Science as an open enterprise

June 2012

THE
ROYAL
SOCIETY

Science as an open enterprise
The Royal Society Science Policy Centre report 02/12

Issued: June 2012 DES24782

ISBN: 978-0-85403-962-3

Requests to reproduce all or part of this document should be submitted to:

The Royal Society
Science Policy Centre
6 – 9 Carlton House Terrace
London SW1Y 5AG

T +44 20 7451 2500
E science.policy@royalsociety.org
W royalsociety.org

Cover image: *The Spanish Cucumber E. Coli.* In May 2011, there was an outbreak of a unusual Shiga-Toxin producing strain of E.Coli, beginning in Hamburg in Germany. This has been dubbed the 'Spanish cucumber' outbreak because the bacteria were initially thought to have come from cucumbers produced in Spain. This figure compares the genome of the outbreak E. Coli strain C227-11 (left semicircle) and the genome of a similar E. Coli strain 55989 (right semicircle). The 55989 reference strain and other similar E.Coli have been associated with sporadic human cases but never large scale outbreak. The ribbons inside the track represent homologous mappings between the two genomes, indicating a high degree of similarity between these genomes. The lines show the chromosomal positioning of repeat elements, such as insertion sequences and other mobile elements, which reveal some heterogeneity between the genomes. Section 1.3 explains how this genome was analysed within weeks because of a global and open effort; data about the strain's genome sequence were released freely over the internet as soon as they were produced. This figure is from Rohde H *et al* (2011). *Open-Source Genomic Analysis of Shiga-Toxin–Producing E. coli O104:H4.* New England Journal of Medicine, 365, 718-724. © New England Journal of Medicine.

Science as an open enterprise: open data for open science

Contents

Contents

Membership of Working Group

The members of the Working Group involved in producing this report are listed below. The Working Group formally met five times between May 2011 and February 2012 and many other meetings with outside bodies were attended by individual members of the Group. Members acted in an individual and not a representative capacity and declared any potential conflicts of interest. The Working Group Members contributed to the project on the basis of their own expertise and good judgement.

Chair	
Professor Geoffrey Boulton OBE FRSE FRS	Regius Professor of Geology Emeritus, University of Edinburgh
Members	
Dr Philip Campbell	Editor in Chief, Nature
Professor Brian Collins CB FREng	Professor of Engineering Policy, University College London
Professor Peter Elias CBE	Institute for Employment Research, University of Warwick
Professor Dame Wendy Hall FREng FRS	Professor of Computer Science, University of Southampton
Professor Graeme Laurie FRSE FMedSci	Professor of Medical Jurisprudence, University of Edinburgh
Baroness Onora O'Neill FBA FMedSci FRS	Professor of Philosophy Emeritus, University of Cambridge
Sir Michael Rawlins FMedSci	Chairman, National Institute for Health and Clinical Excellence
Professor Dame Janet Thornton CBE FRS	Director, European Bioinformatics Institute
Professor Patrick Vallance FMedSci	President, Pharmaceuticals R&D, GlaxoSmithKline
Sir Mark Walport FMedSci FRS	Director, the Wellcome Trust

Review Panel

This report has been reviewed by an independent panel of experts before being approved by the Council of the Royal Society. The Review Panel members were not asked to endorse the conclusions and recommendations of the report but to act as independent referees of its technical content and presentation. Panel members acted in a personal and not an organisational capacity and were asked to declare any potential conflicts of interest. The Royal Society gratefully acknowledges the contribution of the reviewers.

