


Data Descriptors: Providing the necessary information to make data open, discoverable and reusable.

 blogs.lse.ac.uk/impactofsocialsciences/2014/10/22/data-descriptors-open-discoverable-nature/

10/22/2014

Data need to be more than just available, they need to be discoverable and understandable. [Iain Hrynaszkiwicz](#) introduces Nature's new published data paper format, a Data Descriptor. Peer-review and curation of these data papers will facilitate open access to knowledge and interdisciplinary research, pushing the boundaries of discovery. Some of the most tangible benefits of open data stem from social and interdisciplinary sciences as these fields require effective cross-disciplinary communication.



Dengue fever (officially, human dengue virus infection) can cause headaches, pain behind the eyes, nausea, vomiting, and kills thousands. The World Health Organization says that more than 2.5 billion people – 40% of the world's population – are now at risk from dengue. It is spread by mosquitos and there is no vaccine for the virus. Reporting of cases is of inconsistent quality, and can be biased by difficulties in diagnosis, limited resources for diagnostic testing, and the varying reporting capacities of national health systems.

The virus is a global public health challenge which Dr Simon Hay and his team sought to battle – with data. They have collected the largest database of (8,309) human dengue virus occurrences, derived from peer-reviewed literature and case reports as well as informal online sources, with entries dating from 1960 to 2012. To make this data more easily reusable, they published a [Data Descriptor](#) in Nature Publishing Group's new journal *Scientific Data*, which describes all data collection processes in full, as well as geo-positioning, database management and quality-control procedures.

A Data Descriptor is a peer-reviewed article that describes and links to scientifically valuable datasets. It is citable (so researchers get more credit for their work) and is designed to make datasets more discoverable, interpretable and reusable. A Data Descriptor provides information needed for interpretation; links through to one or more trusted data resources where data files, code or workflows are stored; fulfils a significant part of funders' data management requirements; and uses open licenses that enable reuse. Many publishers, including offerings from Springer/BioMed Central, Elsevier, Wiley, Faculty of 1000 and Ubiquity Press, offer some form of data-driven publication – sometimes called data papers or data notes, rather than Data Descriptors, but all largely with similar aims of increasing the visibility of datasets in the peer-reviewed literature.

While it is possible to meet some community and funder requirements and expectations to share data by simply depositing data in repositories, to better enable publication of more reproducible and reliable research, data needs to be more than just available. Data need to be published with an open license; they need to be discoverable; they need to be understandable, and have indicators of quality. These last two points require particular effort – a peer-review and curation process focused on the data, not just the accompanying narrative – and this is a key role of data journals and Data Descriptors.

Open access to knowledge (data and papers) inherently facilitates interdisciplinary research and pushes the boundaries of discovery. The publication of Data Descriptors is still not standard in the sciences, but is growing, and at [Scientific Data](#) – and [Palgrave Communications](#) – we are seeking to engage more with interdisciplinary researchers and the quantitative social sciences.

Publication of Data Descriptors fits with the concept “intelligently open data” (a phrase coined in the Royal Society's [Science as an Open Enterprise](#) report), as the focus is on reuse as well as availability. And one could argue that some of the most tangible benefits of open data, for the wider public and community, stem from social and interdisciplinary sciences: data on social or demographic associations with poverty; on access to and effectiveness

of healthcare; on energy usage; on environmental/ecosystem changes; and economic analyses which can inform many far-reaching government policies. These examples are inherently cross-disciplinary.

One could envisage the dengue fever dataset gathered by Dr Hay's team being combined with other social science data to explore possible links between social factors and spread of disease – which would be facilitated by the ease with which it can be discovered, reused, integrated with other data, and understood. It could, also, then be used with more confidence because of the journal's peer-review and curation process.

Patrick Dunleavy recently pointed out on [this blog](#) how big data are erasing boundaries between physical and social sciences. Nik Rose at KCL has also launched the [Urban Brain Lab](#) which looks at the relations between sociological and neurobiological sciences. Neuro-imaging is an area that is showing increasing need for, and interest in, data sharing (see for example [this](#) functional magnetic resonance (fMRI) dataset in *Scientific Data*) and collaborative publication. All areas of science should in principle benefit from utilising Data Descriptors, although they are likely particularly relevant to fields generating large numbers of datasets. Researchers' ability to generate data outstrips our ability to analyse, report, and discuss these analyses in traditional research papers – and Data Descriptors can, also, provide an outlet for datasets without previous traditional research papers associated with them.

There is a lot that can be gained through more interdisciplinary knowledge sharing. For example, in data sharing best practice, it is often the biomedical areas that are perceived to be leading the debate. But many life and medical researchers could learn a lot of good practice for data management and sharing from social scientists – for example, the [UK Data Service](#) and archive of social science data is an excellent resource with established practices for handling confidential/personally identifiable information. Similarly, the [Harvard Dataverse](#) project was born out of the social sciences (the Institute for Quantitative Social Sciences.) For all these reasons, we are promoting interdisciplinary research in open access forums through *Palgrave Communications*, and *Scientific Data*. There is intrinsic value in open access to articles and open data, but we should remember recognised that they are means rather than the end – a means to conduct and communicate better research and discovery.

On November 14, we are hosting a half-day conference on [Publishing Better Science through Better Data](#), and one of our goals for that conference will be to promote interdisciplinary sharing of good practice. The event sold out rapidly, but if you can't join us in person then you can follow the debate on social media using #scidata14.

Note: This article gives the views of the authors, 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

Iain Hrynaszkiewicz is Head of Data and HSS Publishing and Nature Publishing Group and Palgrave Macmillan, where his responsibilities include developing new areas of open research publishing and data policy, such as [Scientific Data](#), [Palgrave Communications](#), and open access monographs. Iain previously worked at Faculty of 1000 and BioMed Central, and has led various initiatives and published various articles related to data sharing, open access, open data and reproducible research.

- Copyright © The Author (or The Authors) - Unless otherwise stated, this work is licensed under a Creative Commons Attribution Unported 3.0 License.