An Introduction To Open Access Data Sharing

The open access movement is gaining momentum both here in the UK and internationally. ***Emma Fenton*** *explains the principles of open access data sharing.*

This report serves as a follow-up to the report on open access that was produced by the IES in July 2012. It will investigate the principles and practice of data sharing as part of the open access movement within the publishing industry.

The introduction of open access datasets has led to the emergence of novel applications, for example apps and websites reliant on open access data. Examples are given later on in the article.

As open access is becoming an increasingly popular method for publishing scientific articles, the focus of the debate has shifted to include the provision of original datasets alongside scientific articles to enable other researchers/professionals to utilise the data either for subsequent publications or to check the reliability of the original research results through a process called 'data-mining'.

Data mining

The progress of digital data storage and acquisition of technology has caused a proliferation of large databases of research. This has occurred across all aspects of human life, from the everyday (supermarkets recording transactional information) to the more academic (such as databases of genetic sequences). This new wealth of data has promoted the interest in trawling them for data that might be of value, either to the owner of the database or to the wider world[1].

Data mining is the 'process of analysing data from different perspectives and summarising it into useful information'[7]. It allows researchers to find patterns or correlations across multiple, large, complex datasets.

Data mining comprises five major elements:

- Extract, transform, and load transaction data onto the data warehouse system;
- Store and manage the data in a multidimensional database system;
- Provide data access to practitioners/policy-makers;
- Analyze the data by application software; and
- Present the data in a useful format, such as a graph or table

From Frand (2012)

Data mining generally evaluates data that have been collected with some other purpose in mind, it is this characteristic that separates data mining from traditional statistical analysis, for which data are collected to answer specific questions that are explained by statistics. For **this** reason, data mining is often called "secondary" data analysis.[2]

What is open data?

Open data is information that is available for anyone to use, for any purpose, at no cost. It can create opportunity for organisations to make more robust decisions, uncover cost savings and get to know customers better. Open data is defined as data that meets the following criteria:

- Accessible (ideally via the internet) at no more than the cost of reproduction, without limitations based on user identity or intent;
- In a digital, machine -readable format for interoperation with other data; and
- Free of restriction on use or redistribution in its licensing conditions.

-Technology Strategy Board

Many of the specialisations within the environmental sciences have evolved over the past century from those based on discrete, small-scale and short-term, specifically-focused projects conducted by individual research teams to include interconnected, large-scale, long-term and multi-disciplinary projects that enable the utilisation of multiple datasets.[3]

The current data-archiving landscape

Despite funding organisations increasingly requiring the publication of datasets alongside scientific articles, most scientific data are stored in private files rather than institutional repositories and effectively become lost when the researchers move on.[4]

> "Because research ... increasingly depends on the availability and sophisticated analysis of large data sets ... We cannot afford for access to scientific knowledge to become a luxury, and the results of publicly-funded research in particular should be spread as widely as possible"
>
> - Neelie Kroes: European Commissioner responsible for the Digital Agenda

An Introduction To Open Access Data Sharing

This can lead to situations where vast amounts of research funds are spent year on year on 'new' research while existing datasets remain unpublished and underutilised.

Traditionally there have been relatively few incentives for researchers to share their data.[3] The publication process was typified by researchers "gathering [their] own data and publishing the distilled results in peer-reviewed journals".[3] Data sharing was not seen as part of this process.

The UK Government is releasing over 8000 datasets from all central government departments as well as a number of other public sector bodies and local authorities. The aim is for the data to be used to build useful applications that help society or investigate how effective different policies have been over time.

Why is data-sharing beneficial?

Creating large, internationally-renowned institutional repositories will ensure that data are not lost to future researchers and will promote appropriate citation and acknowledgement of their use.[1] It is also assumed that such repositories can 'enhance and accelerate' scientific advancement by enabling multiple uses of datasets and preventing duplication of efforts.[1 5 6]

There is evidence from specific fields, such as climate science, that the provision of open access data is increasing the impact of research by making it available, not only to researchers, but also to practitioners such as resource managers and policy-makers who seek to use data to inform their decisions and develop strategies.[7] Climate changes scientists need to integrate these large and varied datasets from completely different disciplines in order to understand the current situation and provide help in creating and predicting the impact of government and institutional policies.[10]

More recently, cases of scientific misrepresentation such as 'climate-gate' have emphasised the importance of reliability of data and experiment reproducibility as a built-in peer-review system[1234]. Furthermore, the lack of access to data is seen as an obstacle to scientific advancement, particularly for interdisciplinary fields or internationally-focused research[2].

