
# 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


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. |

# The purpose and practice of science

Scientists aspire to understand the workings of nature, people and society and to communicate that understanding for the general good. Governments worldwide recognise this and fund science for its contribution to knowledge, to national economies and social policies, and its role in managing global risks such as pandemics or environmental degradation.[1] The digital revolution is pervasively changing science and society. This report is concerned with its impact on fundamental processes that determine the rate of progress of science and that enable the effective communication of scientific results and understanding. It recommends how these processes must adapt to novel technologies and evolving public expectations and political culture.

## 1.1 The role of openness in science

Much of the remarkable growth of scientific understanding in recent centuries is due to open practices; open communication and deliberation sit at the heart of scientific practice.[2] Publishing scientific theories, including experimental and observational data, permits others to scrutinise them, to replicate experiments and to reuse data to create further understanding. It permits the identification of errors and for theories to be rejected or refined. Facilitating sustained and rigorous analysis of evidence and theory is the most rigorous form of peer review. It has made science a self-correcting process since the first scientific journals were established: the *Journal des Sçavans* in France and *Philosophical Transactions of the Royal Society* in England (Box 1.1). Scientific journals made vital contributions to the explosion of scientific knowledge in the seventeenth and eighteenth centuries,[3] and permitted ideas and measurements to be more readily corroborated, invalidated or improved. They also communicated the results of research to a wider audience, who were in turn stimulated to contribute further ideas and observations to the development of science.

### Box 1.1 Henry Oldenburg: the scientific journal and the process of peer review[4]

Henry Oldenburg (1619-1677) was a German theologian who became the first Secretary of the Royal Society. He corresponded with leading scientists across Europe, believing that rather than waiting for entire books to be published, letters were much better suited to the quick communication of facts or new discoveries. He invited people to write to him - even laymen, who were not involved with science but had discovered some item of knowledge.[5] He no longer required that science be conveyed in Latin, but in any vernacular language. From these letters the idea of printing scientific papers or articles in a scientific journal was born. 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.

1 *Typical Statements from national academy websites - Royal Society*: to expand the frontiers of knowledge by championing the development and use of science, mathematics, engineering and medicine for the benefit of humanity and the good of the planet. *US National Academy of Science*: a society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the public good. *Chinese Academy of Sciences*: striving to accomplish world-class science and to continuously make fundamental, strategic and forward-looking contributions to national economic construction, national security and social sustainable development by strengthening original scientific innovation, innovation of key technologies and system integration.

2 Classically elaborated in: Polanyi M (FRS), *The Republic of Science,* Minerva 38, 1-21.

3 Shapin S (1994). *A social history of truth: civility and science in seventeenth-century England*. University of Chicago Press: Chicago.

4 Klug A (2000). *Address of the President, Sir Aaron Klug, O.M., P.R.S., Given at the Anniversary Meeting on 30 November 1999*, Notes Record Royal Society: London, 54, 99-108.

5 Boas Hall M (2002). *Henry Oldenburg: Shaping the Royal Society*. Oxford University Press: Oxford.

## 1.2 Data, information and effective communication

Before going further, it is important to define terms and understand the principles that underlie effective communication. There is sometimes confusion between data, information and knowledge. This report uses them as overlapping concepts, differentiated by the breadth and the depth of the explanation they provide about a phenomenon. **Data** are numbers, characters or images that designate an attribute of a phenomenon. They become **information** when they are combined together in ways that have the potential to reveal patterns in the phenomenon. Information yields **knowledge** when it supports non-trivial, true claims about a phenomenon. For example, the numbers generated by a theodolite measuring the height of mountain peaks are data. Using a formula, the height of the peak can be deduced from the data, which is information. When combined with other information, for example about the mountain's rocks, this creates knowledge about the origin of the mountain. Some are sceptical about these distinctions, but this report regards them as a useful framework for understanding the role of data in science.

Raw and derived data have different roles in scientific analysis, and should be further distinguished from their associated metadata. Raw data are measured data, for example, daily rainfall measurements over the course of years, this can then be averaged to estimate mean annual rainfall, which is derived data.

To be interpretable, data usually require some contextual information or **metadata**. This should include information about the data creator, how the data were acquired, the creation date and method, as well as technical details about how to use the dataset, how the data have been selected and treated, and how they have been analysed for scientific purposes. The preparation of metadata is particularly onerous for complex datasets or for those that have been subjected to mathematical modelling. But metadata are indispensible for reproducing results.

Mere disclosure of data has very little value *per se*[6]. Realising the benefits of open data requires a more *intelligent openness*, one where data are effectively communicated. For this, data must fulfil four fundamental requirements, something not always achieved by generic metadata. They must be accessible, intelligible, assessable and usable as follows:

a. Accessible. Data must be located in such a manner that it can readily be found. This has implications both for the custodianship of data and the processes by which access is granted to data and information.

b. Intelligible. Data must provide an account of the results of scientific work that is intelligible to those wishing to understand or scrutinise them. Data communication must therefore be differentiated for different audiences. What is intelligible to a specialist in one field may not be intelligible to one in another field. Effective communication to the non-scientific wider public is more difficult, necessitating a deeper understanding of what the audience needs in order to understand the data and dialogue about priorities for such communication.

c. Assessable. Recipients need to be able to make some judgment or assessment of what is communicated. They will, for example, need to judge the nature of the claims that are made. Are the claims speculations or evidence based? They should be able to judge the competence and reliability of those making the claims. Are they from a scientifically competent source?[7] What was the purpose of the research project and who funded it? Is the communication influenced by extraneous considerations and are these possible sources of influence identified?[8] Assessability also includes the disclosure of attendant factors that might influence trust in the research. For example, medical journals increasingly require a statement of interests from authors.

