Data documents

Joachim Schöpfel, Dominic J. Farace, Hélène Prost, Antonella Zane, Birger Hjorland

▶ **To cite this version:**

Joachim Schöpfel, Dominic J. Farace, Hélène Prost, Antonella Zane, Birger Hjorland. Data documents. Knowledge Organization, 2021, 48 (4), pp.307-328. 10.5771/0943-7444-2021-4-307 . hal-03500944

HAL Id: hal-03500944
https://hal.univ-lille.fr/hal-03500944

Submitted on 22 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Data Documents†

Joachim Schöpfel*, Dominic Farace**, Hélène Prost***,
Antonella Zane****, Birger Hjørland*****

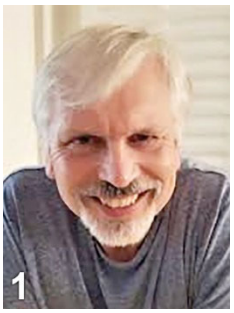
*39 rue de la Paix d'Utrecht, 59000 Lille, France, < joachim.schopfel@univ-lille.fr>

** Javastraat 194-HS, 1095 CP Amsterdam, Netherlands, <dominic.farace@textrelease.com>

***43 rue de Saurupt, 54000 Nancy, France, <helene.prost007@gmail.com>

**** Biblioteca Biologico-Medica "Antonio Vallisneri", Centro di Ateneo per le Biblioteche, Università degli Studi di Padova, Viale Giuseppe Colombo, 3, 35131, Padova, Italy, <antonella.zane@unipd.it>

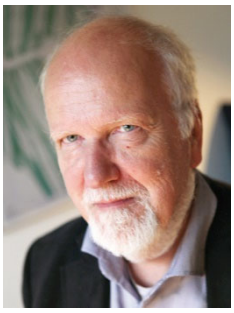
***** University of Copenhagen, Faculty of Humanities, Department of Communication, South Campus, building 14, 2. Floor, Karen Blixens Plads 8, 2300 Copenhagen S, Denmark, < birger.hjorland@hum.ku.dk>



Joachim Schöpfel is associate professor in information sciences at the University of Lille, France, and researcher at the GERiiCO laboratory. He is consultant in scientific communication at the Ourouk consulting bureau, Paris. His research field is scientific information and academic publishing, research data, open science. He has been working at the French Institute of Scientific and Technical Information (CNRS) from 1991 to 2008, at last as the director of the library and document delivery services. He holds a PhD in psychology from the University of Hamburg.

Dominic Farace is head of GreyNet International and director of TextRelease, an independent information bureau specializing in grey literature and networked information. He holds degrees in sociology from Creighton University (BA) and the University of New Orleans (MA). His doctoral dissertation in social sciences is from the University of Utrecht. After six years heading the Department of Documentary Information at the Royal Netherlands Academy of Arts and Sciences, Farace founded GreyNet, Grey Literature Network Service in 1992. He has since been responsible for the International Conference Series on Grey Literature (1993-2020). In this capacity, he also serves as program and conference director, managing editor of the Conference Proceedings, and editor of *The Grey Journal*. He likewise provides workshops and training seminars.

Hélène Prost is an information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiiCO research laboratory (University of Lille). She is interested in empirical library and information sciences and statistical data analysis. She participates in research projects on evaluation of collections, usage analysis, grey literature, research data management and open access, and she is the author of several publications.



Antonella Zane has a doctor's degree on Earth Sciences and ten years of research background in petrology and archaeometry. After research, she started working at the University of Padova (Padua) Library System in 1998. Since 2002 she has been collaborating with the Digital Library Group of the Institute of Information Science and Technologies (ISTI) in Pisa, a partner of DELOS, the Network of Excellence on Digital Libraries. From April 2011 to September 2013, she was the Padova Library System resource person for European-funded programs. In the same period, she was the project manager of the Linked Heritage team at the University of Padova. From 2011 to 2018, with the role of head of the Digital Library, she has coordinated a working group of librarians involved in digital services and in supporting researchers on Open Access issues with a special focus on research data repositories. Currently, Antonella is the head of the Antonio Vallisneri Bio-medical Library of the University of Padova and a member of the Italian Open Science Support Group (IOSSG).

Birger Hjørland holds an MA in psychology and PhD in library and information science. He is professor emeritus in knowledge organization at the Department of Information Studies, University of Copenhagen (formerly Royal School of Library and Information Science) since 2001

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

and at the University College in Borås 2000-2001. He was research librarian at the Royal Library in Copenhagen 1978-1990, and taught information science at the Department of Mathematical and Applied Linguistics at the University of Copenhagen 1983-1986. He is a member of ISKO Scientific Advisory Council, the editor-in-chief of the *ISKO Encyclopedia of Knowledge Organization* and a member of the editorial boards of *Knowledge Organization*, *Journal of the Association for Information Science and Technology* and *Journal of Documentation*. His *h*-index on 2020-05-30 is 49 in Google Scholar and 29 in Web of Science. *Wikipedia* has an entry about him.

Schöpfel, Joachim, Dominic Farace, Hélène Prost, Antonella Zane and Birger Hjørland. 2021. "Data Documents." *Knowledge Organization* 48(4): 307-328. 63 references. DOI:10.5771/0943-7444-2021-4-307.

Abstract: This article presents and discusses different kinds of data documents, including data sets, data studies, data papers and data journals. It provides descriptive and bibliometric data on different kinds of data documents and discusses the theoretical and philosophical problems by classifying documents according to the DIKW model (data documents, information documents, knowledge documents and wisdom documents). Data documents are, on the one hand, an established category today, even with its own data citation index (DCI). On the other hand, data documents have blurred boundaries in relation to other kinds of documents and seem sometimes to be understood from the problematic philosophical assumption that a datum can be understood as "a single, fixed truth, valid for everyone, everywhere, at all times"[1]

Received: 27 July 2020; Revised: 18 August 2020; Accepted: 21 August 2020

Keywords: data, document formats, DIKW model, bibliography

† Derived from the article titled "Data documents" in the ISKO Encyclopedia of Knowledge Organization, Version 1.0 published 2020-07-16. Article category: Document types, genres and media.

1.0 Introduction: data and data documents

The relation between data documents and knowledge organization primarily concerns metadata: how data documents should be described, indexed, and classified. Secondarily, the concept of data document represents a kind of knowledge organization represented by different document types.

As Furner (2016, 288) expressed, the relation between data and documents seems often to be confused in the literature:

> There is little consensus on the precise nature of the conceptual relationship between "data" and "document." The default position appears to be the view that all documents are in some sense made up of data [...]. The position I wish to develop in this paper, however, is that it is not in fact the case that documents are made up of data. On the contrary, it is the other way round: datasets are made up of documents.

Furner's view is supported by data indexing practices, such as Clarivate Analytics' *Data Citation Index* (DCI), which were introduced in 2012.[2] The "data" indexed in DCI are of four kinds (Clarivate Analytics n.d.):

- *data repositories*, which consist of data studies and data sets and provide access to the data. 411 repositories were indexed 'DT=repository' by DCI on April 16, 2020.
- *data studies* which are descriptions of studies or experiments with the associated data used in the study. Includes serial or longitudinal studies over time. 1,221,993 items data studies were indexed 'DT=data study' by DCI on April 16, 2020.
- *datasets* consisting of a single or coherent set of data, or a data file, provided by the repository as part of a collection, data study, or experiment. 9,349,330 items were indexed 'DT= data set' by DCI on April 16, 2020.
- *software*: A computer program or package in source code or compiled form, which can be installed on another machine and used to support and analyze research. 119,389 items were indexed 'DT=software' by DCI on April 16, 2020.[3]

Records are organized in hierarchy (repository > data study > data set / software): data sets are linked to their parent data study, and these are linked to repositories. Remark that all four kinds are explicitly named "document" types by the DCI. DCI also provides descriptive "data types" taken "as is" directly from the metadata provided by the repository. Therefore, the data types are varied and not normalized or unified, and there is no index, where they can be searched (but searches can be limited by these uncontrolled data types). In a search made 2020-04-27 the five most frequent data types associated to data studies were: FOLKTALE (222,301 records), NUCLEOTIDE SEQUENCING INFORMATION (110,111 records), FILESET (76,974 records), PROCESSED DATA (49,759 records) and RAW DATA (45,653 records).

