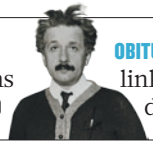
COMMENT

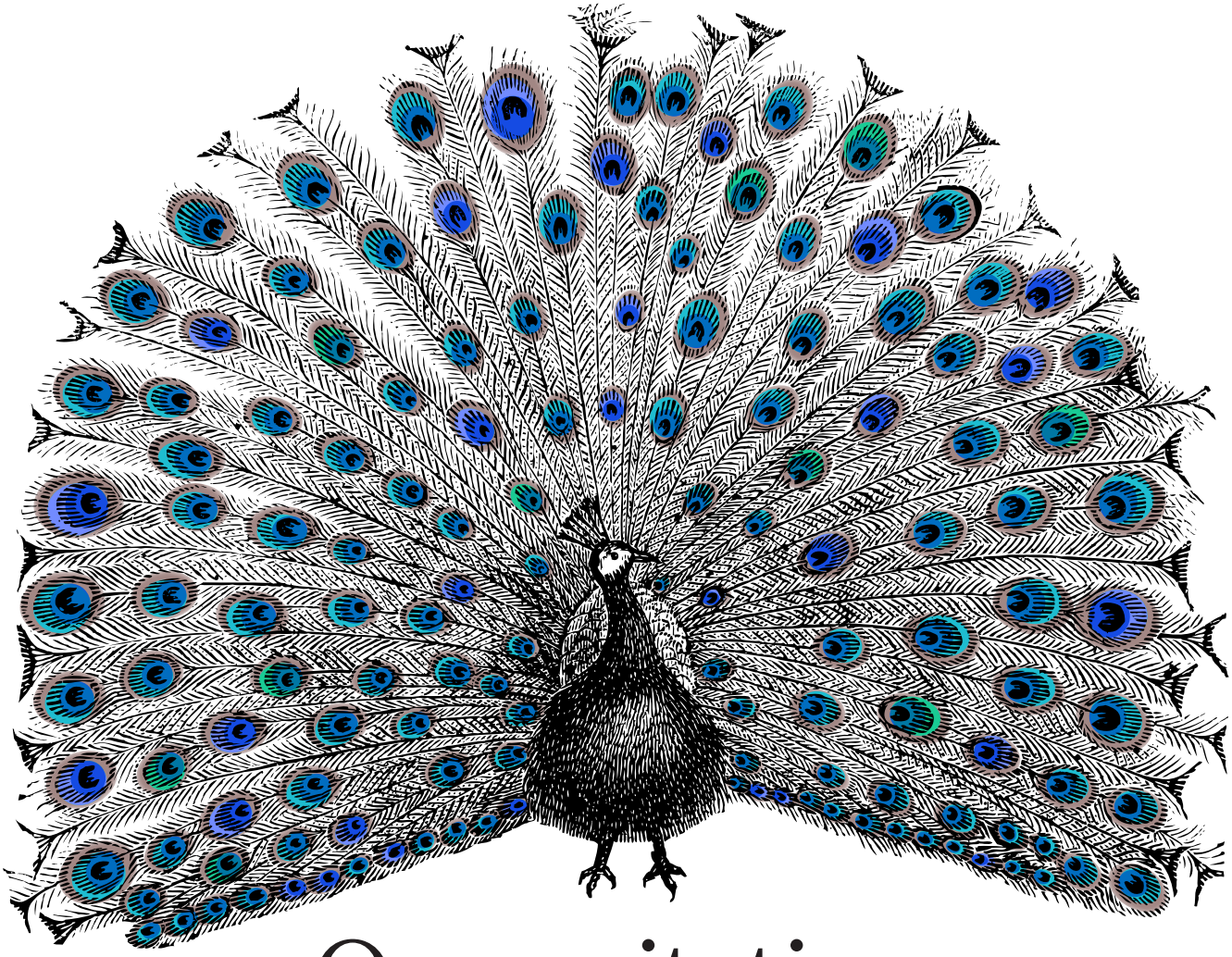
IMPACT Make data findable, shareable and citable urges Mark Hahnel **p.298**

TECHNOLOGY What could make our ever-expanding cities smarter? **p.299**

PHYSICS A celebration of Einstein's contributions to quantum theory **p.300**



OBITUARY David Barker, who linked early life with chronic disease, remembered **p.304**



LYNEA/SHUTTERSTOCK

Open citations

Make bibliographic citation data freely available and substantial benefits will flow, says **David Shotton**, director of the Open Citations Corpus.