Professor John Pethica FRS	Vice President, Royal Society
Professor Ross Anderson FREng FRS	Security Engineering, Computer Laboratory, University Of Cambridge
Professor Sir Leszek Borysiewicz KBE FRCP FMedSci FRS	Vice-Chancellor, University of Cambridge
Dr Simon Campbell CBE FMedSci FRS	Former Senior Vice President, Pfizer and former President, the Royal Society of Chemistry
Professor Bryan Lawrence	Professor of Weather and Climate Computing, University of Reading and Director, STFC Centre for Environmental Data Archival
Dr LI Janhui	Director of Scientific Data Center, Computer Network Information Center, Chinese Academy of Sciences
Professor Ed Steinmueller	Science Policy Research Unit, University of Sussex

Science Policy Centre Staff

Jessica Bland	Policy Adviser
Dr Claire Cope	Intern (December 2011 – March 2012)
Caroline Dynes	Policy Adviser (April 2012 – June 2012)
Nils Hanwahr	Intern (July 2011 – October 2011)
Dr Jack Stilgoe	Senior Policy Adviser (May 2011 – June 2011)
Dr James Wilson	Senior Policy Adviser (July 2011 – April 2012)

Summary

The practice of science

Open inquiry is at the heart of the scientific enterprise. Publication of scientific theories - and of the experimental and observational data on which they are based - permits others to identify errors, to support, reject or refine theories and to reuse data for further understanding and knowledge. Science's powerful capacity for self-correction comes from this openness to scrutiny and challenge.

Drivers of change: making intelligent openness standard

Rapid and pervasive technological change has created new ways of acquiring, storing, manipulating and transmitting vast data volumes, as well as stimulating new habits of communication and collaboration amongst scientists. These changes challenge many existing norms of scientific behaviour.

The historical centrality of the printed page in communication has receded with the arrival of digital technologies. Large scale data collection and analysis creates challenges for the traditional autonomy of individual researchers. The internet provides a conduit for networks of professional and amateur scientists to collaborate and communicate in new ways and may pave the way for a second open science revolution, as great as that triggered by the creation of the first scientific journals. At the same time many of us want to satisfy ourselves as to the credibility of scientific conclusions that may affect our lives, often by scrutinising the underlying evidence, and democratic governments are increasingly held to account through the public release of their data. Two widely expressed hopes are that this will increase public trust and stimulate business activity. Science needs to adapt to this changing technological, social and political environment. This report considers how the conduct and communication of science needs to adapt to this new era of information technology. It recommends how the governance of science can be updated, how scientists should respond to changing public expectations and political culture, and how it may be possible to enhance public benefits from research.

The changes that are needed go to the heart of the scientific enterprise and are much more than a requirement to publish or disclose more data. Realising the benefits of open data requires effective communication through a more intelligent openness: data must be accessible and readily located; they must be intelligible to those who wish to scrutinise them; data must be assessable so that judgments can be made about their reliability and the competence of those who created them; and they must be usable by others. For data to meet these requirements it must be supported by explanatory metadata (data about data). As a first step towards this intelligent openness, data that underpin a journal article should be made concurrently available in an accessible database. We are now on the brink of an achievable aim: for all science literature to be online, for all of the data to be online and for the two to be interoperable.

New ways of doing science: computational and communications technologies

Modern computers permit massive datasets to be assembled and explored in ways that reveal inherent but unsuspected relationships. This data-led science is a promising new source of knowledge. Already there are medicines discovered from databases that describe the properties of drug-like compounds. Businesses are changing their services because they have the tools to identify customer behaviour from sales data. The emergence of linked data technologies creates new information through deeper integration of data across different datasets with the potential to greatly enhance automated approaches to data analysis. Communications technologies have the potential to create novel social dynamics in science. For example, in 2009 the Fields medallist mathematician Tim Gowers posted an unsolved mathematical problem on his blog with an invitation to others to contribute to its solution. In just over a month and after 27 people had made more than 800 comments, the problem was solved. At the last count, ten similar projects are under way to solve other mathematical problems in the same way.

Not only is open science often effective in stimulating scientific discovery, it may also help to deter, detect and stamp out bad science. Openness facilitates a systemic integrity that is conducive to early identification of error, malpractice and fraud, and therefore deters them. But this kind of transparency only works when openness meets standards of intelligibility and assessability - where there is intelligent openness.

