Data Documentation for: Sustaining Dryad: Reconfiguring relationships and “thinking like a business” to maintain open data infrastructures

Dorothea Strecker (corresponding author)[1]
dorothea.strecker@hu-berlin.de
0000-0002-9754-3807

Kathleen Gregory[2]
0000-0001-5475-8632

[1] Humboldt-Universität zu Berlin, Berlin School of Library and Information Science
[2] Leiden University, Centre for Science and Technology Studies (CWTS)

Version 1.0 – August 27, 2026

Abstract

Open data infrastructures underpin how open research is practiced across communities. While critical, the longevity of such infrastructures is far from certain, as they grapple with challenges related to technologies, dominant ideologies, and continued funding streams. This dataset comprises materials of a mixed-methods study investigating the financial evolution of Dryad, an established open data infrastructure developed to facilitate data sharing and reuse. The study combines semi-structured interviews with Dryad staff and board members and a quantitative descriptive analysis of the data collection and its use to show how Dryad has worked toward financial sustainability.

Keywords: research data, open science, open infrastructures, data repositories, funding models, sustainability

Citing the Data Set

Strecker, Dorothea ; Gregory, Kathleen (2026) Data Documentation for: Sustaining Dryad: Reconfiguring relationships and "thinking like a business" to maintain open data infrastructures. Version 1.0: DOI: 10.5281/zenodo.19310438

Licenses

This dataset was compiled as part of a mixed-methods study investigating the financial evolution of Dryad, an established open data infrastructure developed to facilitate data sharing and reuse. We combine semi-structured interviews with Dryad staff and board members and a quantitative descriptive analysis of the data collection and its use to show how Dryad has worked toward financial sustainability.

Data collection and processing

1.1 Interviews

We conducted 14 one-hour, online semistructured interviews between April and August 2025 with current and past Dryad staff and members of the Board of Directors. Interviews lasted approximately 60 minutes and were transcribed for analysis. To recruit participants, a relationship was first developed with Dryad senior staff, who agreed to support the study. Participants were then recruited based on recommendations and snowball sampling. Recruitment stopped once we had spoken with representatives from all professional roles at Dryad who were willing to participate in this study. In sum, we spoke with two full-time curators and technologists, eight board members (split equally between publishing and institutional representatives), and three (former) staff. We also spoke with one employee at another OI with extensive experience of Dryad. Speaking with these individuals enabled us to collect enough data to answer our research questions and achieve sufficiency in our data (LaDonna et al., 2021; Vasileiou et al., 2018).
Participants were invited to bring a document to serve as a discussion prompt; these included blogposts, reports, and policy documents. Participants explained the importance of these prompts and their relationship to sustainability. Other interview questions addressed different factors influencing the sustainability of open data infrastructures, identified from the literature, e.g. funding, governance, communities, and relationships to other repositories. Questions were tailored to match participants' expertise. ATLAS.ti ("ATLAS.ti", 2022) was used to support a combination of deductive and inductive coding and reflexive thematic analysis (Braun & Clarke, 2006, 2019).
We developed the codebook iteratively in rounds of discussion with both authors. After initial deductive code development, both authors coded a sample of three transcripts; codes were then further discussed. The authors then re-coded the set of transcripts and had a further discussion. The first author then coded the remaining transcripts; the second author read the remaining transcripts to feed into further discussion and iteration. Deductive codes were based on literature about sustaining knowledge and data infrastructures as well as the interview protocol. Deductive codes related to governance (e.g. community governance; board organization); funding (e.g. descriptions of the new fee structure); Dryad's relation to other repositories; and community needs and engagement. Inductive codes included reconfigurations of particular relationships and business thinking.

1.2 Quantitative descriptive analysis

For the quantitative descriptive analysis, DOIs were retrieved from the Dryad API on 2025-06-24.[1] At that time, 64.775 DOIs were accessible through the API. Metadata of datasets (author last names, publication year, publisher name and resource type) were added via the DataCite API on 2025-07-11.[2] The resulting dataset (*dryad-metadata.csv*) includes metadata for 64.775 Dryad datasets.
Three sources were combined to identify research outputs referencing Dryad DOIs:

- Data Citation Corpus v3.0 (2025-02-01): Dryad DOI strings were matched with the "dataset" column - 46.873 references (DataCite & Make Data Count, 2025)
- DataCite REST API - relationships section (2025-07-11): "citations" were retrieved from the API[3]; these include research outputs referring to a Dryad DOI with one of these relationship types: Cites, References, IsSupplementedBy - 54.355 references
- OpenAlex (2025-07-04): Dryad DOIs were passed as an argument to the "cites" filter[4] - 492 references

As Figure 1 shows, this approach generated overlap - some references are identified by more than one source. Therefore, references to Dryad DOIs were deduplicated.

