

Editorial: *Dark Data and the Cost to Open Science*

Open science is sometimes presented as an aspiration. For a data journal, it is an operating condition. A finding that cannot be connected to the data, metadata, code, and provenance needed to examine it is not fully available to the scientific record. The article may be visible while the evidence beneath it remains in the dark.

Heidorn used the term *dark data* for data that are not easily found by potential users—data that, because they are poorly indexed or stored, are likely to be underused and eventually lost [1]. In contemporary engineering and computer science, darkness has several familiar forms: Files left on personal devices; data described in an article but never deposited; broken or short-lived links; undocumented simulation outputs; proprietary training or evaluation sets with no auditable surrogate; and the deceptively reassuring statement that data are “available upon request.”

Darkness should not, however, be treated as a synonym for every legitimate restriction. Personal, clinical, security-sensitive, culturally governed, or commercially protected data may be unsuitable for unrestricted release. Such data can still be findable and responsibly governed: A persistent record can describe what exists, why access is restricted, who may apply, which conditions apply, and which materials permit independent validation. A restriction with durable metadata and a functioning access process is categorically different from a dataset whose location, ownership, and conditions of reuse are unknown. The FAIR principles similarly distinguish accessibility from universal, anonymous download; they emphasize persistent identification, metadata, clear access protocols, interoperability, and reuse conditions [2].

WHAT THE EVIDENCE SHOWS

The scale of the problem is measurable, but it cannot be summarized responsibly by one universal percentage. Studies examine different disciplines, publication years, definitions of availability, and levels of completeness. Their estimates should therefore be read as evidence from specified samples rather than as a global census of scientific data.

One of the clearest cross-disciplinary studies examined the most critical data underlying 875 articles published in *Nature* and *Science* during 2000–2019 [3]. Before the investigators contacted authors, an average of 54.2% of articles had their

critical data fully available; the range across nine disciplines was 33.0%–82.8%. Full availability rose to 69.5% after follow-up. Yet among 310 requests to authors, data were obtained in only 39.4% of requests on average, while 19.4% were declined. The study’s conclusion is difficult to avoid: “available upon request” is not a durable substitute for repository deposit.

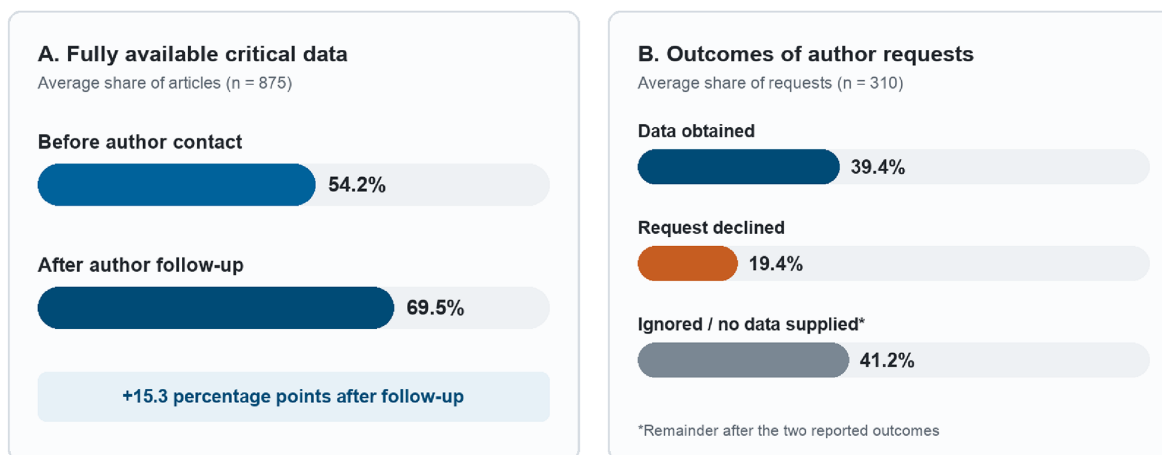
Time compounds the problem. Vines et al. [4] requested datasets associated with 516 papers published between two and 22 years earlier. Among articles for which authors reported the status of their data, the odds that a dataset was still extant fell by 17% for each additional year since publication. The odds of finding a working e-mail address for a first, last, or corresponding author fell by 7% per year. These are declines in *odds*, not percentage-point losses and not a universal half-life. In the broader sample studied by Tedersoo et al. [3], estimated annual loss rates were lower—3.5% for initially available data and 5.9% for data effectively available upon request—but the direction was the same. When preservation depends on an author’s inbox or hardware, access decays.