---

6 O'Neill O (2006). *Transparency and the Ethics of Communication*. In *Transparency: The Key to Better Governance?* Heald D & Hood C (eds.). Proceedings of the British Academy 135. Oxford University Press: Oxford.

7 Only an expert is really likely to be able to make this judgement; this represents one of the important functions of peer review. The non-expert, which will include the vast majority of the population, including professional scientists from other scientific domains, has to rely on peer review.

8 It is essential that there are clear statements about possible conflicts of interest. There is nothing wrong with a conflict of interest *per se*. What is important is that conflicts of interest are declared in a transparent fashion.

d. Usable. Data should be able to be reused, often for different purposes. The usability of data will also depend on the suitability of background material and metadata for those who wish to use the data. They should, at a minimum, be reusable by other scientists.

Responsibility for effective communication lies with the recipient as well as the data provider. Understanding what must be accessible, what is intelligible and what kind of assessment and reuse are going to occur requires input from both parties. In some cases, this is simple: clinical trial regulators – the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK - have well defined rules for the data that must accompany any application for trials in order for the regulator to grant a licence for that trial. But providing the same data for a different audience can prove much more difficult. A support group for patients who could be treated by a new drug might be interested in research data, but understanding what can be responsibly released is trickier. Intelligent openness is a response to the varying demands on different sorts of data from diverse research communities and interest groups. This report showcases where this has been successful - usually through decentralised initiatives where specific demands and uses of data are well understood.

### 1.3 The power of intelligently open data

The benefits of intelligently open data were powerfully illustrated by events following an outbreak of a severe gastro-intestinal infection in Hamburg in Germany in May 2011. This spread through several European countries and the US, affecting about 4000 people and resulting in over 50 deaths.[9] All tested positive for an unusual and little-known Shiga-toxin–producing E. coli bacterium. The strain was initially analysed by scientists at BGI-Shenzhen in China, working together with those in Hamburg, and three days later a draft genome was released under an open data licence.[10] This generated interest from bioinformaticians on four continents. 24 hours after the release of the genome it had been assembled. Within a week two dozen reports had been filed on an open-source site dedicated to the analysis of the strain.[11] These analyses provided crucial information about the strain's virulence and resistance genes – how it spreads and which antibiotics are effective against it.[12] They produced results in time to help contain the outbreak. By July 2011, scientists published papers based on this work. By opening up their early sequencing results to international collaboration, researchers in Hamburg produced results that were quickly tested by a wide range of experts, used to produce new knowledge and ultimately to control a public health emergency.

There is great value in making individual pseudonymised patient data from clinical trials available to other medical scientists provided that the privacy of individuals can be reasonably protected. It allows suspicions of scientific fraud to be examined using statistical techniques. It helps eliminate incomplete reporting of results in peer reviewed journals, and it facilitates more meta-analyses based on raw data rather than on summary results. The power of this approach has recently been demonstrated with a meta-analysis – incorporating information from 95,000 patients – of the effects of aspirin in the prevention of cardiovascular disease. The study confirmed the benefits of aspirin for those with established heart conditions. But it questioned whether adverse effects, like an increase risk of bleeding, might outweigh the more modest benefits for those who do not already suffer from these problems.[13]

9 World Health Organisation (2011). *Outbreaks of E. coli 0104:H4 infection*. Available at: http://www.euro.who.int/en/what-we-do/health-topics/emergencies/international-health-regulations/outbreaks-of-e.-coli-o104h4-infection

10 BGI used a Creative Commons zero licence, waiving all rights to the work worldwide under copyright law. They also assigned it a Digital Object Identifier, providing permanent access to the analysis: http://datacite.wordpress.com/2011/06/15/ehec-genome-with-a-doi-name/

11 GitHub (2012). *E. coli O104:H4 Genome Analysis Crowdsourcing*. Available at: https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki

12 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. Available at: http://www.nejm.org/doi/full/10.1056/NEJMoa1107643#t=articleTop

13 Antithrombotic Trialists Collaboration (2009). *Aspirin in the primary and secondary prevention of vascular disease: meta-analysis of individual participant data from randomised controlled trials*. Lancet, 373, 1849-1860.

Recent developments at the OPERA collaboration at CERN illustrate how data openness can help in the scrutiny of scientific results. The OPERA team fired a beam of muon neutrinos from CERN to the Gran Sasso National Laboratory, 730 km away in central Italy. In September 2011, and to the surprise of the experiment's scientists, the neutrinos seemed to travel faster than the speed of light – understood to be a universal speed limit.[14] Hoping for ideas to explain this apparent violation of physical law CERN opened the result to broader scrutiny, uploading the results in unprecedented detail to the physics pre-print archive, arXiv.org. More than 200 papers appeared on arXiv.org attempting to debunk or explain the effect. A large group of papers focused on the technique used to time the neutrinos' flight path. On 23 February 2012, the OPERA collaborators announced two potential sources of timing error.[15] There was a delay in the stop and start signals sent via GPS to the clock at Gran Sasso due to a faulty fibre optic cable, and there was a fault inside the master clock at Gran Sasso. It was announced in June 2012 that attempts to replicate the original result with four separate instruments at Gran Sasso found that neutrinos respected the universal speed limit, confirming the suspected experimental error.

