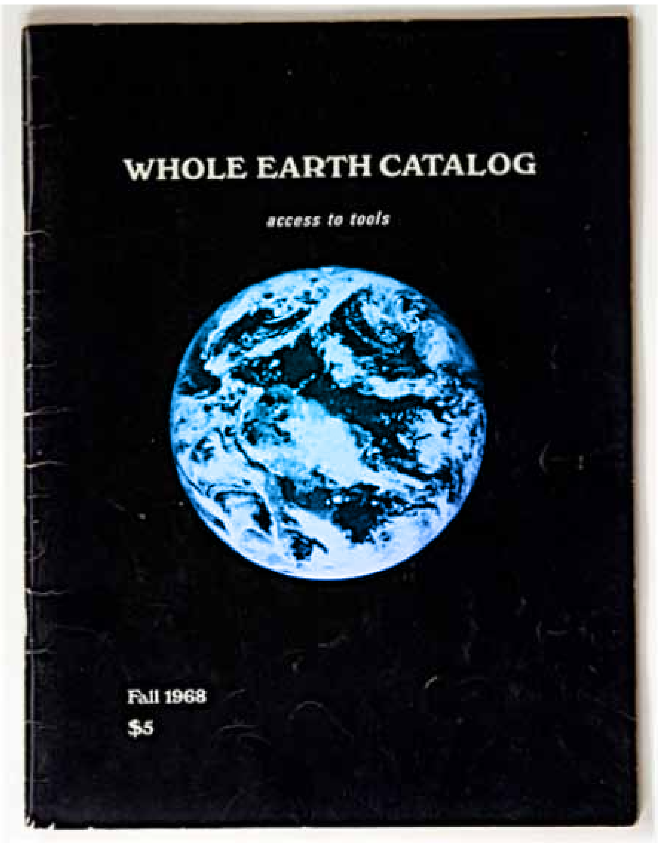
e science TALK
e-ScienceBriefings
Talking about e-science

Open Data,Open Science

Data: unavoidably expensive?

In the late 1960s and early 1970s, falling costs of integrated circuits meant the computer was making a transition from being a tool available to only the very few to one available to the many. As the potential for computers to be used to create, store and transmit ideas and information became apparent, technological evangelist Stewart Brand, publisher of The Whole Earth Catalog, identified a duality in the nature of digital data: "Information wants to be free. And information wants to be very expensive." [1]



The ease with which we are able to share information using computers shows how 'free' it can be. Compared to the expenses of print, the monetary cost of publishing information using the web is virtually nothing. There often remain costs associated with gaining access to scientific data on the Web, however. Sometimes, information is expensive. Scientists have long built careers by sharing their data – and staking a claim on it – through publishing it in prestigious scientific journals. Such journals often command high subscription fees even on the Web, restricting the flow of data to all but the very wealthy.

Slowly however, a quiet revolution has been gaining momentum: open access publishing, open science, and open data. The first is a change in the publishing model to one more suited to the age of the Web; the second, a change in how scientists connect with society – their major funders through taxation. Open data is even more revolutionary. Being able to share data more quickly and easily will accelerate the pace of scientific progress, and help scientists to solve the pressing problems of the 21st century: climate change, energy security and feeding the population. Open data is about sharing it freely – that means without restriction more than without monetary cost

Neelie Kroes, European Commissioner for Digital Agenda' – *"Sharing data, and having the forum to openly use and build on what is shared, are essential to science. They fuel the progress and practice of scientific discovery. That's why scientists have long sought out new tools and new ways to share their knowledge."*

When does free mean free?

The best things in life may be free, but that does not necessarily mean without cost. The word has two distinct meanings in relation to ownership, but they are often conflated. There is free, as in cost-free or gratis – you don't have to pay for it (although you may have to accept advertising). Then there is free as in libre – it comes with freedoms that allow the work to be adapted and reused. For open data and open science, libre is more important.

