Round table discussion

Publishing research data on a professional basis

Toby Green[a]

As Pisani & AbouZahr have identified, there are many obstacles to the publishing of data: social (incentives for researchers to make the effort to publish), financial (having adequate financing to cover short-term publishing and long-term curation costs), and technical (standards and systems).[1] This paper looks at some of the technical challenges of publishing data professionally and describes the discoverability and citability benefits that follow.

Let's take it as read that publishing research data is a "good thing," that researchers are as willing to publish data as they are research papers and funding is in place to make them available online in the long run. Job done? Well, no, not by a long chalk.

Just as loading a journal article onto a web site somewhere isn't the same as publishing it properly, so the same is true for data. To be as discoverable and as citable as research articles, data sets need to be published using an infrastructure that is compatible with research articles. It is not enough that data sets hang like dongles off a research article; they need to be discoverable and citable in their own right – just like a journal article. This means the metadata must be compatible with existing bibliographic management and citation systems like Ref Works® and CrossRef®. Users will expect search engines, abstracting and indexing services and library catalogues to reference data sets, so, for example, librarians will need MARC (MAchine-Readable Cataloging) records.

Is this overkill? Well, the Organisation for Economic Co-operation and Development (OECD) doesn't think so. OECD publishes more than 390 data sets as stand-alone objects, as well as thousands of data sets as supplemental data to its books and journal articles. Sub-sets of the data sets are also posted on the web as stand-alone objects too. So it is no surprise that, in the absence of good discovery metadata and systems, the number one complaint from users is the challenge of finding a relevant data set. They know the data is there, but they can't find it – even with Google's help.

To solve this problem, OECD's Publishing Division has spent the past three years grappling with the challenge of how to publish these many thousands of data objects so that users can not only find the data they need, but can then cite and manage the data sets using the same tools that they already use to manage journal articles or book chapters. The first result was a white paper,[2] first released in March 2009, which described this challenge and proposed a set of metadata schema for databases in their own right, sub-sets of databases and supplemental data.

More significantly, was the launch of OECD iLibrary, OECD's new publishing platform, in July 2009. OECD iLibrary[3] hosts all OECD books, working papers, journals and data sets in a seamless manner. OECD iLibrary puts the white paper's proposed bibliographic schema for data objects into practice. Search for "health data" and the search results include data sets, book chapters – even individual tables found inside books.

[a] Organisation for Economic Co-operation and Development, 2 rue André Pascal, 75775 Paris Cedex 16, France.
Correspondence to Toby Green (e-mail: toby.green@oecd.org).

OECD's data sets can now be discovered more easily and they can be cited as simply and as easily as a research article using the downloadable citation provided. Later in 2010, librarians will be supplied with MARC records and the bibliographic records for OECD data sets will be shared with discovery platforms like RePEc (Research Papers in Economics)[4] – the world's largest collection of economics grey literature – enabling visitors to find data objects alongside working papers and journal articles. Imagine being able to discover and cite data sets as easily and as simply as journal articles. Imagine no more. ∎

Competing interests: None declared.

References

1. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ* 2010;88:462–6.
2. Green T. *We need publishing standards for datasets and data tables.* Paris: Organisation for Economic Co-operation and Development; 2010. Available from: http://dx.doi.org/10.1787/603233448430 [accessed 1 March 2010].
3. OECD iLibrary [Internet site]. Paris: Organisation for Economic Co-operation and Development; 2009. Available from: www.oecd-ilibrary.org [accessed 19 May 2010].
4. RePEc (Research Papers in Economics) [Internet site]. Available from: www.repec.org [accessed 1 March 2010].