TOOLBOX

THE TROUBLE WITH REFERENCE ROT

Computer scientists are trying to shore up broken links in the scholarly literature.

ILLUSTRATION BY THE PROJECT TWINS

BY JEFFREY M. PERKEL

The scholarly literature is meant to be a permanent record of science. So it is an embarrassing state of affairs that many of the web references in research papers are broken: click on them, and there's a fair chance they will point nowhere or to a site that may have altered since the paper referred to it.

Herbert Van de Sompel, an information scientist at the Los Alamos National Laboratory Research Library in New Mexico, quantified the alarming extent of this 'link rot' and 'content drift' (together, 'reference rot') in a paper published last December (M. Klein *et al. PLoS ONE* **9,** e115253; 2014). With a group of researchers under the auspices of the Hiberlink project (http://hiberlink.org), he analysed more than 1 million 'web-at-large' links (defined as those beginning with 'http://' that point to sites other than research articles) in some 3.5 million articles published between 1997 and 2012. The Hiberlink team found that in articles from 2012, 13% of hyperlinks in arXiv papers and 22% of hyperlinks in papers from Elsevier journals were rotten (the proportion rises in older articles), and overall some 75% of links were not cached on any Internet archiving site within two weeks of the article's publication date, meaning their content might no longer reflect the citing author's original intent — although the reader may not know this.

Hyperlinks to web-at-large content were present in only one-quarter of the 2012 scholarly articles, but some four-fifths of those papers that did contain a link suffered from reference rot, the team found — that is, at least one reference to web-at-large content was either dead or not archived. Van de Sompel terms the situation "rather dramatic". Because the content of servers can change, or they can 'go dark' or change hands, researchers following up links to online data sets, software or other resources might have nowhere to turn. "You've lost a trace to the evidence that was used in the research," he says.

SNAPSHOTS OF THE WEB

Fortunately, online archiving services, such as the Internet Archive, make it possible for researchers to store permanent copies of a web page as they see it when preparing their manuscripts — a practice Van de Sompel recommends. He urges researchers to include their cached link and its creation date in their manuscripts (or for publishers to take a snapshot of referenced material when articles are submitted). The Harvard Law School Library in Cambridge, Massachusetts, has developed a web-archiving service called Perma.cc (https://perma.cc): enter a hyperlink here and the site spits back a new hyperlink for a ▶

page that contains links to both the original web source and an archived version.

Van de Sompel and others have in the past few weeks rolled out a complementary approach. It relies on a service that Van de Sompel has co-developed called Memento, which he dubs "time travel for the web". The Memento infrastructure provides a single interface for myriad online archives, allowing users access to all of the saved versions of a given web page. This infrastructure could potentially allow access to web-at-large links in any scholarly article, even if the linked sites go down. Publishers would have to incorporate a small piece of extra computer code in their articles, and the standard single weblinks would have to be replaced with three pieces of information — the live link, a cached link and its creation date — all wrapped in Van de Sompel's proposed machine-readable tags.

STORAGE BLOCK

Van de Sompel says that he is "unbelievably enthusiastic" about the team's approach. But the solution depends on the cooperation of authors and publishers — who may be disinclined to help. Another issue is that web-page owners who hold copyright over content can demand that archives remove copies of it. They can also disallow archiving of their sites by including a file or line of code that prevents computer programs from 'crawling' over or capturing content — and many do. If Perma.cc, for instance, encounters such an exclusion code, it preserves the content in a 'dark archive'; to access a web page in a dark archive, the reader must contact a library participating in the Perma.cc project and request to see the site.

Scholarly articles that are behind a paywall routinely exclude such crawling, too — although publishers have introduced the DOI system to ensure that scientists can confidently cite a persistent hyperlink to the right version of an online research article, even if the publisher changes its local web addresses. (In January, however, the system that redirects DOI links went down, showing that it is not immune to failure.) Publishing companies also guard against link rot by automatically preserving articles in archives; the articles can be released if the company folds.

