geoscience Data Journal; JRR = International Journal of Robotics Research; JN1 = Neuroinformatics; JPO = PLOS ONE; JSO = SpringerOpen journals; JUP = Ubiquity Press journals.

### Table 5. Data set information explicitly promoted by journals via guidelines and templates.

| JBM | JCC | JDP | JES | JEC | JFI | JGD | JGS | JRR | JN1 | JPO | JSD | JSO | JUP |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |

**Note.** Journals are identified as follows: JBM = BioMed Central journals; JCC = Chemistry Central journals; JDP = Dataset Papers in Science; JES = Earth System Science Data; JEC = Ecology; JFI = F1000 Research; JGD = Genomics Data; JGS = Geoscience Data Journal; JRR = International Journal of Robotics Research; JN1 = Neuroinformatics; JPO = PLOS ONE; JSO = SpringerOpen journals; JUP = Ubiquity Press journals.
dedicated writing tool, the Pensoft Writing Tool (PWT; Smith et al., 2013). This is an online tool that supports collaborative production of a data paper. It is still based on a template approach (there are many templates, including a data paper template), yet it guides the authors step by step in properly filling in the template sections by making it possible to select automatically 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). This tool is specifically conceived to support the production of metadata for data sets of primary biodiversity data to be published through the GBIF network. However, the tool is also equipped with a facility for automatically generating a data paper manuscript from the data set metadata. The author is requested only 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 data set metadata and regenerating the manuscript thus to improve simultaneously the data paper and the data set description.

Observations. From this analysis it emerges that the only information that the editors require to be necessarily specified in all data papers is the data set’s availability. The lack of a core, shared set of information to characterize data sets is a strong limitation if data journals and data papers are expected to promote the real use of the data sets that are their subject. It is fundamental to develop a shared, open, flexible, and rich data characterization framework that should be used across disciplines and across the boundaries of the community for a data journal. 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 data sets possible, no matter the primary domain for which such data are collected. Moreover, it should be as open and flexible as possible to fit easily within diverse scenarios and domains. The availability of such a shared framework may allow 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, such as 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, such as automatically extracting information from the data set for supporting the compilation of the data paper.

How Data Journals Promote Data Set Availability

A data paper is always associated with a data set. The solution of implementing this association by submitting supplementary files with the article is being progressively discontinued in favor of publishing the data sets in an appropriate repository and creating the association through a link between the two artefacts. Usually, journals propose a list of “recommended trusted data repositories” or “qualified data repositories” where data sets should be deposited. These are intended as repositories meeting certain criteria. From our research it has emerged that usually such repositories must fulfill the following basic requirements to be considered qualified: (a) They must be accredited, that is, internationally or institutionally recognized; (b) they must guarantee long-term availability of the data sets and permanent access; (c) the data sets in the repositories must have a unique digital object identifier, a DOI, that must be included in the related data paper; and (d) the data sets in the repositories must be available free of charge and without any barriers, except for a possible registration to obtain a free login.

The recommended repositories may include institutional repositories, such as UCL Discovery; national repositories, such as the British Atmospheric Data Centre (BADC) or DANS-EASY; and international repositories, such as DataVerse repositories (King, 2007), Dryad (White, Carrier, Thompson, Greenberg, & Scherle, 2008), Figshare, PANGAEA, or Zenodo. Some of these repositories are discipline specific, such as Worldwide Protein Data Bank, whereas others offer generic data hosting as a service, such as Dryad, Figshare, and Zenodo.

The author guidelines provided by F1000 Research, for example, recommend that “for some data types, 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 in which no suitable, more specialized public data repository exists, such as GBIF for species-by-occurrence data and taxon checklists or GenBank for genomic 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 a few exceptions to this rule; for example, the journal Ecology requires that data sets are published in the Ecological Archives, a proprietary repository of the Ecological Society of America for publishing material associated with their journals. Another exception is the journal GigaScience that recommends data set deposition in the GigaDB repository (Sneddon, Li, & Edmunds, 2012). This repository primarily hosts data and

---

1 UCL Discovery http://eprints.ucl.ac.uk/
2 British Atmospheric Data Centre (BADC) http://badc.nerc.ac.uk
3 Data Archiving and Networking 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
tools associated with articles in *GigaScience*, but it also includes data sets 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*, which promotes a “hybrid” model, that is, data sets 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.