Improvements in the sophistication of data-mining techniques mean that new research can combine multiple datasets from diverse but related disciplines[3]. It also means that the data from initially narrow-focussed studies can be used to inform research far beyond their original analysis[4].

The availability of so many complex datasets is also changing the way that scientists approach their subject. It has opened up careers that do not practice science as a traditional 'wet lab' profession, rather they treat science in terms of data modelling and forecasting.

Case study: Genomics

The proliferation of online genomics databases came about as a result of two main factors: firstly, software was developed that enabled the management of genomic data online; and secondly publishers grew increasingly unwilling to carry on editing and printing the growing quantity of gene sequences that were being isolated. Nowadays registering gene sequences and sharing them online is standard procedure and is recognises as 'fostering one of the greatest scientific revolutions in the past century'[3].

Case study: International Council for Exploration of the Sea (ICES)

This collaboration dates back to 1902 when a group of scientists took the opportunity to share their data to enable better understanding of fish distribution, oceanography and the marine ecosystem across geographical and research boundaries. This initial information exchange was driven very much by science and research rather than politics but it led to the signing of the 1964 ICES Convention in Copenhagen that finally solidified the ICES as an official advisory board to add value to national research efforts[8].

Case study: Meteorology

Data sharing within the meteorology community started back in 1873 when the International Meteorological Organisation began to coordinate international weather forecasts[8]. This data sharing has enabled scientists to map global weather systems and examine the global climate as a whole rather than as a series or poorly-integrated geographical units.

Case study: Open Data in Government

As of 2012 the UK government has made over 8000 of their datasets available, this has opened up possibilities, not just for research but also for commercial uses. For example, throughout the summer, the Environment Agency collects detailed scientific data on the cleanliness of bathing waters across England and Wales. This has now been used under the Open Government Licence to produce a tangible, interactive map so that members of the public can explore the data themselves.

> "Without the infrastructure that helps scientists manage their data in a convenient and efficient way, no culture of data sharing will evolve."[8]
>
> - Stefan Winkler-Nees

An Introduction To Open Access Data Sharing

Challenges for open access data sharing

Technological

New technological solutions will be needed to allow scientists to share their data efficiently and in a meaningful format[15] .

The data need to be presented in an understandable and usable format to ensure maximum transferability between projects[5]. This must include a standard terminology[1 10], integration of datasets[1678], as well as ensuring that the legal aspects of data sharing, such as intellectual property rights etc are covered [2]. Many researchers understand the importance of data sharing but the lack of sector-specific repositories and well-established metadata standards does not enable them to do so.

Provenance

The repositories that will house the data must be sufficiently robust that researchers are able to track the original provenance of data. This will be especially important for policy and management decisions that are taken as a result of utilising data shared via an open access repository[1].

Financial cost

Maintaing and updating large databases is not just time consuming , there is also a substantial financial cost to procuring and running the hardware needed to store the data.

Time

The time costs for researchers to make their data available open access are significant. Both in uploading the data but also in the requests for assistance that may follow after the release of the data. Time may also be spent rebutting future re-analysis of the data and defending the original research findings.[4]

There is also the effort that it takes to format, document and release the data itself before it is displayed in order that it be as useful as possible.

Social and Cultural

Human Behaviour Change studies have highlighted the difference between *intention* and *action* in a number of different settings and the same challenges occur in open access data sharing[1].

Some studies suggest that in order to promote data sharing it must be incentivised, perhaps through a revision of standard industry impact factor-calculating metrics[26].

There is also an additional concern that the original conclusions of the research may be challenged after re-analysis by a subsequent research group. This could be down to errors in the original research, misinterpretation or misunderstanding of the data or the development of more refined research techniques after the original publication[4].

It is now possible, with the advent of sophisticated data-mining techniques, that data miners could be able to identify extra relationships in the data that could interfere with the original researcher's planned investigation agenda[4].

There are a variety of ways in which science centres try and avoid this issue. For example, the Sanger Institute sequence genomes in order to study diseases that have an impact on health globally. Every time a genome is sequenced at the centre it is automatically uploaded into a free, online database that can be used by researchers across the world.

