SCIENCE AS A PUBLIC ENTERPRISE: OPENING UP SCIENTIFIC DATA

Hovedforedrag på fagsymposium på Universitetsmuseet i Bergen den 30. november 2011

ved professor Geoffrey Boulton, University of Edinburgh, UK

Professor Øyvind Østerud, preses i Det Norske Videnskaps-Akademi, holdt åpningstalen.

Kommentarer ved professor emeritus Ole Didrik Lærum, Universitetet i Bergen, professor Truls Norby, Universitetet i Oslo og professor Inger Sandlie, Universitetet i Oslo. Dette ble etterfulgt av en debatt.

Seminaret kom i stand gjennom et samarbeid mellom Royal Society i London, Det Norske Videnskaps-Akademi og Universitetet i Bergen.

The first "open science" revolution

Henry Oldenburg, a German theologian, who became the first Secretary of the newly formed Royal Society in London in 1661, was an inveterate correspondent. He communicated with the leading scientists of Europe, believing that rather than waiting for entire books to be published, letters were much better suited to quick communication of new facts or discoveries. He invited people to write to him with ideas about science, even laymen who were not involved with science but had discovered some item of knowledge. It was

Geoffrey Boulton.
Foto: Guri Gunnes Oppegård

from these letters that the idea of printing scientific "papers" or "articles" in a scientific journal was born, which could be in any vernacular language rather than needing to be in Latin, the traditional language of scholarship. In creating the Philosophical Transactions of the Royal Society in 1665, he wrote:

> *"It is therefore thought fit to employ the [printing] press, as the most proper way to gratify those [who] ... delight in the advancement of Learning and profitable Discoveries [and who are] invited and encouraged to search, try, and find out new things, impart their knowledge to one another, and contribute what they can to the Grand Design of improving Natural Knowledge ... for the Glory of God ... and the Universal Good of Mankind."*

Oldenburg also initiated the process of peer review of submissions by asking three of the Society's Fellows who had more knowledge of the matters in question than he to comment on submissions prior to making the decision about whether to publish.

Historians such as Shapin[1] have argued that this development, and the advent of other journals that were to follow, made vital contributions to the explosion of scientific knowledge that occurred in the seventeenth and eighteenth centuries. They permitted ideas and measurements to be more readily corroborated, invalidated or improved. They communicated the results of scientific inquiry to an audience beyond the gentlemanly clique through which science had hitherto developed, and that was in turn enabled to contribute further ideas and observations to its subsequent development. Since then, open publication of scientific ideas and the data on which they are based has been the bedrock on which the scientific edifice has been built. It permits the logic of a published argument to be scrutinised, the underlying data to be replicated or invalidated, and the resultant hypotheses to be tested and either confirmed, discredited or improved. Such openness deters and exposes error, poor practise or fraud, and is essential to the self-correcting capacity of science, the source of its greatest strength as a route to trustworthy knowledge.

1. See Stephen Shapin, *A social history of truth : civility and science in seventeenth-century England*, Chicago : University of Chicago Press, 1994.

The practice and power of open science

These processes are the staple of conventional ways of working and of the means by which scientists' contributions are recognised, their careers progressed and the credibility of their work judged. The tempo of a research career is one of research grants applied for, grants allocated, work done, papers submitted for publication, and citations of published papers recorded, with cycles of proposal and publication overlapping each other in time. Although the tempo of an individual's work and the progress of their careers is generally slow, more open, collaborative modes of behaviour can stimulate remarkably rapid achievement. This was strikingly illustrated in 2011 during an outbreak of gastro-intestinal infection in Hamburg in Germany, which spread to affect about 4000 people in Europe and North America, and resulted in over 50 deaths. Samples were made available by a Hamburg laboratory and scientists at BGI-Shenzen in China, working together with those in Hamburg, analysed the strain and three days after the outbreak, released a draft genome under an open data licence. This openness and the communication of results by scientists on four continents created new knowledge about the strain's virulence and resistance genes that rapidly helped control a public health emergency. It was example of open science at its best.

In contrast, data and information hoarding can be a severe brake on progress. An extreme example being that of the Chamberlen family, who kept the discovery of the obstetrics forceps secret for more than 100 years in order to protect their midwifery business.[2]

A way in which the collective intelligence of the scientific community can be mobilised through open collaboration was illustrated in 2009 when Tim Gowers, a Fields medallist mathematician, set out a difficult and unsolved mathematical problem on his blog, together with his ideas about it and an invitation for others to contribute to its solution. Comments erupted. Within 37 days, 27 people made more than 800 substantive comments, with tentative ideas being rapidly developed or discarded. Together they solved not only the core problem, but a harder generalization of it. Gowers commented that the process "had seemed like driving a car whereas ordinary research was like pushing it".

