Data papers as a new form of knowledge organization in the field of research data

Joachim Schöpfel, Dominic Farace, Hélène Prost, Antonella Zane

Data Papers as a New Form of Knowledge Organization in the Field of Research Data

Joachim Schöpfel

University of Lille, GERiiCO laboratory

Associate professor in information sciences, ORCID 0000-0002-4000-807X

joachim.schopfel@univ-lille.fr

Dominic Farace

GreyNet International, Amsterdam

Director of the Grey Literature Network Service, ORCID 0000-0003-2561-3631

dominic.farace@textrelease.com

Hélène Prost

CNRS, GERiiCO laboratory

Information engineer, ORCID 0000-0002-7982-2765

helene.prost007@gmail.com

Antonella Zane

University of Padova, Library System

Head of the Antonio Vallisneri Bio-medical Library, ORCID 0000-0001-7218-6068

antonella.zane@unipd.it

Abstract

Data papers have been defined as scholarly journal publications whose primary purpose is to describe research data. Our survey provides more insights about the environment of data papers, i.e. disciplines, publishers and business models, and about their structure, length, formats, metadata and licensing. Data papers are a product of the emerging ecosystem of data-driven open science. They contribute to the FAIR principles for research data management. However, the boundaries with other categories of academic publishing are partly blurred. Data papers are (can be) generated automatically and are potentially machine-readable. Data papers are essentially information, i.e. description of data, but also partly contribute to the generation of knowledge and data on its own. Part of the new ecosystem of open and data-driven science, data papers and data journals are an interesting and relevant object for the assessment and understanding of the transition of the former system of academic publishing.

Keywords

Data papers, research data, knowledge organization, open science, data journals, FAIR principles, academic publishing

Titre

Les articles de données comme nouvelle forme d'organisation des connaissances dans le domaine des données de recherche

Résumé

Les articles de données ont été définis comme des publications de revues scientifiques dont l'objectif principal est de décrire les données de recherche. Notre enquête fournit davantage d'informations sur l'environnement des documents de données, c'est-à-dire les disciplines, les éditeurs et les modèles économiques, ainsi que sur leur structure, leur longueur, leurs formats, leurs métadonnées et leurs licences. Les articles de données sont un produit de l'écosystème émergent de la science ouverte axée sur les données. Ils contribuent aux principes de FAIR pour la gestion des données de recherche. Cependant, les frontières avec les autres catégories de publications scientifiques sont en partie floues. Les articles de données sont (peuvent être) générés automatiquement et sont potentiellement lisibles par machine. Les articles de données sont essentiellement des informations, c'est-à-dire des descriptions de données, mais ils contribuent aussi en partie à la production de connaissances et de données par eux-mêmes.

Mots clés

Articles de données, données de recherche, organisation des connaissances, science ouverte, revues de données, principes FAIR, publication scientifique

INTRODUCTION

In the context of open science, an increasing volume of research data are made available on Internet, contributing to the so-called big data of science. New tools, methods and infrastructures have been developed for the dissemination, processing, analysis and preservation of research data. Data papers are part of them.

Data papers are a young species of academic publishing. In 2006, Pärtel stated for the field of ecology that "*until now (...) very few data papers have appeared*" (p.99). In fact, most of the data papers or papers about data papers have been published since 2008 and 2009 [1]. Yet, as Smith (2011) reminds, "*the concept has actually been around for quite a while (even if) the older journals that date from the print era tend to be not particularly useful in the modern environment*" (p.16). In fact, one (the first?) data journal (*Journal of chemical and engineering data* from ACS) was already launched in 1956 (see the timeline in Garcia-Garcia et al. 2015).

The simplest definition is that data papers focus on "*information on the what, where, why, how and who of the data*" rather than original research results (Callaghan et al. 2012, p.112).

Data papers have been defined as "*a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal*" [2]. They are published in specific data journals like *Data in Brief* (Elsevier) and *Scientific Data* (Nature), or in regular academic journals with special sections for data papers, like *BMC Research Notes GigaScience* (Oxford University Press) and *PLoS One.* Most data papers are published in journal platforms; yet, some are (also or exclusively) published on data repository platforms [3]. Unlike usual research papers, the main purpose of data papers is to describe datasets, including the conditions and context of their acquisition and their potential utility, rather than to report and discuss results. Also, it is generally assumed that data papers are short papers with up to 4 pages.

In the "classical" research paradigm, the focus is on articles presenting results while research data are useful for the validation of published research findings. Data papers invert the roles, insofar the paper's main function is to inform about and link to research data on data repositories, contributing to their findability and reusability. Are data papers complementary to research papers, or will they replace them, as a seamless and direct way of providing access to research results?

Also, traditional knowledge organization makes a clear distinction between research results (datasets), the analysis and discussion of these results (papers) and the description (cataloguing, abstracting and indexing) of those datasets and papers. This emerging category of data papers appears to challenge this clear distinction, interlinking datasets, papers and metadata, blurring boundaries, changing priorities and modifying the basic purpose of academic publishing.

Built on an overview of recently published studies, the following study produces an empirical update on the publishing of data papers: the number and development of data papers and journals, the country and language of publications, the platforms and publishers, as well as the business models. The purpose of our paper is to analyse data papers as a new tool of scientific communication and to produce insight on their contribution to the organization of scientific knowledge via questions pertaining to the production and the functions of data papers:

- How are they "written"?
- Which is the link with data repositories, metadata and other papers?
- Which is the (potential and real) part of automatic or semi-automatic production
- Which is the part of human added value?

- Which degree of standardization, which link between metadata formats and the data journals' author guidelines?
- In which way are data papers related to the so-called "FAIR Guiding Principles for scientific data management and stewardship" (Wilkinson et al. 2016)?
- Do they just improve the referencing of datasets on repositories, or do they fulfil other roles?
- Are data papers "written by machines" and meant *in fine* to be "read by machines"?

The paper will conclude with a conceptual approach to data papers as part of the organization of knowledge based on research data, in the context of open science.

1 – LITERATURE OVERVIEW

1.1 Definitions and functions

An increasing number of journal editors announce the launch of a new section with data papers. They put forward different objectives, even if the main purpose is similar: to inform about research data and to foster their accessibility and reuse. Three examples among others illustrate the diversity of goals:

- The objective of *The International Journal of Robotics Research* is "*to facilitate and encourage the release of high-quality, peer-reviewed datasets to the (...) community*" (Peter & Corke 2009, p.587).
- *Studies in Family Planning* tries to promote "*interdisciplinary research and integrative analyses by making accessible to researchers, policymakers, students, and donors data that may be useful in answering critical questions of interest to (...) readers*" (Friedmann et al. 2017, p.291).
- The French journal of information and communication sciences *RFSIC* invites data papers to describe the scientific process, methods and tools that result in research data in a Bruno Latour perspective, "*since they never just magically appear*" (Le Deuff 2018, §2).

The publisher Pensoft describes a data paper as "*a scholarly journal publication whose primary purpose is to describe a dataset or a group of datasets, rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments in support of the data, as found in a conventional research article*" (Penev et al. 2012).

The term remains ambiguous. For instance, Bordelon et al. (2016) define data papers as "*papers that present, analyze, or use data obtained with the respective facilities*" (i.e. observatories) (p.1). Pärtel (2006) consider data papers as a kind of "abstracts" that aim to collect, organize, synthesise, and document data sets of value in a given field; only the abstract appears in a data journal (or the data paper section of a regular journal) while the data and metadata are available through a field-specific data repository on the Internet. For Penev et al. (2012), their purposes are three-fold: "*to provide a citable journal publication that brings scholarly credit to data publishers; to describe the data in a structured human-readable form; (and) to bring the existence of the data to the attention of the scholarly community*". At first sight, data papers, in spite of their common general purpose, appears to belong to a rather heterogenous and dissimilar new kind of documents. Our study will reveal, nevertheless, more common features, such as the fundamental structure.

