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

Joachim Schöpfel (University of Lille)
Dominic Farace (Greynet International)
Hélène Prost (CNRS)
Antonella Zane (University of Padova)

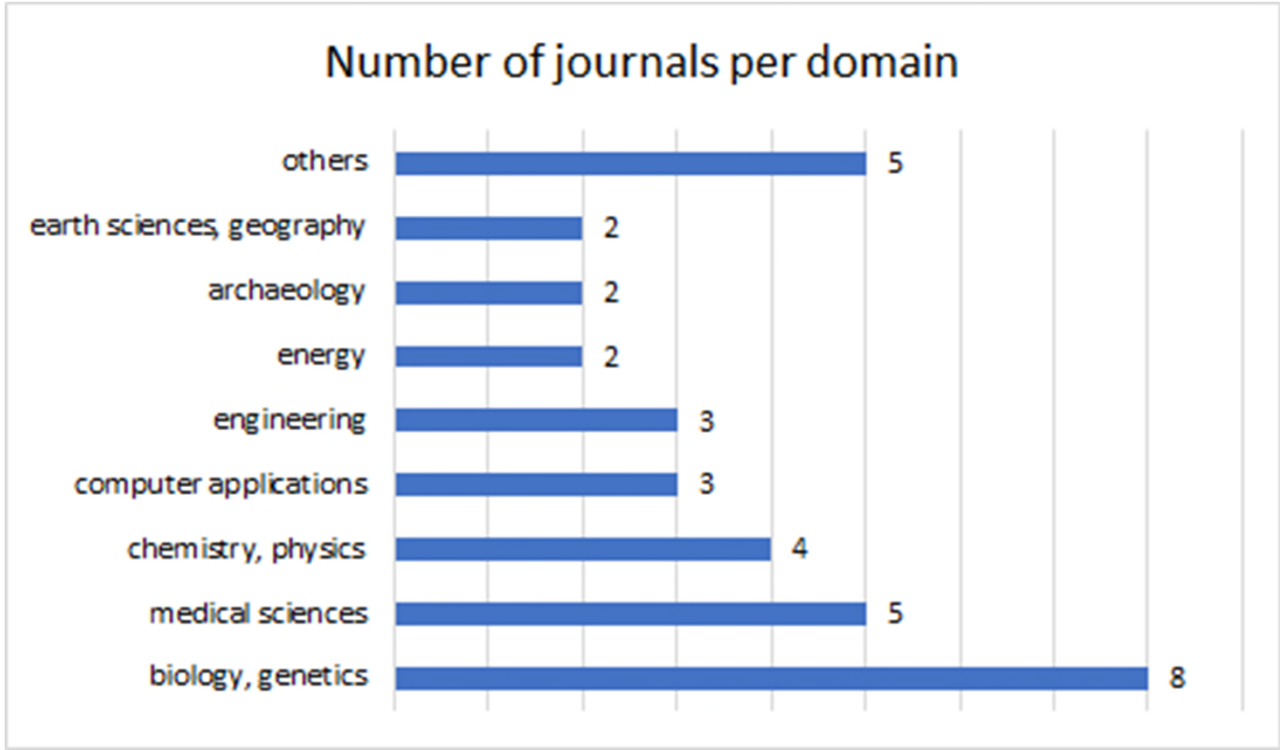
Twenty-First International Conference on Grey Literature
"Open Science Encompasses New Forms of Grey Literature"

