DOI: https://doi.org/10.14324/000.learn.12

Case Study 11

Why Open Data?

Author: Professor Geoffrey Boulton (University of Edinburgh and President of the International Council for Science's Committee on Data for Science and Technology)
Email: G.Boulton@ed.ac.uk

11.1. THE DIGITAL REVOLUTION

At about the turn of the millennium, the global volume of data and information that was stored digitally overtook that stored in analogue systems on paper, tape and disc. The result has been a digital revolution, with the global data acquisition rate now 40 times greater (35×1000^7 bytes) than 10 years ago, still accelerating and driven in part by the massive reduction in the cost of digital storage. In 2003, the human genome was sequenced for the first time. It had taken 10 years and cost $4billion. It now takes 3 days and costs $1,000.

The unprecedented rate that we are able to acquire, store, manipulate and instantaneously communicate vast amounts of digital data and information has profound implications for all fields of science and scholarly research as well as for economies and societies. It is crucial that these implications are explored to the maximum effect by the research and scholarly communities in all parts of the world. Part of the opportunity lies in exploiting "Big Data", where enormous fluxes of data stream into computational and storage devices, often from a great diversity of sensors and sources; in "Linked Data", where semantic linking between different datasets opens opportunities for eliciting much deeper meanings (of great potential relevance for many global challenges such as infectious disease, disaster risk reduction and migration); in the myriad opportunities that arise from blending the physical and digital realms through the "Internet of Things"; and in the powerful but problematic potential of machine learning. The fundamental benefits derived from these approaches are in elucidating patterns and relationships that have previously been beyond our capacity to resolve and both to characterize and to simulate the dynamics of complex systems.

11.2. SCIENCE[1] AS AN INHERENTLY OPEN ENTERPRISE

Openness has been the bedrock on which modern science has been built. The rules of the game were established in the late seventeenth century, when scientific ideas began to be published in open journals rather than hidden in the private correspondence of gentlemen. A further crucial step was the requirement by journal editors that truth claims must be accompanied by the evidence (the data) on which they were based. This permitted others to attempt replication of the observational or experimental evidence and to scrutinise the logic of the proposed relationship between evidence and concept. Failure on either count indicated error. It is a process termed "self correction" by historians of science, tellingly characterised by Arthur Koestler in writing: "The progress of science is strewn, like an ancient desert trail, with the bleached skeletons of discarded theories that once seemed to possess eternal life". If there is a scientific method, this is it, the power of the negative. Albert Einstein characterised it as: "No amount of experimentation can prove me right. A single experiment can prove me wrong."

[1] The word science is used here to mean the systematic organisation of knowledge that can be rationally explained and reliably applied. It is used, as in most languages other than English, to include all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines.

11.3. THE BRIGHT SIDE

Like all revolutions that have not yet run their course, it is often difficult to distinguish reality and potential from hype. But powerful, real discoveries have now emerged in the elucidation of previously unsuspected patterns and relationships. In genomics, rapid sequencing and advanced computing power permit systematic testing of relationships between genetic variations and specific traits and diseases, rather than using trial and error, with profound implications for medicine, agriculture, the production of biofuels and the process of drug discovery. The advent of the modern computer has long permitted simulation of the dynamics of highly coupled complex systems, their sensitivity to small variations in initial conditions and their capacity to produce "emergent behaviours" that were not evident from their individual components. We can now add to this by the use of big, linked data to characterise complexity, and by iterating between characterisations and simulation, to follow and forecast the evolution of complex systems, as is now done in modern high-resolution weather forecasting. Only however if data is routinely made "intelligently open" (accessible, intelligible, assessable and re-usable),[2] can the full benefit of such approaches be realised.

11.4. THE DARK SIDE

However, the vast and complex data volumes that many scientists are now able to access also challenge the open approach required for self-correction. This arises from the difficulty of making such data sets open to scrutiny, together with the metadata, the computer code used in analysis, and the logic of any "learning machine" used in the process. It is hardly surprising that many of us fail this standard, or have succumbed to the temptation to keep our data under wraps so that it can be milked again for further publications. A current debate in the *New England Journal of Medicine*[3] about the rights and wrongs of openness in medical research epitomises this conflict; between the public interest in openness and the interests of scientists' careers in maintaining data ownership. Moreover, the recent attempts to replicate the results of highly regarded papers, in areas as diverse as pre-clinical oncology, social psychology and economics, with replication rates never exceeding 25%, illustrate the consequences of not rigorously presenting all the data and metadata. Without this, self-correction cannot work. If we are to maintain the credibility of the scientific process, we need to regard absence or inadequate presentation of data and metadata as scientific malpractice and to re-establish standards of reproducibility for a data-rich age. Without this we run the risk of the digital explosion overwhelming the processes that ultimately maintain scientific rigour.