1.2 Data journals

A first survey on data journals was conducted by Candela et al. (2015), with a sample of 116 data journals published by 15 different publishers. They distinguished 7 "pure" data journals publishing only data papers and 109 "mixed" data journals publishing any typology of paper including data papers. The most represented subjects (in terms of number of journals) were Medicine (53%), Biochemistry, Genomics and Molecular Biology (26%), and Agricultural and Biological Sciences (16%). They identified only 9 data journals in social sciences and humanities (8%). Their results show a recent and slowly developing landscape (the average number of data papers per journal is <10), with conceptual, structural and terminological diversity (they identified 10 different terms assigned for data papers). Also, there is no consensus about the usual content, the only section present in all data papers being the data availability (location, accessibility), followed by information about the provenance of the dataset. Most of the data journals perform some kind of traditional peer review to guarantee a certain level of the papers' quality but also to assert some quality of the datasets, in terms of utility and reusability; only few journals adopted an "open peer review". Most journals are published in open access, with an average APC [4] amount of 1,300 euros.

The Grey Journal, published by Textrelease (Amsterdam) is one of those "mixed" data journals. Initially a regular journal with papers from international conferences and original research articles, *The Grey Journal* started to publish a collection of data papers in 2017. This collection was born out of an 'Enhanced Publications Project' fueled by the FAIR Data Principles (Farace et al. 2018). The main pillars for this collection are the International Conference Series on Grey Literature, the research data that is created and archived within this framework (actually 37 datasets housed in the Dutch data repository DANS), and the existence of a flagship journal for the publication of the data papers. A standardized template is provided to ensure the identity and longevity of the collection and to guide prospective authors and researchers in submitting a data paper.

The template consists of five sections each of which has a note field providing examples and/or a maximum word count. The fields are labelled as follows: overview, methods, data description, potential reuse, and references. Currently, 7 of GreyNet's 37 datasets are supported by a data paper (19%). Yet, even on a small scale this data paper collection illustrates an operational and functional ecosystem of open science constructed year after year with five main elements, i.e. an academic community, original research within this community, conferences, a journal, and a data repository. In this emerging framework, data papers gain their particular relevance, different from regular articles.

1.3 Features and metadata

Yet, other aspects appear challenging the idea of a clear distinction between data papers and regular papers. Li et al. (2019) conducted a content analysis with 82 data papers from 16 journals to investigate what information they describe regarding the methods to create and manipulate the data objects (i.e. "data events"). For Li and his colleagues, even if they have distinct features from research articles, data papers are "*nevertheless created under similar conditions*", and they reveal "*functional overlaps*" between both categories, related to the narratives of data events (natural language) and to their composition which is "*inevitably situated in the specific epistemic communities*". Their main function is to improve the findability of published datasets and, through enriched metadata description, to foster their reusability.

Metadata are constitutive for data papers. Candela et al. (2015) produced a conceptual map of the data paper (see figure 1). They insist not to confuse the data paper's content, its metadata, and the datasets' metadata. "*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" (p.1752).

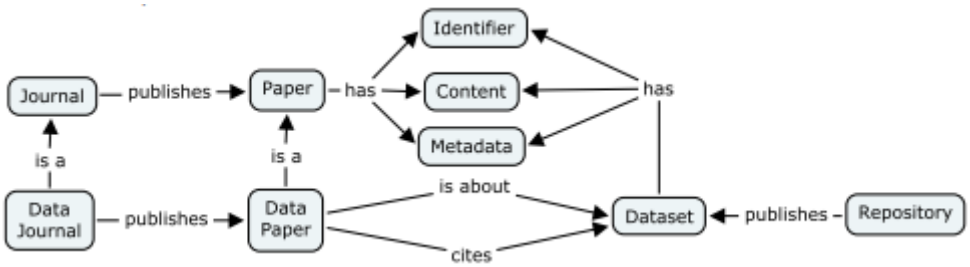

Figure 1. Data papers concept map (from Candela et al. 2015 p.1752)

The link between data papers and the metadata of research data is essential because both have similar functions (describe data, define accessibility, (re)usability, and content. Insofar data papers are about deposited datasets and insofar deposits require metadata, data papers can be (partly) derived from existing metadata.

Chavan & Penev (2011) describe a tool that "*facilitates conversion of a metadata document into a traditional manuscript for submission to a journal*" (p.7) for biodiversity resource datasets. The human contribution is minimal if the metadata is standardized (with controlled vocabulary), exhaustive, and of sufficient quality: "*Once the metadata are completed to the best of the author's ability, a data paper manuscript can be generated automatically from these metadata using the automated tool (...) The author checks the created manuscript and then submits it for publication in the data paper section through the online submission system of an appropriate (...) journal*" (p.7).

This kind of generated data papers can be further enhanced in different ways, such as "*describing fitness for use of data resources (which) will increase the usability, verifiability and credibility of those resources*", persistent identifiers, an "*interpretive analysis of the data (which) could include taxonomic, geospatial or temporal assessment of data and its potential of integration with other types of data resources*" or the inclusion of "*a taxonomic checklist and/or the data themselves*". Data papers represent a highly standardized type of publication, with a standard structure and a content which is largely defined in terms of metadata formats (such as DataCite Metadata Schema) and identifiers for datasets, persons etc. (such as DOI and ORCID).

1.4 Production and processing

In fact, Chavan & Penev (2011) describe an integrated workflow of data repositories and journal platforms, requiring shared standards and formats. Senderov et al. (2016) provide an example of this data paper generation in the field of biodiversity. Their workflow relies on three key standards (RESTful API's for the web, Darwin Core and EML) and imports metadata into the ARPHA writing tool (AWT). In other words, and more generally spoken, "*the boundary between a workflow tool, a data store, and a publishing platform blurs*" (de Waard 2010, p.9).

But are data papers produced only for machines? No, according to Li et al. (2019) who are convinced that "*as a genre built upon natural languages, data papers are primarily a human-readable document, much less designed for reproducing data workflows in computational approaches*" (p.18). Both are complimentary, rather than competitive.

In her review of data papers, Reymonet (2017) compares data papers and data management plans (DMP). Indeed, as the expected structure of such an article may be based on the items provided when preparing a DMP, Reymonet suggests a tool (or workflow) to export selected items of DMPs in order to prepare or generate a data paper.

A general assumption is that data papers, like regular papers, are peer reviewed, implying some kind of quality control and selection. This means, too, that metadata of research data (and, indirectly, the datasets themselves) become object of scientific evaluation which "*contributes to the popularity of data papers in increasingly more scientific fields*" (Li et al. 2019 p.2, see also Costello et al. 2013). For the same reason, data papers contribute to the trustworthiness of research data. For example, Elsevier's *Chemical Data Collections* invites authors to submit data papers because this "*ensures that your data (...) is actively peer reviewed (...)*" [5]. As cited above, Pärtel (2006) mentions that data papers were about "*data sets of value in a given field*" which implies a selection by the authors themselves, upstream of the writing of data papers and of peer reviews, even if the criteria of selection remain uncertain.

1.5 Critics and outlook

Similar to most cited authors, Smith (2011) states that data papers "*are like traditional research papers in some aspects: they are formally accepted, they are peer-reviewed, they are citable entities*" but then adds that "*in other respects they are very different from traditional research articles because they are not about the research, they are about the data*" (p.15). And this exactly is the main reason for some more critical voices, expressing concerns about the real demand by society and research, about the additional workload for authors and peer reviewers, and about the motivation of scientists to share their data. The underlying idea is that scientists should (and mostly do) publish about results, not about data.

Other arguments against data papers are their price (APCs) and the slow uptake, at least initially. "*To address professional recognition and data quality control, there are viable alternatives to the data paper (such as the) implementation of a joint data-publishing and -archiving policy by databases and journals (...) instead of popularizing a new kind of publication, it is more important to improve current peer-review processes and the operating policies and integration of journals and databases*" (Huang et al. 2013, p.5). Huang's critic may be specific for a given field of research (here, biodiversity) but should be taken into account for a general understanding of the future development of data papers.

Nevertheless, data journals and data papers appear to be here to stay. The French national plan for open science recommends "*as part of its government support for journals (...) the adoption of an open data policy associated with articles and the development of data articles and data journals*" (MESRI 2018, p. 6-7). While data papers become a legitim (mainstream) part of the landscape of academic publishing, only few studies provide empirical or conceptual elements of an answer to the question of how exactly data papers contribute to the organization of scientific knowledge, compared to regular research articles.

2 – METHODOLOGY

In order to analyse specific features of data papers, we established a representative sample of data journals, based on lists from the European FOSTER Plus project [6], the German wiki forschungsdaten.org hosted by the University of Konstanz [7] and two French public research organizations [8]. The complete list consists of 82 data journals, i.e. journals which publish data papers. They represent less than 0,5% of academic and scholarly journals. For each of these 82 data journals, we gathered information about the discipline, the global business

model, the publisher, peer reviewing etc. The analysis is partly based on data from ProQuest's Ulrichsweb database, enriched and completed by information available on the journals' home pages.