2. Moore, Wendy "Keeping mum," *BMJ* 334, no. 7595 (March 31, 2007): 698-a, doi:10.1136/bmj.39157.514815.47

Why Change? A second open science revolution?

The last half-century has seen a breath-taking increase in scientific discovery. We have put satellites into orbit and probed the universe; we have discovered the chemical structure of living organisms and learned to manipulate it; we have been able to read Earth's history in minute detail from ice sheet and ocean cores; we have improved human and animal health through increasingly large epidemiological studies; and we are able to watch football matches, as they happen, on the far side of the planet. Many of these achievements have been based on new ways of collecting, storing, manipulating and transmitting data and information in ways that far surpass anything previously dreamed of.

But the information and communications technologies that have underpinned these developments have also created problems and challenges for science that make some of the conventional norms of scientific process both problematic and limiting. Computational and data storage technologies threaten the crucial principle of scientific openness and reproducibility, but they also offer the means, if appropriately deployed, of solving them.

Publishing results and the maintenance of self-correction

The norm of, say, 30–40 years ago, was to publish a paper in which the argument was routinely accompanied by a complete description of the data on which it depended together with the experimental method, an assessment of uncertainties and details of the metadata required to validate, repeat or reuse the data. This pattern has now been fundamentally disrupted in many areas of science, where new tools for data acquisition and digital storage and manipulation have created a data deluge that is so great and so complex that no print journal could reproduce it. Data has become detached from the published scientific conclusions that depend upon it, such that the two vital complementary components of the scientific endeavour – the idea and the evidence – are too frequently pulled apart. If the principle that the data underlying a scientific argument must be accessible for rigorous analysis and replication is to be maintained, ways must be found to knit these two components together again.

Although digital technologies have created the problem, they also carry the solution. The objective must be to have the data lodged in a publicly accessible, curated database, ideally accessed by the click of a mouse on a

live link in the published paper. It is a process that is increasingly mandated by scientific journals, although the compliance rate is still low, but increasing.

But it is possible to go much further. The books and journals on library shelves are immutable, fixed by the date of publication. Many modern databases are now dynamic, able to be automatically up-dated, corrected and refined as new data is acquired. Live links to such data in electronic articles not only permit an evolving database to be interrogated but also permit the reader to manipulate the data whilst reading the article. The article needs no longer be fossilised on the day of publication, but can be part of a flexible and evolving research package. We should strive for a state where all of the science literature is online, all of the data is online and where the two interoperate.

Simulation of complexity

Computer simulation has become an essential tool of modern science. It is argued that it represents a third paradigm of science, to add to the classical duo of theory and experiment. It has permitted for the first time the behaviour of truly complex systems to be analysed in a rigorous fashion. The simulation is analogous to a physical experiment, but one conducted with mathematical equations rather than physical entities, and therefore with the capacity to undertake experiments that nature does not permit. The output of the simulation is analogous to the output of a physical experiment, and can be regarded as data, though not to be confused with the original physical measurement data on which most models rely. It is therefore important that simulations that support the arguments in published papers or reports are accessible for scrutiny for exactly the same reasons as argued for more conventional data. We need to be able to scrutinise the workings of the simulation, as exemplified by its code – indeed the science often lies in the code with the published paper merely an advert for the science – we need access to the output, or the means to recreate it.

Openness to whom? The challenge of citizen engagement

The importance of science in the modern world is now so great that citizens need to make informed judgements about it if they are to exercise democratic oversight of decisions made in their name. Science must be, and

be seen to be, a public enterprise, rather than a private one conducted behind the closed laboratory doors. Growing numbers are averse to accepting *ex cathedra* statements from scientists, whilst ubiquitous digital media offer a powerful means for the public to interrogate, question and re-analyse scientific priorities, evidence and conclusions. The scientific community must adapt and respond to such legitimate demands from citizens. After all, they are the ones who pay for public science.

A necessary pre-requisite for public communication of scientific knowledge is that it should not simply disclose conclusions, but must communicate the reasoning and evidence that underlie them. There is however a difficult dilemma. Whereas a major principle of science is "don't trust anything except the evidence", many scientific analyses of issues that are of public importance or concern require very high levels of expertise. Consequently, trust in science and the processes of science are in fact necessary if informed, democratic consent is to be gained for public policies that depend on difficult or uncertain science.