11.5. ADAPTING TO CHANGE

Information and knowledge have always been essential drivers of human material and social progress, and the technologies by which knowledge is stored and communicated have been determinants of the efficiency of these processes. The digital revolution is a world historical event as significant as Gutenberg's invention of moveable type, and certainly more pervasive. A crucial question for the research and scholarly community is the extent to which our current habits of storing and communicating data, information and the knowledge derived from them are fundamental to creative knowledge production and its communication for use in society, irrespective of the supporting technologies, or whether many are merely adaptations to an increasingly outmoded paper/print technology. Do we any longer need expensive commercial publishers as intermediaries in the communication process? Do conventional means of recognising and rewarding research achievements militate against creative collaboration? Has pre-publication peer review ceased to have a useful function? These are non-trivial questions that need non-trivial responses.

Both individuals and institutions need to adapt. The recently published Accord on Open Data[4] sets out principles and responsibilities. It advocates a normative principle at the level of individuals:

> *"Publicly funded scientists have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They should make such data openly available to others as soon as possible after their production in ways that permit them to be re-used and re-purposed."*

and an operational principle that:

> *"The data that provide evidence for published scientific claims should be made concurrently and publicly available in an intelligently open form. This should permit the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations."*

A positive reaction to the Accord from the International Union of Crystallography[5] included an even stronger clarion call to action:

> *"We urge the worldwide community of scientists, whether publicly or privately funded, always to have the starting goal to divulge fully all data collected or generated in experiments."*

Such statements from the global research community about the open ethos of scientific inquiry, and its relevance to the need of humanity to use ideas freely, should be echoed by universities as part of their traditional role in preserving, re-assessing and creating knowledge and communicating it, in questioning received wisdom rather than blandly regurgitating it. They are also important in combating a countervailing trend towards the privatisation of knowledge, of which some universities are part, by succumbing to injunctions to see themselves largely as instruments of national wealth creation, where intellectual output is marketable property rather than public good. In contrast, the technologies at our fingertips have a key enabling potential for "open science", in which publicly funded science is done openly, its data are open to scrutiny, its results are available freely or at minimal cost, and results and their implications communicated more effectively to a wide range of stakeholders. Moreover scientific knowledge 'producers' should cease to think of knowledge 'users' as passive information receivers, or at best as contributors of data to analyses framed by scientists, but potentially as respected allies in the co-framing of issues and the co-production of actionable knowledge[6].

2 The Royal Society (2012), *Science as an open enterprise*. Royal Society: https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/; accessed 5 February 2017.

3 STAT: https://www.statnews.com/2016/08/10/data-sharing-science-nejm/; accessed 5 February 2017.

4 Science International 2015: Open Data in a Big Data World; available at www.science-international.org; accessed 5 February 2017.

5 ICUr: *Open Data in a Big Data World: A position paper for crystallography*; available at http://www.iucr.org/iucr/open-data, accessed 5 February 2017.

6 Hackmann, Heide and Boulton, Geoffrey: *Science for a sustainable and just world: a new framework for global Science policy?* UNESCO World Science Report 2015, pp. 12-14; available at UNESCO: http://unesdoc.unesco.org/images/0023/002354/235406e.pdf; accessed 5 February 2017.

11.6. INFRASTRUCTURES FOR OPEN DATA

Whilst universities must respond to these ethical challenges in their own ways, they must also respond to the need to manage their data in ways that they believe to best reflect their mission. Several years ago, rigorous data management was seen by many universities merely as a cost, as an "unfunded mandate". Increasing numbers of universities now see open data as a necessary part of their future and plan to position themselves to exploit the opportunities that it offers. Some of the essential principles of good research data management have now been established as a result of hard won experience[7] [8], many of which are shared in this volume. The "hard" infrastructure of high performance computing or cloud technologies and the software tools needed to acquire and manipulate data in these settings are only part of the problem. Much more problematic is the "soft" infrastructure of national policies, institutional relationships and practices, and incentives and capacities of individuals. For although science is an international enterprise, it is done within national systems of priorities, institutional roles and cultural practices, such that university policies and practices need to accommodate to their national environment. The iceberg figure reflects this (figure 11.1). The easy part is the visible part comprising the hardware and software tools required by a national open data system and any consents required for data use. Below the surface lie issues of process and organization. What is the ecology of the national research system? Do funders recognize and respond to the open data imperative? And is there adequate support for data management, data science advice and training? Then there are the people. Do they have the skills required to exploit the potential of the digital revolution? Are there incentives for researchers to make their data intelligently open? And does the mindset of a researcher accept the ethos of the first principle in the Accord?