Some data journals are presented as "pure" data journals *stricto sensu*, i.e. journals which publish exclusively or mainly data papers. We identified 28 journals of this category (34%). For each journal, we assessed through direct search on the journals' homepages (information about the journal, author's guidelines etc.) the use of identifiers and metadata, the mode of selection and the business model, and we assessed different parameters of the data papers themselves, such as length, structure, linking etc.

The results of this analysis are compared with other research journals ("mixed" data journals) which publish data papers along with regular research articles, in order to identify possible differences between both journal categories, on the level of data papers as well as on the level of the regular research papers. Moreover, the results are discussed against concepts of knowledge organization.

3 – Results

Four of the 28 data journals have ceased, and two have merged. All of them are published online while 9 have still a print version. One data journal is a report series.

3.1 Research disciplines

Most data journals are from STEM domains, in particular from life and medical sciences, including genetics (see figure 2). Only four journals publish data from humanities (psychology, archaeology) and social sciences. One data journal covers a large range of disciplines from sciences (*Scientific Data* by Nature), another is open for all topics in social sciences and humanities (*Research Data Journal for the Humanities and Social* Sciences by Brill).

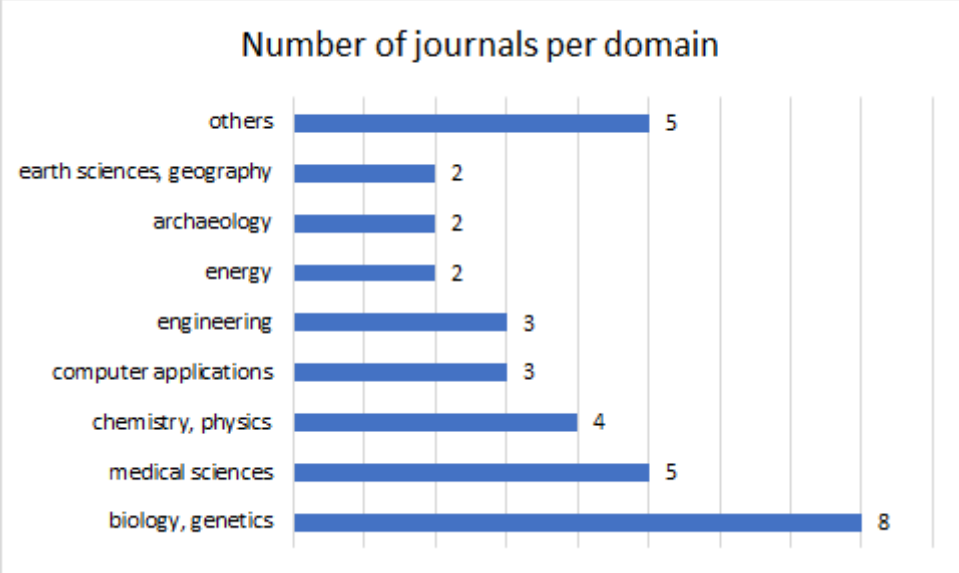

Figure 2. Number of data journals per domain (N=28)

The 5 data journals with papers on data from arts, social sciences and humanities represent 18% of all "pure" data journals. In terms of articles (see below), they represent less than 4% of all data papers published in data journals, with estimated 400-450 papers, mostly in archaeology.

3.2 Publishers

Except for Taylor & Francis, all big five academic publishers (Elsevier, Springer-Nature, Wiley-Blackwell, Taylor & Francis and SAGE) have their own data journals. Five data journals are published by Elsevier (from which two are published by Academic Press, an imprint of Elsevier, two others merged), two by Wiley, one by Springer-Nature and one by SAGE.

Other data journals are published or hosted by newcomers, especially by open access publishers such as Ubiquity Press (3 journals), BioMed Central (2 journals) Hindawi, MDPI, Copernicus Publications, Pensoft or Faculty of 1000, by smaller publishing houses like Brill or De Gruyter (Sciendo) or by learned societies or university presses (AIP, ACS, Wageningen…).

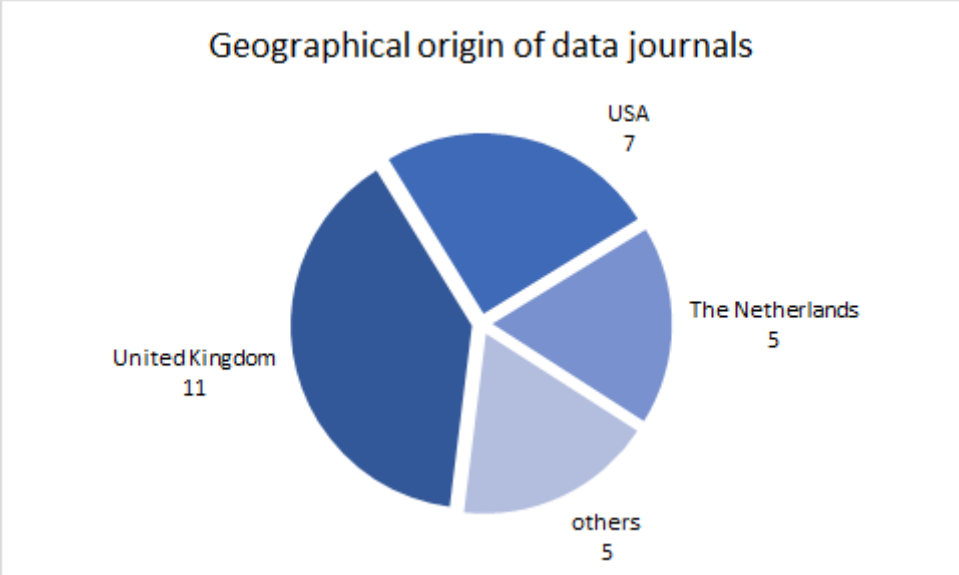

Figure 3. Geographical origin of data journals

Most of the data journals are published in three countries, i.e. the United Kingdom, the United States and The Netherlands. The other journals are from Bulgaria, Switzerland, Germany and Poland (figure 3). All are published in English, only one journal also publish papers in another language, Dutch (*Research Data Journal for the Humanities and Social Sciences*).

3.3 Business models

Most of the data journals are "young" products, with a short history. Only seven journals have been launched before 2000. The other 21 journals have been launched during the last ten years, from 2008 on, and especially in 2013 (7 journals) and 2014 (5 journals). Four journals have ceased or are suspended.

At least one part of the data journals are considered as good or high quality journals. 11 data journals are indexed by Clarivate Analytics, 8 by Elsevier's Scopus database. 16 journals are referenced in the international Directory of Open Access Journals (DOAJ).

The overall number of data papers published by these data journals is approximately 11,500, with large differences, ranging from some papers up to more than 3,500. The median number, however is rather low, with 97 (figure 4).