There are studies that suggest that open data can increase a published paper's profile. An examination of 85 cancer microarray clinical trials showed that publicly available data was associated with a 69% increase in citation of the original trial publication, independent of journal impact factor, date of publication or the author's country of origin.[16]

### 1.4 Open science: aspiration and reality

Much of today's scientific practice falls short of the ideals of intelligent openness reflected in section 1.3. A lot of science is unintelligible beyond its own specialist discipline and the evidential data that underpins scientific communications is not consistently made accessible, even to other scientists. Moreover, although scientists do routinely exploit the massive data volumes and computing capacity of the digital age, the approach is often redolent of the paper age rather than the digital age. Computer science pioneer Jim Gray, took a dim view of his fellow researchers: "When you go and look at what scientists are doing, day in and day out, in terms of data analysis, it is truly dreadful. We are embarrassed by our data!"[17]

There are important issues that need to be resolved about the boundaries of openness, which are addressed in chapter 3. Should the boundary of open science be coincident with the divide between publicly and privately funded science? Are legitimate commercial interests in the exploitation of scientific data, information and knowledge invariably favoured by restriction or invariably appropriate; or can openness be economically beneficial or socially desirable in some sectors? How are privacy and confidentiality best maintained? And do open data and open science conflict with the interests of privacy, safety and security?

Open science is defined here as open data (available, intelligible, assessable and useable data) combined with open access to scientific publications and effective communication of their contents. This report focuses on the challenges and opportunities offered by the modern data deluge and how a culture of open data and communication can, with some exceptions, maximise the capacity to respond to them.

But the last decade has seen substantial moves towards free online public archives of journal articles, such as PubMed Central and arXiv.org. Nearly 34,000 scientists from 180 nations signed a letter in 2000 asking for an online public library that would provide the full contents of the published records of research and scholarly discourse in medicine and the life sciences. This led to the launch of an open access journal from the Public Library of Science (PLoS) in 2003. Researchers funded by The Wellcome Trust must allow their papers to be put in the PubMed Central repository.

14 CERN (2011). *Press Release: OPERA experiment reports anomaly in flight time of neutrinos from CERN to Gran Sasso*. Available at: http://public.web.cern.ch/press/pressreleases/Releases2011/PR19.11E.html

15 Reich E S (2012). *Timing glitches dog neutrino claim: Team admits to possible errors in faster-than-light finding*. Nature News, 483, 17. Available at: http://www.nature.com/news/timing-glitches-dog-neutrino-claim-1.10123

16 Piwowar H A, Day RS, Fridsma DB (2007). *Sharing detailed data is associated with increased citation rate*. PLoS ONE, 2, 3, e308.

17 Gray J (2009). *A transformed scientific method*. In: The Fourth Paradigm. Hey T, Tansley S & Tolle K (eds.). Microsoft Research: Washington.

13 of the 26 European Research Area countries that responded to a recent survey have national or regional open access policies.[18] Sweden has a formal national open access programme, OpenAcess.se[19], to support open access journals and repositories. Iceland has a national licence that allows free access to a wide range of electronic journals for any citizen with a national ISP address. Recent attempts to curtail the open access policies of the US Government research funders through a proposed *Research Works Act* (House Resolution 3699) were discontinued as a consequence of a campaign by the scientific community.[20]

What this report states in section 1.3 about the power of open data can also be said about the idea of an open primary scientific literature, including full and immediate access for all to published research papers. New text-mining technologies (3.1.1) and developments in multidisciplinary research would be empowered by that removal of subscription barriers. There are global policy and political signals that this is not only scientifically desirable but ultimately inevitable. However, publishers who add value to the literature do so through selectivity, editing for scientific accuracy and comprehensibility, adding metadata and hosting data in ways that most users find valuable or even essential. These activities have substantial costs associated with them. For this reason, in order to replace a subscription funded model of publication, the costs of publication will need to be replaced by charges to authors that are borne by researchers' funders or employers. Developing the primary literature's open accessibility (and reusability through appropriate licensing), while also doing financial justice to its quality a nd integrity, is a thorny challenge faced by policy-makers worldwide In the UK this is being addressed on behalf of the government by the Finch working group.[21]

## 1.5 The dimensions of open science: value outside the science community

In what context would the UK, or any other state, make a decisive move towards more open data? Where do the benefits lie? Is there a risk that it might benefit international scientific competitors that are more restrictive in their release of data, without a complementary benefit to the initiating state? How might openness influence the commercial interests of science-intensive companies in that state? And, how might this affect public and civic issues and priorities?

### 1.5.1 Global science, global benefits

It is important to recognise that science published openly online is inevitably international. Researchers and members of the public in one country are able to test, refute, reinforce or build on the results and conclusions of researchers in another. New knowledge published openly is rapidly diffused internationally, with the result that the knowledge and skills embedded in a national science base are not merely those paid for by the taxpayers of that state but also those absorbed from the wider international effort, of which it is a part.[22] Simply relying on the science of others is not an option. The greater the strength of the home science base, the greater its capacity to absorb and benefit from science done elsewhere.[23] Scientists whose capacities and talents are nurtured through national programmes are readily welcomed into international networks, where they are able to acquire early knowledge of emerging science within the networks. Such openness to international collaboration stimulates creativity, spreads influence and produces early awareness of innovations, no matter where they originate, that can be applied in the home context. National funding brings both national and global benefits from international interaction.

18 European Commission, European Research Area Committee (2011). *National open access and preservation policies in Europe*. Available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-report-2011_en.pdf

19 Open Access.se (2012). *Scholarly publishing*. Available at: http://www.kb.se/OpenAccess/Hjalptexter/English/

20 At the time of publishing, over 12,000 researchers have signed the 'Costs of Knowledge' boycott of Elsevier journals. Available at: http://thecostofknowledge.com/

21 Dame Janet Finch chaired an independent working group on expanding access to published research finding, including representation from the Royal Society. More details available at: http://www.researchinfonet.org/publish/wg-expand-access/

22 Griffith R, Lee S & Van Reenan J (2011). *Is distance dying at last? Falling home bias in fixed-effects models of patent citations.* Quantitative Economics, Econometric Society, 2, 2, 211-249, 07.