The consequences are especially visible in computational research. In a review of 400 empirical articles from the AAAI and IJCAI conference series, Gundersen and Kjensmo [5] found that none documented every variable in their reproducibility framework; articles typically documented only 20%–30% of the evaluated experiment, data, and method variables. A later commentary on medical artificial intelligence argued that insufficient methodological detail and absent algorithm code undermined the scientific value of a prominent study [6]. Data alone are therefore not enough. In data-driven research, the reproducible object commonly includes data, preprocessing, code, parameters, software dependencies, and a traceable account of how the reported result was produced.

There are also constructive examples. The Integrated Resource for Reproducibility in Macromolecular Crystallography was created to preserve raw diffraction experiments that conventional structural databases did not retain. At publication, it held 2920 diffraction experiments corresponding to 5767 datasets and approximately 1.2 million images [7]. The lesson is not that one repository model fits every discipline. It is that data cease to be dark when a community

Evidence on data availability in Nature and Science (2000–2019)

Tedersoo et al. (2021): 875 articles; 310 author contacts when critical data were not accessible



Note: The panels have different denominators and cannot be combined as slices of one population.

Source: Tedersoo et al., *Scientific Data* 8, 192 (2021), doi:10.1038/s41597-021-00981-0.

FIG. 1. Two distinct results reported by Tedersoo et al. [3]. Left: Average full availability of critical data among 875 *Nature* and *Science* articles before and after author follow-up. Right: Average outcomes among 310 requests; the 41.2% remainder is inferred from the reported 39.4% obtained and 19.4% declined. The denominators differ and therefore must not be combined into a single pie chart.

supplies durable infrastructure, sufficient metadata, and an intelligible route from a published claim to its underlying observations.

THE COST OF LEAVING DATA DARK

The first cost is lost scrutiny. Without the evidentiary record, reviewers and readers cannot determine whether an apparent effect depends on an exclusion, preprocessing choice, mislabeled observation, or implementation detail. Reproduction is not a ceremonial rerun of a script; it is a practical means of locating error and testing whether a result is robust to inspection.

The second cost is duplicated effort. Data that cannot be discovered are functionally indistinguishable from data that never existed to the next researcher planning an experiment. Measurements may be recollected, simulations rerun, and failures repeated. This waste is not confined to storage costs. It consumes research time, public and private funding, scarce samples, participant effort, and sometimes energy-intensive computation.

The third cost is selection bias in the scientific record. Positive or polished outcomes are more likely to appear in articles, whereas failed trials, null results, intermediate artifacts, and inconvenient cases often remain on local storage. The visible literature can then make a method appear more reliable or a design space more thoroughly explored than the underlying work warrants. A data description can counter that bias by making a well-curated dataset valuable even when it does not support a dramatic headline.

The fourth cost is weakened cumulative science. Data become more valuable when they can be combined, compared,

and reinterpreted. That value depends on identifiers, provenance, licenses, common formats, and metadata rich enough for a future researcher—or a machine—to decide whether the data are fit for a new purpose. The FAIR principles were designed around precisely this transition from nominal availability to meaningful reuse [2].

FROM DATA-AVAILABILITY STATEMENTS TO DATA AVAILABILITY

Publishers, editors, reviewers, institutions, and authors each control part of the route by which data become durable scientific objects. Five practices should become routine.

First, Deposit Before Acceptance

Data supporting the central claims of an article should be placed in a trusted disciplinary or general-purpose repository, assigned a persistent identifier, versioned, and accompanied by an explicit license or terms of use. A personal website or project folder may be useful for dissemination, but it is not a preservation strategy.

Second, Make the Availability Statement Executable

The statement should identify the repository record and files; state which version supports the article; explain any embargo or restriction; name the access authority; and specify how a qualified user can apply. “Available upon reasonable request” should be reserved for exceptional, justified circumstances and should include a durable access process rather than reliance on one corresponding author’s continued availability.