Software, media and data can be provided gratis, but may still be restricted by various levels of copyright, preventing or limiting a user's freedom to reproduce, reuse or adapt the work. For open data to work, the data must be 'freed' using a strongly permissive form of licence (see 'Licensing Open Data'). Software, media and data that are provided libre are also usually provided gratis – but actually they don't have to be.

[1] *'What the Dormouse Said', John Markoff, 2005, Penguin Books*

Jenny Molloy, Coordinator, Open Science Working Group – *"Good science should be reproducible, but in many fields (not all) it is often impossible for other researchers to repeat and critically assess analyses without access to raw data. If researchers make a scientific claim they should ensure that the data is openly available to back up that assertion in a form that is reuseable by their peers for both independent analysis and inclusion in meta-analyses."*

Publishers: Opening Access

Open Access is an important step forward for open science, because it completely turns the old model of academic journal publishing on its head. Instead of charging universities and research institutes large sums to access scientific papers, Open Access publications charge authors a nominal fee (starting at around £500) per paper to publish their research, which is then made freely available over the Web. This means it is as accessible to scientists everywhere, for example in poverty-stricken regions, as well as the general public who essentially fund the research. Policy makers, governments, funding bodies and charities welcome the move because it sets the global stage for international innovation.

BMC: A Model for Open Access Publishing

BioMed Central (BMC) was set up by entrepreneur Vitek Tracz in 2000 in response to the shifting publishing landscape brought about by the advent of the Web. In the run-up to the millennium, requests for print journals were declining as scientists began to instead demand greater access to online publication repositories. The number of journals being published was meanwhile increasing, just as university libraries and research institutes struggled to afford the rising costs of subscriptions.

Tracz realised that scientists were prepared to pay to share their data – researchers generally want access to their work to be gratis, as long as they get credit for it. In that way, their main currency – their reputation – would have a wider reach. A greater global reach, arguably, than it would have if their work was restricted to only those that can afford to pay to see it. Restricting its publishing to online only has helped BMC to be profitable and has served as an example of how the Open Access model can work in a commercial environment.

Tagging open data is important to keep data searchable in the future (PDDL)

Iain Hrynaszkiewicz, Publisher, Open Science at BioMed Central – *"Publishing data and software online, whether included with or linked to journal articles, greatly increases the value and reproducibility of reported research. Publishers should embrace open data in response to scientists' needs and to drive innovation, but more efficient and reliable science is the ultimate goal. Open data is a way to help achieve that. Copyright is messy with respect to data and at BioMed Central we are working on implementing explicit public domain dedication of published data, to better facilitate data integration and reuse without legal barriers."*

Directory of Open Access Journals

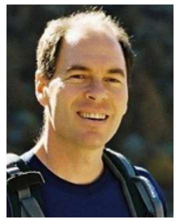
The Directory of Open Access Journals (doaj.org) now lists nearly 7000 journals that are available at zero cost. However, one barrier to truly open science is that even in Open Access, publishers can choose to impose restrictions on the use of the content they publish. Only around a fifth – including BMC, Public Library of Science (PLoS), and a number of smaller publishers – allow reuse and adaptation in the libre model. BMC (and PLoS) journals are covered by a Creative Commons Attribution licence (CC-BY), ensuring scientists get credit for their freely distributable works. Announced at the Open E-Infrastructures for Open Science, hosted by ALLEA (ALL European Academies), UK biomedical foundation, the Wellcome Trust and the World Bank have similar initiatives to open up the work that they fund.