23 Royal Society (2011). *Knowledge, Networks and Nations*. Royal Society: London.

There is growing international support for open data. In 1997, the US National Research Council argued that "full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research."[24] In 2007, the OECD published a set of *Principles and Guidelines for Access to Research Data from Public Funding.*[25] A 2009 report by the US National Academies of Science recommends that: "all researchers should make research data, methods, and other information integral to their publicly reported results publicly accessible, in a timely manner, to allow verification of published findings and to enable other researchers to build on published results, except in unusual cases in which there are compelling reasons for not releasing data. In these cases, researchers should explain in a publicly accessible manner why the data are being withheld from release."[26] A 2010 report by the European Commission's High Level Expert Group on Scientific Data called on the Commission to accelerate moves towards a common data infrastructure.[27]

As the distribution of scientific effort changes in an increasingly multi-polar world, with rising scientific powers such as China, India and Brazil and the growth of scientific efforts in the Middle East, South-East Asia and North Africa,[28] many have signed up to the principles of open data through membership of the International Council of Science (ICSU). In addition, international collaboration that depends on the open data principle is increasingly supported by inter-governmental funding or funding from international agencies. Such collaboration focuses on matters of global concern such as climate change, energy, sustainability, trade, migration and pandemics. The OECD Global Science Forum Expert Group on Data and Research infrastructure for the Social Sciences will produce a report in Autumn 2012 recommending ways that the research community can better coordinate the data collection that is vital for global responses to these global concerns.

Improvements in connectivity and alternatives to internet access, such as the International Panel on Climate Change's DVD data distribution for climate datasets,[29] have made a difference in access to research in the developing world. But access to publication still remains problematic in nations with an emerging science base[30]. Many such countries are unable to afford the huge cost of subscription to international journals, a cost which even large institutions in developed countries struggle with. This seriously hinders their ability to carry out research based on up-to-date knowledge and to train future scientists. The rise of open access publication has gone some way to alleviating this issue. The Research4Life program[31] is a public-private partnership between three United Nations agencies, two universities and major commercial publishers that enable eligible libraries and their users to access peer-reviewed international scientific journals, books and databases for free or for a small fee.

There are also understandable difficulties in ensuring access to data from developing countries. Whereas some are developing open access journals (for example the journal *African Health Sciences*[32]), others are uneasy at the prospect that those with greater scientific resources will benefit overseas interests, to the detriment of home researchers. For example, Indonesia ceased providing access to their flu samples in 2007 because of worries that more scientifically developed countries would create flu vaccines based on their data, with no benefit to Indonesia. This policy was reversed only after the World Health Organisation put in place protocols for equitable access to vaccines and medicines in future pandemics.[33]

---

24 US National Research Council (1997). *Bits of power*. US National Research Council : Washington.
25 OECD (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD Publications: Paris.
26 National Academy of Science (2009). *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. National Academy of Science: Washington.
27 European Commission (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data*. Final report of the High Level Expert Group on Scientific Data. Available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
28 Royal Society (2011). *Knowledge, Networks and Nations*. Royal Society: London.
29 Modelle & Daten (2008). *Order Data on DVD*. Available at: http://www.mad.zmaw.de/projects-at-md/ipcc-data/order-data-on-dvd/
30 Chan L, Kirsop B & Arunachalam S (2011). *Towards open and equitable access to research and knowledge for development*. Public Library of Science Medicine: San Francisco.
31 Hinari, Oare, Ardi, Agora (2012). *Research4Life*.Available at: http://www.research4life.org/
32 African Journals Online (2012). *African Health Sciences*. Available at: http://www.ajol.info/index.php/ahs
33 World Health Organisation (2011). *Pandemic influenza preparedness Framework*. World Health Organisation: New York. Available at: http://whqlibdoc.who.int/publications/2011/9789241503082_eng.pdf

There are some cases where the boundaries of openness continue to restrict international access. National security concerns in the US have led to an attempt to restrict the export of software incorporating encryption capabilities commonly employed in other OECD countries. This has created a complex system for ascertaining whether or not an export licence is required. The US National Academies of Science argued in 2009[34] that these processes are excessively restrictive, and exemptions for research may be strengthened as a result. However, legitimate concerns about national security will continue to restrict open data between countries.

### 1.5.2 Economic benefit

Science plays a fundamental role in today's knowledge economies. The substantial direct and indirect economic benefits of science include the creation of new jobs, the attraction of inward investment and the development of new science and technologybased products and services. The UK has a world leading science base and an excellent university system that play key roles in technology enabled transformations in manufacturing, in knowledge based business and in infrastructural developments.[35]

The Royal Society's 2010 report, *The Scientific Century: Securing Our Future Prosperity*, distilled two key messages. First, science and innovation need to be at the heart of the UK's long term strategy for economic growth. Second, the UK faces a fierce competitive challenge from countries that are investing on a scale and speed that the UK will struggle to match.[36]

In parallel, there is ever more emphasis on the power of data in our future economy. An analysis of UK data equity estimated it is worth £25.1 billion to UK business in 2011. This is predicted to increase to £216 billion or 2.3% of cumulative GDP between 2012 and 2017. But a majority of this (£149 billion) will come from greater business efficiency in data use. £24 billion will come from the expected increase in expenditure on data-driven R&D.[37]

Governments have recognised the potential benefits of opening up data and information held by them to allow others to build on or utilise the information. In 2004 the UK Government's Office of Public Sector Information began a pilot scheme to use the Semantic Web (see section 2.1.4) to integrate and publish information from across the Public Sector.[38] This led to a UK Open Government Data project and in 2009 the creation of the data.gov.uk site - a single point of access for all Government non-personal public data.[39] Some public service information, such as live public transport information, became available in mid-2011; and in December of the same year, as part of the *UK Strategy for Life Sciences*,[40] the Prime Minister announced a change to the NHS constitution to allow access to routine patient data for research purposes, including by healthcare industries developing new products and services. The aim is to use data to boost investment in medical research and in digital technology in the UK, particularly by UK based pharmaceutical firms. London's Tech City (Box 1.2) promises to cement the link between open data and economic growth in the UK.

