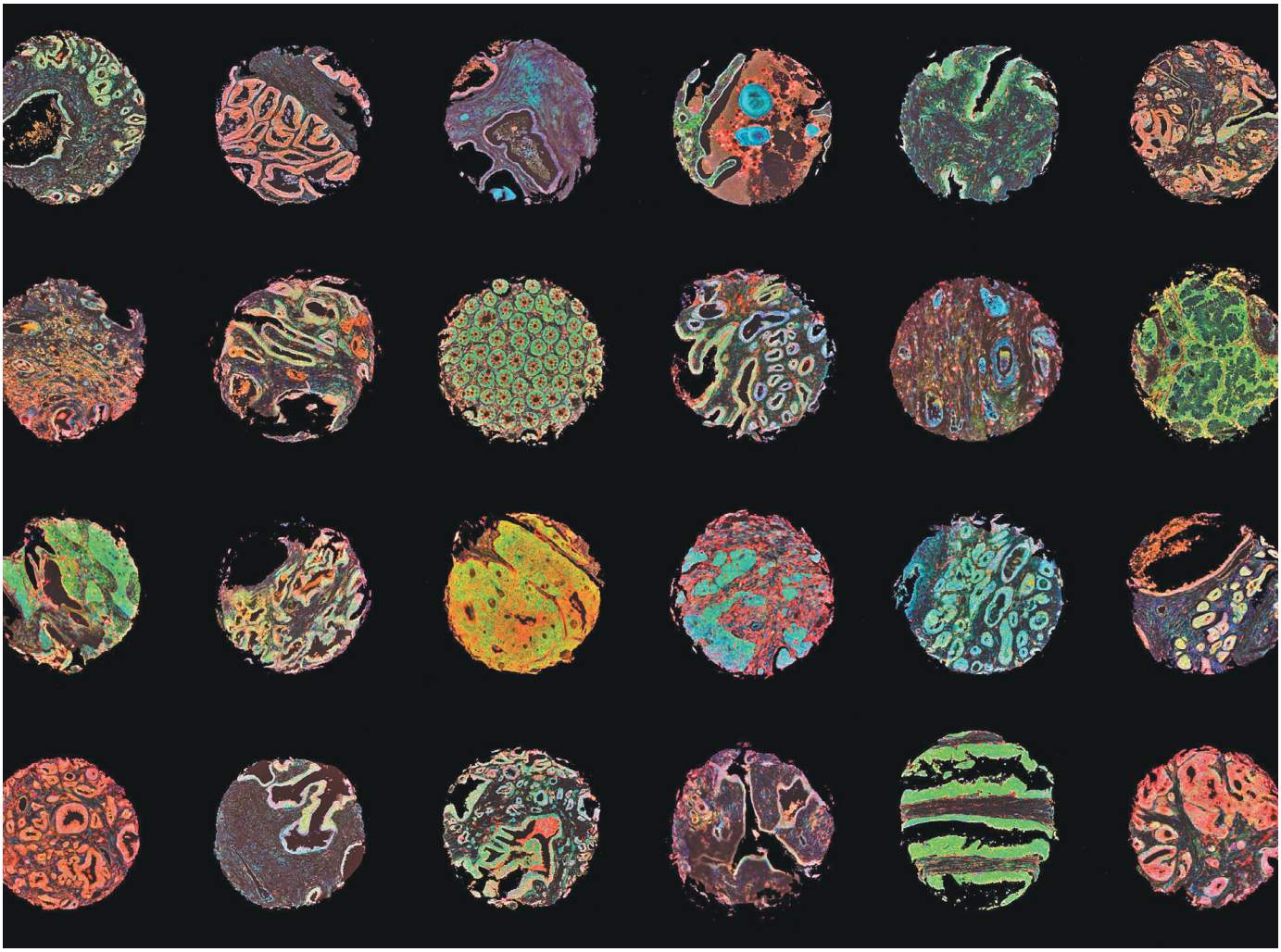


A. AHMED, J. PENDJIKY AND M. MILLAR (CC BY)

**Microscopy images of normal and cancerous human tissue — when reused by others, the impact of the collected data is multiplied.**

# Credit data generators for data reuse

To promote effective sharing, we must create an enduring link between the people who generate data and its future uses, urge **Heather H. Pierce** and colleagues.

Much effort has gone towards crafting mandates and standards for researchers to share their data[1–3]. Considerably less time has been spent measuring just how valuable data sharing is, or recognizing the scientific contributions of the people responsible for those data sets. The impact of research continues to be measured by primary publications, rather than by subsequent uses of the data.