When Heather Piwowar set out in May last year to investigate whether making research data publicly available increased the citation rates of articles[1], she never anticipated the difficulties. Piwowar, co-founder of ImpactStory[2], and who is based in Vancouver, Canada, was at the time a postdoc at Duke University in North Carolina. Lacking institutional access to Scopus, Elsevier's database of scholarly citations, she eventually obtained access through a research-worker agreement with Canada's National Science Library. But this required her to be fingerprinted to obtain a police clearance certificate because she had lived in the United States. "I wasted days trying to access the citation data required for my study," she told me. "It was just ridiculous." Piwowar needed to analyse citation counts for 10,000 articles, but the other major citation source, the Thomson Reuters Web of Science, did not at the time support queries using PubMed's unique identifier numbers. She explains: "Had there been open citation data, I could have written my own script!" ▶

IMPACT
A *Nature* special issue
nature.com/impact

Steven Greenberg, a neurologist at Harvard Medical School in Boston, Massachusetts, had a similar experience when he set about revealing how hypotheses can be converted into 'facts' simply by repeated citation[3]. Greenberg had manually to construct and analyse a citation network that contained 242 papers, 675 citations and 220,553 distinct citation paths that were relevant to a particular hypothesis. Had those citation data been readily accessible online, he would have been saved considerable effort. Research practice suffers because access to citation data is currently so difficult.

In this open-access age, it is a scandal that reference lists from journal articles — core elements of scholarly communication that permit the attribution of credit and integrate our independent research endeavours — are not readily and freely available for use by all scholars.

To rectify this, citation data now need to be recognized as a part of the commons — those works that are freely and legally available for sharing — and placed in an open repository. To that end, since 2010 I have led a project funded by two small grants totalling £132,000 (US$212,000) from Jisc (www.jisc.ac.uk), a UK information technology research and development funding organization, to establish and develop the Open Citations Corpus (OCC). The OCC is a fledgling repository for open scholarly citation data that is now seeking sustainable funding to become a cornerstone of the digital research infrastructure that supports the academic enterprise.

CLOSED SHOP

Although alternative metrics for impact and esteem are being developed[4], direct citation remains a keystone indicator of the significance of an output (see page 298). Scholarly communication involves the flow of information and ideas through the citation network, and analysis of changes in the network over time can reveal patterns of communication between scholars and the development and demise of academic disciplines. Such information is central to scholarly endeavour. It is also fundamental to good decision-making about research investment and strategy, to facilitate innovation, and to promote growth and prosperity, particularly in light of the increasingly international nature of research collaborations[5].

The most authoritative sources of scholarly citation data are the Thomson Reuters Web of Science, which grew from the Science Citation Index created by US scientist Eugene Garfield in 1964, and which was originally published by the Institute for Scientific Information (ISI); and its main commercial rival, Elsevier's Scopus, released in 2004. Both have wide coverage of the leading literature, but because neither is complete, they are widely regarded as complementary[6].

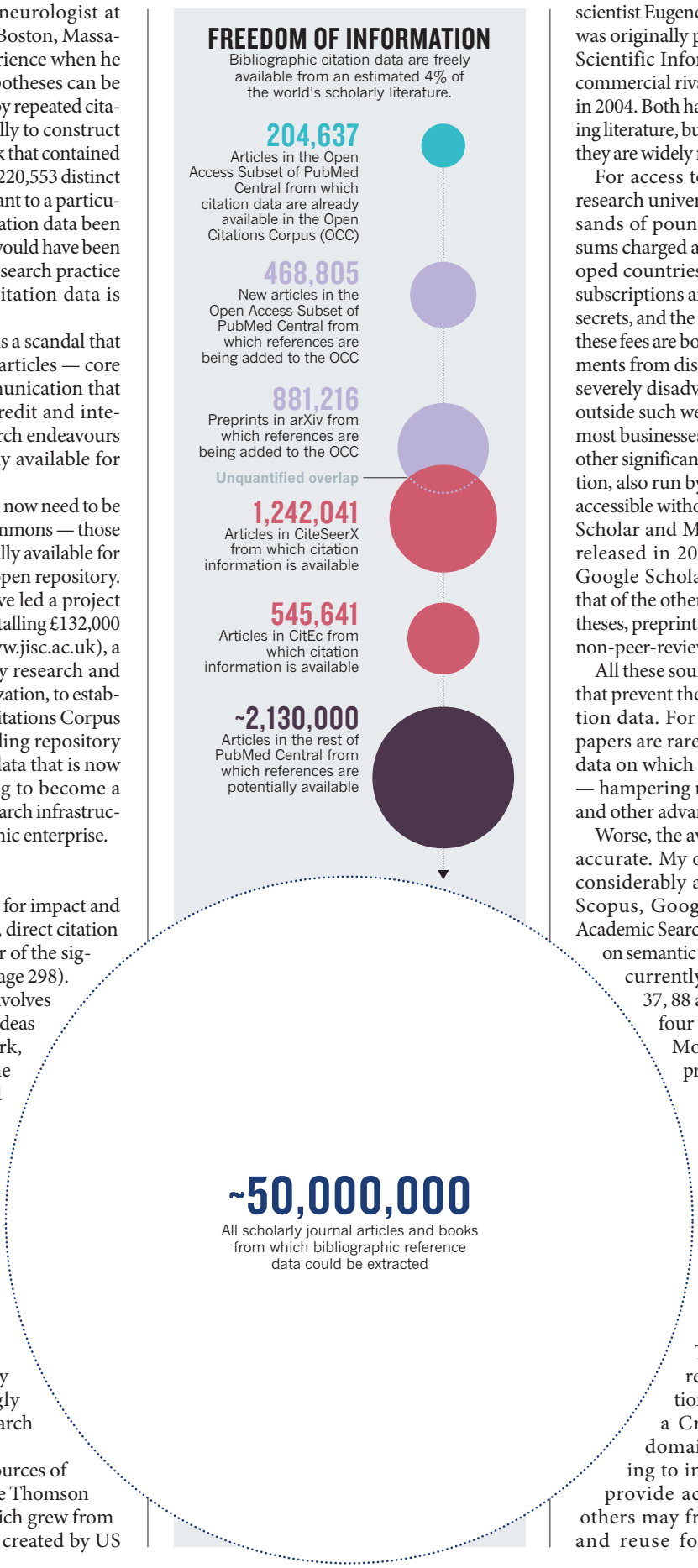
For access to these two resources, UK research universities each pay tens of thousands of pounds a year[6], with equivalent sums charged at institutions in other developed countries. The exact values of these subscriptions are closely guarded industrial secrets, and the university librarians who pay these fees are bound by confidentiality agreements from disclosing them. This high cost severely disadvantages all those who work outside such wealthy institutions, including most businesses and the general public. The other significant sources of citation information, also run by commercial companies but accessible without subscriptions, are Google Scholar and Microsoft Academic Search, released in 2004 and 2009, respectively. Google Scholar's coverage is wider than that of the others, because it includes books, theses, preprints, technical reports and other non-peer-reviewed 'grey' literature.