Questions and Methodology

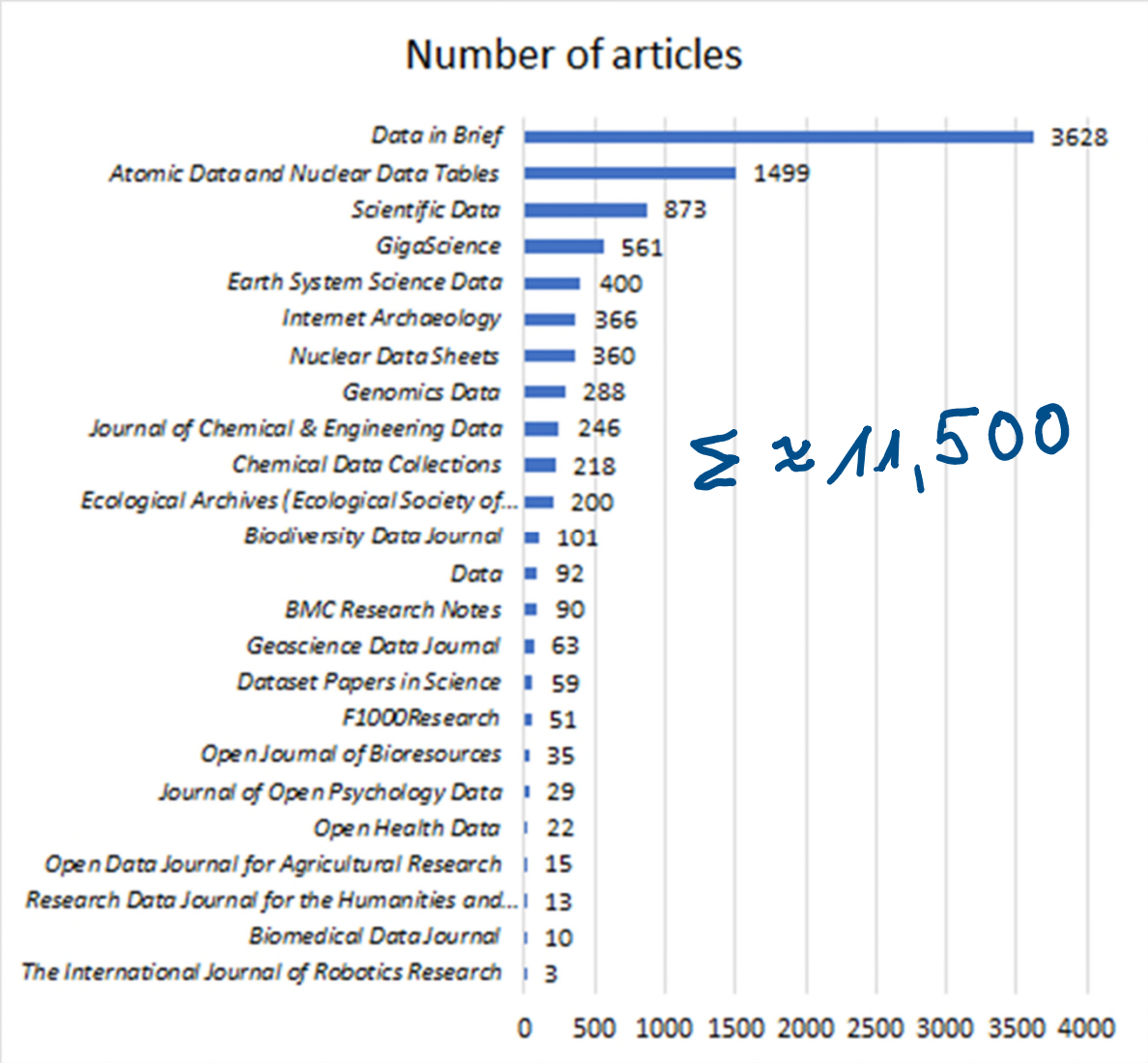
- Data papers are a young species of academic publishing
- The simplest definition is that data papers focus on "*information on the what, where, why, how and who of the data*"
- The purpose of our paper is to analyse data papers as a new tool of scientific communication and to produce insight on their contribution to the organization of scientific knowledge via questions pertaining to the production and the functions of data papers, eg:
  - How are they "written"?
  - Which is the link with data repositories, metadata and other papers?
  - Which is the (potential and real) part of automatic or semi-automatic production
  - Which is the part of human added value?
- Literature overview and bibliometric study on data journals
  - Sampling: FOSTER Plus, forschungsdaten.info, INRA, CIRAD
  - 82 data journals, with 28 « pure » data journals