To incentivize the sharing of useful data, the scientific enterprise needs a well-defined system that links individuals with reuse of data sets they generate[4]. To further this goal, the Association of American Medical Colleges (where H.H.P. and A.D. work) and the Multi-Regional Clinical Trials Center at Brigham and Women's Hospital and Harvard Medical School (where E.S. and B.E.B. work), along with *The New England Journal of Medicine*, convened a 2018 workshop of representatives from 50 organizations to discuss and validate such a system. The workshop included major journals, funders, data-citation groups and academic centres (see Supplementary Information, Participant list) and was preceded by numerous meetings.

Here we propose a system for leveraging existing initiatives and infrastructure to track the use, reuse and impact of scientific data through the consistent adoption of unique identifiers. Our system begins when researchers deposit a data set that they have generated. It then links every use and published analysis of that data set back to the original researchers (see 'Virtuous cycle').

## HIDDEN GEMS

When data are reused, the impact of their collection is multiplied. Sharing is commonplace in several disciplines, including ▶

genomics, neuroscience, geoscience and astronomy, reinforced by mandates and infrastructure. And an increasing number of studies reuse shared data. Data from a project on galactic star formation, led by astronomer Alyssa Goodman at Harvard University and her colleagues, has so far been downloaded hundreds of times. And since the Cancer Genome Atlas (TCGA) was founded in 2006, it has amassed genomic information on more than 20,000 samples of primary cancer and matched normal tissues. These data continue to be used in ways that could not have been anticipated at the outset. For instance, Julie Dunning Hotopp, a microbiologist at the University of Maryland in Baltimore, interrogates TCGA data sets to find evidence of bacterial DNA integration in the human genome.

Data reuse also happens on smaller scales. For example, neuroscientist Sam Nastase at Princeton University in New Jersey tweeted in February to celebrate a publication by another team "based entirely on one of [his] shared datasets". The paper used neuroimaging data he had deposited with the OpenNeuro repository, and tackled questions he had "never fully developed". Nastase learned of the reuse from automated citation alerts he had set up for his published papers on the scientific social-networking site ResearchGate and Google Scholar. A more sophisticated tracking system would detect when such publications involved data reuse, rather than citation to the conclusions of the previous paper. It would enable the original data generators to be credited formally.

## DATA CITATION

Mandates and recommendations alone can go only so far to promote effective data sharing. Academic recognition and promotion depend on researchers getting credit for their work. Calls for increased data sharing, including by the International Committee of Medical Journal Editors (ICMJE), recommend that researchers be recognized for these contributions, implicitly acknowledging that recognition is not the default. The original researchers might well be reluctant to put in the time for data curation if they feel their data are unlikely to be useful to others. A robust system for tracking and crediting data reuse would demonstrate when shared data show value beyond the original publication.

Authors and publishers do not have consistent approaches to indicate that a data set that forms the basis of a scientific conclusion in one paper was generated by other researchers. Simply citing the paper that first described and analysed a data set fails to capture the connection to specific researchers or the data's importance to the scholarship. Granting authorship to the generators of data used in a subsequent paper may misstate researchers' contributions and responsibilities, especially if the data generators were not involved in the subsequent analysis. Some have suggested that granting authorship and utilizing the CRediT taxonomy, which defines 14 roles for specific author contributions, would fill this gap (see go.nature.com/2hz9mjb). However, that taxonomy, although useful in other ways, is not set up to link data sets with data generators or to track data reuse.

One way to bridge these gaps is to establish that the data themselves are a valuable scientific output, separate from their analysis in a published manuscript. As part of this effort, a dedicated community has worked to establish guidelines for citing research data. The aim is to encourage scientists to archive data and cite the data set itself when it serves as the basis for specific scientific conclusions, as well as when it is used in reanalyses by others[5]. A series of roadmaps[6–8] lay out how publishers and repositories should ensure that data are cited properly, and encourage making data sets findable, accessible, interoperable and reusable (FAIR)[9] (see page 27).

Non-profit organizations, such as DataCite and Crossref, already provide the technical services necessary for data citation. They enable data repositories to assign unique persistent identifiers (PIDs) to data sets and provide infrastructure to link those identifiers to publications. One important international initiative is Make Data Count, which aims to bring together both data citations and repository-usage metrics in an early effort to establish data as a valuable research output, separate from conventional publications.

What is still missing is an established way for researchers, institutions and funders to link useful data sets to the researchers responsible for them, and an understanding of why this link is crucial.

## PROPOSED SOLUTION

To connect the use of a data set with its originators, both data sets and individual researchers must have PIDs (see Supplementary Information, Data tracking process). Ideally, each individual scientist would obtain a unique ORCID identification number and associate that with every data set they deposit. Repositories would issue PIDs for the data sets (such as digital object identifiers, or DOIs) and connect those to one or more ORCID identifiers. Journals would require the data set PIDs to be cited in every submitted manuscript (both primary and subsequent analyses). That system would allow data generators, academic leaders, funders, scientometricians and others to track the data in searchable databases. The processes for generating and recording each of these PIDs has been well defined, but they are not yet connected.