We have so far presented four kinds of data documents all of which have been connected to DCI and thus to data indexing as opposed to the more traditional indexing of literatures, which has been covered by traditional bibliographic databases for a very long time, and by the *Science Ci-*

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

tation Index since 1964, which today is available with other citation indexes in Clarivate Analytics' platform *Web of Science* (WoS).[4]

WoS has two document types related to data: data paper[5] and database review.[6]

1. "Data Paper: A scholarly publication describing a particular dataset or collection of datasets and usually published in the form of a peer-reviewed article[7] in a scholarly journal. The main purpose of a data paper is to provide facts about the data (metadata, such as data collection, access, features etc.) rather than analysis and research in support of the data, as found in a conventional research article".
   6621 documents in WoS were assigned the document type (DT) category "data paper" on March 21, 2020. (All 6621 were also assigned DT=Article[8]).
2. "Database Review: a critical appraisal of a database, often reflecting a reviewer's personal opinion or recommendation. Refers to a structured collection of records or data that is stored in a computer system."
   (WoS has no document type named "database", but other bibliographic databases, such as MEDLINE have.[9])

Another kind of data document is the data handbook, which exists in many forms and sizes, and are often published under other titles such as "statistical handbook" or "statistical yearbook".[10]

A prominent example of a data handbook is *CRC Handbook of Chemistry and Physics* with the nickname: *The Rubber Bible* (Rumble 2019). This is a one volume book updated annually (now in its 100th edition and today available in print as well as online). By contrast Gmelin (1924-1997)[11] is the most comprehensive handbook of inorganic chemistry ever published (more than 400 volumes), and today only published as *the Gmelin Database*. Such electronic databases are probably the dominant medium of present-day data handbooks.

A further kind of data document is the data journal (presented in Section 4 below). This brings the number of different kinds identified up to the following nine:

- data repository
- data study
- data set
- software
- data paper
- database
- database review
- data handbook
- data journal

What about data themselves? Furner (2016, 288) claimed that data could not exist without documents:

> In this case, "document" is the primary concept: if documents did not exist, data could not; even though documents do exist, data need not.

Can this be true? A datum such as "the melting point of lead is 327,5 °C" or "there are 3 books on my table just now" can exist even if it is not recorded, and is not made up of documents. One definition of data is provided by Kaase (2015, 830) "Data are information on units of analysis or observation"[12]. It is hard to follow Furner all the way; data (information on units of analysis) seems to exist independently of documents. There are, however, two reasons why this may be relatively unimportant for data and information sciences and for data management:

1. Buckland (1991, 351) wrote "Whatever information storage and retrieval systems store and retrieve is necessarily 'information-as-thing.'" Or, translated to data: the only thing that can be managed are data recorded as documents.
2. Spang-Hanssen (2001, 128-9) found that data need to be accompanied by other information, that can only be found in documents:

   > "Information about some physical property of a material is actually incomplete without information about the precision of the data and about the conditions under which these data were obtained. Moreover, various investigations of a property have often led to different results that cannot be compared and evaluated apart from information about their background. An empirical fact always has a history and a perhaps not too certain future. This history and future can be known only through information from particular documents, i.e. by document retrieval. The so-called fact retrieval centers seem to me to be just information centers that keep their information sources, i.e., their documents, exclusively to themselves."

Therefore, overall, the conclusions of Furner seem right: documents, not data, is the primary concept for data- and information sciences. However, in many cases, data exist only in highly dynamic databases, where it may not be possible to reconstruct the exact data at a given former time, and, in some contexts, the very meaning of "data" may be very different from one time to another.[13]

Two kinds of data documents: data journal and data paper are more fully described below in Sections 4 and 5 of the present article (and kinds of emergent document types are mentioned in Section 5.6). Database reviews will not be discussed, while 'Database' is planned to be an independent article.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

2.0 Documents classified according to the data-information-knowledge hierarchy

In information science the so-called knowledge pyramid (or DIKW model) is a suggestion of three or four layers of knowledge:

- Data
- Information
- Knowledge
- Wisdom (only in certain versions of the model)

It is further assumed that these layers form a pyramid (there is much data, less information, lesser knowledge, and little wisdom):

Figure 1. The Knowledge Pyramid (Wikimedia Commons, by Longlivetheux, CC-BY-SA).

It seems obvious to discuss this model because, if one of its categories, data, is represented with its own document types (data documents), this seems to indicate that each category in the knowledge pyramid corresponds to specific document types: if data are published in data documents, where, then, are the other categories published?

The DIKW model has been discussed by a comprehensive literature, including Ackoff (1989), Frické (2007), Rowley (2007) and Zins (2007).[14] The model is popular, but also seriously criticized for being based on problematic assumptions. Frické (2019, 40) thus argues "From a logico-conceptual point of view, DIKW seems not to work". Frické claimed that the problem of the model is, among other things, a problematic philosophy of empiricism and inductivism. Data from measuring instruments, for example, need to be understood on the background of the theories, on which the instruments are constructed, therefore theories and knowledge are essential to inform us of what the surface indications of the instruments are telling us about a reality beyond the instruments themselves. This means that just as the data contribute providing knowledge, knowledge contributes providing data.

Since some documents are called "data documents" the question follows: what are the alternatives? Are there "data documents", "information documents", "knowledge documents" and "wisdom documents" as separate forms of documents?

As a preliminary thought experiment, we could suggest that data sets represent the data level, that single empirical studies represent information documents, that forms of knowledge aggregation and syntheses (like systematic reviews or encyclopaedia articles) represent knowledge documents, and that high-level theoretical and philosophical analyses represent the wisdom level.

However, this hypothetical classification is problematic because the categories (data, information, knowledge, and wisdom) are not conceptually distinct. For example, a table of bakers in a given town, showing which are open, and which are closed at a certain time, (a) provides data on opening hours, but it also (b) informs users, and (c) users so informed have knowledge about their openings (according to the philosophical meaning: *P* knows that *X*). Therefore, it seems not possible to say that data documents just carry data (as opposed to information, knowledge, or wisdom). Another reason why the hypothetical model in Table 1 is problematic is that documents may be used differently. A "wisdom document" may, for example, by a given user be used exclusively for extracting some data and therefore, from this point of view, it represents a data document.

Following Hanson (1958) and Kuhn (1962) the view that observations are theory-laden has flourished. Related to this philosophical movement, a growing number of researchers have also indicated that the idea of data as something "given" is problematic, including Jensen (1950, ix), Manovich (2001, 224), Bowker (2005, 184) and Gitelman and Jackson (2013). If data are not given but carefully constructed in a process that involve theoretical decisions, the idea of a sharp demarcation between data and knowledge becomes problematic. By emphasizing the constructed na-

"data documents"	Presenting data sets
"information documents"	Empirical studies, e.g. based on the IMRAD structure[15]
"knowledge documents"	Systematic reviews,[16] encyclopaedia articles and handbook chapters
"wisdom documents"	Papers presenting high-level theoretical, philosophical analyses and historical analysis

Table 1. Hypothetical classification of documents according to the DIKW hierarchy.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

ture of data, researchers influenced by this philosophy open an important perspective for studying data: how their construction influence how they should be interpreted and the purposes for which they may be used.

The characteristic of data papers as documents which are expected to provide simple descriptions of facts (data) as opposed to research articles which are expected to provide insight, understanding, interpretations, hypotheses etc. 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 seen as part of their core function.

Our conclusion of this section is that classification of documents according to the DIKW model seems not feasible, raising difficult questions about the identity and attributes of data documents.