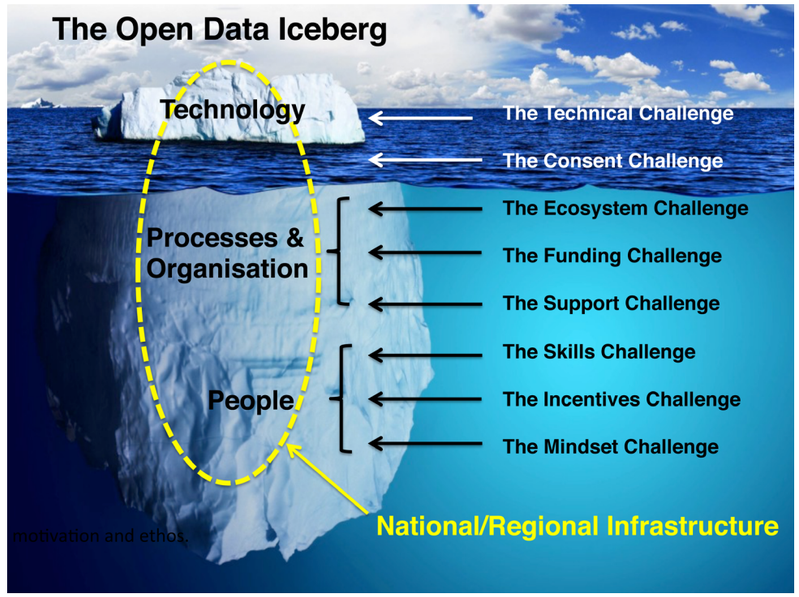

Figure 11.1

There are, however, important developments in support of open data beyond the confines of the university, with which universities can engage to their considerable benefit, if only to relieve themselves of the burden of being data management islands. "Open data platforms" are currently being developed where the needs of users are matched with hardware/software provision and data managerial skills, and created within individual disciplines (e.g. US National Institute of Health[9], Elixir-europe programme[10]) or multi-national geographic regions (e.g. Open Science Platform for Africa; Latin America and the Caribbean Platform; European Open Science Cloud).

11.7. A DECADAL VISION

Nearly two decades ago, Tim Berners-Lee proposed that datasets that relate to the same or related phenomena could be semantically linked in ways that integrate different perspectives,[11] and thereby offer much deeper understanding than merely using the web as a means of retrieving documents. Such a semantic web for science has the potential to integrate data from many sources to gain insight into complex relationships. It could, for example, be a means of integrating data from the natural and social sciences that are highly relevant to many complex global challenges; or of integrating data from the "internet of things", where almost any device with its own power source is able to acquire non-trivial information about its environment. Such a development is impeded by two barriers: a failure of many disciples to define their own vocabularies and ontologies, which impedes the efficiency with which they are able both to locate and use data relevant to their own discipline; and a failure to adhere to standards that enable inter-operability between disciplines. A strategic initiative is currently being launched by the International Council for Science's Committee on Data for Science and Technology, together with international science unions and associations, in the form of a Commission on Data Standards for Science to tackle these two major issues. It has great potential not only to enhance scientific understanding, but also the way that science is able to engage with the wider public in a more truly open science. This will require a major, decadal effort from across the science community, and could prove to be a profound step that will fundamentally change the way that science is done in the 21st century, through an unprecedented capacity to integrate data from disparate disciplines in ways that profoundly increase the potential of science to address major global challenges.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654139.

7 CODATA, 2015: *Current Best Practice for Research Data Management Policies*; available at http://dx.doi.org/10.5281/zenodo.27872; accessed 5 February 2017.

8 *LERU Roadmap for Research Data*: https://www.fosteropenscience.eu/content/leru-roadmap-research-data; accessed 5 February 2017.

9 NIH: https://www.ott.nih.gov/nih-ott-open-data-initiative; accessed 5 February 2017.

10 ELIXIR: www.Elixir-europe.org; accessed 5 February 2017.

11 Berners-Lee, Tim; Hendler, James; Lassila, Ora: 'The Semantic Web'. *Scientific American Magazine*, 17 May 2001; available at https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf; accessed 5 february 2017.