

Submission to the Review of the National Innovation System

Professor Justin Zobel

29 April, 2008

Summary. Open access publishing, of research data and research papers, is becoming a critical component of an effective research environment. Having materials in the public domain is, arguably, a strong stimulus to innovation and to dissemination of knowledge; and current common practices, of keeping materials private – either in the medium term or indefinitely – are a significant obstacle to successful, high-impact research.

The topic of open access (OA) is currently in wide debate. Arguments for different forms of OA are being made in many forums, and have led to some definite steps in the direction of widespread OA. These include:

- Creation of numerous OA journals;
- Mandates from some journals that work can only be refereed after data or code, or both, are in an OA repository;
- Tying of research grants to commitments to immediately place all data gathered in the course of the research in the public domain;
- Development of large, even vast research projects whose aim is to gather data that is to be made freely available;
- Expectations from some research funding bodies in the EU and the US that researchers put accepted publicly funded papers in OA repositories within a year of publication;
- The widespread practice of individual researchers placing their papers in open repositories or on their websites, legally or otherwise, either at time of submission or time of publication;
- Acceptance by some publishers that individuals will self-publish their work online.

There are several illustrations of the success of OA. A high-profile example is the NCBI's GenBank, which published the data gathered by the Human Genome Project. GenBank continues to grow, and the NCBI is now host to dozens of OA databases of biological data. This data has become critical to research in biomedicine; a great many current biomedical research projects would struggle to proceed without these resources. Some research, such as many instances of development of medical treatments based on linkage of genetics and disease, could not take place without access to such OA data.

Another example is provided by the TREC data, which was, initially, several gigabytes of documents first collected in 1992, designed for testing of search engines. This data is not

fully OA, but was widely available in the relevant research community, and led to a revolution in the field of search. Up to that time research prototypes had almost universally been tested on toy data sets, and the state of the art was not strong. After distribution of the TREC data and the evaluation of systems that this data allowed, it quickly became clear that some well-established approaches to search were not valid, and numerous innovative new search techniques appeared. These techniques underpinned the first web search engines, which began to appear in 1994–5, a couple of years after Mosaic first popularised the web. It is well possible that, without the spur of the TREC data, practical web search – and the enabling of access to information that this allowed – would have been significantly delayed.

OA has had a role in development of computer software since the 1970s, and is one of the main engines of development of software today, primarily through open source. While to some extent the topic is tangential to the topic of this submission, open source software development is worth noting as it has provided a rapid channel for delivery of research. For example, some of the key tools widely used in bioinformatics, such as Bioconductor, are a product of the open source culture; there are probably thousands of tools of this kind. More widely, the web itself is a result of open source, as are many of the other applications of the internet, such as email protocols and the content of Wikipedia.

The advantages of OA publishing of papers are obvious. Results become immediately available, without the restrictive gatekeeping (and, in some cases, poor search tools) provided by for-profit publishers, whose vested interests are in some respects in conflict with the interests of the community at large. OA publishing is relatively easy to provide – a collection of all research papers ever published might occupy just a few tens of terabytes, possibly much less. If papers are freely or cheaply available, the savings to the academic community are considerable: the labour of writing, editing, and reviewing papers is provided free by academics, and the cost of publishing online is minimal. Currently, a typical university library spends millions in access fees.

The advantages of OA publishing of research data are, for the research community as a whole, also obvious. Individual research data sets may be too small for some investigations; pooled, much more can be extracted and learnt. Experts in data analysis can make new interpretations missed by the authors of a study. Data gathered for one purpose can have great value when analysed from another perspective. Openness encourages high standards in research. The need for OA repositories would lead to more systematic preservation of data, which today is too often lost or maintained in ad hoc form such as spreadsheets on individuals' computers.

As the GenBank example illustrates, OA allows research that would otherwise be impossible. A recent example is the Human Variome Project, being led out of Melbourne. Should this project be successful, so that there is widespread linking of data on human genetic variation and its known effects, there will be opportunities for dramatic research discoveries that existing restrictions on data access simply would not allow.

There are significant practical obstacles to OA, for example:

- Existing models of publishing and academic credit will to some extent be undermined (however, new mechanisms are developing that help address these issues);
- Data volumes can be huge;

- Publishing of data must be in line with confidentiality provisions and may lead to variation in how data is gathered.

A significant issue is the attitudes in some disciplines; for example, in some areas the owners of databases expect authorship on papers that use their data. Such expectations need to be addressed (and not simply dismissed), but current practice is not realistic in the context of projects that may link thousands of data collections. Another issue is that authors may fear that early release of data may prevent them from being the discoverers of the knowledge that the data contains; again, such issues need to be addressed, while noting that this is an example of conflict between personal desire and public good. Both of these issues could, for example, be addressed by mechanisms that allow citations to data; recent proposals for microattribution are a step in this direction.

While there are obstacles, however, they are not insuperable. Moreover, there are good reasons to set aspirational goals, such as encouraging authors to make papers freely available at time of publication and requiring researchers to provide data on request. Positive change can be created by providing practical mechanisms for OA, and helping the research community to move towards an OA culture.

There are also risks to not adopting OA. It is being adopted elsewhere in the world, and we should be participants in what is becoming a mainstream trend in science. In some fields, it may be that OA data rather than papers will become the primary object of reference and citation. OA reduces the cost of access to research and the cost of undertaking research, and can improve research impact and research quality. A culture of OA creates incentives for institutions to responsibly curate their data, and recognise it for what it is – a critical community asset.

One hears telling slogans on this topic: that publicly funded data belongs to the public; that the person most able to gather data may not be the person most able to analyse it; that the results in some papers are but a few drops from a barrel of data. Were we starting a research culture from scratch, it would, I think, be difficult to make a compelling case for current practice. It is an artefact of history – effective mechanisms for sharing data and open repositories have only appeared in the last few years – and we should look forward to universal adoption of open practices that give the greatest scope to discovery, to invention, and to creation of new knowledge.

Justin Zobel
Senior Principal Researcher, NICTA VRL; President, CORE