Openness to the public must be audience-sensitive and recognize a diversity of demands from citizens. We should also recognise an increasingly numerous body of engaged "citizen scientists" that wish to dig deeply into the scientific data relating to a particular issue, and that is developing an increasingly powerful "digital voice". Some are highly sceptical of research conclusions in issues that interest them, often asking tough and illuminating questions. Others have effectively become members of particular scientific communities, for example in astronomy, proteomics and climate science, by dint of their rigorous and valuable observations and measurements, and are formally involved in scientific projects. The growth of this citizen science movement could be a major shift in the social dynamics of science; blurring the professional/amateur divide and changing the nature of the public engagement with science.

A fourth paradigm: data-intensive science

The power of modern computers permits very large datasets to be explored in ways that yield inherent but unsuspected relationships, with hypotheses being constructed after identifying relationships in a dataset. It is an approach that has permeated several disciplines, most prominently to date in the life sciences where bio-informatics seeks patterns in molecular biological data, in contrast to computational biology, which simulates how bio-

logical systems work, and where both together are being used in pioneering ways to more efficient processes of drug discovery. It has been argued that this "*data-led science*" is a fourth paradigm for science, a new way of using data that can dramatically enhance progress in data-rich areas. Such has been its impact in the life sciences that the community curated *Metabase* lists more than 2000 biological databases ranging in content from species distributions and migration patterns to genetics and proteomics, and which have the potential to be linked up to create an integrated mapping of biological systems.

Where specific databases have live links to other, cognate databases, there opens up the possibility of identifying much deeper relationships between phenomena. Two decades on from the creation of the web, technologies are emerging that focus not on harvesting pages of information on a topic but on linking data from different sources that promise a deeper integration of data to create new information. Such linked data technologies also provide a means of making greater sense of the data deluge by assessing and comparing an ever-increasing number of data sets and identifying ways of re-using them.

Many areas of basic science are lagging behind their commercial peers in deducing meaning from data. An example of its power is the way that the Google search engine, when linked to public health data, can give much earlier warning than conventional reporting procedures of the seasonal flu epidemics that annually cause up to half a million deaths worldwide.

Making it happen

Exploiting the potential of these technologies and seizing the opportunities that they offer will require four things. Firstly the recognition that greater benefit is to be derived from opening and sharing data than in maintaining them as a private preserve. Secondly that researchers observe a set of shared principles and standards for permanent data identifiers, for links to other relevant data sources and metadata about linkage methods and data quality and context. Thirdly it is important that a cohort of "data scientists" is developed that will be crucial to successful management of digital data systems. They need to be mathematically adept and trained in informatics disciplines. If the science library is to evolve from a store of immutable printed texts to a flexible dynamic resource, they will be the librarians. And fourthly, those that employ scientists, those that fund them, those that pub-

lish their work, the learned societies that articulate the priorities and values of disciplines, and most of all scientists themselves, need to adapt the way they pursue their enterprises.

The boundaries of openness

But for many, myself included, there are boundaries to openness related to legitimate commercial interests, privacy and security, all of which pose difficult technical and ethical or existential problems. It is in the public interest of states that fund research to derive economic benefit from it. Many business models depend upon commercial data confidentiality, and although it is too simplistic to think that the open/confidential boundary coincides with the boundary between public and private funding of research, the boundary is a complex and fuzzy one. Research in the social and health sciences often depends heavily on the use of personal data. But how should we balance personal privacy against the public good, in the diagnosis of some diseases for example, when we now know that it is mathematically impossible to anonymise personal data? Similarly, how do we balance the danger of bio-terrorism that might arise from knowledge of processes by which the H5N1 bird flu virus might cross the species boundary to humans against the benefits to public health that might flow from understanding how this process might take place?

Such issues have been at the heart of a Royal Society study into open data and open science which is intended for publication later this year. It is an issue that is increasingly engaging the attention of the scientific world. European academies will discuss the possibility of a joint statement on open data later this month. It is an important issue for the International Council for Science. Discussions have taken place with the US and Chinese Academies of Science, and a joint Royal Society/Norsk Videnskaps-Akademi seminar on the topic took place in November 2011. The difficulty of promoting processes that will support and not fossilise exciting new developments should not however be underestimated. We should remember Sydney Brenner's comment, that "a modern computer hovers between the obsolescent and the non-existent!"

Geoffrey Boulton is Regius Professor Emeritus of Geology at the University of Edinburgh and Chair of the Royal Society's Work Group on "Science as an Open Enterprise". He also chairs its science policy committee.