---

34 National Academies of Science (2009). *Beyond 'Fortress America': National Security Controls on Science and Technology in a Globalized World.* National Academy of Sciences: Washington. Available at: http://www.nap.edu/catalog.php?record_id=12567#description

35 Government Office for Science (2010). *Technology and Innovation Futures: UK Growth Opportunities for the 2020s*. BIS: London. Available at: http://www.bis.gov.uk/assets/bispartners/foresight/docs/general-publications/10-1252-technology-and-innovation-futures.pdf

36 The Royal Society (2010). *The Scientific Century: Securing Our Future Prosperity*. Royal Society: London. Available at: http://royalsociety.org/policy/publications/2010/scientific-century/

37 CEBR (2012). *Data equity: unlocking the value of big data*. Available at: http://www.cebr.com/wp-content/uploads/1733_Cebr_Value-of-Data-Equity_report.pdf

38 Shadbolt N, O'Hara K, Salvadores M & Alani H (2011). eGovernment. In *Handbook of Semantic Web Technologies*. Domingue J, Fensel D & Hendler J (eds.). Springer-Verlag: Berlin. 840-900. Available at: http://eprints.ecs.soton.ac.uk/21711/

39 Berners-Lee T & Shadbolt N (2009). Put in your postcode, out comes the data. The Times: London. Available at http://eprints.ecs.soton.ac.uk/23212/

40 BIS (2011). *UK Strategy for Life Sciences.* BIS. Available at: http://www.bis.gov.uk/assets/biscore/innovation/docs/s/11-1429-strategy-for-uk-life-sciences

### Box 1.2 London's Tech City

In November 2010, the Prime Minister announced that the Government would be investing in the existing cluster of technology companies in East London to create a world-leading technology centre. The ambition is that the existing 'silicon roundabout' would be extended eastwards into the redeveloped areas around the Olympic Park to create the largest technology park in Europe - an environment where the next Apple or Skype could come out of the UK.

A year on, the government added an Open Data Institute[41] to the cluster, funded to exploit and research open data opportunities with business and academia. This brought open data into the centre of the government's flagship technology initiative. There was also support for a new collaboration between Imperial College London, University College London and Cisco. This three-year agreement to create a Future Cities Centre, focuses on four areas: Future Cities and Mobility, Smart Energy Systems, the Internet of Things and Business Model Innovation.

Influential international examples of the success of these strategies come from the USA, where government funded datasets have been proactively released for free and open reuse in order to generate economic activity. For example, the US National Weather Service puts its weather data into the public domain, and this is believed to be a key driver in the development of a private sector meteorology market estimated to exceed $1.5 billion.[42] In an attempt to capture some of this same value and impetus, it was announced in 2011 that the UK Met Office and the Land Registry will make data available under an open licence. The UK Met Office is also currently working with partners including IBM, Imperial College Business School and the Grantham Institute for Climate Change at Imperial College London to enhance sharing and access to Met Office data. Box 1.3 details how opening up earth surface information has created new opportunities in different ways on both sides of the Atlantic.

41 Berners-Lee T & Shadbolt N (2011). *There's gold to be mined from all our data*. The Times: London. Available at: http://eprints.ecs.soton.ac.uk/23090/

42 Spiegler D B (2006). *The Private Sector in Meteorology- An Update*. Available at: http://www.ametsoc.org/boardpges/cwce/docs/DocLib/2007-07-02_PrivateSectorInMeteorologyUpdate.pdf

### Box 1.3 Benefits of open release: satellite imagery and geospatial information

NASA Landsat satellite imagery of Earth surface environment, collected over the last 40 years was sold through the US Geological Survey for US$600 per scene until 2008, when it became freely available from the Survey over the internet.[43] Usage leapt from sales of 19,000 scenes per year, to transmission of 2,100,000 scenes per year. Google Earth now uses the images. There has been great scientific benefit, not least to the Geological Survey, which has seen a huge increase in its influence and its involvement in international collaboration. It is estimated to have created value for the environmental management industry of $935 million per year, with direct benefit of more than $100 million per year to the US economy, and has stimulated the development of applications from a large number of companies worldwide.

Since 2009, the UK's detailed national geological information has been available online for free.[44] This includes detailed baseline gravity and magnetic data-sets and many tens of thousands of images, including of the UK offshore hydrocarbon cores. 3D models used by the British Geological Survey (BGS) are also available. The BGS have developed an iGeology mobile app, where a user can zoom in on their current location and view their environment in overlain geological maps, giving details of bedrock, ice age deposits and old city maps. More detailed descriptions can be found by following links to the BGS Lexicon rock name database. Since 2010 it has been downloaded over 60,000 times from 56 countries.

Following the UK's lead, the European Commission has recently launched a wide-ranging open data initiative[45] which it expects will generate €140billion a year of income.[46] The Commission will open its own stores of data through a new portal, establish a level playing field for open data across Europe, and contributing €100 million to research into improving data handling technologies. The Commission has signalled that it hopes to back up these plans with an update to the 2003 Directive on the reuse of public sector information.