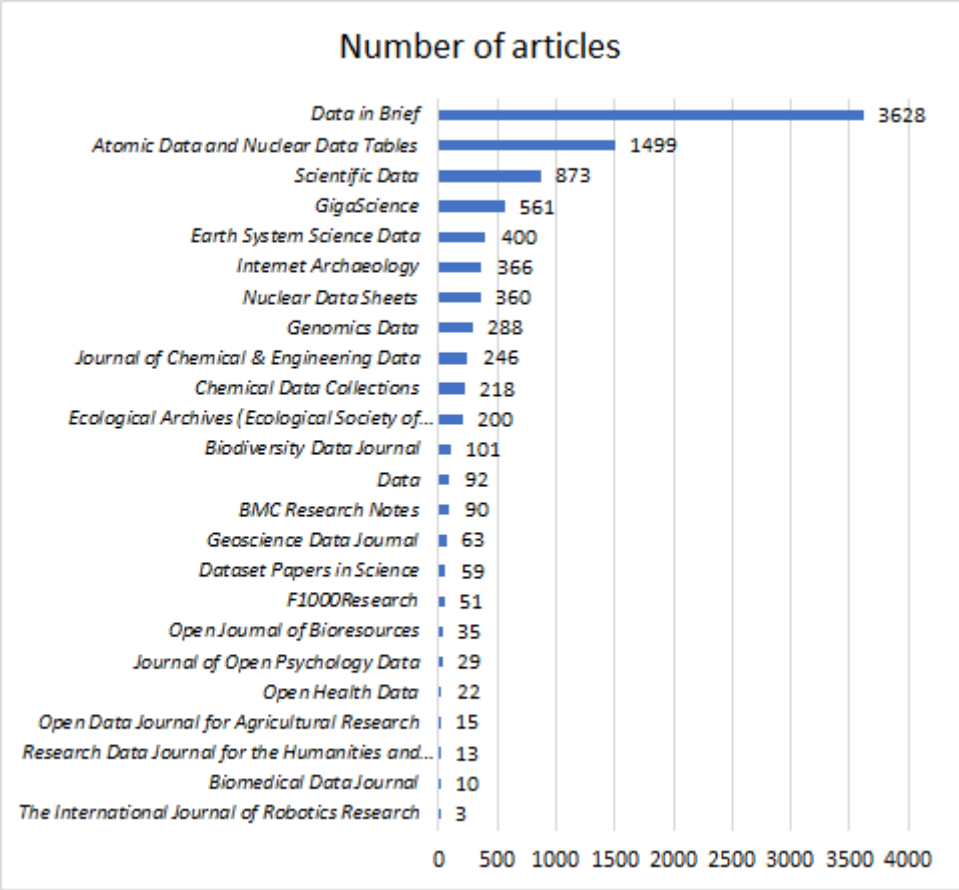

Figure 4. Number of data papers per journal (with best estimates)

In terms of volume, Elsevier's *Data in Brief* is by far the most important data journal, followed by Elsevier's "old" *Atomic Data and Nuclear Data Tables* (launched in 1979) and *Scientific Data*, a NatureResearch journal from Springer Nature. Together, these three journals represent more than half of the data papers published in pure data journals.

The major business model is OA Gold, mostly with APCs (19) but also without (2). 4 journals are hybrid, and only one journal is available through the traditional subscription model (figure 5).

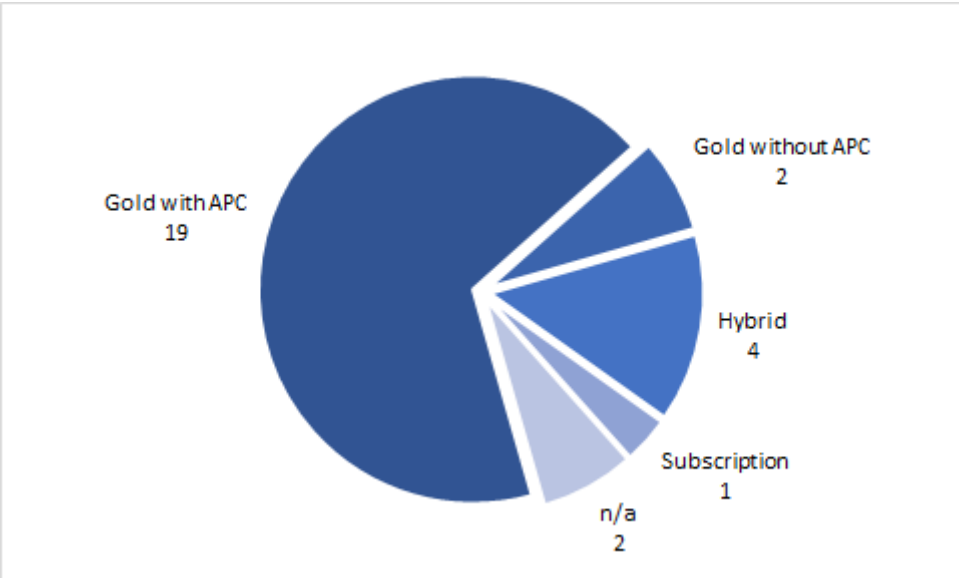

Figure 5. Business models

In this small sample there is no "diamond OA journal" without subscription and APCs. In other words, 25 journals (89%) are OA journals or allows OA publishing, and in 23 journals (82%) authors have to pay for OA.

All data journals covering arts, social sciences and humanities are OA journals, all with APCs.

3.4 Licensing

21 data journals disseminate data papers with an open license, most often a CC-BY license, sometimes together with a public domain license (CC0) or the more restrictive CC-BY-NC-ND or CC-BY-NC-SA licenses (no commercial re-use).

Elsevier proposes (also) its own user license.

Only one journal does not propose an open license for the dissemination of the data papers but keeps the full copyright (*Journal of Physical and Chemical Reference Data*).

3.5 Selection

Except for one title (*European Power Watch*) all data journals explicitly inform about some kind of formal selection procedure. Often the information for authors just mention "peer review" but six describe the selection as a single-blind review process where the identities of the reviewers are not disclosed to the author(s). One journal applies a "quick peer review" with focus on the data value and potential re-use but does not explain who does the peer review and how long it takes (*Chemical Data Collections*).

5 data journals apply some kind of innovative open peer review, either as an option (if required) or for all submitted papers. Yet, this term has different meanings:

- the reviewers are suggested (and known) by authors (*F1000Research*);
- community peer review (*Biodiversity Data Journal*);
- interactive public peer review (*Earth System Science Data*).

The last procedure is particular interesting: all referee and editor reports, the authors' response, as well as the different manuscript versions of the peer-review completion (post-discussion review of revised submission) will be published if the paper is accepted [9].

3.6 Structure and length

We already mentioned that it is generally assumed that data papers are short texts, up to 4 pages. In fact, this is only partly true. In this sample, only 5 journals require short papers, limited to 4-6 pages or maximal 3,000 words. Most journals do not limit the length of submitted papers or make the usual recommendations (6-10 pages, or maximal 6,000 words). One journal only accepts short abstracts (*Ecological Archives*), while others publish papers well beyond the length of regular papers, up to 20 or 30 or even 100 pages, including detailed data descriptions, illustrations (figures) or data tables like *Atomic Data and Nuclear Data Tables*. On the other hand, data journals in the field of arts, social sciences and humanities publish generally shorter data papers.

No results, no discussion, no conclusion: usually the data journal guidelines for authors contain these or similar recommendations, like Elsevier's *Data in Brief* which asks authors to "*avoid using words such as 'study', 'results', and 'conclusions'*" [10]. Quite different, the *Atomic Data and Nuclear Data Tables* guidelines leave it to the authors whether or not to include results, discussion, and conclusion to the description of the data.

Nearly all journals require or suggest a particular structure, and some of them provide a template with mandatory sections. However, there is no standard structure. Instead of a generally accepted succession of sections, data papers are made of three constitutive elements, i.e. an introduction with information about the context and the rationale, a more or less detailed description of the datasets with specifications (sometimes formalized as disciplinary or generic metadata of data, such as the DataCite Metadata Schema or the DDI [11]), and a section of materials and methods, instrumentation, on the production of the data and procedures, sometimes extended to experimental designs and calculation (figure 6).

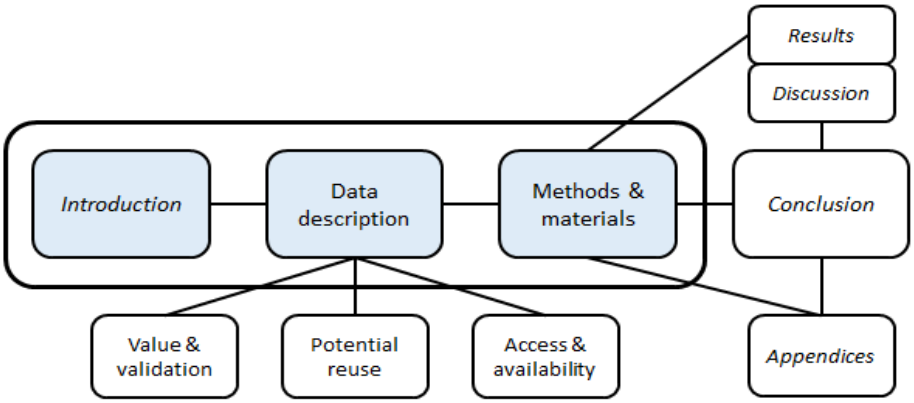

Figure 6. Sections of a data paper

The figure presents a core structure with three central sections (in blue), with other, optional or peripheral sections, some of them similar to regular papers (in italics), others characteristic for data papers, such as:

- Value & validation: information about the (potential or real) value of the datasets and the quality control (validation), like peer review, automatic procedures (technical validation) etc.
- Potential reuse: information about potential usage, about reuse and the potential interest for scientists or other users.
- Access & availability: information about the address of datasets (repository, URL) and the availability, including access and reuse rights and limitations; this part may include implementation details, about the availability of source code and requirements, and about the availability of supporting data and materials.

