THE FOUNDATION
FOR SCIENCE AND
TECHNOLOGY

DEBATE AND ROUND-TABLE DISCUSSION SUMMARY

Maximising the use of public data – should research and publically acquired data be made more accessible?

Held at The Open Data Institute and The Royal Society on 10th July, 2013

The Foundation is grateful for the support for this meeting from
the Financial Services Knowledge Transfer Network, Lloyd's of London and The Michael John Trust

The hash tag for this debate is #fstopendata .

Chair:	**The Earl of Selborne GBE FRS** Chairman, The Foundation for Science and Technology
Speakers: **(Evening)**	**Professor Geoffrey Boulton OBE FRS FRSE** Chair, The Royal Society Inquiry into Science as an Open Enterprise **Professor Sir Nigel Shadbolt FREng** Chairman and Co-Founder of the Open Data Institute **The Rt Hon David Willetts MP** Minister of State for Universities and Science, Department for Business, Innovation and Skills
Panellist:	**Professor Sheila M Bird OBE FRSE** Programme Leader, MRC Biostatistics Unit, Institute of Public Health, Cambridge

PROFESSOR GEOFFREY BOULTON summarised the conclusions of the inquiry he chaired for The Royal Society – Science as an open enterprise: open data for open science[1]. He began the evening by talking about the process by which science is done and the changes in this process consequent on the availability of much large volumes of data.

He referred to Henry Oldenburg, first Secretary of The Royal Society, who in the 1660s encouraged the Royal Society to published his extensive scientific correspondence in the vernacular (not in Latin), so that scientific work could be made available widely. He argued that concepts needed support from evidence and data and this data must be made available. This approach supported the scientific revolution in the next few hundred years. We need the same sort of revolution now that very large volumes of data of all sorts are available.

The past ideal for scientific papers was that they gave sufficient information for the results to be replicated by another researcher, so that they could confirm, or perhaps deny, the results that were claimed. Only a minority of papers nowadays are reproducible, as studies have shown. No or insufficient data may be provided or the "metadata" – the description of the data – may be unsatisfactory or missing.

Yet there are huge amounts of data, with all sorts of interesting linkages. How they can be integrated is a challenge.

There may also be a change in the scientific approach: in the past one proposed a hypothesis, and then collected data to support or deny that hypothesis. Now one may need to start with the data, and see what it tells us. But this requires good "informatics" to access and process that data, and good statistics to analyse it.

We need to persuade scientists that it is in their interests to share data rather than hug it to their own chests. But there are examples of successes. After a recent outbreak of e-coli in Hamburg 20 laboratories in four continents analysed the genetic origins with rapid and great success. An unsolved mathematical problem posed by Sir Tim Gowers was solved by contributions from

[1] http://royalsociety.org/policy/projects/science-public-enterprise/report/

about 25 people in about 30 days after he posed the problem on his blog site[2].

There is a problem, however, in that researchers look for credit for their own work. How does none credit collaborative research work?

Another element coming from the availability of open data is for interested amateurs to make their own contributions, the concept of "Citizen Science".

There have been examples recently of scientific fraud, the invention of data, the cherry-picking of favourable data, the non-publication of refutations of previously claimed results. It should be considered malpractice to not make available data that supports any claim.

We need, however, intelligent openness, which requires the data and its metadata to be accessible, intelligible, assessable and reusable.

But openness must be sensitive also to public opinion. There are boundaries to openness, though these boundaries are fuzzy. People do not expect their medical records or their tax returns to be publicly available, and it may be impossible to guarantee anonymity in a supposedly anonymised data set.

Changes in mindsets may be necessary. Scientists may need to change to a more open approach, rather than keeping their data private. Learned societies may need to influence their members toward this more open mindset. Universities may need to recognise and accept the cost of open data. Funders may need to accept the costs. Publishers may need to be persuaded that is an advantage to make the research data supporting published papers available.

Recent actions have moved in the right direction: a statement by the G8 at the June 2012 summit, the creation of the Research Data Alliance (funded by the US, Australia and the EU) and of the UK Research Sector Transparency Board. The Royal Society is also considering setting up an open data Forum of some kind.

Three main steps are needed:

(a) Doing science openly
(b) Making open data available, and
(c) Providing open access to publications

The aspirations should be that all scientific literature should be online, preferably freely available, and that all data, whether supporting the publication or not, should, within limits also be available online.

PROFESSOR NIGEL SHADBOLT spoke next on "The Power of Open: the fifth paradigm". He started by giving examples of how openness had contributed greatly to responding to societal challenges. In Haiti after the earthquake in 2010, many individuals collaborated in real time and on the spot to produce detailed post-earthquake maps to help the emergency services. In Kenya the real-time reporting of acts of violence around the recent election helped to contain the violence. In the past the statistical work of Florence Nightingale reduced the number of deaths from disease in the Crimean War and John Snow's map of a cholera outbreak in Soho had identified the source.