To homogenize the data deposition strategy, journal editors have introduced a Joint Data Archiving Policy (JDAP), which requires that supporting data for the papers that they publish, be they traditional papers or data papers, be publicly available via appropriate public archives. In some cases, journals establish special arrangements with repositories that offer their service to authors with payment of a fee. For example, *BMC Bioinformatics* authors can obtain a complimentary subscription to LabArchives, with an allotment of 100 MB of storage. LabArchives is “an Electronic Laboratory Notebook which will enable scientists to share and publish data files in situ.” Data files linked to published articles are provided with DOIs and remain persistently available. The journal guidelines remind authors that “use of LabArchives or similar data publishing services does not replace preexisting data deposition requirements, such as for nuclear acid sequences, protein sequences, and atomic coordinates.”

### Observations

The analysis conducted so far highlights that editors are embracing the publication of new types of journal articles, but they do not consider the publication of data as part of their own mission. On the contrary, there is currently fast growth of specialized data repository providers who envisage that scientific data publication and maintenance are a promising new market. In parallel, data sets are progressively emerging as “first-class entities”; they have a value per se, so their publication is no longer a side effect of the publication of a scientific article. In the future this new role played by data sets may completely change the panorama and also partially reduce the motivations that have pushed the introduction of data papers. As a matter of fact, data set 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 data sets by actual users. However, although innovative and enhanced data publication and citation practices will be put into 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 played neither by data sets with their metadata nor by research papers because they have an information payload that is unique (Callaghan, 2013).

### How Data Journals Support Data Set Citation

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

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

The approaches that data journals promote for citation of the data sets are given in Table 6. Journals tend to have a section specifically dedicated to report on data availability. In some cases this section is the abstract, as in *Earth System Science Data* and *Genomic Data*. In some cases (see “Reference format” in Table 6), journals have developed the format that authors should use to refer to data in a dedicated section (e.g., *BioMed Central* journals). Often, journals require that data sets be included in the reference list of the paper, relying on DCC recommendations (Ball & Duke, 2012) or on the DataCite recommendations (Starr & Gastl, 2011). With regard to the use of persistent identifiers, for obvious reasons, almost all the journals promote this mechanism as the one to use for accessing data sets in an

**TABLE 6. Data set citation approaches promoted by journals for data papers.**

| JBM | JCC | JFP | JES | JEl | JFD | JGD | JGR | JG1 | JP1 | JPO | JSD | JSO | JUP |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |

**Note.** Journals are identified as follows. JBM = BioMed Central journals; JCC = Chemistry Central journals; JFP = Dataset Papers in Science; JES = Earth System Science Data; JEl = Ecology; JFD = F1000 Research; JGD = Genomics Data; JGR = Geoscience Data Journal; JG1 = International Journal of Robotics Research; JG1 = Neuroinformatics; JG1 = Pensoft journals; JP1 = PLOS ONE; JG1 = Scientific Data; JSO = SpringerOpen journals; JUP = Ubiquity Press journals.

---

9 Joint Data Archiving Policy (JDAP) [http://datadryad.org/pages/jdap](http://datadryad.org/pages/jdap)

10 Electronic Laboratory Notebook—LabArchives [http://www.labarchives.com/](http://www.labarchives.com/)
unequivocal manner. In the majority of cases, journals recommend DOIs in accordance with data repositories that used to assign DOIs to deposited data sets. The identifier is often displayed as a URI.

**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 data set citation. If data papers should be the artefact to cite in research papers willing to cite the data sets they use, then it is fundamental that journals contribute to the development and diffusion of emerging attempts to systematize and standardize data citation practices, such as the CODATA data citation principles (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013) and the Force11 data citation principles (Force 11 Data Citation Synthesis Group, 2014).

**How Data Journals Guarantee Data Set 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 often discussed; for example, Solomon (2007) examined the role of peer review for journals and the benefits resulting from new review modes, and Lee, Sugimoto, Zhang, and Cronin (2013) highlighted the bias affecting this process.

The peer review concept when applied to data journals 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 data set peer review, we analyzed the practices that data journals have in place for the review of data papers.

Almost all the data journals we have analyzed perform peer review to some extent. The differences emerge when analyzing (a) the review process approaches, that is, whether they are open, semiopen, or closed, and (b) the features that reviewers are requested to assess. A summary of the differences is given in Table 7.

Most data journals adopt the conventional scheme of closed peer review, namely, prepublication, anonymous, private peer review. Few journals adopt open peer review to promote fairness and objectivity and to reduce publication time, namely, *Earth System Science Data* and *F1000 Research*. *Earth System Science Data* uses a two-stage approach. After submission, manuscripts are published on the web as “discussion papers” after a light review performed by a dedicated editor to evaluate the paper, except for its scientific content (e.g., if it fits with the journal scope), and to suggest minimal technical corrections such as typing errors. The paper remains at this stage for a 8-week period during which the community can review and discuss it. Every discussion paper receives at least two referee comments. After the public discussion phase, the authors are requested to reply publicly to comments and to produce a revised manuscript; if approved by the editor, it is finally published in the journal. In the case of *F1000 Research*, after passing through 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, who will have to judge whether the work seems scientifically sound by assigning it one of three different statuses, approved, approved with reservations, or not approved. All status and referees’ comments are published along with the manuscript, or actually a version of it. In fact, this process leads to the publication of multiple versions of the manuscript, which is “accepted” only when it has been assigned 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 such as PubMed, Scopus, and Google Scholar.