They simply ask in return for making the data freely available that researchers do not investigate aspects of the data that are being actively researched by the Sanger Institute themselves.

In certain circumstances, release of research data is restricted by people other than the researchers. For example, in the US, many hydrological data are restricted by state and national administration. If subsequent researchers are to gain access to these data at all, it is often years after the research was conducted, as the results are perceived to be of national strategic importance (i.e. documenting resources or agricultural capabilities)[6].

An Introduction To Open Access Data Sharing

Potential pitfalls of open access data sharing

As with all behavioural change studies, research has shown that *intention* to share data often does not translate into a researcher actually sharing data. This is likely to be compounded by the fact that researchers often withhold information about how frequently they themselves do not share data, whilst reporting the frequency with which they discover data being withheld by others[3].

Whilst the benefits of dataset sharing are open to the wider research community, much of the burden and cost falls to the publishing researcher themselves and some suggest that the benefit to the publishing researcher it outweighed by the potential costs as outlined above[4].

The technology of data-sharing is also a complicated issue. If not well-managed the data could become corrupted, lost or disordered, rendering it useless[4]. A quick glance at the data-sharing practices of certain research bodies throws up yet another challenge - the practices and capabilities of difference scientific disciplines can vary widely, even within a single discipline, which minimises the usefulness of the shared data[7].

Conclusions and views of the IES

It is clear that, despite some challenges yet to be addressed, open access data sharing has the potential to revolutionise research practices, particularly for interdisciplinary fields such as the environmental sciences. With the advent of a new type of treatment of scientific data, it is likely that universities will have to respond by updating their curricula to reflect the novel skills that scientists will need in the future, for example, it is lilely that that scientists in the future will need to learn more data handling and programming skills than are traditionally taught in science degree programmes currently.

The IES believes that the limitations associated with restricting data have the potential to impede scientific progress. Whilst some institutions champion open access data sharing, many lag behind. Fully open access data repositories would be an ideal solution for the environmental sciences as well as the wider academic community. The IES will endeavour to promote this change as the open access movement becomes more widespread.

The IES supports the requirement for environmental data to be made publicly available. The interdisciplinary nature of the environmental sciences means that open access data publishing could open up avenues of research previously inaccessible to environmental scientists.

Sources

1 Hand et al., 2001. Principles of Data Mining. The MIT Press: Cambridge, Massachussetts, USA.

2 Frand, J. 2012. Data Mining: What is Data Mining [online]. Available from: https://www.anderson.ucla.edu/faculty/Jason.frand/teacher/technologies/palace/datamining.htm. [Accessed: 31st October 2012].

3 Reichman et al., 2011. Challenges and Opportunities of Open Data in Ecology. Science. 331, pp.703-705.

4 Klump et al., 2006. Data Publication in the Open Access Initiative. Data Science Journal. 5, pp.79-83.

5 Piwowar & Chapman. 2007. Public sharing of research data sets: A pilot study of associations. Journal of Informetrics. 4, pp-148-156.

6 Piwowar et al. 2007. Sharing Detailed Research Data is Associated with Increased Citation Rate. Plos ONE. 2 (3) e308.

7 Overpeck et al. 2011. Climate Data Challenges in the 21st Century. Science. 331, pp.700-702.

8 ODE. Ten Tales of Drivers and Barriers in Data Sharing [online]. Available from: https://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/7836_ODE_brochure_final.pdf. [Accessed: 29th October 2012].

9 Youngseek et al. 2012. Institutional and Individual Influences on Scientists' Data Sharing Practices. Journal of Computational Science Education. 3(1), pp.47-55.

10 Science Commons. 2012. Protocol for Implementing Open Access Data [online]. Available from: https://sciencecommons.org/projects/publishing/open-access-data-protocol. [Accessed: 29th October 2012].

About the Institution of Environmental Sciences (IES)

The IES is a visionary organisation leading debate, dissemination and promotion of environmental science and sustainability. We promote an evidence-based approach to decision and policy making.

We are devoted to championing the crucial role of environmental science in ensuring the well-being of humanity now and in the future.

Contact

Institution of Environmental Sciences
34 Grosvenor Gardens
London
SW1W 0DH
T: +44 (0)20 7730 5516
E: enquiries@ies-uk.org.uk