All these sources have licence restrictions that prevent the re-publication of their citation data. For this reason, bibliometrics papers are rarely permitted to publish the data on which their conclusions are based — hampering reuse, validation of findings and other advantages of open data.

Worse, the available citation data are not accurate. My own citation record differs considerably across the Web of Science, Scopus, Google Scholar and Microsoft Academic Search. For example, a 2009 paper[7] on semantic publishing that I co-authored currently has citation counts of 22, 37, 88 and 16, respectively, in these four databases. Which to trust? More worryingly, an earlier protein-crystallography paper[8] has three separate entries in the Web of Science, with citation counts of 59, 19 and 0, respectively. In my view, this calls into question the reliability of the Thomson Reuters Impact Factor, which is based on such counts.

A SOLUTION

The OCC, as an open repository of scholarly citation data made available under a Creative Commons public domain dedication, is attempting to improve matters. It aims to provide accurate citation data that others may freely build upon, enhance and reuse for any purpose, without



restriction under copyright or database law.

We began building the OCC in mid-2010, and released the first version in mid-2011. This prototype provided open access to reference lists from the 204,637 articles that then comprised the Open Access Subset of PubMed Central (OA–PMC), containing 6,325,178 individual references to 3,373,961 unique papers. Despite its small size, this corpus contains references to about 20% of all the biomedical literature indexed in PubMed that had been published between 1950 and 2010, including all highly cited papers in every biomedical field. Available at http://opencitations.net, the OCC is structured to enable the information to be easily integrated with similar information from elsewhere — the data are encoded as Linked Open Data using the SPAR (Semantic Publishing and Referencing) Ontologies[9] and the latest Semantic Web standards.

Other open citations resources exist. The two main ones are CiteSeerX (citeseerx.ist.psu.edu), which contains around 13,500,000 references from 1,242,041 articles, primarily in computer science; and CitEc (Citations in Economics; citec.repec.org), which contains 13,544,970 references from 545,641 documents. Together, these resources and the OCC have the references from some 1,980,000 articles — a mere 4% of the estimated 50 million articles that have been published (see 'Freedom of information').

We are currently revising the OCC data model, improving its hosting infrastructure and expanding its coverage, both by updating the OA–PMC holdings, which have more than doubled since the initial ingest to 672,442 articles, and by ingesting citation data from the 881,216 preprints in the arXiv server, thus adding citations in mathematics and the 'hard' sciences to augment the initial biomedical coverage. Future work will include integration with CiteSeerX, harvesting dataset-to-article references from the Dryad Digital Repository, and extracting references from the pre-digital 'legacy' literature that is poorly represented in other citation repositories. This applies particularly to fields in which such literature is both well organized and of enduring value — notably astronomy, and biodiversity and biological taxonomy.

