The open data imperative

The information revolution of recent decades is a world historical event that is changing the lives of individuals, societies and economies and with major implications for science, research and learning. It offers profound opportunities to explore phenomena that were hitherto beyond our power to resolve, and at the same time is undermining the process whereby concurrent publication of scientific concept and evidence (data) permitted scrutiny, replication and refutation and that has been the bedrock of scientific progress and of 'self-correction' since the inception of the first scientific journals in the 17th century. Open publication, release and sharing of data are vital habits that need to be redefined and redeveloped for the modern age by the research community if it is to exploit technological opportunities, maintain self-correction and maximize the contribution of research to human understanding and welfare.

The birth of modern science

GEOFFREY BOULTON

Regius Professor of Geology Emeritus
University of Edinburgh

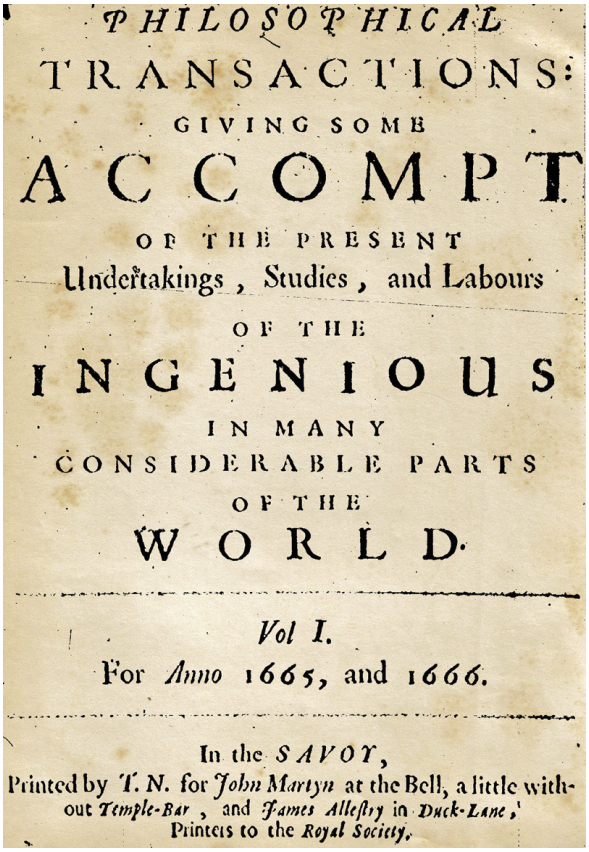
Open science is not new. Its principles were established in the early years of the European Enlightenment and have proved to be fundamental to modern science as the most reliable way of gaining knowledge. They were embodied by the first openly published scientific journals, which historians of science regard as having been germinal to the scientific revolutions of the 18th and 19th centuries.[1]

The first scientific journal, first published in 1665, and still being published, was the *Philosophical Transactions of the Royal Society* (see Figure 1). It was the brainchild of Henry Oldenburg, the first secretary of the newly formed Royal Society, and an inveterate correspondent on matters scientific. Rather than keep his correspondence private, he thought it would be a good idea to publish it, and persuaded the new Society to do so by creating the journal. He required of his correspondents that to be published, their concept must be accompanied by the evidence (the data) on which it was based. This permitted others to scrutinize the logical relationship between concept and evidence, replication of experiments and observations and reuse of the data. Such openness to scrutiny has proved to be the most powerful form of peer review, ultimately much more important than pre-publication peer review. It came to be seen as the basis of 'scientific self-correction', with openness to refutation as a key building block in the progress of science and the construction of scientific knowledge.

A technological revolution and its consequences

> "…openness to scrutiny has proved to be the most powerful form of peer review …"

Recent decades have seen a major technological revolution of historical proportions that has created an unprecedented explosion in the human capacity to acquire, store and manipulate vast volumes of data and information and to instantaneously communicate them globally, irrespective of location. It has produced fundamental changes in human, social and economic behaviour and has implications for research and learning that are, for example, far more profound and pervasive than those of the earlier, analogous revolution in human communication, that of Gutenberg's invention of the printing press. It is a revolution that has pervaded science and scholarship and the way that they are undertaken, but as in all such revolutions, many aspects of behaviour of both individuals and institutions remain adapted to an earlier technological era. The bound book and journal are still the basic tools of the trade of many of our libraries and of those that staff them. Both tend to be consolidated in single large



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

Figure 1. Henry Oldenburg, first secretary of the Royal Society, who launched the first and most enduring scientific journal, the *Philosophical Transactions of the Royal Society*, and the title page of its first volume. Oldenburg also invented 'peer review' by asking two Fellows of the Society to review submitted work and give him advice on whether it should be published.

edifices, even though in science at least, electronic access anywhere, any time, is the norm, dispersed support from appropriately trained e-librarians is the need, and few scientists now darken the door of a conventional library.