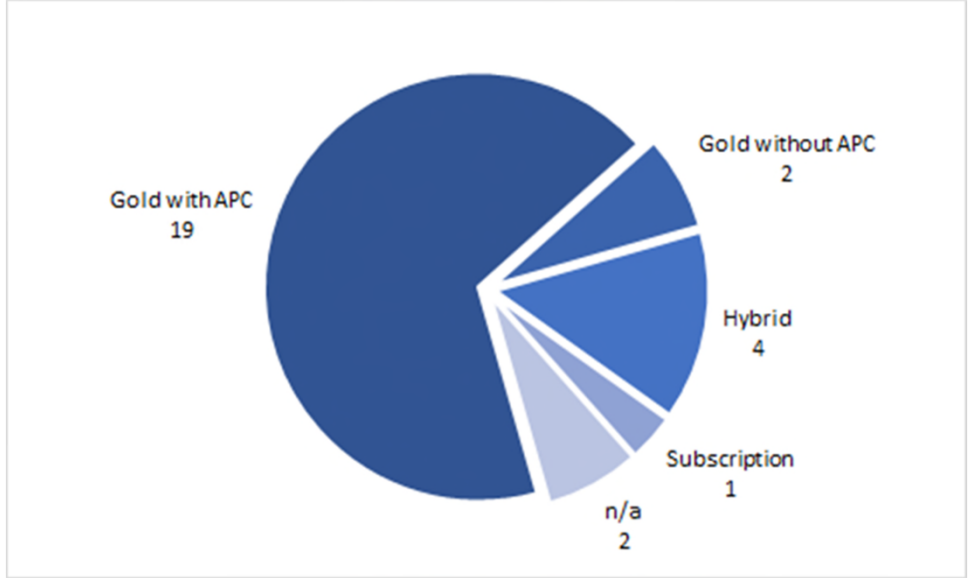
Disciplines and Publishers

- Most data journals are from STEM domains, in particular from life and medical sciences, including genetics
- Except for Taylor & Francis, all big five academic publishers (Elsevier, Springer-Nature, Wiley-Blackwell and SAGE) have their own data journals.
- Other data journals are published or hosted by newcomers, especially by OA publishers such as Ubiquity Press, BioMed Central, Hindawi, MDPI, Copernicus Publications, Pensoft or Faculty of 1000.



Business Models, Selection, and Licensing





21 data journals disseminate data papers with an open license, most often a CC-BY license.
All data journals are peer reviewed.
5 data journals apply open peer review:

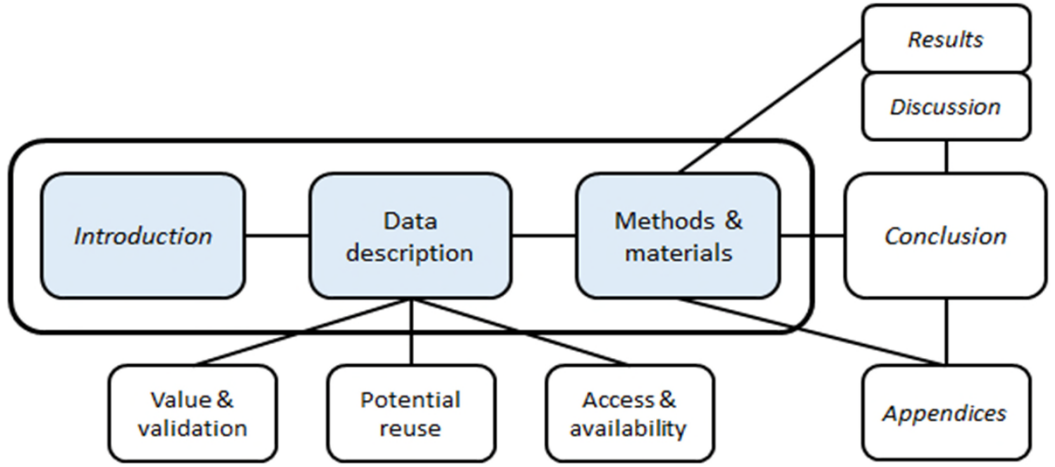
- the reviewers are suggested (and known) by authors (*F1000Research*);
- community peer review (*Biodiversity Data Journal*);
- interactive public peer review (*Earth System Science Data*).

Metadata, Identifiers, Linking

- Most journals ask for some general and usual information, compliant with the Dublin Core format, such as author, organisation, title etc.
- 26 journals publish the data papers with a DOI (93%).
- 5 journals include the author identifier ORCID (18%).
- Most of them recommend if not 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.
- Some papers contain downloadable data

Length and Structure

- It is generally assumed that data papers are short texts, up to 4 pages. In fact, this is only partly true.
- Most journals do not limit the length of submitted papers or make the usual recommendations (6-10 pages, or maximal 6,000 words).
- Up to 100 pages…



- A core structure with three central sections (in blue)
- Other, optional or peripheral sections
  - some similar to regular papers (in italics)
  - others characteristic for data papers (in white)

A New Ecosystem

- Data papers 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.
- **Reuse rights:** most data journals allow publishing with an open license, often with generous reuse and remixing rights (e.g. CC-BY license and/or CC0 waiver).
- **Findability:** the editorial model of data journals requires standard identifiers for the datasets, e.g. DataCite's DOI, to guarantee (and increase) the findability of datasets; they also attribute DOIs to their own data papers, creating a kind of cross-linked DOI system between data papers and datasets.
- **Interconnectedness:** perhaps the most relevant aspect is the integration of data journals and papers in a complex structure of open access journal platforms and data repositories, academic communities, research projects, conferences etc. Interconnectedness requires interoperability between platforms and infrastructures but is more than technology, formats and standards, insofar it means new ways of doing science, including research management, research environment, workflows etc.