Some journals adopt a semiopen approach; they make available the prepublication history of accepted papers, including the initial submission, reviews, and revisions. This approach is used, for example, by some of the BioMed

---

**TABLE 7. Data journal peer review approaches and criteria.**

| Criteria                  | JBM | JCC | JDP | JES | JF1 | JGD | JGS | JRR | JNi | JPa | JPO | JSD | JSO | JUP |
|--------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Open process             | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| Semiopen process         | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| Close process            | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | n/a | ✓   | ✓   | ✓   | ✓   | ✓   |
| Data paper specific      | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   |
| Manuscript quality       | 3   | 3   | n/a | 5   | 3   | 12  | 7   | 2   | n/a | 29  | 6   | 4   | n/a | 3   |
| Consistency              | 1   | 1   | n/a | 3   | 3   | 1   | 2   | 3   | n/a | 9   | 5   | 1   | 1   | 1   |
| Data quality             | 2   | 2   | n/a | 5   | 6   | 6   | 4   | 4   | n/a | 8   | 3   | n/a | 3   | 2   |
| Data reusability         | 1   | 1   | n/a | 5   | 3   | 9   | 5   | 2   | n/a | 2   | 5   | 2   | 4   | 1   |
| Utility                  | n/a | n/a | n/a | 3   | 4   | n/a | 6   | 2   | n/a | 4   | 2   | 1   | n/a | n/a |

*Note. Journals are identified as follows. JBM = BioMed Central journals; JCC = Chemistry Central journals; JDP = Dataset Papers in Science; JES = Earth System Science Data; JF1 = F1000 Research; JGD = Genomics Data; JGS = Geoscience Data Journal; JRR = International Journal of Robotics Research; JNi = Neuroinformatics; JPa = Pensoft journals; JPO = PLOS ONE; JSD = Scientific Data; JSO = SpringerOpen journals; JUP = Ubiquity Press journals.*
Central journals such as *BMC Anesthesiology* and *BMC Cancer*.

In the case of *Biodiversity Data Journal*, in addition to the “community” peer review (a closed 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 analyzing existing evaluation criteria, we have identified the following five classes of criteria.

- Quality of manuscript, the conventional criteria for assessing manuscript writing, clarity, organization, adherence to template.
- Consistency between data paper and data set(s), criteria for assessing the effectiveness of the data paper content as a means for accessing the data set(s).
- Data quality, criteria for assessing the methodologies leading to the production of the data set(s).
- Data reusability, criteria for assessing the actual reusability of the data set(s).
- Utility and contribution of data, criteria for assessing the potential of the data set(s) for the community.

A detailed picture of the emphasis each journal associates with the key classes is given in Table 7, in which the number of questions in the guidelines is reported by class. Most questions fall into the manuscript quality class. Significant numbers of questions are dedicated to assessing information on quality and reusability of the data set(s) reported in the paper. For instance, Ubiquity Press journals include among the top criteria methods and reuse sections.

**Observations.** Our perspective on data set(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 in which the data are expected to be used; usually this is the “fitness for use,” and there is no standard for data quality that is globally accepted. (b) In some cases data sets 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) Assessing data quality is a challenging task that is fundamentally different from evaluating a scholarly paper; often it is not feasible for a human to analyze data as a reviewer does for a paper, because data frequently are large in volume and complex in structure. From the analysis of the data paper review processes, it emerges that most existing journals rely on a traditional review approach that remains almost hidden from the data paper audience; for example, review reports are known to paper-related authors, reviewers, and editors only. Moreover, in only a few cases, applied review strategies are specialized to deal with data papers, and, in the majority of the cases, assessing the quality of these artefacts is conceived as equivalent to assessing the quality of a traditional paper. Actually, data papers are a particular kind of paper, so dedicated quality assessment criteria should instead be defined. These criteria should be oriented to evaluating the capacity of the artefact to provide its readers with the minimal set of information needed to allow an effective reuse of the data set(s) associated with the paper. Moreover, disclosure of the review process should be encouraged as a means for proving evidence of the data paper quality.