Enabling change

Successful exploitation of these powerful new approaches will come from six changes: (1) a shift away from a research culture where data is viewed as a private preserve; (2) expanding the criteria used to evaluate research to give credit for useful data communication and novel ways of collaborating; (3) the development of common standards for communicating data; (4) mandating intelligent openness for data relevant to published scientific papers; (5) strengthening the cohort of data scientists needed to manage and support the use of digital data (which will also be crucial to the success of private sector data analysis and the government's Open Data strategy); and (6) the development and use of new software tools to automate and simplify the creation and exploitation of datasets. The means to make these changes are available. But their realisation needs an effective commitment to their use from scientists, their institutions and those who fund and support science.

Additional efforts to collect data, expand databases and develop the tools to exploit them all have financial as well as opportunity costs. These very practical qualifications on openness cannot be ignored; sharing research data needs to be tempered by realistic estimates of demand for those data. The report points to powerful pathfinder examples from many areas of science in which the benefits of openness outweigh the costs. The cost of data curation to exacting standards is often demonstrably smaller than the costs of collecting further or new data. For example, the annual cost of managing the world's data on protein structures in the world wide Protein Data Bank is less than 1% of the cost of generating that data.

Communicating with citizens

Recent decades have seen an increased demand from citizens, civic groups and non-governmental organisations for greater scrutiny of the evidence that underpins scientific conclusions. In some fields, there is growing participation by members of the public in research programmes, as so-called citizen scientists: blurring the divide between professional and amateur in new ways.

However, effective communication of science embodies a dilemma. A major principle of scientific enquiry is to "take nobody's word for it". Yet many areas of science demand levels of skill and understanding that are beyond the grasp of the most people, including those of scientists working in other fields. An immunologist is likely to have a poor understanding of cosmology, and vice versa. Most citizens have little alternative but to put their trust in what they can judge about scientific practice and standards, rather than in personal familiarity with the evidence. If democratic consent is to be gained for public policies that depend on difficult or uncertain science, the nature of that trust will depend to a significant extent on open and effective communication within expert scientific communities and their participation in public debate.

A realistic means of making data open to the wider public needs to ensure that the data that are most relevant to the public are accessible, intelligible, assessable and usable for the likely purposes of non-specialists. The effort required to do this is far greater than making data available to fellow specialists and might require focussed efforts to do so in the public interest or where there is strong interest in making use of research findings. However, open data is only part of the spectrum of public engagement with science. Communication of data is a necessary, though not a sufficient element of the wider project to make science a publicly robust enterprise.

The international dimension

Does a conflict exist between the interests of taxpayers of a given state and open science where the results reached in one state can be readily used in another? Scientific output is very rapidly diffused. Researchers in one state may test, refute, reinforce or build on the results and conclusions of researchers in another. This international exchange often evolves into complex networks of collaboration and stimulates competition to develop new understanding. As a consequence, the knowledge and skills embedded in the science base of one state are not merely those paid for by the taxpayers of that state, but also those absorbed from a wider international effort. Trying to control this exchange would risk yet another "tragedy of the commons", where myopic self-interest depletes a common resource, whilst the current operation of the internet would make it almost impossible to police.

Qualified openness

Opening up scientific data is not an unqualified good. There are legitimate boundaries of openness which must be maintained in order to protect commercial value, privacy, safety and security.

The importance of open data varies in different business sectors. Business models are evolving to include a more open approach to innovation. This affects the way that firms value data; in some areas there is more attention to the development of analytic tools than on keeping data secret. Nevertheless, protecting Intellectual Property (IP) rights over data are still vital in many sectors, and legitimate reasons for keeping data closed must be respected. Greater openness is also appropriate when commercial research data has the potential for public impact - such as in the release of data from clinical trials.