FAIR Principles

- Along with metadata, data papers contribute to the compliance with FAIR principles.
- In particular to the two principles of findability and reusability, insofar they help people (and machines) 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. Data papers are a new infrastructure of refinement and dissemination of the metadata fuel.

Blurred Boundaries

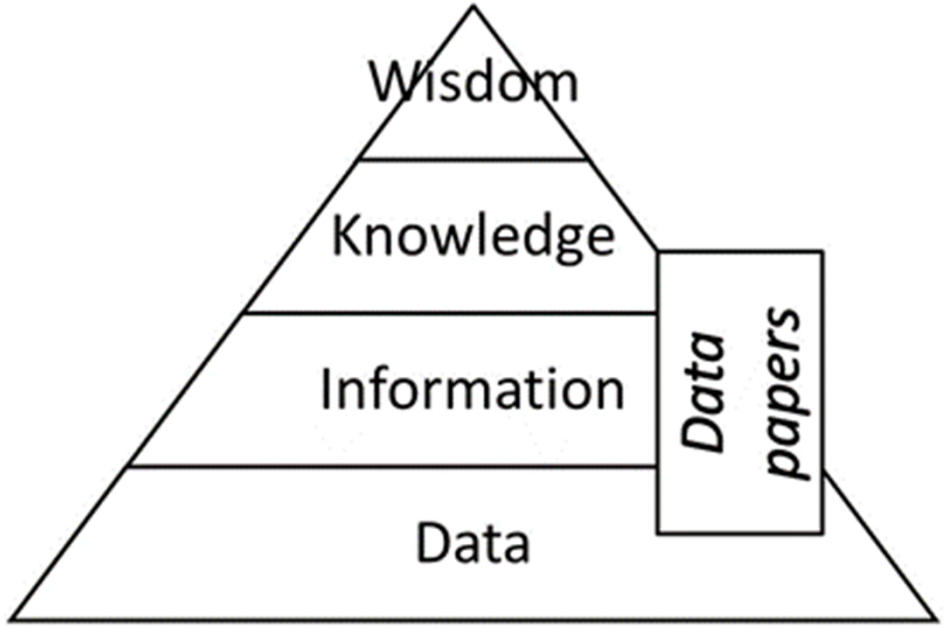
- The boundaries with other categories of academic publishing are partly blurred, especially with regular research papers.
- The specific identity of data papers is mainly defined in opposition with regular research papers. The reality is different. There is a lot of divergence and heterogeneity which can be described on four levels.
- Data journals also accept other articles.
- Data papers are published in other journals.
- Data papers are more than simple data papers.
- There are other emerging types of articles, similar to but not identical with data papers.
  - 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*).
  - 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*).

Who is writing? Who is reading?

- Data papers are (can be) generated automatically and are potentially machine-readable; yet, the human contribution (still) appears vital in terms of intellectual property and richness of content.
- Automatic generation: the Pensoft workflow as well as the INRA tool, reveal the potential of automatic generation of data papers, but also its requirements and limits. Automatic generation of data papers requires a high degree of standardization and interoperability between data repositories, text processing tools and journal platforms, especially regarding metadata formats and identifiers.
- Machine readability: 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.

Data, Information, Knowledge

- Data papers are essentially information, i.e. description of data (as defined by the DIKW model) but also partly contribute to the generation of knowledge and data on its own.



Definition

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

Further Questions

- Monitoring: how can the indexing of data papers be improved in order to facilitate their identification and follow-up?
- Business models: what is the risk of predatory publishing of data journals and data papers?
- Disciplines: are data papers as relevant in arts, social sciences and humanities as in life sciences, chemistry etc.? Should their data papers be published in large and multidisciplinary data journals, together with STM, or should they have their own data journals?
- Ecosystem: more case studies are needed on specific links between research data management, academic publishing, and the production and dissemination of data papers, in a given environment and community (equipment, discipline, structure…).
- Evaluation: how do scholars get credit for publishing data papers?

- Will data journals remain part of the research ecosystem or not? Perhaps they will not.
- However, it seems probable that the number of data papers will continue to grow and gain importance
  - perhaps (probably) not via data journals but via increasing hybridization of research journals and journal platforms,
  - perhaps even through the merging of journal and data platforms.

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

THANK YOU !

Contact
joachim.schopfel@univ-lille.fr