EC Consultation on open research data

Response from the Open Knowledge Foundation

How can we define research data and what types of research data should be open?

Research data is extremely heterogeneous, and would include (although not be limited to) numerical data, textual records, images, audio and visual data, as well as custom-written software, other code underlying the research, and pre-analysis plans. Research data would also include metadata – data about the research data itself – including uncertainties and methodology, versioned software, standards and other tools. Metadata standards are discipline-specific, but to be considered 'open', at a bare minimum it would be expected to provide sufficient information that a fellow researcher in the same discipline would be able to interpret and reuse the data, as well as be itself openly available and machine-readable. Here, we are specifically concerned with data that is being produced, and therefore can be controlled by the researcher, as opposed to data the researcher may use that has been produced by others.

When we talk about open research data, we are mostly concerned with data that is digital, or the digital representation of non-digital data. While primary research artifacts, such as fossils, have obvious and substantial value, the extent to which they can be 'opened' is not clear. However, the use of 3D scanning techniques can and should be used to enable the capture of many physical features or an image, enabling broad access to the artifact. This would benefit both researchers who are unable to travel to visit a physical object, as well as interested citizens who would typically be unable to access such an item.

By default there should be an expectation that all types of research data that can be made public, including all metadata, should be made available in machine-readable form and open as per the Open Definition[1]. This means the data resulting from public work is free for anyone to use, reuse and redistribute, with at most a requirement to attribute the original author(s) and/or share derivative works. It should be publicly available and licensed with this open license.

This expectation is even stronger when the research has been paid for by taxpayer, which results in a moral responsibility to ensure long term preservation of data and to maximise opportunities for reuse by other researchers. This responsibility includes the obligation to support the long term interoperability and protection of research data.

[1] http://opendefinition.org

When and how does openness need to be limited?

The default position should be that research data should be made open in accordance with the Open Definition, as defined above. However, while access to research data is fundamentally democratising, there will be situations where the full data cannot be released; for instance for reasons of privacy.

In these cases, researchers should share analysis under the least restrictive terms consistent with legal requirements, and abiding by the research ethics as dictated by the terms of research grant. This should include opening up non-sensitive data, summary data, metadata and code; and providing access to the original data available to those who can ensure that appropriate measures are in place to mitigate any risks.

Access to research data should not be limited by the introduction of embargo periods, and arguments in support of embargo periods should be considered a reflection of inherent conservatism among some members of the academic community. Instead, the expectation should be that data is to be released before the project that funds the data production has been completed; and certainly no later than the publication of any research output resulting from it.

How should the issue of data re-use be addressed?

Data is only meaningfully open when it is available in a format and under an open license which allows re-use by others. But simply making data available is often not sufficient for reusing it. Metadata must be provided that provides sufficient documentation to enable other researchers to replicate empirical results.

There is a role here for data publishers and repository managers to endeavour to make the data usable and discoverable by others. This can be by providing further documentation, the use of standard code lists, etc., as these all help make data more interoperable and reusable. Submission of the data to standard registries and use of common metadata also enable greater discoverability. Interoperability and the availability of data in machine-readable form are crucial to ensure data-mining and text-mining of the data can be performed, a form of re-use that must not be restricted.

Arguments are sometimes made that we should monitor levels of data reuse, to allow us to dynamically determine which data sets should be retained. We refute this suggestion. There is a moral responsibility to preserve data created by taxpayer funds, including data that represents negative results or that is not obviously linked to publications. It is impossible to predict possible future uses, and reuse opportunities may currently exist that may not be immediately obvious. It is also crucial to note the research interests change over time.

Where should research data be stored and made accessible?

Each discipline needs different options available to store data and open it up to their community and the world; there is no one-size-fits-all solution. The research data infrastructure should be based on open source software and interoperable based on open standards. With these provisions we would encourage researchers to use the data repository that best fits their needs and expectations, for example an institutional or subject repository. It is crucial that appropriate metadata about the data deposited is stored as well, to ensure this data is discoverable and can be re-used more easily.

Both the data and the metadata should be openly licensed. They should be deposited in machine-readable and open formats, similar to how the US government mandate this in their Executive Order on Government Information[2]. This ensures the possibility to link repositories and data across various portals and makes it easier to find the data. For example, the open source data portal CKAN[3] has been developed by the Open Knowledge Foundation, which enables the depositing of data and metadata and makes it easy to find and re-use data. Various universities, such as the Universities of Bristol and Lincoln, already use CKAN for these purposes.

How can we enhance data awareness and a culture of sharing?

Academics, research institutions, funders, and learned societies all have significant responsibilities in developing a culture of data sharing. Funding agencies and organisations disbursing public funds have a central role to play and must ensure research institutions, including publicly supported universities, have access to appropriate funds for longer-term data management. Furthermore, they should establish policies and mandates that support these principles.

Publication and, more generally sharing, of research data should be ingrained in the academic culture, and should be seen as a fundamental part of scholarly communication. However, it is often seen as detrimental to a career, partly as a result of the current incentive system set up by by universities and funders, partly as a result of much misunderstanding of the issues.

Educational and promotional activities should be set up to promote the awareness of open access to research data amongst researchers, to help disentangle the many myths, and to encourage them to self-identify as supporting open access. These activities should be set up in recognition of the fact that different disciplines are at different stages in the development of the culture of sharing. Simultaneously, universities and funders should explore options for creating incentives to encourage researchers to publish their research data openly. Acknowledgements

[2] http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

[3] http://ckan.org

of research funding, traditionally limited to publications, could be extended to research data and contribution of data curators should be recognised.

References

1. The Open Definition http://opendefinition.org/okd/
2. Open Data Licenses http://opendefinition.org/licenses/#Data
3. The Panton Principles for Open Data in Science http://pantonprinciples.org/
4. Open Economics Principles: http://openeconomics.net/principles/