There is a balance to be struck between creating incentives for individuals to exploit new scientific knowledge for financial gain and the macroeconomic benefits that accrue when knowledge is broadly available and can be exploited creatively in a wide variety of ways. The small percentage of university income from IP undermines the rationale for tighter control of IP by them. It is important that the search for short term benefit to the finances of a university does not work against longer term benefit to the national economy. New UK guidelines to address this are a welcome first step towards a more sophisticated approach.

The sharing of datasets containing personal information is of critical importance for research in the medical and social sciences, but poses challenges for information governance and the protection of confidentiality. It can be strongly in the public interest provided it is performed under an appropriate governance framework. This must adapt to the fact that the security of personal records in databases cannot be guaranteed through anonymisation procedures.

Careful scrutiny of the boundaries of openness is important where research could in principle be misused to threaten security, public safety or health. In such cases this report recommends a balanced and proportionate approach rather than a blanket prohibition.

Recommendations

This report analyses the impact of new and emerging technologies that are transforming the conduct and communication of research. The recommendations are designed to improve the conduct of science, respond to changing public expectations and political culture and enable researchers to maximise the impact of their research. They are designed to ensure that reproducibility and self-correction are maintained in an era of massive data volumes. They aim to stimulate the communication and collaboration where these are needed to maximise the value of data-intensive approaches to science. Action is needed to maximise the exploitation of science in business and in public policy. But not all data are of equal interest and importance. Some are rightly confidential for commercial, privacy, safety or security reasons. There are both opportunities and financial costs in the full presentation of data and metadata. The recommendations set out key principles. The main text explores how to judge their application and where accountability should lie

Recommendation 1

Scientists should communicate the data they collect and the models they create, to allow free and open access, and in ways that are intelligible, assessable and usable for other specialists in the same or linked fields wherever they are in the world. Where data justify it, scientists should make them available in an appropriate data repository. Where possible, communication with a wider public audience should be made a priority, and particularly so in areas where openness is in the public interest.

Although the first and most important recommendation is addressed directly to the scientific community itself, major barriers to widespread adoption of the principles of open data lie in the systems of reward, esteem and promotion in universities and institutes. It is crucial that the generation of important datasets, their curation and open and effective communication is recognised, cited and rewarded. Existing incentives do not support the promotion of these activities by universities and research institutes, or by individual scientists. This report argues that universities and research institutes should press for the financial incentives that will facilitate not only the best research, but the best communication of data. They must recognise and reward their employees and reconfigure their infrastructure for a changing world of science.

Here the report makes recommendations to the organisations that have the power to incentivise and support open data policies and promote data-intensive science and its applications. These organisations increasingly set policies for access to data produced by the research they have funded. Others with an important role include the learned societies, the academies and professional bodies that represent and promote the values and priorities of disciplines. Scientific journals will continue to be media through which a great deal of scientific research finds its way into the public domain, and they too must adapt to and support policies that promote open data wherever appropriate.

Recommendation 2

Universities and research institutes should play a major role in supporting an open data culture by: recognising data communication by their researchers as an important criterion for career progression and reward; developing a data strategy and their own capacity to curate their own knowledge resources and support the data needs of researchers; having open data as a default position, and only withholding access when it is optimal for realising a return on public investment.

Recommendation 3

Assessment of university research should reward the development of open data on the same scale as journal articles and other publications, and should include measures that reward collaborative ways of working.

Recommendation 4

Learned societies, academies and professional bodies should promote the priorities of open science amongst their members, and seek to secure financially sustainable open access to journal articles. They should explore how enhanced data management could benefit their constituency, and how habits might need to change to achieve this.

Recommendation 5

Research Councils and Charities should improve the communication of research data from the projects they fund by recognising those who could maximise usability and good communication of their data; by including the costs of preparing data and metadata for curation as part of the costs of the research process; and by working with others to ensure the sustainability of datasets.

Recommendation 6

As a condition of publication, scientific journals should enforce a requirement that the data on which the argument of the article depends should be accessible, assessable, usable and traceable through information in the article. This should be in line with the practical limits for that field of research. The article should indicate when and under what conditions the data will be available for others to access.

