Open data sharing and reuse

The session described here was on 'Open data sharing and reuse' and is summarised by the session chairs, Dominic Dixon (Research Librarian) and Dr Sacha Jones (Research Data Manager) at the Office of Scholarly Communication, Cambridge University Libraries.

Have you wondered how research data is used after it has been shared publicly as open data? What are some of the impacts of sharing data and of its subsequent reuse by others? Are there ethical factors to consider? Does the researcher or research group who shared their data openly benefit in any way from its reuse? What are the essential properties of a reusable dataset? This session on 'Open data sharing and reuse' explored these questions and more via presentations delivered by a panel of University of Cambridge researchers from various fields. They included: **Professor Richard (Rik) Henson**, Deputy Director of the MRC Cognition and Brain Sciences Unit, Professor of Cognitive Neuroscience at the Department of Psychiatry and President of the British Neuroscience Association; **Professor John Suckling**, Director of Research in Psychiatric Neuroimaging in the Department of Psychiatry and chair of the University of Cambridge Research Ethics Committee; **Dr Mihály Fazekas**, Assistant Professor at the Department of Public Policy, Central European University, and scientific director of an innovative think-tank at the Government Transparency Institute; and **Professor Simon Deakin**, Professor of Law in the Faculty of Law and Director of the Centre for Business Research.

All speakers discussed challenges and concerns around data sharing, including how and when to share. Rik asks, "Why wait until publication?" to share research data, and perhaps consider publishing a data paper where a dataset is celebrated in its own right, without the narrative of a traditional article. Researchers are often concerned about scooping but there's little evidence of this and it may be a "paper tiger". There's an additional fear that data sharing will expose errors in work but as Rik noted, "I think we just need to get over our egos and accept that everyone makes errors". One particular challenge can be to control what people (or bots) do with your data, but researchers have a choice over where to share (e.g., which repository to choose) and how to license their data. Something that was implicit in all talks, and stated explicitly by Simon, is that the benefits of sharing data openly vastly outweigh the costs.

Sharing data deriving from research involving human participants is understandably complex due to data protection regulations (e.g., GDPR), obtaining informed consent, and the challenge of anonymising datasets, particularly those containing qualitative data. Participants need to be informed about how their data will be used, so the message is that data sharing needs to be planned far in advance, even at the gestation of the project idea. It is important to be aware of the repository options; for example, if managed/controlled access to data is required then hear about the set-up at MRC CBU discussed by Rik, or the UK Data Service for sensitive qualitative data, as highlighted by Simon. John discusses the import and export of datasets from an ethical perspective, giving two examples from the biomedical and social sciences with a focus on secondary data use. He says that these examples illustrate just how far in the future you might need to think when considering how

your data might be reused by others: it is “a lot better to ask for permission from all the stakeholders in these studies than it is to ask for their forgiveness”.

Data must be shared well for both researchers and society to reap the benefits. To do this, select an appropriate repository, adhere to any ethical/legal requirements, follow discipline-specific standards and make your data FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable). A key element of the latter is data documentation, an issue raised repeatedly during this session. Sharing the data alongside any associated code and detailed information about the data will enable it to be reused effectively and mitigate against misuse. Mihály discusses sharing the Digiwhist project data, which has been reused by academia, policy, civil society and the media, and emphasises this: “Every time I put out bits and pieces of my data and code that was not clear, I just kept on receiving the same question over and over again. So actually, it’s in your own best interests to document your work fully because then it is a lot more efficient for you”. Providing data about data is part of being completely transparent about the research process and results, enabling others to understand exactly what was done and to build on it. In some fields, this is an essential part of research reproducibility and replicability. As another example, Simon describes sharing the CBR Leximetric datasets – currently, the 2nd most downloaded dataset in Apollo and 8th of all UK institutional repositories – where not only the data were shared but also the methodology and an extensive codebook.

In both examples, being transparent in this way has led to wider reuse of these data and many citations of the data and associated publications. The benefits of FAIR data sharing and data reuse certainly do not rest solely in the number of resulting citations. Ethical and transparent research leads to credible research and researchers, enhancing reputations and quality of outputs. These are elements that all speakers highlighted in their talks. To end on a quote from Simon about the outcome of sharing data and of its subsequent reuse: “It’s been a very very positive experience for us”.

We’re always happy to receive any questions or comments you may have about data sharing and reuse. You can contact us at info@data.cam.ac.uk and see our Research Data website for more information.

Materials

- OR at Cam: Open Data sharing - YouTube

Additional resources

University of Cambridge School of Clinical Medicine guidance on secondary data use and related ethical considerations, discussed by Professor John Suckling.

The Digiwhist project website discussed by Dr Mihály Fazekas. The Digiwhist project is also one of the University’s research projects highlighted on the University of Cambridge global impact map.

Video of a previous talk by Professor Simon Deakin for OpenConCam 2016 talk on ‘Open Access and Knowledge Production 0 “Leximetric” Data Coding’.

The FAIR principles are outlined by Wilkinson *et al*. (2016) in *Scientific Data* – "The FAIR Guiding Principles for scientific data management and stewardship". There is also a useful guide for researchers on how to make your data FAIR.

Visit the University of Cambridge Research Data website for information on research data management, data sharing and guidance on depositing data into Apollo, the institutional repository. The site also hosts the University of Cambridge Research Data Management Policy framework, which is relevant to all research staff and students.