Publishers and authors together must now take steps to ensure that the PID for data underlying a publication is always included as citation metadata, both for the original publication and any subsequent uses of the data set (see Supplementary Information, Action items). Publishers must also make sure that these citations are deposited in Crossref, a central, searchable and open system. If an article does not cite a data set PID that Crossref can recognize, the link between the data generator, the publication and the data set will break down. ▶

## WANTED

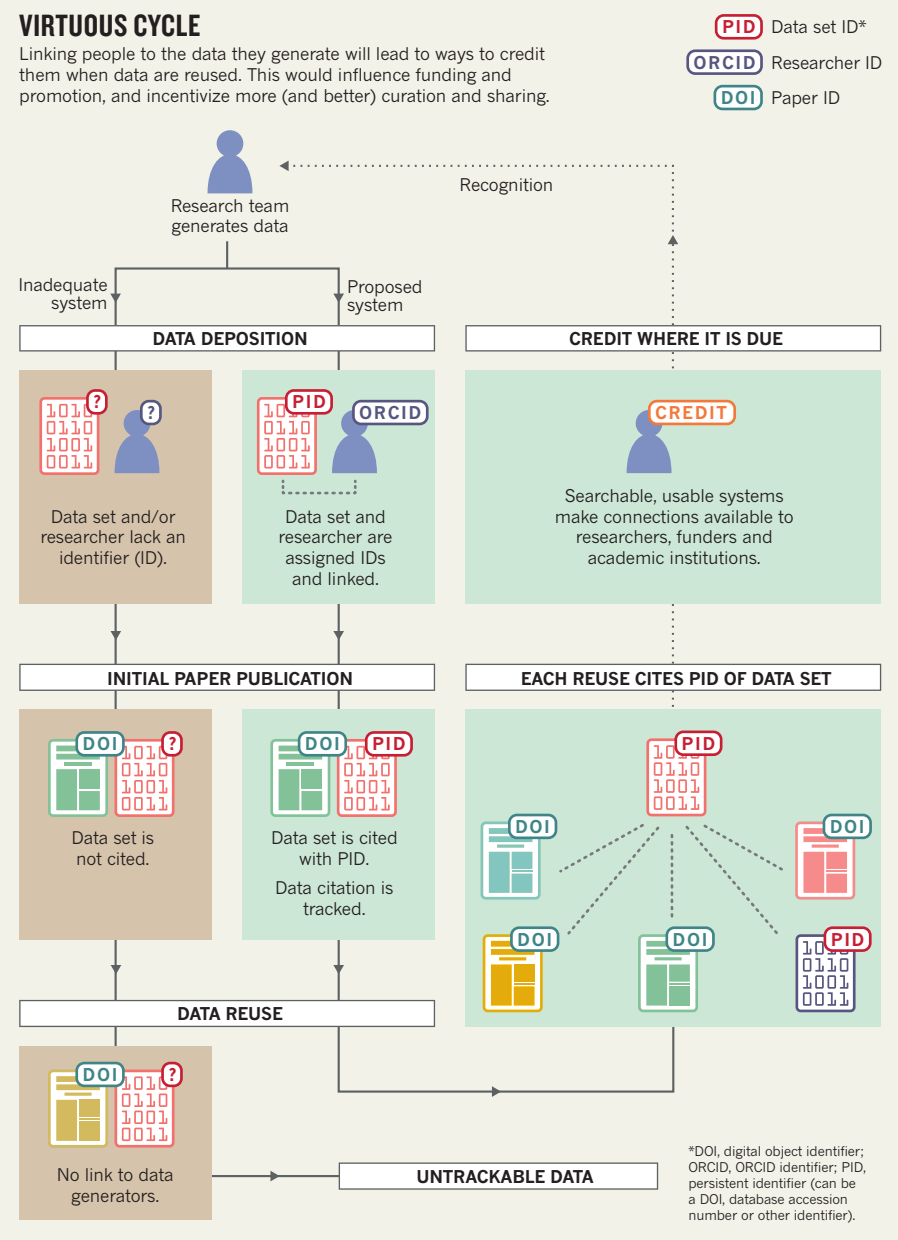
### *Data reuse metrics*

Hiring and promotion committees know that publications are easily described accomplishments, but that they are not the only ways in which researchers advance knowledge. The struggle is how to measure other outputs.

If a system linked data sets to individuals and reliably tracked the subsequent uses of those data, would institutions incorporate these metrics into the promotion process? "The answer is an unambiguous 'yes'," says Antony Rosen, vice-dean for research at Johns Hopkins School of Medicine in Baltimore, Maryland. "Having an objective method to assess the uses of data would give faculty additional ways to communicate the contributions of their work."

