# Research datasets need to be easy to find if they are to achieve their potential impact.

blogs.lse.ac.uk/impactofsocialsciences/2014/03/18/research-datasets-potential-impact/

The impact of research data is now of definitive importance for universities, funders and disciplines themselves. Similar to the wave of preprint repositories established for journal articles that helped preserve access to papers in disciplines not otherwise catered for, we are now seeing a steady stream of institutional data repositories emerging. **Alex Ball** *provides the background for a Jisc project looking to establish a national data registry to make finding relevant data across disciplines and repositories quicker and easier.*
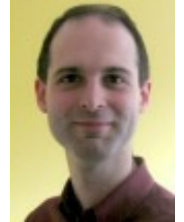
In 2012, an article in the Harvard Business Review proclaimed that the sexiest job of the 21st Century was that of data scientist. It had a point. There is something compelling about the idea of mashing together vast quantities of data and pulling out new and unexpected insights. Indeed, the late Jim Gray hailed it as the Fourth Paradigm of science.

Meanwhile, over the past decade or so, a movement has been growing to restore the place of research data in academic discourse. There are lots of arguments driving this. One is that if public money paid for it, the public should have access to it. Perhaps more striking is the notion of trust. In 2011, Diederik Stapel was found to have fabricated the data underlying 30 peer-reviewed papers in clinical psychology, and he was by no means the first. No wonder that the Royal Society, in its report Science as an Open Enterprise, argued that reporting conclusions without making the underlying data open to scrutiny was tantamount to malpractice.

These ideas are coming together: as funders require more data to be shared, and as the perception of secondary data usage transitions from 'second class' to 'sexy', the next piece of the puzzle to fall into place is scholarly credit. REF 2014 explicitly allowed datasets to be submitted as evidence of research quality. Initiatives such as DataCite are making it easier to cite datasets as first class research outputs, while Thomson Reuter's Data Citation Index makes such citations easier to measure. The impact of research data is starting to matter as never before.

Many researchers are already well catered for, in terms of data sharing infrastructure. In the UK we have data centres specialising in environmental, space science, social science and humanities data, and internationally there is a system of World Data Centres with various specialities. But for others there is little or no support available, so how do they fulfil their data sharing requirements? And how can they benefit from opening their data to greater impact? Well, the EPSRC Expectations regarding research data are that institutions should fill the gap.

Image credit: Franck BLAIS, Flickr (CC BY-SA)

In many ways what we are seeing happen with research data parallels what happened a decade or so ago with journal paper preprints. There were already subject-based preprint archives such as arXiv and Cogprints working around certain problems with the traditional publication model. When journal subscription prices (unrelatedly) began to soar, though, we saw a wave of institutions setting up their own preprint repositories. This helped preserve access to papers in disciplines not otherwise catered for, and drove up their potential impact.

After a few pioneer efforts, we are now seeing a steady stream of institutional data repositories appearing. But in terms of impact, data repositories are at a disadvantage compared to their preprint counterparts. There are mature

systems in place to ensure the visibility of journal papers – reference lists, abstract and indexing services, current awareness alerts – to which preprint repositories can act as a sort of access backup via search engines like Google Scholar and Microsoft Academic Search. For data, though, such things are in their infancy, and the best way of finding data of interest is to go to a data archive and search for it. This is fine if you are working in a field with a national data centre or an international data portal. It is not so fine if you aren't, or if you're working across disciplinary boundaries. That's a lot of data repositories you have to check.

In recognition of this, Jisc is looking at setting up a national data registry. This is not, I have to stress, a giant data repository. It will not hold research data. Rather, it is a sort of union catalogue that would allow you to search the holdings of the UK's data centres and repositories all in one go. This means that researchers would be able to deposit their dataset wherever it would be best looked after, without having to worry whether anyone would think to look for it there.

There are other potential uses for the registry. If a funder like the EPSRC wants to keep track of the data outputs from the research it funds, it would save a lot of time all round if they could get the information they need from the registry, rather than having to process a lot of forms. And when REF 2020 comes around, university administrators might find the registry helpful for tracking down data outputs that have been deposited outside the institution's own repository.

At the moment, the registry is still in its pilot phase. We (the Digital Curation Centre and UK Data Archive) are using the software developed for Research Data Australia to test the waters here in the UK, but there's a long way to go before the vision is realised and the final product may look very different. We're presenting our progress at the 2014 Jisc Digital Festival so if you're interested it would be good to see you there. Otherwise you can keep up to date with the latest news from the Jisc Research Data Registry project page.

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Alex Ball works for the UK Digital Curation Centre, and is based in UKOLN Informatics at the University of Bath. He has written guidance on a wide range of digital curation and data management issues, including data citation and curating CAD models. He is co-moderator of the Dublin Core Science and Metadata Community and co-chair of the Research Data Alliance Metadata Standards Directory Working Group.