3.0 Disciplinary issues and citation patterns of data documents

The challenges of data publishing vary greatly among disciplines (see Beckles et al. 2018 about disciplinary data publication guides). The disciplinary distribution of data papers can be illuminated by WoS. The disciplinary scattering in WoS for the top 25 categories are shown in Table 2: on March 21, 2020 6621 documents in WoS were assigned the document type (DT) category "data paper" (all 6621 were also assigned DT=Article).[17] Also the number of citations of the most cited data paper in the category and, for comparation, the number of citations of the most cited journal article in the same subject category are shown.

Table 2 shows a concentration of papers in one WoS category, multidisciplinary sciences (containing more than 85% of the data papers in the database), which makes the subject scattering rather unclear. The table also shows that data papers in all subject categories tend to have few citations, both in absolute numbers, and compared to journal articles. However, since data papers represent a relative new document type, they have not been able to collect as many citations as have articles. Therefore, in Table 3, one year (2017) has been selected to provide a fairer comparison of citations for the two document types:

Table 3 demonstrates that given equal number of years to be cited, data papers still are exceptionally low cited compared to articles. Perhaps this low rate of citations is due to a tendency of scientific papers to cite data sets directly?

In Table 4 is shown the number of data sets indexed in different subject categories (just the top 25 WoS categories). In addition, the number of citations they have received has been added for comparison.

The low number of citations of data sets is remarkable, as is the information that the social sciences dominate in this regard. We shall return to this finding below. First, the corresponding numbers are shown (top categories only) for data repositories and for data studies in Table 5 and Table 6:

The citation distributions of different kinds of data documents are somewhat surprising given which subject fields are generally considered most data intensive. In physics, the particle-collision events in CERN Large Hadron Collider near Geneva in Switzerland generate around 15 petabytes of data annually (Marx 2013, 255). In astronomy, the construction of the Large Synoptic Survey Telescope (LSST) in Chile is designed to produce about 15 terabytes of raw data per night and 30 petabytes over its 10-year survey life (Murray 2017). Compared to these amounts of data, Marx (2013, 257) found biology to have arrived later in the big science field and having relatively smaller amounts of data, and we may add that the social sciences provide even smaller amounts (corresponding to the much smaller number of data journals in these fields, cf. Section 4). All this is not visible in the citation data given in Tables 2-6. We cannot explain why this is the case, but a possible reason could be that these huge amounts of data are primarily used in a more direct way compared to their use in the scientific literature (works such as Edwards 2010 about climate data provides hints of such use of big data).

4.0 Data journals

A first survey on data journals was conducted by Candela et al. (2015), with a sample of 116 data journals published by fifteen different publishers. They distinguished seven "pure" data journals publishing only data papers and 109 "mixed" data journals publishing any kind 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 nine data journals in social sciences and humanities (8%). A more recent study from Schöpfel et al. (2019) confirms the preponderance of medical and life sciences while only four data journals (from 28) publish data from the 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).

All big five academic publishers (Elsevier, Springer-Nature, Wiley-Blackwell, Taylor & Francis and SAGE) have their own data journals. Other data journals are published or hosted by newcomers, especially by open access publishers such as Ubiquity Press, BioMed Central, 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,

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

Top WoS categories	Records	% of 6621	Citations of most cited paper	Citations of most cited journal article (all dates)
MULTIDISCIPLINARY SCIENCES	5689	85.924	1,134	248,721
METEOROLOGY ATMOSPHERIC SCIENCES	305	4.607	554	19,593
GEOSCIENCES MULTIDISCIPLINARY	300	4.531	554	4,915
GENETICS HEREDITY	152	2.296	37	45,689
BIOLOGY	130	1.963	85	33,357
COMPUTER SCIENCE INFORMATION SYSTEMS	121	1.828	94	17,229
BIODIVERSITY CONSERVATION	113	1.707	44	4,846
ECOLOGY	104	1.571	56	30,808
ONCOLOGY	43	0.649	40	16,729
ZOOLOGY	29	0.438	5	8,620
CARDIAC CARDIOVASCULAR SYSTEMS	19	0.287	34	10,575
BIOTECHNOLOGY APPLIED MICROBIOLOGY	17	0.257	32	24,028
FORESTRY	15	0.227	11	3,045
PHARMACOLOGY PHARMACY	12	0.181	5	22,707
BIOCHEMICAL RESEARCH METHODS	11	0.166	32	210,830
BIOCHEMISTRY MOLECULAR BIOLOGY	11	0.166	7	341,151 (highest cited article in database!)
ENVIRONMENTAL SCIENCES	11	0.166	6	12,690
PLANT SCIENCES	11	0.166	36	44,843
ENGINEERING CIVIL	10	0.151	14	3,506
ENGINEERING GEOLOGICAL	10	0.151	14	8,342
GEOGRAPHY PHYSICAL	10	0.151	56	2,327
PSYCHOLOGY MULTIDISCIPLINARY	10	0.151	6	19,639
ROBOTICS	10	0.151	114	3,103
ARCHAEOLOGY	9	0.136	2	1,238
MARINE FRESHWATER BIOLOGY	9	0.136	3	5,324
....[#24-#58]				
[#59] SOCIAL SCIENCES INTERDISCIPLINARY	1	0.015	3	36,366

Table 2. Disciplinary scattering of data papers: the top 25 subject categories in *Web of Science* (WoS).

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

Top WoS categories	Citations of most cited paper (2017)	Citations of most cited journal article (2017)
MULTIDISCIPLINARY SCIENCES	251	2,760
METEOROLOGY ATMOSPHERIC SCIENCES	80	1,007
GEOSCIENCES MULTIDISCIPLINARY	80	258
GENETICS HEREDITY	14	987
BIOLOGY	85	1,808
COMPUTER SCIENCE INFORMATION SYSTEMS	96	514
BIODIVERSITY CONSERVATION	46	229
ECOLOGY	44	416
ONCOLOGY	20	11,819
ZOOLOGY	5	156
CARDIAC CARDIOVASCULAR SYSTEMS	35	3,517
BIOTECHNOLOGY APPLIED MICROBIOLOGY	32	943
FORESTRY	7	73
PHARMACOLOGY PHARMACY	5	558
BIOCHEMICAL RESEARCH METHODS	32	952
BIOCHEMISTRY MOLECULAR BIOLOGY	7	1,808
ENVIRONMENTAL SCIENCES	7	817
PLANT SCIENCES	2	362
ENGINEERING CIVIL	14	377
ENGINEERING GEOLOGICAL	14	107
GEOGRAPHY PHYSICAL	0	145
PSYCHOLOGY MULTIDISCIPLINARY	6	564
ROBOTICS	116	497
ARCHAEOLOGY	0	46
MARINE FRESHWATER BIOLOGY	0	235
....[#24-#58]		
[#59] SOCIAL SCIENCES INTERDISCIPLINARY	0	427

Table 3. Citation data for most cited data papers and journal articles published 2017.

ACS, Wageningen, etc.). Most of the data journals are "young" products and have been launched during the last ten years, from 2008 on.

At the end of 2019, the overall number of data papers published by these data journals is estimated[18] to be approximately 11,500, with large differences, ranging from some papers up to more than 3,500, with a rather low median number (ninety-seven).

Following Schöpfel et al. (2019), some of the data journals are considered as good or high-quality journals: from 28 "pure" data journals, eleven are indexed by Clarivate Analytics, eight by Elsevier's Scopus database while sixteen journals are referenced in the international Directory of Open Access Journals (DOAJ).