Information about access and availability may also be part of the appendices, like acknowledgements, references, competing interests, author roles and information, rights and permissions, or even peer review comments.

As mentioned above, some data journals allow or invite sections about results of data analysis, together with a discussion of these results and an outlook on further research, very similar to the usual structure of scientific articles and blurring the frontiers between both types of papers.

A last aspect: no invitation or guidelines were found concerning machine-based generation and/or automatic processing of data papers. Apparently, the publishers' platforms do not support automatic ingestion of text files (via FTP of repository metadata or similar) but require manual deposits of manuscripts and authorship. Of course, this requirement does not

exclude partly or complete machine-based generation of data papers upstream of the human deposit of manuscripts.

3.7 Metadata and identifiers

Two types of metadata must be distinguished regarding data papers, i.e. metadata of the described datasets, and metadata of the data papers themselves.

- Metadata of datasets: as mentioned above, some data journals requires a detailed and formalized description of datasets, in a format which potentially compliant with metadata. But only few journals insist on a specific standard. Two examples: *Ecological Archives* expects strict adhesion to the metadata content standards derived from a set of generic metadata descriptors published by the Ecological Society of America (Michener et al. 1997); the metadata set should be sent to the editor as a separate text file. *Genomics Data* requires compliance with an internal standard for data description with eight fields [12]. Both formats have in common that they are community-specific, disciplinary metadata standards. A third example is quite different, generic and limited to the datasets' identifiers: *Scientific Data* requires an ISA-Tab [13] metadata text file where the DOI of all datasets are mentioned.
- Metadata of data papers: most journals ask for some general and usual information, compliant with the Dublin Core format, such as author, organisation, title etc. *F1000Research* recommends XML Schema, Xlink, MathML, or the NLM Journal publishing DTD (JATS) [14].

26 journals publish the data papers with a DOI (93%), and 5 also include the author identifier ORCID (18%). Also, most of them recommend if not require a standard identifier (DOI) or at least a stable address for the described datasets.

3.8 Linking

All data papers provide information about the availability of the described datasets, mostly together with an address (URL), but they do it in different ways:

- usually in a special section of the paper with a statement on data access and availability,
- in an appendix which contains a declaration with data availability and address,
- in the abstract,
- as part of the metadata.

Some papers contain downloadable data; others require that the described datasets should be deposited in one or a shortlist of recommended repositories.

4 – Discussion

4.1 A new ecosystem

Compared to former studies, the number of data journals and papers appears to increase slowly, on a low level. Garcia-Garcia et al. (2015) identified 20 pure data journals; four years later, our sample consists of 28 data journals and not all are still active and even pure (see below). 28 journals represent less than 0.01% of the academic and scholarly serials (source: Scopus). Arts, social sciences and humanities are nearly missing (2 journals in 2015; 4 in 2019). The number of data papers progressed at a faster pace, from 846 in 2013 (Candela et

al. 2015) to an estimated number of 11,500 data papers in 2019. Yet, this volume represents roughly 0.4% of the overall number of articles published in 2017 (source: Scopus).

Also, the interest of data papers and journals is not their volume but the fact that they clearly are a product of the emerging ecosystem of data-driven open science. Four aspects characterise this embeddedness in the new environment:

Business model: The dominant business model (gold OA with APCs) is different from the traditional and still prevailing serials landscape, and it appears already compliant with the requirements of the new plan S [15].

Reuse rights: most data journals allow publishing with an open license, often with generous reuse and remixing rights (e.g. CC-BY license and/or CC0 waiver).

Findability: the editorial model of data journals requires standard identifiers for the datasets, e.g. DataCite's DOI, to guarantee (and increase) the findability of datasets; they also attribute DOIs to their own data papers, creating a kind of cross-linked DOI system between data papers and datasets.

Interconnectedness: perhaps the most relevant aspect is the integration of data journals and papers in a complex structure of open access journal platforms and data repositories, academic communities, research projects, conferences etc. Interconnectedness requires interoperability between platforms and infrastructures but is more than technology, formats and standards, insofar it means new ways of doing science, including research management, research environment, workflows etc.

A fifth aspect, i.e. evaluation and selection, is already visible but still in transition and not dominant. Data journals replace the usual evaluation and selection procedure (double-blind peer review) by partly open single-blind peer review and, for already one out of five journals, by some kind of open peer review, including innovative community peer review and interactive public peer review. They can also contribute to the assessment of data value through the follow-up of citations (Belter 2014).

4.2 FAIR principles and beyond

Most data journals have never been produced as traditional serials but are a pure (and young) product of this new ecosystem of open access, open (and big) data, and new forms of selection and dissemination. This makes them particular, different from other academic journals. And this makes them also particularly interesting for the requirements of the so-called "*FAIR Guiding Principles for scientific data management and stewardship*" (Wilkinson et al. 2016). Their data papers contribute to these principles in different ways, in order to improve the findability, accessibility, interoperability, and reuse of research data, e.g. [16]:

- **Findable**
  - *F2. Data are described with rich metadata:* data papers enrich existing metadata of datasets.
  - *F4. (Meta)data are registered or indexed in a searchable resource:* the enriched metadata are registered, indexed and preserved on the data journal platform.
- **Accessible**
  - *A2. Metadata are accessible, even when the data are no longer available:* the accessibility of metadata published via data papers does not depend on the datasets' accessibility in a data repository.
- **Interoperable**
  - *I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation:* at least some of the data journals insist on the

application of formal, standard language (vocabularies) for the description of datasets. As a minimum, they reproduce the data repositories' own formal dataset representation.
    - *I3. (Meta)data include qualified references to other (meta)data:* data papers can (and usually do) provide links to other related resources, e.g. research papers, institutional affiliations, similar or related datasets, etc.
- **Reusable**
    - *R1.1. (Meta)data are released with a clear and accessible data usage license:* as mentioned above, most data papers are published with an open license; whenever the data paper is derived from the original metadata, this license may depend on the repository's initial licensing and reuse rights.
    - *R1.2. (Meta)data are associated with detailed provenance:* one of the main functions of data papers is to provide detailed knowledge about where the data came from, who to cite, who generated or collected it and how has it been processed (workflow).

Along with metadata, data papers contribute to the compliance with FAIR principles, in particular to the two principles of findability and reusability, insofar they help people (and machines [17]) finding datasets and inform about the provenance and reuse rights. Additionally, data papers contribute to another aspect, beyond the FAIR principles, i.e. the evaluation of the datasets' quality and value.

In the context of open science, metadata has been considered fuel for economy (Neuroth et al. 2013). As a new vector of communication of metadata on research data, data papers can be defined as a kind of pipeline for this fuel. Yet, as they also add value to metadata, through contextual information, evaluation, new identifiers etc., they are not only pipelines but also refineries, more or less specialised, more or less standardized. To stay with the fuel metaphor, data papers are a new infrastructure of refinement and dissemination of the metadata fuel.

Regarding knowledge organization, two aspects require attention and further investigation:

Standardization: the quality of data papers depends for much on the quality of the metadata of the underlying datasets; and this means, on controlled terminologies, on standard formats, well-defined elements etc. One example is the International Geo Sample Number (IGSN) designed to provide an unambiguous globally unique persistent identifier (PID) for physical samples (specimens) and to facilitate the location, identification, and citation of physical samples used in research [18]. The development of data papers and data journals should (will) be accompanied by further work on standards, by academic communities, publishers, information professionals and knowledge practitioners.

Specialisation: to be relevant and useful, metadata standards should be as compliant as possible with the specific requirements and features of scientific communities, disciplines, methods, tools and equipment. This specialisation, however, tends to limit their interoperability between different domains, infrastructures, information systems… and their interest and usefulness for interdisciplinary research, discovery tools etc. One solution to this problem could be described by "as specific as possible, as generic as necessary", an approach which would apply a kind of ad-hoc-compromise for each particular situation, resulting in many different formats more or less specific, and more or less generic. Another, perhaps more realistic approach would be to accept (and support) two (or more) different standards for each dataset and each data paper, one generic (like, for instance, the DataCite metadata schema), the other specific, depending on the particular domain, method, tool etc.