Deriving macroeconomic estimates for the extent to which research data is a driver of economic development is problematic. The most detailed estimate of the value to an economy of opening up scientific information comes from an analysis of the effects of open access on Australian public sector research. This suggests that a one-off increase in accessibility to public sector R&D ("the proportion of R&D stock available to firms that will use it" and "the proportion of R&D stock that generates useful knowledge") produces a return to the national economy of AUD$ 9 billion (£7 billion) over 20 years.[47]

43 Parcher J (2012). *Benefits of open availability of Landsat data*. Available at: www.oosa.unvienna.org/pdf/pres/stsc2012/2012ind-05E.pdf
44 British Geological Survey (2012). *What is OpenGeoscience?* Available at: http://www.bgs.ac.uk/opengeoscience/home.html
45 European Commission: Information Society (2012). *Public Sector Information - Raw Data for New Services and Products* Available at: http://ec.europa.eu/information_society/policy/psi/index_en.htm
46 European Commission (2011). *Review of recent PSI studies*. European Commission: Brussels. Available at: http://epsiplatform.eu/content/review-recent-psi-re-use-studies-published
47 Houghton J & Sheehan P (2009). *Estimating the Potential Impacts of Open Access to Research Findings*. Economic Analysis & Policy, 29, 1, 127-142.

### 1.5.3 Public and civic benefit

Public and civic benefits are derived from scientific understanding that is relevant to the needs of public policy, and much science is funded for this purpose. Recent decades have seen an increased demand from citizens, civic groups and non-governmental organisations for greater scrutiny of the evidence underpinning scientific conclusions, particularly where these have the potential for major impacts on individuals and society. The Icelandic initiative that opens up academic articles to all citizens (1.4) is an overt move to make scientific work more accessible to citizens. Over the last two decades, the scientific community has made a major effort to engage more effectively with the public, particularly in areas that this report describes as public interest science (areas of science with important health, economic and ethical implications for citizens and society such as climate science, stem cell research or synthetic biology) and to stimulate the involvement of amateurs in science[48] in areas such as astronomy, meteorology and ornithology. However, effective openness to citizens in ways that are compatible with this report's principles of effective communication (1.2) demands a considerable effort.

Public dialogue workshops with representative public groups recently undertaken by Research Councils UK, with the support of this report's inquiry[49], produced a set of principles for open research that could help guide this effort. The members of the public involved were content that researchers and funders oversee open data practices in most cases. When there is a clear public interest (defined by the participants almost exclusively in terms of affects on human health and the environment), the groups wanted ethicists, lawyers, NGOs and economists involved as well. None in the dialogue group were among the growing number of people interested in exploring data for themselves but they were clear that those data should be discoverable for those who wish to explore them.

Governments have also made moves in the direction of greater transparency with the evidence used in their decision making and in assessing the efficiency of public policies. This reflects the view that "Sunlight is…the best of disinfectants"[50] - that greater transparency combats corruption and improves citizens' trust in government. This report stresses a similar point for the governance of science, but emphasising intelligent openness - intelligible and assessable communication - rather than transparency as mere disclosure. The Freedom of Information Act (FoIA) 2000 created a public right of access to information held by public authorities, which include universities and research institutes. Responses to FoI requests can too easily lead to the dumping of uninformative data rather than the effective communication of information. Section 2.2 returns to the particular challenges created by FoIA to researchers.

In 2010, the UK government committed itself to "throw open the doors of public bodies, to enable the public to hold politicians and public bodies to account".[51] This meant publishing the job titles of every member of staff and the salaries of some senior officials. It also included a new "right to data" so that government-held datasets could be requested and used by the public, and then published on a regular basis. In 2011, the Prime Minister reemphasised that his "revolution in government transparency"[52] was as much motivated by a drive for public accountability as by the creation of economic value (see Box 1.2). By opening up public service information over the following year, he argued that the government is empowering citizens: making it easier for the public to make informed choices between providers and

48 Public interest tests for release of information appear in the Freedom of Information Act (2000) and the Environmental Information Regulations (2004). Some circumstances that usually exempt a public authority from providing information are not applicable if it is in the interest of the public for that information to be released. Here the concept of public interest science is used in a way that is distinct from, but related to, these uses, to distinguish those areas of scientific research that deserve more public discussion, and support in creating that discussion.

49 TNS BMRB (2012). *Public dialogue on data openness, data re-use and data management Final Report*. Research Councils UK: London. Available at: http://www.sciencewise-erc.org.uk/cms/public-dialogue-on-data-openness-data-re-use-and-data-management/

50 This quotation originates with US Supreme Court Justice Luis Brandeis. For a discussion of transparency as a regulatory mechanism, see Etzoni A (2010). *Is Transparency the Best Disinfectant?* Journal of Political Philosophy, 18, 389-404.

51 HM Government (2010). *The Coalition: our programme for government*. UK Government: London. Available at: http://www.direct.gov.uk/prod_consum_dg/groups/dg_digitalassets/@dg/@en/documents/digitalasset/dg_187876.pdf

52 Number10, David Cameron (2011). *Letter to Cabinet Ministers on Transparency and Open Data*. Available at: http://www.number10.gov.uk/news/letter-to-cabinet-ministers-on-transparency-and-open-data/

hold the government to account for the performance of public services. Research data falls under the remit of this initiative. The UK's Cabinet Office are due to publish their *Right to Data* white paper as this report goes to press.

It is not yet clear how the demands for spending and services data will extend to the products of publicly funded research. Spending and services datasets are usually large, unstructured, uniform datasets, often built for sharing within a department or agency. This is government 'big data' – similar in many ways to the volumes of customer data collected by private companies. Through initiatives like *data.gov.uk*, data is structured so that it is available to everyone through the web, labelled 'broad data'.[53] Research datasets vary from small bespoke collections to complex model outputs. They are used and managed in vastly different ways (2.1.2).

Research data is mostly not big data, and so it is not easily restructured as broad data. Instead, opening up research data in a useful way requires a tiered approach (4.1). Governments around the world are adopting a *data.gov* approach too, including the recent and ambitious Indian *data.gov.in* (Box 1.4). These portals are far from the programmes that characterise intelligently open research - decentralised initiatives where the demands and uses of data are well understood.