It was now recognised that the open publication of the human genome sequence was a public good. The world-wide web would not have been so successful if the protocols had not been made public. The US President had realised that making accurate GPS positioning available was of wide public benefit.

Openness comes in many forms:

Open licences
Open source software
Open standards
Open participation
Open data

All this leads to Open Innovation.

Open data is information that is available for anyone to use, for any purpose, at no cost and should have a licence that says it is open data. Without such a licence the data may not be able to be reused.

He and others had initiated the open approach by setting up a data set based on public data indexed by postcode. When the government was very pleased with it, it was pointed out that most of the data was unlicensed. This resulted in a transformation

[2] http://gowers.wordpress.com/2009/01/27/is-massively-collaborative-mathematics-possible/

of government attitudes towards the release of publically funded data and the setting up of a public data web portal, www.data.gov.uk, which is now flourishing.

There are great advantages in having a national data infrastructure, including mapping, addressing, transport, education, health, environment, science, etc. This all results in good governance, affecting the political, economic, social, research, media and data fields.

A further example was the analysis of a NHS data base that showed doctor's prescribing habits by local area. In certain areas doctors were very much more likely to prescribe proprietary statins rather than the cheaper generic varieties. If all areas had prescribed the cheaper variety, where appropriate, the NHS might have saved about £200 million in a year.

There were many varieties of data sets:

Big or small
Public or private
Open or closed
Personal or non-personal
Anonymous or identified
Transient or patrimonial
Exhaust or core task

Open data presented many challenges. One had to set up a satisfactory infrastructure. One had to consider the quality of the data. Users had to have sufficient data literacy. And there were questions of security and privacy.

This last had its problems. The public in the United States seemed much more sensitive about this. There was, for example, a conflict about publishind research of how a virus could cross species boundaries, which would allow counter-measures to be developed world-wide, but might also allow terrorists to develop a new threat. The publically available recording of ship positions helped shipping safety and insurers, but also helped Somali pirates.

Some experts felt that others would not understand data and could misinterpret analysis. The incumbents felt that their ownership was being challenged. Legislation was not necessarily up to date. And some government data, that should be open was still not made accessible.

Science has progressed through four paradigms; experimental observation, coherent theories drawing on such observations, large scale simulations using computer power, in depth analysis of very large data sets. The next step was to combine the data with the power of everyone to create knowledge for the benefit of society – the fifth paradigm.

The prime example of the firth paradigm is the story of Jack Andraka, a sixteen year old US student, who had studied the literature and data so thoroughly on the web that he had been able to develop a very simple test for the early detection of pancreatic, ovarian and lung cancers.

THE RT HON DAVID WILLETS said that he wished to discuss three points: first, open access to publicly funded research; secondly the provision of the data on which the research was based; and finally the ability of any member of the public being able to contribute to the data and to 'citizen science'.

The public has a right of open access to the findings of any publicly funded research. Much research funded by taxation still had a pay-wall to restrict access to it. The ideal public research programme should have a "gold" target, so that the costs of open access to publication were included in the initial costings and recovered from research grants. The lesser "green" target permitted publishers to recoup their costs by charging, but only for a limited time period. In his view green access was not good enough, even with a short time period.

Work was to being done to negotiate licences for walk-in library access; and also set up a 'gateway to research', a one-stop web site to allow easy access to all publicly funded research and their results. After the first portal was in place, others could produce their own 'apps' to allow different forms of access. A further aspect was promoting 'knowledge exchange'

But, tensions within universities existed. Some universities appeared reluctant to pay to publish research, but this seemed to reflect a conflict between university management and university researchers.

But, behind every research paper published there is a mass of data available, and at some point that data set should be made available, especially if it is publicly funded.

However, one had to be sure that a researcher had to have the first shot at analysing the data they had collected. Some feared that the Freedom of Information Act could require researchers to release their data before publication, but an exemption currently available in Scotland was being extended to the rest of the United Kingdom, to protect the original researchers.

There needed also to be sufficiently high performance computers to allow access to very large data sets, such as were being developed at Edinburgh or Manchester universities.

The privacy regime needed to be right. Whilst individuals could properly opt out of research based on their own medical data, this could cause problems if it was desirable, for example, to investigate whether there was an unusual density of cancer near high voltage power lines. It was not very practicable to get the permission of every resident, and a sample might well be biased.

The security issues and US concern about 'dual use' of data were concerns.

Modern librarianship needed to adapt to this mass of open data, and this was among the subjects discussed by the Research Sector Transparency Board.

Finally, individuals could contribute a great deal, and any ideas about how to help this would be welcomed by the Minister.

PROFESSOR SHEILA BIRD[3], spoke about the good work done by Professors Boulton and Shadbolt and The Royal Society's Inquiry, which had focussed on the 'data storm' that was approaching, also on replicability and credibility; giving credit for collaboration, but warning about the problems of big data.

Much data was not created by scientists but was created by administrators.