Wouter Los, Project Leader of Lifewatch, e-infrastructure for biodiversity research– *"Understanding our environment requires large volumes of data of very different kinds. We assume that we now only capture a few percent of the data that we would like to have available. Open data are crucial, as modern interdisciplinary environmental science cannot deal with limited data sets. The same holds for society. Environmental management is dependent on sufficient and reliable open data."*

OpenAIRE

OpenAIRE
Open Access Infrastructure for Research in Europe

Understanding how central open data is to scientific advancement, 20% of the budget of the European Commission's 7th research framework (FP7) is dedicated to making the science it funds open and accessible. OpenAIRE (Open Access Infrastructure for Research in Europe) is the result: a project set up to establish and operate an electronic infrastructure for handling peer-reviewed articles, enabling researchers to deposit their final peer-reviewed manuscripts and/or post-prints either in an institutional repository or a subject-based repository. It also provides support structures for researchers wanting advice on how to make their research open, which for ERC – European Research Council – projects is often a condition of their grant.



Tim Smith, Collaboration and Information Services Group Leader at CERN – *"The strength of science has always been its open dialogue on the results and conclusions of experiments. Since data is recorded in such volume now that it cannot be communicated effectively via the results tables of scientific papers, data sets themselves need to be accessible independently to ensure the scientific hardening process. Furthermore, sharing can allow more knowledge to be derived from a data set than in the original research."*

The limitations of copyright

The legal tool of copyright, which was introduced to the world through English law in the late 17th and early 18th Centuries so that authors could be sure of a fair income for their work, has perhaps been more of a hindrance than a help in the drive towards open data. Many publishers require a 'transferral of copyright' from scientists wanting to publish with them, including those behind some of the most prestigious journals. Publishers argue that full transferral unburdens scientists from needing to assert their authorship and retain control over their own work. However, copyright can also restrict how data is re-used, and as science becomes increasingly data driven, access to the data sets can be as important as access to the paper itself. Scientists wanting to reuse the material of others, even if they cite it properly, could find themselves in breach of copyright if they do not ask the publisher's permission and pay any necessary fees, which may again affect scientists in poorer countries.

EUDAT

Making sense of open data in the future

As more and more data becomes available publically through the move towards open data, common shared tools are needed, so that scientists can sift through it and make sense of it in the future. One particular mechanism that meets the requirements for data organisation is the application of metadata – data that describes data. Organisations like Open Data Foundation (ODaF) are dedicated to the adoption of global metadata standards and the development of open-source solutions promoting the use of statistical data. To facilitate the sharing of data, the Europe-wide EUDAT project (eudat.eu) implements a secure means of sharing data using persistent identifiers, similar to the way written documents have been given ISBN and now digital object identifiers. Agreeing on such standards will make it easier to share data between disciplines, and to sort through mountains of data decades after it might have been generated.

Making sense of open data in the future

To solve the problem of licensing open data and protecting authorship, one solutions is the concept of "copyleft" – a play on copyright, and the practice of using copyright law to actually keep data open. Richard Stallman, a computer scientist at MIT, created the GNU Public Licence, GPL, after finding that he was legally unable to reuse some of his own code, which he had previously given freely to a corporate developer. The GPL ensures that any software released under it may be used, adapted, changed and freely distributed by its users, but that any copy or derivative work is covered by the same licence. In effect, it uses the legal system of licensing to prevent prospective developers imposing a commercial licence on GPL software-derived work. GPL can be good for some types of software, but is not always appropriate for creative works such as scientific publications.

Of all copyleft licences, perhaps the most well-known are those from Creative Commons (CC), whose Sharealike (CC-SA) licence is perhaps the closest to GPL. A common misconception is that Creative Commons is equivalent to public domain – and that a user can effectively do what they like with it. In fact, CC licences are precisely worded legal documents that use terms from the legal concept of copyright. Open data requires licences that are unrestrictive. For CC, the least restrictive and most appropriate for open data is CC0, which effectively releases a work into the public domain. However, scientists can still be sure of receiving credit for their work because the cultural norm of citation exists separately to the notion of copyright. Further information can be found at *pantonprinciples.org*