The major business model is OA Gold, mostly with article processing charges (APC)[19] but some without. Some journals are hybrid, but only one journal is available through the tra-

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

Top WoS category	Datasets indexed	Citations of most cited dataset
GENETICS HEREDITY	3,603,573	121
BIOCHEMISTRY MOLECULAR BIOLOGY	2,478,940	3
MULTIDISCIPLINARY SCIENCES	2,010,474	26
CRYSTALLOGRAPHY	1,209,714	5
GEOSCIENCES MULTIDISCIPLINARY	674,506	88
GEOGRAPHY	602,826	3
ECOLOGY	400,953	21
MICROBIOLOGY	266,264	13
OCEANOGRAPHY	211,382	64
CHEMISTRY MULTIDISCIPLINARY	193,197	0
SPECTROSCOPY	142,874	3
GEOCHEMISTRY GEOPHYSICS	120,910	34
METEOROLOGY ATMOSPHERIC SCIENCES	105,031	88
BIODIVERSITY CONSERVATION	91,681	8
ENGINEERING MULTIDISCIPLINARY	75,454	14
MATERIALS SCIENCE MULTIDISCIPLINARY	72,361	4
PLANT SCIENCES	65,910	1
SOCIAL SCIENCES INTERDISCIPLINARY	62,120	328

Table 4. Number of data sets indexed by discipline.

ditional subscription model. Many 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).

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 adopt an "open peer review" (suggestion of reviewers, community or interactive public peer review). (See endnote 20 for an example of peer-review instructions.) There have also been critical voices about this activity (e.g., Huang, Hawkins and Qiao 2013).

At the end of 2019, the number of data journals and papers appears to increase slowly, on a low level. Garcia-Garcia, Borrul and Peset (2015) identified twenty pure data journals; four years later, the Schöpfel et al. (2019) sample consists of twenty-eight data journals and not all are still active or even pure (see below). Twenty-eight journals represent less than 0.01% of the academic and scholarly serials (source: Scopus). Arts, social sciences and humanities are nearly non-existent (two journals in 2015, four 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 lies not in their volume but in 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.[21]
- 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

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

WoS categories	Data repositories indexed	Citations of most cited data repositories
GENETICS HEREDITY	65	701
MULTIDISCIPLINARY SCIENCES	63	320
SOCIAL SCIENCES INTERDISCIPLINARY	48	3,198
HEALTH CARE SCIENCES SERVICES	37	765
BIOCHEMISTRY MOLECULAR BIOLOGY	34	137
GEOSCIENCES MULTIDISCIPLINARY	32	26
SOCIOLOGY	29	765
DEMOGRAPHY	25	494
HUMANITIES MULTIDISCIPLINARY	24	2
ECONOMICS	23	494
METEROLOGY ATMOSPHERIC SCIENCES	21	59
ASTRONOMY ASTROPHYSICS	14	567
OCEANOGRAPHY	14	3
ENVIRONMENTAL SCIENCES	13	6
ENVIRONMENTAL STUDIES	13	3
ECOLOGY	12	12
COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS	11	213
HEALTH POLICY SERVICES	10	480
CELL BIOLOGY	8	12
CRYSTALLOGRAPHY	8	44

Table 5. Number of data repositories indexed by discipline and number of citations received by the most cited repository in each category (on 2020-04-26).

WoS categories	Data studies indexed	Citations of most cited data studies
GENETICS HEREDITY	545,315	643
MULTIDISCIPLINARY SCIENCES	506,820	13
HUMANITIES MULTIDISCIPLINARY	287,992	4
SOCIAL SCIENCES INTERDISCIPLINARY	281,816	1,595
BIOCHEMISTRY MOLECULAR BIOLOGY	77,320	6
CHEMISTRY ORGANIC	34,805	1
SPECTROSCOPY	34,805	1
BIODIVERSITY CONSERVATION	29,273	86
ECOLOGY	26,056	10
EVOLUTIONARY BIOLOGY	22,914	10
GEOSCIENCES MULTIDISCIPLINARY	16,078	4
ENVIRONMENTAL SCIENCES	12,733	4
TOXICOLOGY	8,988	4
MICROSCOPY	5,970	1
SOCIOLOGY	5,671	213
MARINE FRESHWATER BIOLOGY	5,228	3
CRYSTALLOGRAPHY	5,006	1
PHYSICS ATOMIC MOLECULAR CHEMICAL	4,125	4
ECONOMICS	3,760	383

Table 6. Number of data studied indexed by WoS categories and number of citations received by the most cited data study in each category.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

also attribute DOIs to their own data papers, creating a kind of cross-linked DOI system between data papers and datasets (see below).
– 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 as 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, already, for one out of five journals, by a 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).

5.0 Data papers

5.1 Definition and developments

Data papers are authored and citable articles in academic or scholarly journals. They are mostly, but not necessarily, peer reviewed. Their 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 of facilitating the findability, availability and reuse of research data; they are part of the research data management and crosslinked to data repositories.

In the context of open science, an increasing volume of research data is made available on the Internet, contributing to the big data of science. New tools, methods and infrastructures have been developed for the dissemination, processing, analysis, and preservation of research data. Data papers, along with the other data documents mentioned in Section 1 are part of them. Table 7 shows the development in number of data papers as covered by WoS Core collection and all databases by year.

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, 112). Another definition describes data papers 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".[22]

Data papers 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* (Springer), *GigaScience* (Oxford University Press) and *PLoS One*. Most data papers are published on journal platforms; yet some are (also or exclusively) published on data repository platforms.[23]

Unlike standard 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 four pages.

In the "classical" research paradigm, the focus is on articles presenting results while research data are useful for the

Year	Core	All
2020	80	80
2019	1,950	1,953
2018	1,938	1,943
2017	1,099	1,104
2016	1,096	1,103
2015	389	390
2014	27	98
2013	29	30
2012	13	163
2011		7
2006		1
All	6,621	6,872

Table 7. DT=Data Paper in WoS in April 2020.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

validation of published research findings. Data papers invert the roles, insofar as the paper's main function is to inform about and link to research data on data repositories, contributing to their findability and reusability.

Also, traditional knowledge organization makes a relatively 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. The 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.

5.2 Functions and objectives

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" (Newman and Corke 2009, 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, Psaki and Bingenheimer 2017, 291).
- The French journal *Review 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), Pärtel (2006) considers the data paper as a kind of "abstract" that aims 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, appear to belong to a rather heterogeneous and dissimilar new kind of documents. Nevertheless, there are more common features with regular articles, such as the fundamental structure.

As we saw in Section 2, data papers provide other functions than just describing data sets. For this reason, data papers do not just improve the referencing of datasets on repositories but fulfil other roles. Their profile can perhaps 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 and Griesemer 1989) on the frontline between academic publishing and data driven research. Our analysis confirms the statement that data papers are like traditional research papers in some respects but quite 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 an relatively independent research data practice.

5.3 Structure and contents

As already said, it is generally assumed that data papers are short texts, up to four pages. In fact, this is only partly true. In the survey from Schöpfel et al. (2019), only five journals require short papers, limited to four to six pages or maximal 3,000 words. Most journals do not limit the length of submitted papers or make the usual recommendations (six to ten pages, or maximal 6,000 words). One journal only accepts short abstracts while others publish papers well beyond the length of regular papers, up to twenty or thirty or even 100 pages, including detailed data descriptions, illustrations (figures) or data tables.

No results, no discussion, no conclusion: usually the data journal guidelines for authors contain these or similar recommendations; others, however, 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. Yet, 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*[24] or the *Data Documentation Initiative* (DDI),[25] and a section of

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

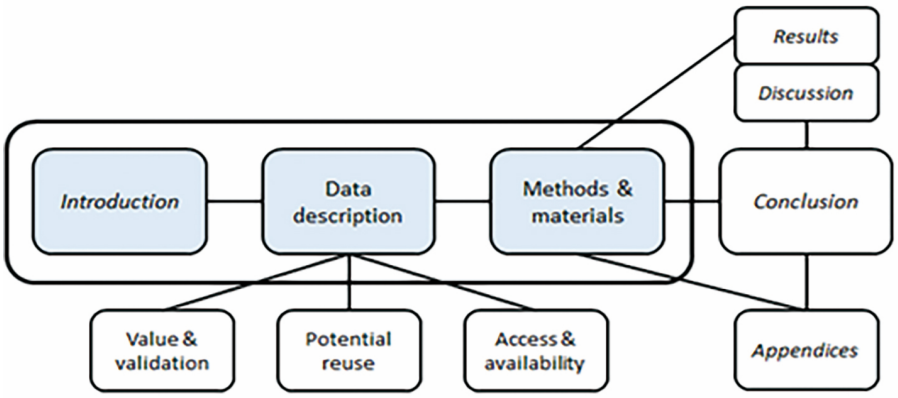


Figure 2. Sections of a data paper.

materials and methods, instrumentation, on the production of the data and procedures, sometimes extended to experimental designs and calculation (Figure 2).