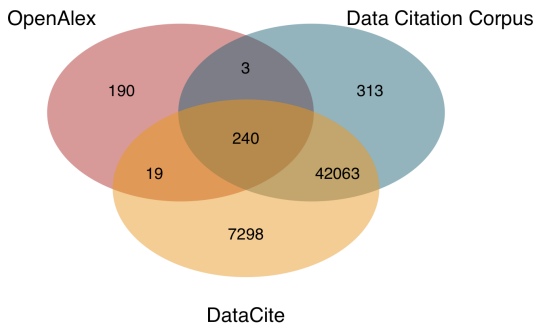

Figure 1: Distribution of references across the three sources

In addition, DOIs of all referencing objects were checked; 318 were removed because the DOI didn't resolve (HTTP error 404). Metadata of referencing research outputs (author last names, publication year, publisher name and resource type) were added via the DataCite and crossref APIs on 2025-07-11. The resulting dataset (*dryad-references.csv*) includes 50.126 references to Dryad datasets.
Subject information for *dryad-metadata.csv* was processed to include only labels

[1] Dryad API: https://datadryad.org/api; last accessed: 2026-03-04
[2] DataCite API: https://support.datacite.org/docs/api; last accessed: 2026-03-04
[3] DataCite - Consuming Citations and References: https://support.datacite.org/docs/consuming-citations-and-references; last accessed: 2026-03-04
[4] OpenAlex API documentation: https://docs.openalex.org/how-to-use-the-api/api-overview; last accessed: 2026-03-04

assigned from the OECD Fields of Science Classification (FOS) - FOS is the most widely used controlled vocabulary in the sample. These subject labels were aggregated at the highest level of the FOS classification. Subject labels were deduplicated for each dataset.
Publisher names in *dryad-references.csv* were harmonized and the relationship between the Dryad dataset and the referencing research output was categorized. After checking publisher names of referencing research outputs, 15 mentions of ”JSTOR” were clarified and 6 missing values were added. Publisher imprints were then grouped using a dataset compiled by the German Kompetenznetzwerk Bibliometrie (Haupka et al., 2025).
The type of relationship between the Dryad dataset and the referencing research output was determined by comparing author last names. If any of the last names overlapped, the relationship was categorized as “data use”, and otherwise as “data reuse”. This approach has been used in previous research to differentiate data use and reuse (Cohen et al., 2025; Strecker et al., 2025).

Description of the data files

2.1 interview-protocol.pdf

The file *interview-protocol.pdf* contains the interview protocol for the interviews.

2.2 codebook.pdf

The file *codebook.pdf* contains the codebook for the qualitative analysis.

2.3 dryad-metadata.csv

The file *dryad-metadata.csv* includes metadata of Dryad datasets. It comprises 64.775 rows (each row corresponding to a dataset) and the following columns:

- **doi** DOI of the dataset
- **title** title of the dataset
- **publication_year** year the dataset was published
- **fos_subject_label** subject label at the highest level of the FOS classification, multiple values separated by ” / ”

2.4 dryad-references.csv

The file *dryad-references.csv* includes references to Dryad datasets. It comprises 50.126 rows (each row corresponding to a reference) and the following columns:

- **doi_to** DOI of the dataset

- **doi_from** DOI of the referencing object
- **publication_year_to** year the dataset was published
- **publication_year_from** year the referencing object was published
- **publisher_from** publisher of the referencing object
- **use_type** relationship between the dataset and the referencing object - either *use* or *reuse*

References

ATLAS.ti. (2022). https://atlasti.com

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Cohen, A., Iarkaeva, A., Ivanovic, B., Nachev, V., & Bobrov, E. (2025). Investigating the Reuse of Biomedical Research Data Using the Data Citation Corpus. https://nbn-resolving.org/urn:nbn:de:bsz:16-heidok-363891

DataCite & Make Data Count. (2025). Data Citation Corpus Data File. https://doi.org/10.5281/zenodo.14897662

Haupka, N., Culbert, J., Donner, P., Jahn, N., Lenke, C., Mayr, P., Meier, A., Mittermaier, B., Scheidt, B., Stahlschmidt, S., & Taubert, N. (2025). OPENBIB: Selected curated open metadata based on OpenAlex. https://doi.org/10.5281/zenodo.15308680

LaDonna, K. A., Jr, A. R. A., & Balmer, D. F. (2021). Beyond the Guise of Saturation: Rigor and Qualitative Interview Data. *Journal of Graduate Medical Education*, *13*(5), 607–611. https://doi.org/10.4300/JGME-D-21-00752.1

Strecker, D., Soltau, K., & Bach, F. (2025). How Are Research Data Referenced? The Use Case of the Research Data Repository RADAR. *Research Data Management: Challenges in a Changing World*, 146–159. https://doi.org/10.11588/heibooks.1652

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, *18*(1), 148. https://doi.org/10.1186/s12874-018-0594-7