Let go of your data

Taking 'upon request' out of data availability statements in papers.

Sometimes scientists just want to see the data. In 1695, Sir Isaac Newton wrote an exasperated letter to the British Astronomer Royal John Flamsteed, whose data on lunar positions he was trying to get for more than half a year[1]. In that letter Newton declared that "these and all your communications will be useless to me unless you can propose some practicable way or other of supplying me with observations … I want not your calculations, but your observations only."

Flamsteed refused, precipitating lifelong animosity between the two men. Surprisingly, modern data sharing may require almost as much effort as Newton had to endure. 93% of the papers published in the first ten months of 2019 in *Nature Materials* declare that at least some of the data are available only from authors upon request. Needless to say, e-mailing authors for data is neither a sustainable nor a reliable way of getting it[2].

Access to the data on which publication results rely should be reliable for many reasons. It facilitates reproducibility in science, at the time when most researchers believe that science faces a reproducibility crisis[3]. It enables transparency for the research that is often supported by taxpayer funds. It accelerates new materials discovery[4] and allows for secondary research such as meta-analysis or machine learning on large datasets aggregated from multiple experiments. It gives scientists who develop new models or analysis procedures an opportunity to validate them on a much larger dataset than is typically available in their labs. Finally, some truly unique datasets — for example, those obtained at large-scale taxpayer-funded user facilities — should be publicly available after a reasonable embargo period.

Why don't most material scientists do it then? Not for lack of data that would be useful to share. However, preparing and curating data for sharing is an additional effort with few direct incentives. The data-generation process is fragmented: most instruments run proprietary software with closed file formats, and every scientist has their favourite data-processing tool, adding to the chaos. And even if researchers put their data online, they rarely add the crucial metadata — salient properties of the dataset, from its title and description, to the type of the data, measurement techniques used and variables measured, and a link to the article that describes the data — that enable discovery and reuse. Some data of course cannot be shared because it contains sensitive, export-controlled or proprietary information, but in the vast majority of cases it is the incentive and the tools to do it that are just not there.

And yet, sharing data is already commonplace in several disciplines, such as geosciences, genomics and neuroscience[5]. We can look to these disciplines for best practices that are both established, and, in Newton's words, "practicable". Let's discuss some of the key ideas.

First, we should ensure long-term preservation of the shared data. Scientists often publish and annotate data on their labs' websites, but these sites go away, their addresses change, the links become outdated. GitHub is a popular place to publish data, but each file comes with little to no metadata and no explicit reference to the associated paper. More important, there is no guarantee that the directory will look the same a year from now as it does today.

Just as most research publications today have digital object identifiers (DOIs), DOIs or similar identifiers should be used for datasets to provide a persistent and resolvable way to reference the data, even if the data need to be moved to another location. Persistent identifiers enable others to cite your datasets and give you credit.

Data will be of limited use if they are not discoverable, either by generic search engines or search engines for data, such as Google's Dataset Search[6]. Search engines understand datasets better if the page includes critical metadata such as type and format of the dataset, keywords, data originator and funder[7]. Such a requirement does put an additional burden on the researcher; however, many dataset repositories today generate this metadata automatically from the information that the scientists submit when they upload datasets. There are also standalone tools that simplify this task[8]. Similarly, data are useful for analysis only when they are in a structured form — yet we frequently see 'data' shared as PDF files rather than anything that computational tools can process, such as crystallographic information files, raw images, tabular files or simulation trajectories.

There are many data repositories, both generic and materials-science specific, that already provide many of these features. For instance, scientists routinely deposit crystallographic structures to the Cambridge Crystallographic Data Centre and provide accession numbers in the paper. The same is true for sequences, which are deposited in the repositories at the National Center for Biotechnology Information. General-purpose repositories such as Figshare and Zenodo ensure long-term storage[9], create a DOI as part of the submission process and automatically generate metadata for discoverability.

Even though a shift to routine data sharing will not occur overnight, our increasingly digital world makes it almost inevitable. As each discipline develops conventions for data sharing, researchers must think proactively about best practices. Publishers should insist on more than vague data availability statements. Funders should require data sharing and encourage including data curation and publishing costs in proposal budgets. Instrument manufacturers need to create easy mechanisms for exporting raw data to sharable formats that include metadata. Perhaps one of the long-term goals could be development of common data sharing formats, a 'PDF for materials data'. Then, someone reading an article a few years from now will look back at Newton's experience in 1695 as a distant past. ❐

Natasha Noy[1*] and Aleksandr Noy[2]

[1]*Google Research, Google Inc, Mountain View, CA, USA.* [2]*Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA.*

**e-mail: noy@google.com*

Published online: 18 December 2019
https://doi.org/10.1038/s41563-019-0539-5

References

1. Kollerstrom, N. & Yallop, B. D. *J. Hist. Astron.* **26**, 237–246 (1995).
2. Arnold, C. *Science* https://doi.org/10.1126/science.caredit.a1400101 (2014).
3. Baker, M. *Nature* **533**, 452–455 (2016).
4. de Pablo, J. J. et al. *npj Comput. Mater.* **5**, 41 (2019).
5. Pierce, H., Dev, A., Statham, E. & Bierer, B. *Nature* **570**, 30–32 (2019).
6. Dataset Search. *Google* https://go.nature.com/2PQZNkg (2019).
7. Google Search. *Google* https://go.nature.com/2JSoXeI (2019).
8. Structured Data Markup Helper. *Google* https://go.nature.com/34zqa2f (2019).
9. Preservation and continuity of access policy. *Figshare* https://go.nature.com/36G8Lqm (2019).

Competing interests

N.N. leads the Dataset Search project at Google.