Figure 2 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 and 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 and 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, information 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.

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.

5.4 Metadata

Metadata are constitutive for data papers. 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: some data journals require 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. 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[26] 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[27]).

Nearly all data journals publish the data papers with a DOI, and some also include the author identifier ORCID. Also, most of them recommend or require a standard identifier (DOI) or at least a stable address for the described datasets.

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.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

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

The link between data papers and the metadata of research data is essential because both have similar functions, i.e., to 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. Also, data papers are particularly interesting for the requirements of the so-called "FAIR Guiding Principles for scientific data management and stewardship" (Wilkinson et al. 2016) because they contribute to these principles in different ways, in order to improve the findability, accessibility, interoperability, and reuse of research data, e.g.[28]:

- 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 as they help people and machines 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 and specialization.

- 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.[29] 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.
- Also, 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.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

5.5 Production and processing

Li, Greenberg and Dunic (2020) conducted a content analysis with eighty-two data papers from sixteen journals to investigate what information they describe regarding methods used 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.

Chavan and Penev (2011, 7) describe a tool that "facilitates conversion of a metadata document into a traditional manuscript for submission to a journal" 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".

This kind of generated data paper 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).

The integrated workflow of data repositories and journal platforms described by Chavan and Penev (2011) requires shared standards and formats. Senderov, Georgiev and Penev (2016) provide an example of this data paper generation in the field of biodiversity. Their workflow relies on three key standards (RESTful APIs for the web, Darwin Core and EML) and imports metadata into the ARPHA writing tool (AWT). In other words, and more generally spoken, as expressed by de Waard (2010, 9): "the boundary between a workflow tool, a data store, and a publishing platform blurs".

Nevertheless, Schöpfel et al. (2019) could not find any invitation or guidelines 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.

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. Yet, 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. Also, 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. Instead of machine generated data papers we should speak of "machine- (or repository-) assisted writing of data papers".

Are data papers produced only for machines? According to Li, Greenberg and Dunic (2020, 18), the answer is no, as they 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". Both are complementary, rather than competitive. While Candela et al. (2015) highlight the distinction between metadata of datasets, metadata of data papers, and data papers themselves, 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 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.

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

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

popularity of data papers in increasingly more scientific fields" (Li, Greenberg and Dunic 2020, 172; 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".[30] 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.

5.6 Blurred boundaries

Candela et al. (2015) identified ten different terms assigned for data papers, including data descriptor, data note, dataset paper and data in brief. 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.

Similar to most cited authors, Smith (2011, 15) 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".[31] And this is exactly 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.

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 Schöpfel et al. (2019) provide evidence that despite a general definition of data papers and journals, there is much 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. 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 thirty-seven journals, including one data journal and sixteen other journals accepting data papers. The French Agricultural Research Centre for International Development (CIRAD) produced a list with fifty-four academic journals accepting data papers relevant for agricultural science, including the mega-journal *PLoS One*. It is quite impossible to make an estimation of the real number of such mixed data journals and their data papers.
3. Data papers are more than simple descriptions of data sets. Even a superficial analysis of data papers reveals that one part of articles labelled 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"[32], in other words, data papers. But there are other examples:
   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*).[33]

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

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

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 former mentioned IMRAD format as a widely used format of scientific article publishing. 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 kinds of papers; a strategy somewhere between predatory publishing and big data principles based on volume and variety rather than on quality and trustworthiness.

6.0 Conclusion: the need for studies of information ecologies

The study of "the information ecology" is an interdisciplinary endeavour, which traditionally has been studied by information science (with bibliometrics), the sociology of science and the sociology of knowledge, among other fields. Recently, a new member, "data science" has joined the field with its own journals, such as *Data Science Journal*,[34] cf. Mayernik (in press). The UNISIST model (UNISIST 1971) is useful for explaining the concept "information ecology". It can be understood as a sociological model of the scientific and scholarly information ecology, which is modelling the system of actors, institutions, systems and processes in two dimensions: (1) from knowledge producers to users with primary, secondary and tertiary information services (2) via different communication channels (informal, formal published, formal unpublished and tabular channels).

In the original UNISIST model, the tabular channel represented an independent "data channel" in the information ecology in that point in history. Fjordback Søndergaard, Andersen and Hjørland (2003) provided an update and revision of the original model emphasizing (a) the development of the Internet and its information systems (b) the domain specific character of information ecologies and (c) the expansion of the model from natural science and technology to include all scholarly fields. In the revised model the tabular channel was omitted of two reasons (1) tabular data were understood as formally or informally published or unpublished data, which made the separate channel redundant (2) other forms of content (e.g., pictures, sounds, physical objects), may have an equal right to be included (especially considering the humanities). Today, however, different kinds of data documents have become in focus in the data- and information sciences, and there is a need to reconsider their communication channels in an updated model.

The finding in the present paper of the low citations of data documents seems to indicate that data may have a life relatively independent of the scientific literature, but this is an issue that needs to be further explored. This makes it important to reconsider the different document genres in scientific and scholarly communication from the perspective of their different functions. A specific project seems worth mentioning. We saw in Section 1 that DCI provides descriptive "data types", but they were taken directly from repositories without any form of classification and normalization. It seems worth to investigate whether the development of some kinds of controlled vocabularies of kinds of data will improve scholarly communication (or improved if they have already been developed by some repositories).

The most important consideration is, however, that data are not independent on theory and vice versa. We cited Edwards (2010) already in the abstracts (cf. endnote 1) and are here following up. Edwards (2010, *xiii*) wrote:

> ... without models, there are no data. I'm not talking about the difference between 'raw' and 'cooked' data. I mean this literally. Today, no collection of signals or observations—even from satellites, which can 'see' the whole planet—becomes global in time and space without first passing through a series of data models.

And further (280):

> ... the theory-ladenness of data reaches a level never imagined by that concept's originators. At the same time, far from expressing pure theory, analysis models are *data-laden*.[35]

Therefore, the separation of scientific theory in some papers and scientific data in other papers seems problematic,[36] as we already saw in Section 2. This issue of the role of theory in data-centric science has been carefully considered by Leonelli (2016). One of her important conclusions is (138):

> Theory, in all its forms, can always be made to function as a motor or a hindrance to scientific advancements, depending on the degree of critical awareness

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

> with which it is employed in directing research. The fruitfulness of the conceptual scaffolding currently developed to make data travel thus rests on the ability of all stakeholders in biological research to exploit data-intensive methods without forgetting which commitments and constraints they impose on scientific reasoning and practice.

In other words, all kinds of data production, mediation and use always contains theoretical assumptions. The important thing is to make such assumptions explicit and to make them function as a motor for scientific advancement. This requires critical studies of the whole information ecology.

Acknowledgments

The authors thank Dr. Daniel Martínez Ávila for serving as the editor for this article and two anonymous reviewers for valuable suggestions and comments.