The data explosion and our capacity to combine, integrate and analyse data offer powerful new ways of unravelling complexity, improving forecasts of system behaviour and detecting patterns in phenomena that have hitherto been beyond our capacity to resolve. They offer the opportunity to reuse, to combine and to recombine data in ways that deepen these capacities. Exploiting these opportunities will depend upon access to and linking between many data sets, requiring that research data should be made routinely open and readily accessible. It will depend upon developing an ethos of data sharing and facilitating new modes of collaboration that increase the creativity of the scientific enterprise through interaction of many brains and many communities unbounded by institutional walls. These changes would also enable scientific concepts and the evidence that underlies them to be more effectively disseminated through society and in education, in ways that could change the social dynamics of science, contributing towards the evolution of science as a public enterprise rather than one conducted behind closed laboratory doors.

> "… many brains and many communities unbounded by institutional walls."

There is, however, a downside to the 'data explosion', of which we have only recently become aware. Such are the magnitudes of much of the data that provide the evidence for scientific concepts, that traditional habits of rigorous inclusion of data, and the metadata that describes their genesis, in conventionally published work have fallen away in recent decades. As a consequence, science may have been sleepwalking into a crisis of credibility. This was exemplified two years ago by a paper in which the authors reported attempts to replicate the results of 50 benchmark papers in pre-clinical oncology[2]. They succeeded in doing so in only 11% of cases. The failure in 89% of cases reflected in part failures of scientific logic, but in many it reflected the failure to include adequate data or metadata, such that even if the conclusions had been logically

sound, there would have been no means of verifying them. There is now an epidemic of non-reproduceability[3] that is not 'traditional' non-reproduceability because of a faulty experiment, faulty observation or the failure of statistical logic. It is failure because of inadequate data and/or metadata, which denies access to the evidence so that an argument can neither be validated nor invalidated. It undermines the vital principle of scientific self-correction.

Reinventing open science for the 21st century

"Openness … has no value unless it is … 'intelligent openness'"

The open publication, release and sharing of data are vital habits that need to be redefined and redeveloped for the modern age by the research community if it is to exploit technological opportunities, maintain self-correction and maximize the contribution of research to human understanding and welfare. If we are to adapt the norms of research to a new data-rich age, exploit the opportunities and correct the detachment between data and concept described above, it is important that the research community understands the imperative for open data and adopts it as a normative principle.

Openness of itself, however, has no value unless it is what the Royal Society report on open science[4] terms 'intelligent openness'. This means that data and the relevant metadata and computer code that provide the evidence for a published paper must be concurrently available for scrutiny and must be:

- *discoverable* – readily found to exist by online search
- *accessible* – when discovered they can be interrogated
- i*ntelligible* – they can be understood
- a*ssessable* – their provenance and reliability can be assessed
- *reusable* – they can be reused and recombined with other data.

The data generated by publicly or charitably funded research that is not used as evidence for a published scientific concept should also be made intelligently open after a pre-specified period in which originators have exclusive access. Those who reuse data but were not their orginators must formally acknowledge originators. In understanding the logic of these arguments, it is also important to avoid a false dichotomy between doing science and publishing a paper on the one hand and making the data intelligently open on the other. The cost of creating intelligently open data from a research project is an intrinsic part of the scientific process, not an optional extra and should be funded as intrinsic to the research.

Although the default position for data generated by publicly or charitably funded research should be one of 'intelligent openness', there are justifiable limits to openness. These are where commercial exploitation is in the public interest and the sectoral business model requires limitations on openness; in preserving the privacy of individuals whose personal information is contained in databases; where data release would endanger safety (unintended accidents) or security (deliberate attack). However, these instances do not provide justification for blanket exceptions to the default position, and should be argued on a case-by-case basis.

Implementing principles of open research data

It is important that these principles are adopted by funders of research as conditions for continuing support, by universities and institutes as principles of modern research, by publishers as conditions for accepting work for publication. They should also be advocated by the learned societies that articulate the principles and priorities of their disciplines to the research communities that they represent.

Although the principles articulated here should be normative for scientists, major change will hinge on the extent to which researchers see value in acquiescing to them. The principles are most likely to be observed if there are incentives for researchers, their institutions and for users. Adoption by researchers of citation procedures for deposited data could be a simple but powerful incentive for change, in giving them a citable product in addition to a conventional paper. Indeed, there is evidence[5] that in disciplines where data citation has become the norm, the frequency of citation of a valuable data set can far exceed that for the first conventional paper interpreting the data. This is understandable given the creativity involved in devising observations or experiments that reveal or confirm important relationships or phenomena, which can be at least as great as in interpreting the data and writing a paper. It deserves to be given equivalent credit. A published paper could be regarded as an advertisement for the science that is embedded in the data.

"… the frequency of citation of a valuable data set can far exceed that for the first conventional paper interpreting the data."

The impact of digital technologies is not restricted to science, but creates opportunities for the whole range of research and scholarship. In the 'digital humanities' for example, research often entails new methodologies and intellectual strategies that are nonetheless grounded in traditional humanistic foci. The challenge is not only to the use of data that are born digital, but also to large bodies of text, as well as visual, aural, audiovisual, sensory, neurological and even kinaesthetic forms of information.