4.3 Blurred boundaries

The specific identity of data papers is mainly defined in opposition with regular research papers (see for instance Penev et al. 2012). The reality is different. The empirical data of our survey provides evidence that despite a general definition of data papers and journals, there is a lot of divergence and heterogeneity which can be described on four levels.

1. **Data journals also accept other articles.** Our survey put the focus on a limited number of academic and scholarly journals indexed by databases or directories as "pure" data journals. Yet, even in this sample some data journals publish regular research articles, reviews, short communications or comments along with data papers, such as *Data* from MDPI and *Earth System Science Data* from Copernicus.
2. **Data papers are published in other journals.** As mentioned above, one limitation of our survey is the focus on supposedly "pure" data journals. However, an increasing number of academic and scholarly journals accept data papers along with regular research papers, usually in a specific section. Pensoft for instance publishes 37 journals, including one data journal and 16 other journals accepting data papers. The French Agricultural Research Centre for International Development (CIRAD) produced a list with 54 academic journals accepting data papers relevant for agricultural science, including the mega-journal *PLoS One* [19]. It is quite impossible to make an estimation of the real number of such "mixed" data journals and their data papers. Pensoft's *Research Ideas and Outcomes* for instance is part of these new "mixed" data journals but published up to now only one data paper, in biosciences.
3. **Data papers are more than simple data papers.** Even a superficial analysis of data papers reveals that one part of articles labeled as "data papers" do not only describe datasets but add data analysis and discussion of results. *Atomic Data and Nuclear Data Tables*, *Dataset Papers in Science* and *Open Archaeology* are three "pure" data journals which explicitly accept data papers with results and discussion of results. This means that a (unknown) part of data papers in fact are more than simple data papers *stricto sensu* because they communicate results of data analysis.
4. **There are other emerging types of articles, similar to but not identical with data papers**. "Pure" and "mixed" data journals are open for other categories of articles which are neither traditional journal items (research articles, reviews, comments etc.) nor data papers. Sometimes the difference may be a question of terminology. For instance, *F1000Research* accepts "*brief descriptions of scientific datasets that promote the potential reuse of research data and include details of why and how the data were created*" called "*data notes*" [20] - in other words, data papers. But there are other examples (see also the listed terms in Candela et al. 2015):
    a. Data services paper: "*papers on data services, and papers which support and inform data publishing best practices (including) the development of systems, techniques or tools that enable data analysis, data visualisation, data collection and data sharing (and) processes and procedures used in the development of datasets*" (*Geoscience Data Journal*).
    b. Meta or overlay articles: "*Descriptions of online simulation, database, and other experiments, partnering with digital repositories on 'meta articles' or 'overlay articles', which link to and allow visualisation of the data, thereby adding an entirely new dimension to the communication and exchange of data research results and educational materials*" (*Data Science Journal*) [21].

    These two examples of a new kind of papers are quite different, yet they have in common that they are both linked to research datasets and above all, to the dissemination and reuse of research data which is their main purpose.

The boundaries between data papers and data journals and other categories of scientific communication are partly blurred, not only due to a lack of reference definitions but also due to a large diversity of publishing practices. This may have at least three explanations:

- The publishing of data papers is still in transition. It took some decades to develop and accept the IMRAD format as a standard format of scientific article publishing [22]. The heterogeneous character and blurred boundaries of data papers may reflect the emergence of a young and new, still not well-defined form of scientific communication.
- The described proximity with research communities, the "embeddedness" in an ecosystem defined by disciplines, materials, methodologies, tools, etc. contributes to the heterogeneity of data journals and papers. Data papers necessarily depend on the community-specific way of how data is produced, collected, processed, preserved, reused… and it seems quite natural that they will reflect the diversity of this environment. Perhaps, fuzziness is a core element of the data paper category.
- One part of the new OA journals announces an inclusive editorial policy. Instead of a selective approach and guidelines with explicit limitations, they invite submission of all kind of papers; a strategy somewhere between predatory publishing and big data principles based on volume and variety rather than on quality and trustworthiness.

4.4 Who is writing? Who is reading?

Some of the underlying questions of this study were about the production and use of data papers. How are they written, and are they really "written"? Which is the (potential and real) part of automatic or semi-automatic production, and which is the part of human added value? In fact, are data papers written by machines and to be read by machines?

The answer to these questions is neither yes nor no. As mentioned above, data papers can be at least research data available in data repositories such as Dataverse or others (see the Pensoft workflow, Chavan & Penev 2011). The technology is there. Recently, the French National Institute for Agricultural Research (INRA) updated their Dataverse-based repository including an online tool that partly generated "by machines", i.e. through the exploitation and transformation of metadata on researchers can use to generate data papers from the deposits' DOI, in an open text format compliant with INRA's own data journal or with Elsevier's *Data in Brief*. [23]

Both examples, the Pensoft workflow as well as the INRA tool, reveal the potential of automatic generation of data papers, but also its requirements and limits. Automatic generation of data papers requires a high degree of standardization and interoperability between data repositories, text processing tools and journal platforms, especially regarding metadata formats and identifiers. Our study was not about metadata formats of data repositories and about their degree of standardization. But our study reveals a lack of standardization on the other side, i.e. the journal platforms. Paradoxically, this may be an opportunity for automatic generation and ingestion of data papers; yet it will not be helpful for machine-based exploitation of data papers.

The limits of automatic generation of data papers are twofold. On the one hand, journal platforms still and always require authorship, i.e. intellectual property and institutional affiliation. They do not accept automatic submission of machine-produced data papers. On the other hand, the format of data papers requires rich contextual information that may not be part of the datasets' metadata and must be added by the researchers or data officers. Candela et al. (2015) mention that the metadata is usually selected by both the data journal editor (for the data paper) and the data archive manager (for the dataset) which "*often results*

in proprietary, ad-hoc-solutions". Relevant for our question is the human contribution (selection) and the resulting diversity and specificity.

Candela et al. (2015) also insist on the distinction between metadata of datasets, metadata of data papers, and data papers themselves. Metadata [24] are made for machines, and the main purpose of FAIR principles is to improve machine readability and transfer of research data. Data papers are part of this ecosystem, and they contribute to the automatic processing of research data and related metadata. However, the state of the art and our empirical results (still) reveal human added value, i.e. enhancement of the information produced by metadata, such as potential reuse (value), related datasets and research, and other contextual information useful for the understanding of the described data. But as mentioned above, this can also include more traditional content, like results of data analysis and rich textual discussion of data and results. Another "*human added value*" is the intellectual responsibility and property of the data papers which are always attributed to people (authors) not to machines. Instead of machine generated data papers we should speak of "*machine- (or repository-) assisted writing of data papers*".

So, are data papers written for machines? Penev et al. (2012) insist on the "*human-readability*" even of automatically generated data papers. Rich and less standardized and coded textual discussion, for instance, is probably more aimed at human readers. This of course does not exclude the potential of data papers for automatic exploitation with tools of text and data mining (artificial intelligence). Similar to the generation (writing), this potential depends on the standardization of data papers, including careful coding, and their own metadata, i.e. standardized and well controlled formats and terminology. Probably, the fast development of artificial intelligence will facilitate the automatic production as well as the automatic exploitation of data papers and their metadata. However, so far, we didn't identify any study about this potential which for the moment apparently remains theoretical.

4.5 Data? Information? Knowledge?

Finally, what is the informational status of data papers, compared with the DIKW model of information sciences (Rowley 2007)? What do they carry: data, information, knowledge, or wisdom? Following the usual definitions, the answer seems easy: insofar data papers provide description of data, and insofar information is inferred from data and contained in descriptions (Rowley & Hartley 2008), data papers correspond to the second level of the DIKW pyramid, i.e. information (figure 7). They are not knowledge but contribute to the generation of knowledge. Also, the main purpose of data papers - to facilitate the findability and the reusability of research data - is similar to another general aspect of information, i.e. its immediate usefulness for decisions or actions.

This characteristic of data papers is one major difference with regular research articles which are expected to provide more than simple descriptions of facts (data), i.e. insight, understanding, interpretations, hypotheses etc. However, as mentioned above, the boundaries are partly blurred and some data papers do more than carrying information about data, in particular when they include sections with data analysis results and discussions. So at least partly, data papers also convey knowledge, even if this is not part of their core function.