Notes

1. Quote from Edwards (2010, 283). We return to this quote in the conclusion (and in endnote 35).
2. See Rivalle and Green (2016-2018) and Clarivate Analytics' *LibGuide* about the DCI, accessed March 20, 2020 from http://clarivate.libguides.com/webofscience platform/dci .
3. Software (or "data as software") is also briefly discussed in Kratz and Strasser (2014).
4. WoS is both the name of a platform (which was originally called *Web of Knowledge*), which contains many databases, as well as on a single database, called WoS Core Collection, which basically is an integration of former databases such as *Science Citation Index, Social Sciences Citation Index* and *Arts & Humanities Citation Index*. DCI is part of the platform, but not of the "core collection".
5. ISO 5127: 2017 defined *paper* as synonym for *article*:
   "3.4.1.27.09 paper (2): <document> *text* (3.2.1.05) *document* (3.1.1.38) delivered before an audience or contributed to an edited *volume* (3.4.1.27.12) or a *scientific journal* (3.4.1.28.19)".
   "3.5.8.12 paper (3): <> scientific *article* (3.5.8.06) in a *scientific journal* (3.4.1.28.19)".
   "3.6.3.06 papers, pl; personal papers, pl; private archives, pl: natural *accumulation* (3.2.1.37) of personal or family *documents* (3.1.1.38)".
   Reitz (2004), on the other hand, defined *paper* in a narrow sense: "Paper. [...] Also refers to a brief composition, especially one prepared for presentation by the author at a conference or other professional meeting. Conference papers may be published in proceedings or transactions. They are indexed in PapersFirst, an online database available in OCLC FirstSearch. Compare with article. See also: invited paper."
6. https://images.webofknowledge.com/images/help/WOS/hs_document_type.html.
7. Reitz (2004) defined *article*: "A self-contained nonfiction prose composition on a fairly narrow topic or subject, written by one or more authors and published under a separate title in a collection or periodical containing other works of the same form. The length of a periodical article is often a clue to the type of publication--magazine articles are generally less than five pages long; scholarly journal articles, longer than five pages. Also, journal articles often include a brief abstract of the content. Periodical articles are indexed, usually by author and subject, in periodical indexes and abstracting services, known as bibliographic databases, when available electronically. Compare with column, editorial, and essay. See also: cover story and feature."
8. WoS writes: "An Article is generally published in a journal. A Proceedings Paper is generally published in a book of conference proceedings. Records covered in the two Conference Proceedings indexes (CPCI-S and CPCI-SSH) are identified as Proceedings Paper. The same records covered in the three indexes (SCI-E, SSCI, and A&HCI) are identified as Article when published in a journal." Book Chapter is a separate category in WoS.
9. MEDLINE describes two document types related to data, one of which (data set) is already described from our description of DCI (https://www.nlm.nih.gov/mesh/pubtypes.html):

"Database: Work consisting of a structured file of information or a set of logically related data stored and retrieved using computer-based means."
"Dataset: Works consisting of organized collections of data, which have been stored permanently in a formalized manner suitable for communication, interpretation, or processing."

10. An example of a data handbook is, for example, United Nations' *Statistical Yearbook*: https://unstats.un.org/unsd/publications/statistical-yearbook/.
11. *Gmelin Handbook of Inorganic Chemistry: A User's Guide*, http://www.umsl.edu/~chickosj/202/Gmelin.pdf. Gmelin's handbook differs from *CRC Handbook of Chemistry and Physics* by providing much more information about how the data have been obtained. It may therefore be perceived more like a handbook than specifically a data handbook. This provides an association to Spang-Hanssen's quote. Are data handbooks (and data documents in general) sources "that keep their information sources [...] exclusively to themselves"? Also: Is the *Rubber Bible* a data handbook because (in contrast

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

to Gmelin's handbook) it omits the descriptions on how the data were obtained?

12. A former version of this definition (Kaase 2001, 3251) was: "Data is information on properties of units of analysis", which was discussed by Nielsen and Hjørland (2014).
13. Edwards (2010, 283): "Today's meteorologists understand the meaning of 'data' very differently from meteorologists of earlier generations. The panopticism that ruled meteorology from Ruskin to Teisserenc de Bort and beyond has been slowly but surely replaced by an acceptance of limits to the power of observation. In place of Ruskin's 'perfect systems of methodical and simultaneous observations ... omnipresent over the globe,' 4-D data assimilation augments and adjudicates spotty, inconsistent, heterogeneous instrument readings through computer simulation. Modern analysis models blend data and theory to render a smooth, consistent, comprehensive, and homogeneous grid of numbers — what I have called in this chapter a *data image*, rather than a data set. Meanwhile, global data images from GCM simulations proliferate."
14. Zins (2007) does not discuss the knowledge pyramid, but the concepts of data, information and knowledge among others.
15. 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 Sollaci and Pereira (2004) and *Wikipedia* (2020b).
16. A systematic review is a systematic search for literature relevant to a certain issue also performing evaluations and organization of research in order to make highly qualified decisions based on published "evidence". There is a huge literature on systematic reviews from practical guides to theoretical and philosophical issues, including Booth (2016) and Hammersley (2006).
17. 35 lower ranking WoS categories are not included in Table 2 but can be seen in WoS.
18. "The data are based on the authors counting of the data papers on each journal's website". Compared to the data in Table 7, it suggests a rather selective coverage in the WoS databases (and then indicates that WoS may cover data journals rather selectively).
19. APC: the fee authors or their institutions have to pay (after the acceptance of their papers) to some publishers to be published immediately in open access. The amount of APC varies 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/).
20. An example of peer-review instructions for data papers was provided by Fabrissin (2019): "Data papers aim at describing a dataset made available to a larger community. Data papers are a scientific valuable production and should provide all required information for a large use of the data. The data paper should be completed by a metadata file that describes the dataset, and by the dataset itself, made available in an open repository. Reviewers will carefully consider: *(a)* the quality of the manuscript, *(b)* the quality, completeness and reusability of the dataset, and *(c)* the relevance of the dataset and its potential contribution to the progress of science. They will not review the whole data set, whose quality is under the responsibility of the authors (see our related blog post "*Guidelines for authors: how to share your datasets?*").

    Quality of the manuscript

    Usual criteria for assessing manuscript quality including style, consistency, clarity. The review will address the following questions:

    – Is a DOI provided for the database and is the dataset accessible via the given identifier?
    – Is the metadata template filled out as recommended?
    – Do title and key message accurately reflect the content of the data paper?
    – Is the Data Paper internally consistent, suitably organized and written in proper English?
    – Are relevant non-textual media (e.g. tables, figures) appropriate?
    – Have abbreviations and symbols been properly defined?
    – Is the context of prior research properly described, citing relevant articles and datasets?

    Quality of the dataset

    Although publication of a data paper does not guarantee the overall quality of the dataset, there is a need to check whether suitable and reproducible methods have been used to obtain the data, and whether the data are displayed in a sensible way. The dataset must be a long-term resource, stable, complete, permanent and of good quality. The review will address following questions:

    – Are the data logically and consistently organised? Are they easily readable and usable?
    – Is anything missing in the manuscript or the data resource itself that would prevent replication of the measurements, reproduction of the figures or other representations?
    – Are the methods used to generate the data (including calibration, code and suitable controls) described in sufficient detail and suitable to maintain of integrity of the dataset?
    – Have possible sources of error (including methods, calculation and interpretation) been appropriately addressed in the protocols and/ or the paper?

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

- Are the data consistent internally and described using applicable standards (e.g. in terms of file formats, file names, file size, units and metadata)?
- Does the manuscript provide an accurate description of the data? And how to access them (e.g. link and/or data access policy)?
- Are the methods used to process and analyse the raw data appropriate? Are they sufficiently well documented that they could be repeated by third parties? Accepted formats are: 1) datasets, deposited additionally to scientific datasets in the repository; 2) links to online published papers; 3) a section in the body of the manuscript dedicated to material and method.
- Are the data files complete and match the description in the Metadata?

Utility and contribution of data set

- Does the data resource cover a scientifically important region(s), time period(s) and/or group(s) of taxa to be worthy of a separate publication?
- Is the dataset sufficiently original to merit publication as a data paper?
- Is there any potential of the data being useful in the future?
- Are the use cases described in the data paper consistent with the data presented? Would other possible use cases merit comments in the paper?
- Are all conclusions made in the data paper substantiated by the underlying data?
- Are the depth, coverage, size, and/or completeness of the data sufficient for the types of application or research questions outlined by the authors?"