CC

Open Standards for raw data

Open standards and raw data are fundamental to a functioning and long-lasting open science movement, and it is imperative that standards are agreed upon and adhered to. Data must not only be legally accessible, but also technologically accessible 20 years down the line despite changing software trends. This fate has already befallen some of the earliest digital archives, such as the BBC Domesday Project – an attempt to produce a digital historical record of life in the UK in 1986. The software, stored on laserdisc in the LV-ROM format, would only run on an expanded Acorn BBC Master computer with a specially produced laserdisc media drive. A few years after production, the computers were obsolete and the data was inaccessible. For data to have the best chance of surviving long into the future, it should be in its most 'raw' format. Open standard formats should be free from proprietary ownership, and simple to 'future-proof'. They could include UTF-8, for text and numerical files (.txt files, in other words), PNG for pixel-based images, SVG for vector images (e.g. technical drawings, scalable logos etc.), and Ogg Vorbis and Theora for audio and video. [3]

Virginie Simon, founder & CEO of MyScienceWork – *"MyScienceWork is dedicated to open science. Our platform enables researchers and engineers from all disciplines to communicate, share, and discover. Scientists can use our innovative search engine to access tens of millions of professional articles. By facilitating the accessibility of knowledge and reinforcing scientific communities, we promote accessibility and visibility of research. I think this is only the beginning of the transformation of how scientific data is organized and shared, and a whole new era of open science."*

ScienceSoft: Open Software for Open Science

Much of the most-used technical and scientific software is open source. From statistical software R, used by scientists across a range of fields; biochemistry application DOCK; to programming languages Python, C and Ruby, and technical typesetting package LaTeX, open source software has always played a major role in scientific computation. Open source software development is a democratic, inclusive enterprise, with some of the major projects involving thousands of volunteer programmers. The quality of software developed this way matches and exceeds commercially-developed software: a reminder of the wisdom of crowds; 'that none of us is as smart as all of us'.

So many software packages are available across the various open source repositories that scientists may struggle to find the most appropriate software for the task they have in mind. Often this may be made worse by there being several variants or 'forks' of a project, some available on only certain repositories, stored among a wealth of non-scientific software. The levels of support on user forums may be similarly variable and disparate.

Alberto Di Meglio, ScienceSoft Project Leader – *"Open source software is based on values like transparency, collaboration and availability. These values are also at the base of Open Science. An active, vibrant community of software developers and users contributes to making global scientific research more accessible and reproducible. New scientific and societal challenges are becoming more and more complex. They cannot be addressed anymore just by clever individuals, but by open collaborations on a global scale. Open source software is a fundamental part of this transformation."*

Initiated in December 2001, Sciencesoft builds a virtual software repository and support network coordinated by the European Middleware Initiative in collaboration with the European Grid Infrastructure, StatusLab, iMarine, OpenAIRE and other e-infrastructure projects. It brings together a wealth of software expertise to help research communities to find the software they need. Users can rate software, which provides useful feedback for developers and helps funding agencies to understand the software use of research communities.

R is a scientific programming language used by scientists in all disciplines that has been developed as an entirely open source package.

For more information:

JISC Legal Open Data guide: ***discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf***
Panton Principles on Vimeo with Iain Hrynaszkiewicz: ***vimeo.com/34555054***
Panton Principles: ***pantonprinciples.org***
Data Definition and tagging: ***opendefinition.org***
Semantic web: ***www.w3.org/2001/sw/***
ODaF : ***opendatafoundation.org***
EUDAT: ***eudat.eu***
CC: ***creativecommons.org***
Science Soft: ***sciencesoft.org***
EGI : ***www.egi.eu***
Real Time Monitor: ***rtm.hep.ph.ic.ac.uk***
iSGTW: ***www.isgtw.org***
e-ScienceTalk: ***www.e-sciencetalk.org***
email: ***info@e-sciencetalk.org***

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7
INFSO-RI-260733

[2] *'See Tim Berners-Lee, 'Raw Data Now' at* is.gd/rawdatanow