**Data Journals and Open Access**

Almost all of the data journals analyzed are open access. However, open access usually implies publication costs, and it does not necessarily mean free access or being license free. This section discusses three aspects related to open access practices, (a) the costs associated with the publication of the data paper, (b) the license associated with the data paper, and (c) the license associated with the data set.

Journals publishing data papers tend to have article processing charges (Solomon & Björk, 2012; Van Noorden, 2013). Unlike a previous analysis of open access journals’ publication fees conducted by Kozak and Hartley (2013), in which only a small percentage of journals charged authors for publishing, for data journals this is the standard approach. In the case of “mixed” journals, charges may be specific for data papers, such as *Ecology*, or may be the same for all the types of papers accepted by the journal, such as BioMed Central journals. A detailed description of the costs is given in Figure 8 and Table 8. On average, the article processing charge is approximately €1,300 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, whereas it is approximately €1,360 for “mixed” journals. Journals tend to ask no charges during the launch phase; for example, *GigaScience* (a BioMed Central journal) is not levying charges after the publication of 2 volumes and 16 issues, and *Journal of Systems Chemistry* (a Chemistry Central journal) is still paying its own publication costs after 4 volumes and 12 issues. In some cases, the publication of the data set per se in a repository has a cost, so 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.

For data paper copyright and licenses, 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 nonopen 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 with their open access paper by selecting among a number of Creative Commons licenses, such as *Genomics Data* and *Scientific Data*. However, there is no major difference among the selectable licenses; all of them carry the obligation to attribute the paper properly when using it.
For data set copyright and licenses, the analyzed journals agree on the need to make the data set described in the paper accessible free of charge for noncommercial uses both during the review phase and once the paper has been accepted. Similarly to papers, data set use carries the obligation to give credit to the original author. Among the most common licenses are the CC0 and the Open Data Commons Attribution License. However, there is no guideline on

**FIG. 8.** Article processing charges by journal. Journals are identified as follows. J_BM = BioMed Central journals; J_CC = Chemistry Central journals; J_DSP = Dataset Papers in Science; J_ES = Earth System Science Data; J_EC = Ecology; J_F1000 = F1000 Research; J_GD = Genomics Data; J_GS = Geoscience Data Journal; J_IBJR = International Journal of Robotics Research; J_NI = Neuroinformatics; J_P = Pensoft journals; J_PLOS = PLOS ONE; J_SD = Scientific Data; J_SO = SpringerOpen journals; J_UP = Ubiquity Press journals. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
whether the data set user should give credits by using the
data paper, the data set per se, or both.

Observations. From this overview it emerges that data
journals promote free access to both the data sets and the
papers describing them. However, the costs of this “open-
ness” 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 data sets in a selected
repository that may ask for another fee. These fees are part
of the costs involved in data publication, which is one of
the major barriers affecting the whole data publication
movement (Costas et al., 2013). The need to reduce these
costs and make the whole process of data publication more
efficient is a topic widely discussed and supported by the
stakeholders involved in data publication; for example,
funding agencies are developing specific arrangements and
policies for open access to research data in research
programs.

Conclusions and Prospects

This article has surveyed existing data journals to analyze
their approaches to data publication and to identify the
extent to which they facilitate sharing and reuse of data. Our
investigation shows 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; 23.5% of the existing data papers were published
in 2013. Most data journals (69.82%) are indexed by well-
known professional services, namely, Thomson Reuters
Web of Science. Moreover, conferences have started pro-
moting data papers; for example, the International Sympos-
ium on Biomedical Imaging has just launched a call for
data papers for the ISBI 2014 to be held in China, April 29
to May 2, 2014.11

We analyzed current data paper practices and approaches
implemented by journal editors from the perspective of five
core aspects contributing to data sharing and reuse, how they
promote data set description, how they promote data set
availability, how they support data set citation, how they
guarantee data set quality, and how they promote open
access to data sets. From our observations it emerged that
journal editors (a) do not yet have a shared and consolidated
strategy for promoting an effective data set description
favoring data reuse (e.g., the only data set information that is
promoted by all the journals is availability, with other infor-
mation such as coverage, quality, and “reuse” neglected in
many cases); (b) usually rely on services offered by third-
party data repositories and archives for making data sets
available; (c) are overlooking many of the issues affecting
data set citation by assuming that citation approaches
mirrored from the scientific publication model are sufficient;
(d) have not yet deeply addressed the issues related to data
set 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, in which authors are
requested to pay an article processing fee, and asking to
authors to deposit the data set(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, much work is
needed to reach a common understanding on what these
practices should be to contribute to data sharing and data
reuse. Promoting data publication practices is not a sole
responsibility of journals. Rather, journals are an integral
part of the entire ecosystem underlying scientific communi-
cation infrastructures (Castelli, Manghi, & Thanos, 2013),
so they are called upon to develop innovative, comprehen-
sive, and effective data publication practices in such settings.