21. The plan S gives preference to immediate open access in 100% OA journals, see https://www.coalition-s.org/
22. Source: Global Biodiversity Information Facility, https://www.gbif.org/data-papers
23. *Wikipedia* (2002a) wrote: "There are several distinct ways to make research data available, including:
    - publishing data as supplemental material associated with a research article, typically with the data files hosted by the publisher of the article
    - hosting data on a publicly available website, with files available for download
    - hosting data in a repository that has been developed to support data publication, e.g., figshare, Dryad, Dataverse, Zenodo. A large number of general and specialty (such as by research topic) data repositories exist (Assante et al., 2016). For example, the UK Data Service enables users to deposit data collections and re-share these for research purposes.
    - publishing a data paper about the dataset, which may be published as a preprint, in a journal, or in a data journal that is dedicated to supporting data papers. The data may be hosted by the journal or hosted separately in a data repository."
24. DataCite, https://schema.datacite.org/
25. Data Documentation Initiative, https://www.ddialliance.org/
26. ISA tools, https://isa-tools.org/
27. Journal Publishing Tag Set, https://jats.nlm.nih.gov/publishing/.
28. The description and numbering of the principles follow the GO FAIR list at https://www.go-fair.org/fair-principles/
29. ISGN, http://www.igsn.org/
30. Chemical Data Collections, see https://www.elsevier.com/journals/chemical-data-collections/2405-8300/guide-for-authors
31. Is it possible, even in theory, to distinguish papers that are about research from papers that are about data? Traditional research papers, for example, contains a methodology section about how the data was established, and sound research papers also discuss weakness in the established data. Data, on the other hand – because they are theory-laden – cannot be described in absence from the research in which they were established. The difference between research papers' and data papers' description of data can only be a question of degree: data papers may describe data more detailed and according to other norms compared to research papers.
32. F1000Research, see https://f1000research.com/for-authors/article-guidelines/data-notes
33. Data Science Journal, see https://datascience.codata.org/about/
34. Perhaps the introduction of data science can be said to introduce yet another kind of data documents: Documents of data science.
35. Edwards (2010, 282) quoted Norton and Suppe (2001, 70): "to be properly interpreted and deployed, *data must be modeled*". He further writes (282): "If Norton and Suppe are right, seeking purity in either models (as theories) or data (as unmediated points of contact with the world) is not only misguided, but impossible. (This is opposed to the understanding of data mentioned in the abstract: "a single, fixed truth, valid for everyone, everywhere, at all times")".
36. Li, Greenberg and Dunic (2020) also challenged the degrees to which data papers are a distinct genre compared to research articles.

References

Ackoff, Russell L. 1989. "From Data to Wisdom". *Journal of Applied Systems Analysis* 16(1): 3-9.

Assante, Massimiliano, Leonardo Candela, Donatella Castelli and Alice Tani. 2016. "Are Scientific Data Reposi-

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

tories Coping with Research Data Publishing?" *Data Science Journal* 15. doi:10.5334/dsj-2016-006

Beckles, Zosia, Stephen Gray, Debra Him, Kirsty Merrett, Kellie Snow and Damian Steer. 2018. "Disciplinary Data Publication Guides." *International Journal of Digital Curation* 13(1): 150–60. doi:10.2218/ijdc.v13i1.603

Belter, Christofer W. 2014. "Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets." *PLoS One* 9(3): e92590. doi:10.1371/journal.pone.0092590

Booth, Andrew. 2016. "Searching for Qualitative Research for Inclusion in Systematic Reviews: A Structured Methodological Review." *Systematic Reviews* 5: 74. http://www.systematicreviewsjournal.com/content/pdf/s13643-016-0249-x.pdf

Bordelon, Dominic, Uta Grothkopf, Sylvia Meakins and Michael Sterzik. 2016. "Trends and Developments in VLT Data Papers as Seen Through Telbib." In *Observatory Operations: Strategies, Processes, and Systems VI: 27 June–1 July 2016 Edinburgh, United Kingdom*, edited by Alison B. Peck, Robert L. Seaman and Chris R. Benn. Proceedings of SPIE 9910. Bellingham, WA: SPIE. doi:10.1117/12.2231697.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Buckland, Michael K. 1991. "Information as Thing." *Journal of the American Society for Information Science* 42(5): 351-60.

Callaghan, Sarah, Steve Donegan, Sam Pepler et al. 2012. "Making Data a First-Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres." *International Journal of Digital Curation* 7(1): 107-13. doi:10.2218/ijdc.v7i1.218

Candela, Leonardo, Donatella Castelli, Paolo Manghi and Alice Tani. 2015. "Data Journals: A Survey." *Journal of the American Society for Information Science and Technology* 66(9): 1747-62. doi:10.1002/asi.23358

Chavan, Wishwas and Lyubomir Penev. 2011. "The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science." *BMC Bioinformatics* 12, S2. doi:10.1186/1471-2105-12-S15-S2

Clarivate Analytics. n.d. *LibGuide about the Data Citation Index*. Accessed March 20, 2020, from http://clarivate.libguides.com/webofscienceplatform/dci

Costello, Mark J., William K. Michener, Mark Gahegan, Zhi-Quiang Zhang and Philipp E. Bourne. 2013. "Biodiversity Data Should Be Published, Cited, and Peer Reviewed." *Trends in Ecology & Evolution* 28(8): 454-61. doi:10.1016/j.tree.2013.05.002

De Waard, Anita. 2010. "The Future of the Journal? Integrating Research Data with Scientific Discourse." *Logos* 21(1/2): 7-11. doi:10.1163/095796510X546878

Edwards, Paul. N. 2010. *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Fabrissin, Isabelle. 2019. "Guidelines for Reviewers: How to Review a Data Paper?" *Annals of Forest Science*, Editorial Board Information. Blog post. https://ist.blogs.inrae.fr/afs/2019/08/23/guidelines-for-reviewers-how-to-review-a-data-paper/

Fjordback Søndergaard, Trine, Jack Andersen and Birger Hjørland. 2003. "Documents and the Communication of Scientific and Scholarly Information. Revising and Updating the UNISIST Model." *Journal of Documentation* 59(3): 278-320.

Frické, Martin. 2009. "The Knowledge Pyramid: A Critique of the DIKW Hierarchy". *Journal of Information Science* 35(2): 131-42. doi:10.1177/0165551508094050

Frické, Martin. 2019. "The Knowledge Pyramid: The DIKW Hierarchy". *Knowledge Organization* 46(1): 33-46. Also available in *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. http://www.isko.org/cyclo/dikw

Friedman, Rachel, Stéphanie Psaki and Jeffrey B. Bingenheimer. 2017. "Announcing a New Journal Section: Data Papers." *Studies in Family Planning* 48(3): 291-2. doi:10.1111/sifp.12032

Furner, Jonathan. 2016. "'Data': The Data". In *Information Cultures in The Digital Age: A Festschrift in Honor of Raphael Capurro*, edited by Matthew Kelly and Jared Bielby. Wiesbaden: Springer, 287-306. http://www.jonathanfurner.info/wp-content/uploads/2016/12/Furner-Final-Proof-18.4.16.pdf

Garcia-Garcia, Alicia, Alexandre Lopez Borrul and Fernanda Peset. 2015. "Data journals: eclosión de nuevas revistas especializadas en datos." ["Data Journals: Emergence of New Journals Specializing in Data."] *El Profesional de la Información* 24(6): 845-54. doi:10.3145/epi.2015.nov.17

Gitelman, Lisa and Virginia Jackson. 2013. "Introduction". In *Raw Data' is an Oxymoron*, ed. Lisa Gitelman. Cambridge, MA: MIT Press, 1-14.

Gmelin, Leopold. 1924-1997. *Gmelin Handbook of Inorganic and Organometallic Chemistry* (= *Gmelin Handbuch der anorganischen Chemie*). 8th ed. Compiled by the Gmelin Institute, part of the Max Planck Institute. Berlin: Springer.