Sir Nigel's five paradigms did not necessarily come in his given order, and one may need to go back to test conclusions for big data. He also pointed out that publishing and analysing data could allow the quality of that data to be improved.

One must be careful about any survey with a too low response rate; she would like to see returns of more than 60%. Medical data are different, because most tests are done for the treatment of the patient, who gives consent, not knowing the results. Because there is a strong duty of confidentiality one must then be careful about how this data is used for research. Social science surveys are not the same, because there is no pressure on respondents to reply except for the purpose of the research, and false information does the respondent no harm. Scientific method has nothing to fear from open data. Subjects of studies should also be given access to the protocol for any investigation for which they have given informed consent. Costs of longitudinal studies have increased, and record linkage is a useful approach. But in England & Wales there are long delays in the registration of deaths subject to an inquest which delay epidemiological studies.

Many points that were raised in the questions and workshop discussion confirmed and elaborated on what had already been said by the speakers. Some new point are organised below under separate headings.

What practical difficulties are there?

There were costs in arranging any access to any open data set, but if any data set needed to be regularly updated, as some might usefully be, that could add considerably to the costs.

Having a great deal of data available makes more demands on informatics and on statistics. Within the academic field more credit may be given for new advances in mathematical statistics rather than in the good analysis of new data; this can deter academics from practical statistics. Also, there is a shortage of academic statisticians. Many newly qualified statisticians go into business, especially the financial sector, and not research.

A wider breadth of computer skills may be needed to handle very large data sets. This is a challenge for those who profess the recent subject of informatics. Librarians too may need to improve their statistical skills.

What are the boundaries for open data?

It is difficult to give full details in an anonymised data set without making it

[3] Professor Bird spoke before the Minister arrived.

possible to identify persons individually one way or another. This created a problem for analyses carried out by the census office. Should the office publish a less detailed data set and use that for all analyses, or could the census office use the full data set for its analyses, which would mean that it could not be independently and exactly confirmed?

Many private companies had data which they were not prepared to share. There might sometimes be good commercial reasons for this, but there might be public benefit in aggregate data being made available. Secrecy varied from industry to industry. Although general insurance companies were very protective of their data, Lloyd's had organised the collection of data on catastrophes, which might affect the whole market; so sometime the private sector was able to work collectively towards a common goal.

Other points

Many speakers made points about specific cases: astronomy and meteorology were obvious examples of successful sharing of data across countries. Crystallography was also mentioned. Release of flood extent data meant flood risks could now be analysed more thoroughly, which could give important insights to householders, planners and insurers when considering the response to future flood events.

It was suggested that museums could be useful contributors to open science, but it was pointed out that there was distinction between a museum's collection of tangible objects and a collection of more intangible data.

A statistical problem was that a statistical method like multiple regression can give very good answers overall, but may be rather unreliable at the edges where there are small numbers of cases. A statistician may recognise that standard errors in some areas are much larger than they are in others, but explaining this varying level of uncertainty to the general public, in particular to journalists, can be difficult.

The press likes to emphasise the errors of politicians, and the successes of scientists. A scientific paper that showed that some possible effect was not true was not news. But large data sets might be able to show that what had been taken as an apparently significant result based on a rather small sample, did not apply more generally.

Professor David Wilkie, CBE

TEDx Talk:

Michael Nielsen on Open Science
www.youtube.com/watch?feature=player_embedded&v=DnWocYKqvhw

Useful URLs:

Jack Andraka
http://jackandraka.net

Department for Business, Innovation and Skills
www.bis.gov.uk

Economic and Social Research Council
www.esrc.ac.uk

Financial Services Knowledge Transfer Network
https://connect.innovateuk.org/web/financialservicesktn/overview/-/asset_publisher/ghKKWLt6630Q/content/contact-us

The Foundation for Science and Technology
www.foundation.org.uk

Government Digital Service
http://digital.cabinetoffice.gov.uk

Government Information Economy Strategy
www.gov.uk/government/uploads/system/uploads/attachment_data/file/206944/13-901-information-economy-strategy.pdf

Lloyd's of London
www.lloyds.com

MRC Biostatistics Unit, Institute of Health, Cambridge
www.mrc-bsu.cam.ac.uk

The Open Data Institute
www.theodi.org

Open Knowledge Foundation – G8 Open Data Statement
http://blog.okfn.org/2013/06/14/g8-science-ministers-support-open-data-in-science/

Research Councils UK
www.rcuk.ac.uk

Research Data Alliance
https://rd-alliance.org/node

Research Sector Transparency Board
www.gov.uk/government/policy-advisory-groups/research-sector-transparency-board

The Royal Society
www.royalsociety.org

The Royal Society Inquiry into Science as an Open Enterprise
www.royalsociety.org/policy/projects/science-public-enterprise/report/

Technology Strategy Board
www.innovateuk.org/

The Foundation for Science and Technology
A Company Limited by Guarantee
Registered in England No: 1327814
Registered Charity No: 274727