National and international priorities and trends

There is much to be said for a decisive move to adopt the well-established DataCite[6] process in the UK research system, both as an important process in its own right and as a means of engaging the enthusiasm of the research community. DataCite is a global network (of which the British Library is the UK node) that works to increase the recognition of data as a legitimate, citable contribution to the scholarly record. It provides digital object identifiers (DOIs) for data sets and other non-traditional research outputs, which helps to make data persistently identifiable and citable. Existing processes, reward structures and norms of behaviour that inhibit or prevent data sharing or new forms of open collaboration should, wherever possible, be reformed so that data sharing and collaboration are encouraged, facilitated and rewarded.

A UK Open Research Data Forum was recently convened by the Royal Society as a follow-up to its 2012 report, *Science as an Open Enterprise*. The Forum has decided to maintain itself as a 'ginger group' involving all the essential elements of the UK research community[7] to promote and implement concepts such as those in this article. Although there has been a strong stance on this issue by the UK Government, which has set up a Research Sector Transparency Board[8] to promote and monitor open research data (as part of its open data agenda), research is an inherently international activity and its norms need to be international and adopted internationally. A first step in this was the 2013 statement by G8 science ministers on the importance of open research data, and their agreement to promote it. Other international means of doing so within the science community itself should be through the Global Research Council[9], an appropriate body to agree and promote the principles of open data, and through CODATA[10], a commission of the International Council for Science[11], which is a technically focused body able to identify and promote the open standards that will be necessary in an effective international open data regime. As a bilateral effort to achieve international coherence, a delegation from the UK Open Research Data Forum met with its US counterpart, the Committee on Coherence at Scale for Higher Education[12] , in Washington in April this year to explore ways in which the two communities could work effectively together on the open data issue.

"… research is an inherently international activity …"

However, the greatest influence on the creation of an international open data environment will come from the persuasive impact within the scientific community of scientists in disciplines such as bioinformatics, chemical crystallography, and broad themes such as public health[13], that have already recognized the great value of an international open data regime both to individual scientists and the communal scientific effort.

Open science?

"Our aspiration should be for science to be a public enterprise …"

This article has been concerned with open data, but what of 'open science'? Open science comprises three elements: doing science openly in the way that research priorities are defined and data are collected; releasing intelligently open data, whether those data are held by public authorities, or created by public sector research institutes or universities whose research is supported by public funds; and open access publishing. Open science defined in this way is important for two reasons. First, the impact of science on modern life is so profound that open scrutiny of the evidence that underlies scientific conclusions is a prerequisite for functioning democracies that reflect the choices and mores of their citizens. Second, both the integrity and the efficiency of the scientific process depend on intelligent openness and the speed and rigour with which scientific reasoning is scrutinized, errors identified and new theories put to the test. Our aspiration should be for science to be a public enterprise, not one conducted behind closed laboratory doors. The principle of open data is the bedrock on which such an enterprise rests.

Acknowledgement

Many of the concepts on which this essay is based were developed during work on the Royal Society Report: *Science as an Open Enterprise* (The Royal Society Science Policy Centre report 2/12 June 2012) and on discussions within the UK Open Research Data Forum. The contributions of my fellow authors, members of the Royal Society Science Policy Centre, particularly the assiduous work and perceptions of Caroline Dynes, and members of the Forum are gratefully acknowledged, though any failures of logic or errors of substance are entirely mine.

References

1. Shapin, S, *A social history of truth: civility and science in seventeenth-century England*, 1994, Chicago, University of Chicago Press.
2. Begley, C E, and Ellis, L M, Raise standards for preclinical cancer research, *Nature*, 2012, 483, 533.
3. PubMed Retraction Notices – By Year (2012): http://pmretract.heroku.com/byyear (accessed 15 May 2014).
4. The Royal Society, *Science as an Open Enterprise*. Science Policy Centre publication 2 December 2012, London, Royal Society.
5. Piwowar, H, A, Day R, S, and Fridsma, D, B, Sharing detailed data is associated with increased citation rate, 2007, *PLOS ONE*, 2, 3, e308.
6. DataCite: http://www.datacite.org (accessed 15 May 2014).
7. Royal Society UK Open Data Research Forum: https://royalsociety.org/events/2014/01/open-data-forum/ (accessed 15 May 2014).
8. UK Government Research Sector Transparency Board: https://www.gov.uk/government/groups/research-sector-transparency-board (accessed 15 May 2014).
9. Global Research Council: http://www.globalresearchcouncil.org (accessed 15 May 2014).
10. CODATA: http://www.codata.org (accessed 15 May 2014).
11. International Council for Science: http://www.icsu.org (accessed 15 May 2014).
12. US Committee on Coherence at Scale for Higher Education: http://coherence.clir.org (accessed 15 May 2014).
13. Wellcome Trust 'Spotlight issue': http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues (accessed 15 May 2014).

Professor Geoffrey Boulton FRS
Regius Professor of Geology Emeritus, University of Edinburgh
School of Geoscience, Grant Institute, Kings Buildings, Edinburgh EH9 3JW, UK
Chair, Royal Society report on Science as an Open Enterprise
E-mail: G.Boulton@ed.ac.uk

To cite this article:

Boulton, G, The open data imperative, *Insights*, 2014, 27(2), 133–138; DOI: http://dx.doi.org/10.1629/2048-7754.148