Third, Review the Research Object, Not Only the Prose

At least one reviewer or editor should test the link, inspect the metadata, confirm that the cited version exists, and determine whether the shared materials plausibly support the article's central analyses. For computational work, this includes checking whether code and environment information are present. Community interventions such as the NeurIPS reproducibility program show how artifact policies, checklists, and replication challenges can make these expectations operational [8].

Fourth, Document Legitimate Closure

Ethical and legal constraints should determine the mode of access, not whether the dataset is acknowledged. Public metadata, a data dictionary, provenance, summary statistics where safe, and a controlled-access procedure can make restricted data scientifically legible. Where the source data cannot be released, authors should explain what validation is possible and provide suitable code, synthetic examples, or secure-computing arrangements without implying that these substitutes are identical to the original data.

Fifth, Assign Credit and Responsibility

Data curation is scholarly work. Dataset citations, contributor roles, and data articles create visible credit; clear stewardship responsibilities reduce the chance that access disappears when a student graduates, a grant ends, or a laboratory changes platforms. Evidence from the *Journal of Memory and Language* also shows that policy can change behavior: After an open-data and code policy, data sharing increased by more than 50%, although only 34%–56% of the evaluated articles were reproducible depending on the criterion used [9]. Availability is necessary, but curation and executable analysis remain essential.

A ROLE FOR DATA DESCRIPTIONS

IEEE Data Descriptions exists at the point where these practices become publishable scholarship. A useful data description does more than announce that files exist. It explains how the data were created, what each variable means, which quality controls were applied, how missingness and exclusions were handled, what limitations govern reuse, and how the repository record relates to the article. It supplies the context that turns stored bytes into evidence.

This standard should apply to small as well as large datasets. The long tail of science identified by Heidorn is composed precisely of the many modest, heterogeneous outputs that are least likely to receive dedicated infrastructure

[1]. Their individual scale may be small, but their collective scientific value is large. Publishing them with durable metadata and credit is one way to prevent them from disappearing into personal storage and institutional silos.

The objective is not openness without judgment. It is durable, proportionate, and verifiable access. Some data should be openly downloadable. Some should be available through controlled mechanisms. In every case, readers should be able to discover what exists, understand the conditions governing it, and determine how the claims in a publication can be examined.

Dark data are therefore not merely a technical inconvenience. They are a break in the chain of evidence. Repairing that break requires infrastructure, incentives, careful documentation, and editorial enforcement. The scientific community has already developed the principles and many of the tools. Our task now is to make their use ordinary: To replace promises of availability with tested access, temporary links with persistent records, and hidden research outputs with citable contributions to the scientific record.

STEPHEN MAKONIN , *Editor-in-Chief*
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
e-mail: smakonin@sfu.ca

REFERENCES

- [1] P. B. Heidorn, "Shedding light on the dark data in the long tail of science," *Library Trends*, vol. 57, no. 2, pp. 280–299, 2008.
- [2] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, 2016, Art. no. 160018.
- [3] L. Tedersoo et al., "Data sharing practices and data availability upon request differ across scientific disciplines," *Sci. Data*, vol. 8, 2021, Art. no. 192.
- [4] T. H. Vines et al., "The availability of research data declines rapidly with article age," *Current Biol.*, vol. 24, no. 1, pp. 94–97, 2014.
- [5] O. E. Gundersen and S. Kjensmo, "State of the art: Reproducibility in artificial intelligence," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1644–1651.
- [6] B. Haibe-Kains et al., "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E14–E16, 2020.
- [7] M. Grabowski et al., "A public database of macromolecular diffraction experiments," *Acta Crystallographica Sect. D, Struct. Biol.*, vol. 72, no. 11, pp. 1181–1193, 2016.
- [8] J. Pineau et al., "Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program," *J. Mach. Learn. Res.*, vol. 22, no. 164, pp. 1–20, 2021. [Online]. Available: <https://jmlr.org/papers/v22/20-303.html>
- [9] A. Laurinavichyute, H. Yadav, and S. Vasisht, "Share the code, not just the data: A case study of the reproducibility of articles published in the journal of memory and language under the open data policy," *J. Memory Lang.*, vol. 125, 2022, Art. no. 104332.