Northwestern University in Chicago, Illinois, implemented a 'team scientist' route for faculty members who make substantial contributions to research at the university's medical school. Among those explicitly recognized on this non-tenure track (which extends to full professor) are individuals whose work constitutes and enables important scholarship but might not always result in publications.

Still, without formal metrics, faculty members whose data are widely reused might find it challenging to convince a promotion committee of that value, especially when they are accustomed to considering the importance of publications. As translational researcher Kristi Holmes, director of the Galter Health Sciences Library at Northwestern University and champion of the team-scientist track, explains: "There is a critical need for clear processes to assess data impact." H.H.P. *et al.*




There is likely to be overlap between those credited with generating a data set and the authors of the first scientific paper based on it, but some authors will not be connected to the data set for each of its future uses. It is up to the research teams to decide who should be linked to a data set; those individuals must have had a substantive role in data collection and quality assurance, and must also take ongoing responsibility for its accuracy and integrity. If more than one data set underlies a single publication, different researchers could be tied to each. In some large, long-running studies, such as the Framingham Heart Study run by Boston University and the US National Heart, Lung, and Blood Institute, a data set might be more appropriately linked to an organization than to an individual.

Data sets could also be linked to funders or funding awards, institutions and other organizations through PIDs for those entities. Plans and pilot projects for doing so — including unique identifiers for grants and organizations — are already in place. Such information can also help funders and institutions to track and measure their contributions to science.

## TRICKY BUT POSSIBLE

This scale of change — requiring identifiers for each data set and data generator, linking individual researchers to data sets and ensuring researchers properly cite the data they use— is difficult, but not unprecedented. In 2005, the majority of clinical trials were not registered in any public database. That year, the ICMJE began requiring that any study accepted for publication must have been entered in a public clinical-trial registry before participants had been enrolled. Registration surged, and a similar requirement for US federally funded clinical trials was soon codified into law through the Food and Drug Administration Amendments Act of 2007.

A system in which researchers are regularly recognized for generating data that become useful to other researchers could transform how academic institutions evaluate faculty members' contributions to science (see 'Data reuse metrics'). A connection between data generators and the further analyses of the data sets they share could provide funders with new means for assessing the impact of their grant awards.

Tracking data reuse will also allow the scientific community to ask a variety of questions: how are shared data sets incorporated into new scholarly works? How often are reanalyses used to verify reproducibility instead of attempting to replicate experiments from scratch? How are new scientific questions answered through the reuse of existing data? Does data quality increase through making data sets available to others? Does accounting for data reuse expand the ways in which funders can determine the impact of their research awards?

Together, these steps will transform how academic institutions, funders and society can assess the value of scientific data. The benefits of data sharing require its reuse[10]. When we can measure how data sharing strengthens science, we can reward those contributors whose data drive research forward. ■

**Heather H. Pierce** *is senior director of science policy and regulatory counsel and* **Anurupa Dev** *is lead specialist at the Association of American Medical Colleges in Washington DC, USA.* **Emily Statham** *is project manager and* **Barbara E. Bierer** *is faculty director of the Multi-Regional Clinical Trials Center at Brigham and Women's Hospital and Harvard Medical School in Boston, Massachusetts, USA.*
*e-mail: hpierce@aamc.org*

1. Institute of Medicine. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk* (National Academies Press, 2015).
2. Bauchner, H., Golub, R. M. & Fontanarosa, P. B. *J. Am. Med. Assoc.* **315**, 1238–1240 (2016).
3. Holdren, J. P. *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research* (Office of Science and Technology Policy, 2013).
4. Bierer, B. E., Crosas, M. & Pierce, H. H. *N. Engl. J. Med.* **376**, 1684–1687 (2017).
5. Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles* (ed. Martone, M.) (FORCE11, 2014).
6. Fenner, M. *et al. Sci. Data* **6**, 28 (2019).
7. Cousijn, H. *et al. Sci. Data* **5**, 180259 (2018).
8. Wimalaratne, S. M. *et al. Sci. Data* **5**, 180029 (2018).
9. Wilkinson, M. D. *et al. Sci. Data* **3**, 160018 (2016).
10. Pasquetto, I. V., Randles, B. M. & Borgman, C. L. *Data Sci. J.* **16**, 8 (2017).

**Supplementary information accompanies this article: see go.nature.com/2hjxt5a.**

**CORRECTION**

The Comment ‘Credit data generators for data reuse’ (*Nature* **570**, 30–32; 2019) wrongly located Julie Dunning Hotopp at the University of Maryland in College Park; in fact, she is in Baltimore.