But not all companies are archiving, says David Rosenthal, a staff member at the library of Stanford University in California; analysis of data from a monitoring service called The Keepers Registry shows that "at most 50% of articles are preserved", Rosenthal writes on his blog (go.nature.com/jrwqo4). So for both web-at-large hyperlinks and scholarly articles, the Memento team's mission to solve reference rot may be "excessively optimistic", he says. ■

Jeffrey M. Perkel *is a writer based in Pocatello, Idaho.*

PUBLISHING

'Living figures' make their debut

Published chart integrates data from outside scientists.

BY DALMEET SINGH CHAWLA

In July last year, neurobiologist Björn Brembs published a paper about how fruit flies walk. Nine months on, his paper looks different: another group has fed its data into the article, altering one of the figures.

The update — to figure 4 — marks the debut of what the paper's London-based publisher, Faculty of 1000 (F1000), is calling a living figure, a concept that it hopes will catch on in other articles.

Brembs, at the University of Regensburg in Germany, says that three other groups have so far agreed to add their data, using software he wrote that automatically redraws the figure as new data come in.

His article, written with Julien Colomb, chief executive of the start-up firm Drosophare in Berlin, finds behavioural differences within a strain of fruit fly: the Canton Special, or CS strain (J. Colomb and B. Brembs *F1000Research* **3,** 176; 2014). Although there are substrains, researchers usually regard CS flies as so similar that they do not distinguish between the substrains in their analyses, but Brembs and Colomb report that the flies exhibit three types of walking behaviour. This might betoken other differences in behaviour and therefore confound experiments in which CS flies are used as a control group, he says.

Having sequenced the genomes of the flies, Brembs thinks that the behaviours have a genetic origin and will not be explained away by environmental variations between labs. The addition of data by other labs could help to test whether his theory is correct.

ITERATIVE PUBLISHING

The living figure concept fits within a central tenet of F1000's publishing philosophy, that papers can be continually updated. The online-only open-access site publishes articles immediately with the status 'Awaiting Peer Review', then invites scientists to review them. Authors can then update their articles with new versions. The process is like adding pieces of paper to the top of an existing pile, the publisher says.

Allowing outside researchers to post their data into a paper simply takes the idea a step further, says Rebecca Lawrence, managing director of the publishing platform *F1000Research*. "The idea is that it better mirrors the way science is conducted," she says. Other laboratories' information confirms or challenges the published research in an incremental process. In addition to updating work, living figures may allow systematic reviews to be updated rather than published afresh each time, Lawrence adds. They should also help to address the issue of lack of reproducibility, she argues, because it provides a way for laboratories to release confirmatory data, which can be hard to get published.

"It's a more accessible way for scientists to get the answer."

Of course, by adding data to someone else's article, scientists are giving up the chance to publish a paper of their own — a potential hurdle, because publications are the lifeblood of reputations in academia. But Gregg Roman from the University of Houston, Texas, the first outside author to add data to Brembs's paper after publication, says that he accepts that. "We're sacrificing a bit of recognition," he says, but "it's a more accessible way for scientists to get the answer than if we publish separately".

New contributors' names do, however, appear in the legend of updated figures; and the updated data set and paper get their own DOIs. Alternatively, contributors can choose to gain a formal publication by submitting what *F1000Research* calls a Data Note that links to the original updated paper.

If the new contributors' methods or results differ significantly from the original paper's, then they can publish a Research Note, Lawrence says. They can also request that the original authors update their article. An updated paper would be peer reviewed again.

Lawrence says that many research groups have shown interest in publishing living figures. And the concept could work with traditional pre-publication review, too, notes Peter Binfield, co-founder of the open-access journal *PeerJ*. "As long as the full version history of the article is available, and it's clear which version of the article was reviewed, and in what way, it should be possible to publish updates," he says.

As for Brembs's work, Roman says that his data seem to support the general trend, but with smaller differences between the flies. The question may be resolved only as figure 4 evolves. ■