Hammersley, Martyn. 2006. "Systematic or Unsystematic, Is That the Question? Reflections on the Science, Art, and Politics of Reviewing Research Evidence." In *Public Health Evidence: Tackling Health Inequalities*, edited by Amanda Killoran, Catherine Swann and Michael Kelly. Oxford, UK: Oxford University Press.

Hanson, Norwood Russell. 1958. *Patterns of Discovery*. Cambridge: Cambridge University Press.

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

Huang, Xiaolei, Bradford A. Hawkins and Gexia Qiao. 2013. "Biodiversity Data Sharing: Will Peer-Reviewed Data Papers Work?" *BioScience* 63(1): 5-6. https://academic.oup.com/bioscience/article/63/1/5/241043

ISO 5127: 2017(E). *Information and Documentation: Foundation and Vocabulary*. 2nd edition. Geneva, Switzerland: International Organization for Standardization.

Jensen, Howard E. 1950 'Editorial Note.' In *Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types, and Prospects* by Howard Paul Becker. Durham, NC: Duke University Press, vii–xi.

Kaase, Max. 2001. "Databases, Core: Political Science and Political Behavior." In *International Encyclopedia of the Social and Behavioral Sciences. Vol. 5*, edited by Neil J. Smelser and Paul B. Baltes. Elsevier, Amsterdam, 3251-5.

Kaase, Max. 2015. "Data Bases and Statistical Systems: Political Science (General)." In *International Encyclopedia of the Social & Behavioral Sciences. Vol. 5*. 2nd edition, edited by James D Wright. Amsterdam: Elsevier, 830-5. http://dx.doi.org/10.1016/B978-0-08-097086-8.41019-6

Kratz, John and Carly Strasser. 2014. "Data Publication Consensus and Controversies. Version 3." *F1000Research* 3, no. 94. doi:10.12688/f1000research.4518

Kuhn, Thomas. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Le Deuff, Olivier. 2018. "Une Nouvelle Rubrique pour la RFSIC : Le Data Paper". *Revue Française des Sciences de L'information et de la Communication,* 15. http://journals.openedition.org/rfsic/5275

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

Li, Kai, Jane Greenberg and Jullian Dunic. 2020. "Data Objects and Documenting Scientific Processes: An Analysis of Data Events in Biodiversity Data Papers." *Journal of the Association for Information Science and Technology* 71(2): 172-82. doi:10.1002/asi.24226

Manovich, Lev. 2001. *The Language of New Media*. Cambridge, MA: MIT Press.

Marx, Vivien. 2013. "Biology: The Big Challenges of Big Data." *Nature* 498: 255-60. Retrieved from http://www.nature.com/articles/498255a.pdf

Mayernik, Matthew S. In press. "Data Science as an Interdiscipline: Historical Parallels from Information Science." *Journal of the Society for Information Science and Technology*.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner and Susan G. Stafford. 1997. "Nongeospatial Metadata for the Ecological Sciences." *Ecological Archives* 7(1): 330-42. doi:10.2307/2269427

Murray, Steve. 2017. "The LSST and Big Data Science: A New Kind of Telescope Will Need a New Kind of Astronomer." *Astronomy* December 15, 2017. Retrieved 2018-06-08 from http://www.astronomy.com/news/2017/12/the-lsst-and-big-data-science

Neuroth, Heike, Stefan Strathmann, Achim Osswald and Jens Ludwig (Eds.). 2013. *Digital Curation of Research Data*. Glückstadt: Werner Hülsbusch.

Newman, Paul and Peter Corke. 2009. "Data Papers - Peer Reviewed Publication of High Quality Data Sets." *The International Journal of Robotics Research* 28(5): 587. doi:10.1177/0278364909104283

Nielsen, Hans Jørn and Birger Hjørland. 2014. "Curating Research Data: The Potential Roles of Libraries and Information Professionals." *Journal of Documentation* 70(2): 221-40.

Norton, Stephen D. and Frederick Suppe. 2001. "Why Atmospheric Modelling Is Good Science." In *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, edited by Clark A. Miller and Paul N. Edwards. Cambridge, MA: MIT Press, 67-106.

Pärtel, Meelis. 2006. "Data Availability for Macroecology: How to Get More out of Regular Ecological Papers." *Acta Oecologica* 30(1): 97-9. doi:10.1016/j.actao.2006.02.002

Penev, Lyubomir, Wishwas Chavan, Teodor Georgiev and Pavel Stoev. 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. https://pensoft.net/img/upl/file/DataPaperPoster.pdf

Reitz, Joan M. 2004. *Online Dictionary for Library and Information Science* (ODLIS). Santa Barbara, CA: ABC-CLIO. https://products.abc-clio.com/ODLIS/odlis_p.aspx

Reymonet, Nathalie. 2017. "Améliorer l'exposition des données de la recherche : la publication de data papers". *Archive ouverte en sciences de l'information et de la communication*, Université Paris Diderot. https://archivesic.ccsd.cnrs.fr/sic_01427978/

Rivalle, Guillaume and Bob Green. 2016-2018. *Document Data Citation Index – Descriptive Document*. https://clarivate.libguides.com/ld.php?content_id=45722564

Rowley, Jennifer. 2007. "The Wisdom Hierarchy: Representations of the DIKW Hierarchy." *Journal of Information Science* 33(2): 163-80. doi:10.1177/0165551506070706

Rumble, John (Ed.). 2019. *CRC Handbook of Chemistry and Physics* 100th Edition. Boca Raton: CRC Press.

Schöpfel, Joachim, Dominic Farace, Hélène Prost and Antonella Zane. 2019. "Data Papers as a New Form of Knowledge Organization in the Field of Research Data." *Knowledge Organization* 46(8): 622-38.

Senderov, Viktor, Teodor Georgiev and Lyubomir Penev. 2016. "Online Direct Import of Specimen Records into Manuscripts and Automatic Creation of Data Papers

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.

from Biological Databases." *Research Ideas and Outcomes* 2: e10617+. doi:10.3897/rio.2.e10617

Sollaci, Luciana B. and Mauricio G. Pereira. 2004. "The Introduction, Methods, Results, and Discussion (IMRAD) Structure: A Fifty-year Survey." *Journal of the Medical Library Association* 92(3): 364-7.

Smith, Mackenzie. 2011. "Data Papers in the Network Era." In *Something's Gotta Give: Charleston Conference Proceedings, 2011*, edited by Beth R. Bernhardt, Leah H. Hinds and Katina P. Strauch. West Lafayette, IN: Against the Grain Press, 13-20. doi:10.5703/1288284314871

Spang-Hanssen, Henning. 2001. "How to Teach About Information as Related to Documentation." *Human IT* 5(1): 125-43. https://humanit.hb.se/article/view/168

Star, Susan Leigh and James R. Grisemer. 1989. "Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19(3): 387–420. doi:10.1177/030631289019003001

United Nations Educational, Scientific and Cultural Organization and the International Council of Scientific Unions. 1971. *UNISIST: Study Report on the Feasibility of a World Science Information System*. Paris: UNESCO.

Wikipedia: The Free Encyclopedia. 2002a. "Data Publishing." https://en.wikipedia.org/wiki/Data_publishing

Wikipedia: The Free Encyclopedia. 2020b. "IMRAD." https://en.wikipedia.org/wiki/IMRAD

Wilkinson, Mark D. et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3:160018. doi:10.1038/sdata.2016.18

Zins, Chaim. 2007. „Conceptual Approaches for Defining Data, Information, and Knowledge." *Journal of the American Society for Information Science and Technology* 58(4): 479-93. doi:10.1002/asi.20508

https://doi.org/10.5771/0943-7444-2021-4-307
Generiert durch Birger Hjorland, am 18.12.2021, 14:32:41.
Das Erstellen und Weitergeben von Kopien dieses PDFs ist nicht zulässig.