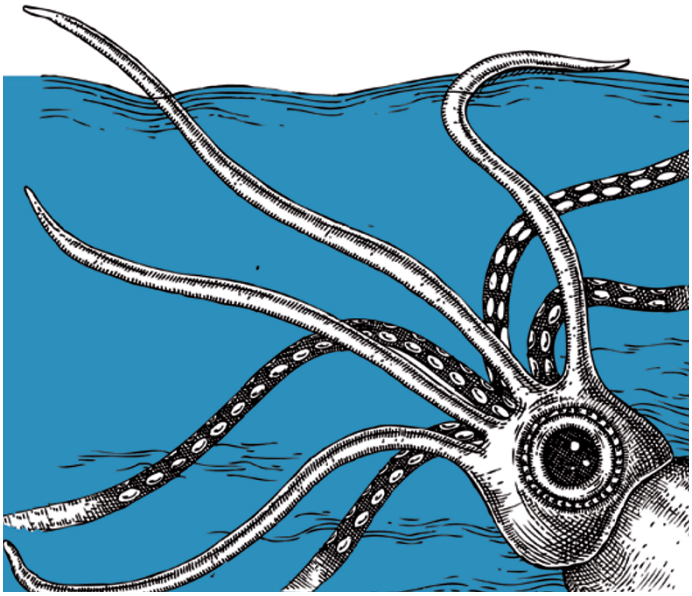
The reuse factor

The reference is not dead — it is exploding to encompass the full spectrum of research outputs from lines of code to video frames, explains **Mark Hahnel**.



PIKA/SHUTTERSTOCK

Researchers are still struggling to find, manipulate and cite research outputs other than published papers. Data-management plans for research — detailing what data will be created and how, and outlining plans for data sharing and preservation — are a core requisite of all grant applications for a long list of US and UK funding agencies. These include the US National Science Foundation, the US National Institutes of Health, NASA, the UK Biotechnology and Biological Sciences Research Council, the UK Medical Research Council and the Wellcome Trust. Funders require grantees to make all their products available in a citable, sharable and discoverable manner. As a result, tools such as the California Digital Library's Data Management Plan and the UK-based Digital Curation Centre's DMPonline have materialized to help optimistic grant applicants to fill in this section.

Even with the help of such tools, data-management plans are being rejected in some grant applications. This is a wake-up call. Researchers should be long past thinking that depositing their data in a file-hosting service such as Dropbox is sufficient. Yet the majority of academics still consider journal articles to be the only valid, formal record of their research — the main currency for credit.

In my view, the current model of grouping non-image files together in supplemental data appended to a paper is laughable. Data are the bedrock of a paper and come in myriad forms; for instance, videos for research on cell motility or biomechanics, or code for climate or epidemiology models. Scholars are increasingly sharing these sorts of raw data through repositories such as the Dryad Digital Repository and GenBank, which allow for the citation of data sets, videos, genetic sequences and other such files that publishers often struggle to accommodate.

In 2011, I set up another such company, figshare, to improve the way that the 'data behind the graph' is disseminated, exploiting visualization tools such as D3.js and Jmol[1]. At figshare, we work with research groups and publishers to make data reusable, reproducible and interactive. (Macmillan Science and Education, the publisher of this journal, is an investor in figshare.)

But the generation of huge numbers of citable research outputs is confusing researchers who are accustomed to citing only papers. The FORCE11 Amsterdam manifesto on data-citation principles, drawn up in 2011 by a community of scholars, librarians, archivists, publishers and research funders, states: "A data citation in a publication should resemble a bibliographic citation and be located in the publication's reference list." A quick look at the most recent journal citations of data held in figshare shows that this recommendation is not enforced by publishers or authors: only one in five cited the data in the reference list; the rest mentioned it in methods or deposition sections.

There are standard protocols for citing static data such as genomes or clinical-trial results[2]. But data is increasingly dynamic, coming from sensors that monitor, for example, geophysical and atmospheric changes in real time. Tracking these feeds in a way that is automated and machine-processable will lead to improved validation and verification, and help to prevent falsification of data that, in areas such as climate-change research, has led to much unnecessary wrangling[3].

Scientists should appreciate that making their research outputs citable enables more of their research to have quantifiable impact. To this end, the Research Data Alliance (RDA) was established in August 2012 by a steering group of funding agencies from the United States, Europe and Australia. The RDA aims to accelerate and facilitate research data sharing and exchange across multiple disciplines that have complicated funder mandates and a need to cite various unconventional research outputs. An RDA working group plans to provide prototypes and examples. For instance, individual cells of a spreadsheet or a few frames of a video can be cited in a way that does not dilute a paper's total number of citations.

Early adopters of open-data science are already seeing the benefits. Computer scientist Titus Brown at Michigan State University in East Lansing, for example, blogged, "My career has already been immeasurably improved by my openness, including posting our software." And the efficiency gains are long overdue.

I believe that publishers should mandate that all the research that goes into forming the conclusions of a paper be made openly available, when ethically possible. Even better would be for raw data to be made available in the paper, as *F1000 Research*, an open-access journal, and PLoS journals already do for their articles. Publishers should also ensure that all citations — to products of all kinds — are included in reference lists, and should make this bibliometric data openly available in a searchable format (see page 295).

Are we witnessing the death of the reference? No, we are seeing the birth of an exponentially larger number of citations, crediting a much wider variety of outputs. The end is nigh for the measuring of impact using only citations to published papers: this is the age of the 'reuse factor'. ■

Mark Hahnel *is founder of figshare.*
e-mail: mark@figshare.com

1. Cowley, M. J., Huch, V., Rzepa, H. S. & Scheschkewitz, D. *Nature Chem.* **5,** 876–879 (2013).
2. Altman, M. & King, G. 'A proposed standard for the scholarly citation of quantitative data' *D-Lib Magazine* (2007); available at http://go.nature.com/sv314e.
3. Heffernan, O. *Nature* **460,** 787 (2009).