When designing data papers in the future it is fundamen-
tal to identify and highlight the added value that a “data
paper” has for data sharing and reuse with respect to rich,
detailed, and curated data set metadata that are managed by
data repositories. Data papers are scientific communica-
tions, so they have the benefits of this kind of communica-
tion with respect to metadata starting from the target
audience to the intended mission and goal. However, scien-
tific communication is potentially affected by the problems
highlighted by Nosek and Bar-Anan (2012b), that is, no,
slow, incomplete, inaccurate, or unmodifiable communica-
tion. To design data papers in the future it is worthwhile to
discuss how each of these problems impacts 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, such as causing a waste of resources by replicating
activities that have already failed. Data identification and
collection, an activity preceding any data paper production,
are intellectually intensive and time-consuming activities,
and the ability to communicate negative results in producing
them is important; this can prevent others from incurring
similar results, and “buggy” data sets can be used to assess
the robustness of certain algorithms when dealing with noisy
data. Slow communication is a problem not affecting data
papers more than traditional papers. Data paper publication
practices are already overcoming this problem, and new
practices can be promoted, such as open peer review, to
reduce communication time further. Incomplete or inaccu-
rate communication happens because authors do not report
everything or because errors are difficult to detect. The lack
of information underlying the production of the data set
that a data paper describes is an issue that can severely
hinder data set reuse (Thanos, 2014). Authors should
describe what is important rather than describing what they
think is important. Unmodifiable communication happens
because, once published, articles are static entities. For data

11IEEE International Symposium on Biomedical Imaging http://
biomedicalimaging.org/2014/
papers and data sets, it is fundamental to guarantee that there is a powerful versioning system allowing researchers to identify the “right” research artefact.

When designing data papers in 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 potential to expand the machinery that the artefact author can use to convey its “content,” namely, the data set description; at data paper access time the user can be provided with a number of graphs automatically produced out from the data set, 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, and different views of the data set 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 (see, e.g., P.A. Lawrence, 2003; Nosek & Bar-Anan, 2012a). It is a responsibility of scientists to assist the rest of the scientific communication realm to remove the barriers affecting it, and this paper represents a contribution in this direction.

Acknowledgments

The work reported was 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 thank M.B. Baldacci (ISTI-CNR) for her valuable support and many helpful comments and the anonymous reviewers for their helpful comments and suggestions.

References

Asher, A., Deards, K., Esteva, M., Halbert, M., Jahnke, L., Jordan, C., . . . Walling, D. (2013). Research data management: Principles, practices, and prospects. Washington, DC: Council on Library and Information Resources.

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

Bardi, A., & Manghi, P. (2014a). Enhanced publications: Data models and information systems. LIBER Quarterly, 23(4). Retrieved from https://liber.library.uu.nl/index.php/ql/article/view/8445/9825

Bardi, A., & Manghi, P. (2014b). A rationale for enhanced publications. LIBER Quarterly. Retrieved from https://liber.library.uu.nl/index.php/ql/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 from https://www.force11.org/white_paper

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

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

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., . . . 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

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/s10799-013-0106-7

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

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

CODOATA-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

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

Douglas, 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, DC: The National Academies Press.

Force 11 Data Citation Synthesis Group. (2014). Joint Declaration of Data Citation Principles—Final. Retrieved from http://www.force11.org/datacitation

Halevi, G., & Moed, H.F. (2012, September). 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., & Tolke, K. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: 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 and Research, 36(2), 173–199. doi: 10.1177/0049124107036660

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Wächter, J. (2006). Data publication in the open access initiative. Data Science Journal, 9(4), 317–320. doi: 10.1007/s12021-010-9084-8

Kirsch, P., Wächter, J. (2006). Data publication in the open access initiative. Neuroinformatics, 9(4), 317–320. doi: 10.1007/s12021-010-9084-8

Kozak, M., & Hartley, J. (2013). Publication fees for open access journals: Different disciplines—Different methods. Journal of the Association for Information Science and Technology, 65(2), 251–262. doi: 10.1002/asi.22988

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Wächter, J. (2006). Data publication in the open access initiative. Data Science Journal, 9(4), 317–320. doi: 10.1007/s12021-010-9084-8

Kozak, M., & Hartley, J. (2013). Publication fees for open access journals: Different disciplines—Different methods. Journal of the Association for Information Science and Technology, 65(2), 251–262. doi: 10.1002/asi.22988