Ideally, references will come directly from publishers at the time of article publication. Most publishers are sympathetic to the idea of putting article reference lists outside the journal-subscription paywall, as they do copyrighted abstracts. We already have agreements with several major journal publishers for the future routine harvesting of reference data. As well as the 'pure' open-access publishers, the references from which are open by definition, the publishers of subscription-access journals include Nature Publishing Group (NPG), Oxford University Press, the American Association for the Advancement of Science (which publishes *Science*), Royal Society Publishing, Portland Press, MIT Press and Taylor & Francis, all of which will make references available either from some or from all of their journals. This represents a small but growing proportion of all the journal articles published in a year.

References will be harvested centrally from CrossRef, the organization that provides digital object identifiers (DOI) for journal articles, to which these publishers already submit article reference lists as participants in its CitedBy Linking service. However, publishers need to indicate their consent in the article metadata for the article's references to be made open (see go.nature.com/x4pzta), because by default references are kept private. No other action is required; it is straightforward and free.

"Ideally, references will come directly from publishers at the time of article publication."

The long-term aim of the OCC is to host citation information for most of the world's scholarly literature, in the arts and humanities as well as the sciences. This will require a major curatorial effort and underpinning technical innovation, on the scale of PubMed, which is run by the US National Library of Medicine.

OPEN SEASON

In an ideal world, publishers would host their own bibliographic and citation data, following the example of NPG (publishers of this journal) — the first and currently only company to make such information available as Linked Open Data, at data.nature.com.

But there are separate benefits to be gained from the aggregation of such data into a single corpus. The OCC will provide integrated access to citation data from a variety of sources, inside and outside traditional scholarly publishing, with clear provenance data. It will expose entity relationships, including article-to-article, article-to-database and database-to-article citations, and will reveal shared authorship and institutional membership, common funding, and semantic relationships between articles, where the data are available.

Once citation data are openly available, useful analytical services can be built, including faceted search-and-browse tools, recommendation and trend identification services, and timeline visualization. Some of these we have already developed in prototype. The OCC's usefulness for calculating citation metrics will, of course, increase in proportion to its expanding coverage.

There is one other service that we think could be of particular benefit to authors and editors — an erroneous reference correction service. About 1% of references in published papers contain errors of varying severity, ranging from the trivial — for example, substitution of 'beta amylase' for 'β-amylase' in the reference title, or the omission of accents in author names — to the more serious, such as errors in the year, volume, page numbers or DOI. The OCC already uses citation-correction methods internally for reference targets that are multiply cited, or for which authoritative bibliographic records can be obtained externally. A similar Web service that could detect errors in uploaded reference lists might significantly reduce the number of mistakes in published papers.

HELP US

So what next? Just over a decade ago, a similar aim for open citation data was held by the Open Citation Project (opcit.eprints.org), a collaboration between Southampton University, UK; Cornell University in Ithaca, New York; and arXiv, that ran between 1999 and 2002. That project developed Citebase, a database of citation information, which its developers described as "the crown jewel of the Open Citation Project". Following the link to citebase.eprints.org today, one gets the message "No website currently exists at this URL."

Making the transition from a promising academic project to a robust sustainable global service is extremely difficult. For the OCC to avoid the fate of Citebase, and instead grow into a comprehensive and trustworthy source of well-curated open citation data serving the entire scholarly community across all disciplines, it requires champions, managers, developers and curators. It also needs genuine collaborations with similar endeavours, a sustained and sizeable income stream from funders, supporters and investors committed to achieving a social good rather than a financial return, direct support from the publishing community, and adoption by a major institution or international organization. Can you help? ■

David Shotton *is director of the Open Citations Corpus and a senior research fellow in the Oxford e-Research Centre, University of Oxford, UK.*
e-mail: david.shotton@oerc.ox.ac.uk

1. Piwowar, H. A. & Vision, T. J. *PeerJ* **1,** e175 (2013).
2. Piwowar, H. *Nature* **493,** 159 (2013).
3. Greenberg, S. A. *Br. Med. J.* **339,** b2680 (2009).
4. Priem, J. *Nature* **495,** 437–440 (2013).
5. Adams, J. *Nature* **490,** 335–336 (2012).
6. Chadegani, A. A. *et al. Asian Social Sci.* **9,** 18–26 (2013).
7. Shotton, D., Portwin, K., Klyne, G. & Miles, A. *PLoS Comput. Biol.* **5,** e1000361 (2009).
8. Shotton, D. M., White, N. J. & Watson, H. C. *Cold Spring Harb. Symp. Quant. Biol.* **36,** 91–105 (1972).
9. Peroni, S. & Shotton, D. *Web Semant.* **17,** 33–34 (2012).