**Box 1.4 Data.gov.in**

The Indian National Data Sharing and Accessibility Policy, passed in February 2012, is designed to promote data sharing and enable access to Government of India owned data for national planning and development. The Indian government recognised the need for open data in order to: maximise use, avoid duplication, maximise integration, ownership of information, increase better decision-making and equity of access. Access will be through *data.gov.in*. As with other *data.gov* initiatives, the portal is designed to be user-friendly and web-based without any process of registration or authorisation. The accompanying metadata will be standardised and contain information on proper citation, access, contact information and discovery.

When compared to the UK's graduated approach and the argument over funding for the original data.gov in the US in 2011, this is an ambitious and fast-paced plan. Their aim is that the government's back catalogue will be online in a year. The policy applies to all non-sensitive data available either in digital or analogue forms having been generated using public funds from within all Ministries, Departments and agencies of the Government of India.

It would be a mistake to confuse the current trend for transparency, by opening up data, with the wider need for trustworthiness. The Research Councils' public dialogue concluded "addressing open data alone is unlikely to have a major impact on governance concerns around research".[54] Those concerns are often more about the motivations of researchers, the rate of the advance of research and when exploitation of research outpaces its regulation.

53 A term used Hendler J (2011). *Tetherless World Constellation: Broad Data*. Available at: http://www.slideshare.net/jahendler/broad-data.
54 TNS BMRB (2012). *Public dialogue on data openness, data re-use and data management Final Report.* Research Councils UK: London. Available at: http://www.sciencewise-erc.org.uk/cms/public-dialogue-on-data-openness-data-re-use-and-data-management/

# Why change is needed: challenges and opportunities

Recent decades have seen the development of extraordinary new ways of collecting, storing, manipulating, and transmitting data and information that have removed the geographical barriers to their movement (Figure 2.1 gives a potted history of key events). Copying digital information has become almost cost free. At the same time, many people are increasingly averse to accepting *ex cathedra* statements from scientists about matters that concern them, and wish to examine and explore the underlying evidence. This trend has been reinforced by new communication channels which, since the world wide web's inception 20 years ago, have become unprecedented vehicles for the transmission of information, ideas and public debate.

The deluge of data produced by today's research has created issues for essential processes at the heart of science. But new digital tools also enable ways of working that some believe have propelled us to the verge of a second open science revolution, every bit as great as that triggered by the invention of scientific journals.[55] Open data, data sharing and collaboration lie at the heart of these opportunities. However, many scientists still pursue their research through the measured and predictable steps in which they communicate their thinking within relatively closed groups of colleagues; publish their findings, usually in peer reviewed journals; file their data and then move on.

This chapter discusses why and how the principle of open data in support of published scientific papers should be maintained in the era of massive data volumes; how open data and collaboration can be the means of exploiting new scientific and technological opportunities; and the extent to which effective open data policies should be part of a wider scientific communication with citizens. Much of the discussion concerns publicly and charitably funded science, but also considers the interface with privately funded science.

55 Nielsen M (2012). *Reinventing discover: the new era of networked science*. Princeton University Press: Princeton.

## Figure 2.1 Gazing back: a recent history of computational and data science[56]


**1960: Hypertext**
*Imagining connectivity in the world's knowledge*
The concept of links between documents begin to be discussed as a paradigm for organizing textual material and knowledge.

**1960: Full-Text Search**
*Finding text without an index*
The first full-text searching of documents by computer is demonstrated.

**1962: Roger Tomlinson**
*Computerizing geographic information*
Roger Tomlinson initiates the Canada Geographic Information System, creating the first GIS system.

**1963: ASCII Code**
*A standard number for every letter*
ASCII Code defines a standard bit representation for every character in English.

**1963: Science Citation Index**
*Mapping science by citations*
Eugene Garfield publishes the first edition of the Science Citation Index, which indexes scientific literature through references in papers.

**1963: Data Universal Numbering System (D-U-N-S)**
*A number for every business*
Dun & Bradstreet begins to assign a unique number to every company.

**1966: SBN Codes**
*A number for every book*
British SBN codes are introduced, later generalized to ISBN in 1970.

**1967: DIALOG**
*Retrieving information from anywhere*
The DIALOG online information retrieval system becomes accessible from remote locations.

**1968: MARC**
Henriette Avram creates the MAchine-Readable Cataloging system at the Library of Congress, defining metatagging standards for books.

**1970s: Relational Databases**
*Making relations between data computable*
Relational databases and query languages allow huge amounts of data to be stored in a way that makes certain common kinds of queries efficient enough to be done as a routine part of business.

**1970-1980s: Interactive Computing**
*Getting immediate results from computers*
With the emergence of progressively cheaper computers, it becomes possible to do computations immediately, integrating them as part of the everyday process of working with knowledge.

**1970—1980s: Expert Systems**
*Capturing expert knowledge as inference rules*
Largely as an offshoot of AI, expert systems are an attempt to capture the knowledge of human experts in specialized domains, using logic-based inferential systems.

**1973: Black-Scholes Formula**
*Bring mathematics to financial derivatives*
Fischer Black and Myron Scholes give a mathematical method for valuing stock options.

**1973: Lexis**
*Legal information goes online*
Lexis provides full-text records of US court opinions in an online retrieval system.

**1974: UPC Codes**
*Every product gets a number*
The UPC standard for barcodes is launched.

**1980s: Neural Networks**
*Handling knowledge by emulating the brain*
With precursors in the 1940s, neural networks emerge in the 1980s as a concept for storing and manipulating various types of knowledge using connections reminiscent of nerve cells.

**1982: GenBank**
*Collecting the codes of life*
Walter Goad at Los Alamos founds GenBank to collect all genome sequences being found.

**1983: DNS**
The Domain Name System for hierarchical Internet addresses is created; in 1984, .com and other top-level domains (TLDs) are named.

**1984: Cyc**
*Creating a computable database of common sense*
Cyc is a long-running project to encode common sense facts in a computable form.