Effective exchange of ideas, expertise and people between the public and private sectors is key to delivering value from research. The economic benefit and public interest in research should influence how and when data, information and knowledge from publicly or privately funded research are made widely available.

Recommendation 7

Industry sectors and relevant regulators should work together to determine the approaches to sharing data, information and knowledge that are in the public interest. This should include negative or null results. Any release of data should be clearly signposted and effectively communicated.

Recommendation 8

Governments should recognise the potential of open data and open science to enhance the excellence of the science base. They should develop policies for opening up scientific data that complement policies for open government data, and support development of the software tools and skilled personnel that are vital to the success of both.

Judging whether data should be made more widely available requires assessment of the public benefits from sharing research data and the need to protect individual privacy and other risks. Guidance for researchers should be clear and consistent.

Recommendation 9

Datasets should be managed according to a system of proportionate governance. This means that personal data is only shared if it is necessary for research with the potential for high public value. The type and volume of information shared should be proportionate to the particular needs of a research project, drawing on consent, authorisation and safe havens as appropriate. The decision to share data should take into account the evolving technological risks and developments in techniques designed to safeguard privacy.

Recommendation 10

In relation to security and safety, good practice and common information sharing protocols based on existing commercial standards must be adopted more widely. Guidelines should reflect the fact that security can come from greater openness as well as from secrecy.

Data terms

Data relationships	Definition
Data	Numbers, characters or images that designate an attribute of a phenomenon.
Information	Data become information when they are combined together in ways that have the potential to reveal patterns in the phenomenon.
Knowledge	Information yields knowledge when it supports non-trivial, true claims about a phenomenon.

Data type	Definition
Big Data	Data that requires massive computing power to process.
Broad Data	Structured big data, so that it is freely available through the web to everyone, eg on websites like www.data.gov
Data	Qualitative or quantitative statements or numbers that are (or assumed to be) factual. Data may be raw or primary data (eg direct from measurement), or derivative of primary data, but are not yet the product of analysis or interpretation other than calculation.
Data-gap	When data becomes detached from the published conclusions
Data-intensive science	Science that involves large or even massive datasets
Data-led approach	Where hypotheses are constructed after identifying relationships in the dataset.
Data-led science	The use of massive datasets to find patterns as the basis of research.
Dataset	A collection of factual information held in electronic form where all or most of the information has been collected for the purpose of provision of a service by the authority or carrying out of any other function of the authority. Datasets contain factual information which is not the product of analysis or interpretation other than calculation, is not an official statistic, and is unaltered and un-adapted since recording.
Linked Data	Linked data is described by a unique identifier naming and locating it in order to facilitate access. It contains identifiers for other relevant data, allowing links to be made between data that would not otherwise be connected, increasing discoverability of related data.
Metadata	Metadata "data about data", contains information about a dataset. This may be state why and how it was generated, who created it and when. It may also be technical, describing its structure, licensing terms, and standards it conforms to.
Open Data	Open data is data that meets the criteria of intelligent openness. Data must be accessible, useable, assessable and intelligible.
Semantic Data	Data that are tagged with particular metadata - metadata that can be used to derive relationships between data.

Intelligent Openness terms	Definition
accessible	Data must be located in such a manner that it can readily be found and in a form that can be used.
assessable	In a state in which judgments can be made as to the data or information's reliability. Data must provide an account of the results of scientific work that is intelligible to those wishing to understand or scrutinise them. Data must therefore be differentiated for different audiences.
intelligible	Comprehensive for those who wish to scrutinise something. Audiences need to be able to make some judgment or assessment of what is communicated. They will need to judge the nature of the claims made. They should be able to judge the competence and reliability of those making the claims. Assessability also includes the disclosure of attendant factors that might influence public trust.
useable	In a format where others can use the data or information. Data should be able to be reused, often for different purposes, and therefore will require proper background information and metadata. The usability of data will also depend on those who wish to use them.