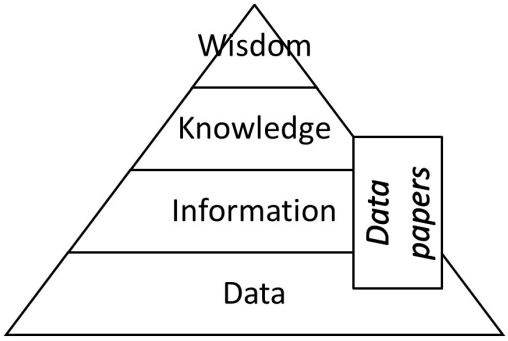

Figure 7. Data papers and the DIKW pyramid

Downside of the pyramid, the boundary to the data level seems equally blurred. Because, as described above, data papers do not only provide information about data but can be exploited as raw data on their own, generating information about research projects, scientific cooperation etc. This means that data papers are also partly data.

For both reasons, data papers do not just improve the referencing of datasets on repositories but fulfil other roles. Their particular information profile can be described in terms of library science, as an original integration (or merging) of writing, cataloguing and indexing, facing major challenges like standards and terminology. Perhaps data papers are a kind of new boundary object (Star & Griesemer 1989) on the frontline between academic publishing and research. Our analysis confirms the statement that data papers are like traditional research papers in some aspects but very different in other respects (Smith 2011). Perhaps data papers are not (only) part of academic publishing but should (also) be considered and assessed as part of research data practice.

CONCLUSION

Data papers have been defined as scholarly journal publications whose primary purpose is to describe research data (facts about data). Yet, the literature overview shows that there is a lack of a generally accepted reference definition of data papers. Likewise, few conceptual studies and empirical evidence are provided. Also, up to now, the success of data papers appears of minor importance and limited to STM disciplines, primarily in the life sciences.

Our survey provides more insights about the environment of data papers, i.e. disciplines, publishers and business models, and about their structure, length, formats, metadata and licensing. Core elements of data papers are the data description and methods and materials; depending on the data journal's policy, other sections are requested or optional, such as value and validation, potential reuse, access and availability, and even results of data analyses and discussion of results.

The discussion section of this study highlights five major aspects of data papers:

1. Data papers are a product of the emerging ecosystem of data-driven open science.
2. They contribute to the FAIR principles for research data management, in particular findability and reusability, and add in some degree to the evaluation of the quality and value of the data.
3. However, the boundaries with other categories of academic publishing are partly blurred, especially with regular research papers.

4. Data papers are (can be) generated automatically and are potentially machine-readable; yet, the human contribution (still) appears vital in terms of intellectual property and richness of content.
5. Data papers are essentially information, i.e. description of data (as defined by the DIKW model) but also partly contribute to the generation of knowledge and data on its own.

As to the two camps human generated vs. machine generated, if a data paper is created by a human – whether or not machine aided, one can speak of knowledge organization. However, if the data paper is solely machine generated it is difficult to attribute this to knowledge organization (excluding any reference to artificial intelligence). The latter is more aligned with automated indexing, cataloguing, and the like.

In relation to the DIKW pyramid, data papers appear between the levels of information and knowledge given that for some people they are not knowledge but only contribute to the generation of knowledge.

However, if one looks at the metadata fields that encompass a full blown data paper – such as the explicit roles of the researchers/authors, the research methods applied, the description of the data, its reusability as well as its limitations, then one may conclude that the data paper provides a fuller understanding of the data/dataset. In itself, the data paper provides a best practice in knowledge organization – if not an example of knowledge generation.

Part of the new ecosystem of open and data-driven science, data papers and data journals are an interesting and relevant object for the assessment and understanding of the transition of the former system of academic publishing. This means that the quality and the usefulness of data papers partly depend on external variables, e.g. the metadata standards of data repositories, their trustworthiness in terms of data quality but also long-term preservation (certification) etc. Therefore, as mentioned above, quality control of data papers (i.e. some kind of peer review) always implies some kind of quality control or evaluation of the datasets themselves and their respective repositories.

Based on our empirical results and former studies, we would suggest the following definition of data papers, keeping in mind the transitional and necessarily provisional character of each conceptual attempt: *Data papers are authored, peer reviewed and citable articles in academic or scholarly journals, whose main content is a description of published research datasets, along with contextual information about the production and the acquisition of the datasets, with the purpose to facilitate the findability, availability and reuse of research data; they are part of the research data management and crosslinked to data repositories*. This definition may not cover all different variants of data papers but will be helpful for a better understanding of what we called “blurred boundaries” and for further investigation.

At this stage, a couple of questions remain open; in particular, the following topics should be addressed:

- Monitoring: how can the indexing of data papers be improved in order to facilitate their identification and follow-up (search engines, databases, data repositories, journal platforms)?
- Business models: what is the risk of predatory publishing of data journals and data papers? Is it different from predatory publishing of regular research papers?
- Disciplines: are data papers as relevant in arts, social sciences and humanities as in life sciences, chemistry etc.? Should their data papers be published in large and multidisciplinary data journals, together with STM, or should they have their own data journals?

- Ecosystem: more case studies are needed on specific links between research data management, academic publishing, and the production and dissemination of data papers, in a given environment and community (equipment, discipline, structure…).
- Evaluation: our study didn't assess whether (and how) scholars get credit for publishing data papers. This, however, will be a key factor for the future development of data papers.

Garcia-Garcia et al. (2015) wondered if data journals will remain part of the research ecosystem or not. Perhaps they will not. However, it seems probable that the number of data papers will continue to grow and gain importance, perhaps (probably) not via data journals but via increasing hybridization of research journals and journal platforms, and perhaps even through the merging of journal and data platforms. In any case, on the boundary between research data management and academic publishing, data papers will continue to provide a highly relevant object for library and information science, especially for the further assessment of the development of academic publishing and knowledge organization in the field of scientific research.

ACKNOWLEDGEMENT

The paper builds on the scientific and professional experience of all authors in the framework of the international Grey Literature Network Service [25] and its international workshops on data papers (Netherlands, Czech Republic, Italy, and the United States), its collection of data papers in the Dutch EASY data repository and its follow-up study of enhanced publications.

DATA AVAILABILITY

The CSV table of the underlying dataset is available in the Dutch repository DANS, at the following address: https://doi.org/10.17026/dans-zk3-jkyb

NOTES

[1] Source: data from Dimensions https://www.dimensions.ai/

[2] Source: Global Biodiversity Information Facility https://www.gbif.org/data-papers

[3] See for instance http://researchdata.cab.unipd.it/122/

[4] Article processing charges: the fee authors or their institutions have to pay (after the acceptation of their papers) to some publishers to be published immediately in open access. The amount of APC is varying between publishers and journals; the average amount research institutions pay per article is about 2,000 euros (see OpenAPC https://treemaps.intact-project.org/apcdata/openapc/).

[5] *Chemical Data Collections*, see https://www.elsevier.com/journals/chemical-data-collections/2405-8300/guide-for-authors

[6] FOSTER portal, see https://www.fosteropenscience.eu/foster-taxonomy/open-data-journals

[7] forschungsdaten.org, see https://www.forschungsdaten.org/

[8] Both in the field of agronomy: INRA https://www6.inra.fr/datapartage/Partager-Publier/Publier-un-Data-Paper and CIRAD http://ou-publier.cirad.fr/formulaire.php

[9] See https://www.earth-system-science-data.net/peer_review/interactive_review_process.html

[10] See https://www.journals.elsevier.com/data-in-brief/about-data-in-brief/how-to-submit-a-data-in-brief-article

[11] DataCite https://schema.datacite.org/ and Data Documentation Initiative https://www.ddialliance.org/

[12] These eight fields are: organism/cell line/tissue; sex; sequencer or array type; data format; experimental factors; experimental features; consent; sample source location.

[13] ISA tools https://isa-tools.org/

[14] Journal Publishing Tag Set https://jats.nlm.nih.gov/publishing/

[15] The plan S gives preference to immediate open access in 100% OA journals, see https://www.coalition-s.org/

[16] The description and numbering of the principles follow the GO FAIR list at https://www.go-fair.org/fair-principles/

[17] The FAIR principles have been initially designed for automatic data processing.

[18] ISGN http://www.igsn.org/

[19] CIRAD, see http://ou-publier.cirad.fr/index.php

[20] *F1000Research*, see https://f1000research.com/for-authors/article-guidelines/data-notes

[21] *Data Science Journal*, see https://datascience.codata.org/about/

[22] IMRAD is a common organizational structure of scientific writing and the usual format of papers on original research published as articles in scientific journals, in particular in empirical sciences but also in other disciplines. It stands for "introduction, methods, results and discussion/conclusion". For more details and references, see https://en.wikipedia.org/wiki/IMRAD

[23] INRA, see https://data.inra.fr/datapartage-datapapers-web/ and https://dataverse.org/blog/data-inra

[24] Metadata considered in the strict sens of the term, i.e. digital data on other digital data.

[25] GreyNet International, Amsterdam; see http://www.greynet.org

REFERENCES

BELTER Christofer W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS One,* March 26, 2014. Disponible sur : https://doi.org/10.1371/journal.pone.0092590 (Consulté le 26/07/2019)

BORDELON Dominic, GROTHKOPF Uta, MEAKINS Sylvia, STERZIK Michael (2016). Trends and developments in VLT data papers as seen through telbib. *Proc. SPIE 9910, Observatory Operations: Strategies, Processes, and Systems* VI, 99102B (15 July 2016). Disponible sur : https://www.eso.org/sci/libraries/SPIE2016/9910-89.pdf (Consulté le 26/07/2019)

CALLAGHAN Sarah, DONEGAN Steve, PEPLER Sam et al. (2012). Making data a first-class scientific output: Data citation and publication by NERC's environmental data centres. *International Journal of Digital Curation,* vol. 7, no. 1, pp. 107-113. Disponible sur : https://doi.org/10.2218/ijdc.v7i1.218 (Consulté le 26/07/2019)

CANDELA Leonardo, CASTELLI Donatella, MANGHI Paolo, TANI Alice (2015). Data Journals: A Survey. *JASIST,* vol. 66, no. 9, pp. 1747-1762. Disponible sur : https://doi.org/10.1002/asi.23358 (Consulté le 26/07/2019)

CHAVAN Wishwas, PENEV Lyubomir (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics,* vol. 12, suppl. 15, S2. Disponible sur : http://www.biomedcentral.com/1471-2105/12/S15/S2 (Consulté le 26/07/2019)

COSTELLO Mark J., MICHENER William K., GAHEGAN Mark, ZHANG Zhi-Quiang, BOURNE Philipp E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, vol. 28, no. 8, pp. 454-461. Disponible sur : https://doi.org/10.1016/j.tree.2013.05.002 (Consulté le 26/07/2019)

DAVIS Grace H., PAYNE Eric, SIH Andrew (2015). Commentary: Four ways in which data-free papers on animal personality fail to be impactful. *Frontiers in Ecology and Evolution,* vol. 3, no. 102, pp. 1-3. Disponible sur : https://doi.org/10.3389/fevo.2015.00102 (Consulté le 26/07/2019)

FARACE Dominic J., FRANTZEN Jerry, SMITH Plato L. (2018). Data Papers are Witness to Trusted Resources in Grey Literature: A Project Use Case. *The Grey Journal,* vol. 14, no. 1, pp. 31–36.

FRIEDMAN Rachel, PSAKI Stéphanie, BINGENHEIMER Jeffrey B. (2017). Announcing a New Journal Section: Data Papers. *Studies in Family Planning,* vol. 48, no. 3, pp. 291-292. Disponible sur : https://doi.org/10.1111/sifp.12032 (Consulté le 26/07/2019)

GARCIA-GARCIA Alicia, LOPEZ BORRUL Alexandre, PESET Fernanda (2015). Data journals: eclosión de nuevas revistas especializadas en datos. *El profesional de la información,* vol. 24, no. 6, pp. 845-854. Disponible sur : https://doi.org/10.3145/epi.2015.nov.17 (Consulté le 26/07/2019)

HUANG Xiaolei, HAWKINS Bradford A., QIAO Gexia (2013). Biodiversity Data Sharing: Will Peer-Reviewed Data Papers Work? *BioScience,* vol. 63, no 1, pp. 5-6. Disponible sur : https://doi.org/10.1525/bio.2013.63.1.2 (Consulté le 26/07/2019)

LE DEUFF Olivier (2018). Une nouvelle rubrique pour la RFSIC : Le Data Paper. *Revue française des sciences de l'information et de la communication,* no. 15. Disponible sur : http://journals.openedition.org/rfsic/5275 (Consulté le 26/07/2019)

LI Kai, GREENBERG Jane, DUNIC Jullian (2019). Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. Preprint accepted by *JASIST*. Disponible sur : https://arxiv.org/abs/1903.06215 (Consulté le 26/07/2019)

MESRI (2018). *National Plan for Open Science*. Paris, Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation. Disponible sur : https://libereurope.eu/wp-content/uploads/2018/07/SO_A4_2018_05-EN_print.pdf (Consulté le 26/07/2019)

MICHENER William K., BRUNT James W., HELLY John J., KIRCHNER Thomas B., STAFFORD Susan G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Archives,* vol. 7, no. 1, pp. 330-342. Disponible sur : https://doi.org/10.2307/2269427 (Consulté le 26/07/2019)

NEUROTH Heike, STRATHMANN Stefan, OSSWALD Achim, LUDWIG Jens (ed.) (2013). *Digital curation of research data*. Glückstadt, Werner Hülsbusch, 2013, 92 p. Disponible sur :

https://univerlag.uni-goettingen.de/bitstream/handle/3/isbn-978-3-86488-054-4/Digital_Curation_SUB.pdf?sequence=1&isAllowed=y (Consulté le 26/07/2019)

NEWMAN Paul, CORKE Peter (2009). Data Papers – Peer Reviewed Publication of High Quality Data Sets. *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 587. Disponible sur : https://doi.org/10.1177/0278364909104283 (Consulté le 26/07/2019)

PÄRTEL Meelis (2006). Data availability for macroecology: how to get more out of regular ecological papers. *Acta Oecologica*, vol. 30, no. 1, pp. 97-99. Disponible sur : https://doi.org/10.1016/j.actao.2006.02.002 (Consulté le 26/07/2019)

PENEV Lyubomir, CHAVAN Wishwas, GEORGIEV Teodor, STOEV Pavel (2012). *Data papers as incentives for opening biodiversity data: one year of experience and perspectives for the future*. Poster. EU BON: Building the European Biodiversity Observation Network. Disponible sur : https://pensoft.net/img/upl/file/DataPaperPoster.pdf (Consulté le 26/07/2019)

REYMONET Nathalie (2017). *Améliorer l'exposition des données de la recherche : la publication de data papers*. Université Paris Diderot. Disponible sur : https://archivesic.ccsd.cnrs.fr/sic_01427978/ (Consulté le 26/07/2019)

ROWLEY Jennifer (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science,* vol. 33, no. 2, pp. 163–180. Disponible sur : https://doi.org/10.1177/0165551506070706 (Consulté le 26/07/2019)

ROWLEY Jennifer, HARTLEY Richard (2008). *Organizing knowledge. An introduction to managing access to information*. Aldershot, Ashgate Publishing.

SENDEROV Viktor, GEORGIEV Teodor, PENEV Lyubomir (2016). Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases. *Research Ideas and Outcomes,* 2:e10617+. Disponible sur : https://doi.org/10.3897/rio.2.e10617 (Consulté le 26/07/2019)

SMITH Mackenzie (2011). Data Papers in the Network Era. In *Charleston Library Conference*. Against the Grain Press, LLC. Disponible sur : http://dx.doi.org/10.5703/1288284314871 (Consulté le 26/07/2019)

STAR Susan Leigh, GRISEMER James R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science,* vol. 19, no. 3, pp. 387–420. Disponible sur : http://dx.doi.org/10.1177/030631289019003001 (Consulté le 26/07/2019)

DE WAARD Anita (2010). The Future of the Journal? Integrating research data with scientific discourse. *Logos,* vol. 21, no. 1-2, pp. 7-11. Disponible sur : http://dx.doi.org/10.1163/095796510X546878 (Consulté le 26/07/2019)

WILKINSON Mark D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3:160018. Disponible sur : https://doi.org/10.1038/sdata.2016.18 (Consulté le 26/07/2019)