**1988: Mathematica**
*Language for algorithmic computation*
Mathematica is created to provide a uniform system for all forms of algorithmic computation by defining a symbolic language to represent arbitrary constructs and then assembling a huge web of consistent algorithms to operate on them.

**1989: The Web**
*Collecting the world's information*
The web grows to provide billions of pages of freely available information from all corners of civilization.

**1990: IMDb**
*Indexing movies*
The Internet Movie Database is launched.

**1991: Gopher**
*Burrowing around the internet*
Gopher provides a menu-based system for finding material on computers connected to the internet.

**1991: Unicode**
*Representing every language*
The Unicode standard assigns a numerical code to every glyph in every human language.

**1991 arXiv.org established:** open access e-print repository for journal articles from physics physics, mathematics, computer science, and related disciplines.

**1993: Tim Berners-Lee**
*A catalog of the web*
Tim Berners-Lee creates the Virtual Library, the first systematic catalog of the web.

**1994: QR Codes**
Quick Response (QR) scannable barcodes are created in Japan, encoding information for computer eyes to read.

**1994: Yahoo!**
Jerry Yang and David Filo create a hierarchical directory of the web.

**1995: CDDB**
*Indexing music*
Ti Kan indexes CDs with CDDB, which becomes Gracenote.

**1996: The Internet Archive**
*Saving the history of the web*
Brewster Kahle founds the Internet Archive to begin systematically capturing and storing the state of the web.

**1997 Launch of SETI@ home** Individuals can provide their computing resources to help in data analysis for the search for extra terrestrial intelligence.

**1998: Google**
*An engine to search the web*
Google and other search engines provide highly efficient capabilities to do textual searches across the whole content of the web.

**2000: Sloan Digital Sky Survey**
*Mapping every object in the universe*
The Sloan Digital Sky Survey spends nearly a decade automatically mapping every visible object in the astronomical universe.

**2000: Web 2.0**
*Societally organized information*
Social networking and other collective websites define a mechanism for collectively assembling information by and about people.

**2001: Wikipedia**
*Self-organized encyclopedia*
Volunteer contributors assemble millions of pages of encyclopedia material, providing textual descriptions of practically all areas of human knowledge.

**2003: Human Genome Project**
*The complete code of a human*
The Human Genome Project is declared complete in finding a reference DNA sequence for every human.

**2004: Facebook**
*Capturing the social network*
Facebook begins to capture social relations between people on a large scale.

**2004: OpenStreetMap**
Steve Coast initiates a project to create a crowdsourced street-level map of the world.

**2004: the UK Government's Office of Public Sector Information** began pilot scheme to use the Semantic Web to integrate and publish information from across the Public Sector.

**2009: Wolfram|Alpha**
www.wolframalpha.com
*An engine for computational knowledge*
Wolfram|Alpha is launched as a website that computes answers to natural-language queries based on a large collection of algorithms and curated data.

56 WolframAlpha (2012). *Timeline of systematic data and the development of computable knowledge*. Available at: http://www.wolframalpha.com/docs/timeline/computable-knowledge-history-6.html

## 2.1 Open scientific data in a data-rich world

### 2.1.1 Closing the data-gap: maintaining science's self-correction principle

Technologies capable of acquiring and storing vast and complex datasets challenge the principle that science is a self-correcting enterprise. How can a theory be challenged or corrected if the data that underlies it is neither accessible nor assessable? The norm for many scientists 30-40 years ago was to publish a paper that included a complete description of an experiment, the resultant data, an assessment of uncertainties and details of the metadata required to validate, repeat or reuse the data. However, the new ways of collecting data have created such a vast data deluge that although it has been the basis of much scientific achievement in many areas it has become so great and so complex that no journal could conceivably publish data in the same way as before (see Box 2.2). A great deal of data has become detached from the published conclusions that depend upon it, such that the two vital complementary components of the scientific endeavour - the idea and the evidence - are too frequently separated. This represents a serious data-gap that is inimical to the rigorous scrutiny to which scientific conclusions should be subject, thereby undermining the principle of self-correction. The principle must be maintained so that the data underlying a scientific argument is accessible for rigorous analysis and replication, and ways must be found to reconnect them.

The ideal would be for the data that supports an argument in a published paper, together with the metadata that makes them comprehensible and usable, to be lodged in a curated database that is accessible via the click of a mouse on a live link in the published paper. This is happening in some fields. For example, over 50% of journals in the -omics (genomics, transcriptomics, metabolomics, proteomics, etc.) and bioinformatics fields mandate that the data that underlies their published papers are submitted to a specified data centre (such as that developed by EBI) and conform to specified data standards. Although the trend of compliance to this imperative is positive, a recent review[57] showed that it is still low (Box 2.1).

57 Alsheikh–Ali A A, Qureshi W, Al-Mallah M H & Ioannidis J P A (2011). *Public Availability of Published Research Data in High-Impact Journals*. PLoS ONE, 6, e24357.

## Box 2.1 Compliance with journal data-sharing policies

Of the 50 highest-impact journals in biomedicine, 22 require public sharing of specific raw data as a condition of publication, a further 22 encourage data sharing without binding instruction, while six of the 50 journals had no published policy for data sharing Notwithstanding these policies, a review of the first ten papers published in each journal in 2009 (500 in all) showed that of 351 papers covered by some data-sharing policy, only 143 fully adhered to that policy. Neglecting to publish microarray data, such as those produced in gene-expression studies, was the most common offence. Only 47 of the papers (9%) had deposited the full raw data online.[58]

| | | Policy of Required Deposition for Types of Data | | | | Policy of Provision of Materials and Methods | | | Full data deposited |
|---|---|---|---|---|---|---|---|---|---|
| Journal | Impact Factor | Microarray | Nucleic Acid | Protein | Macromolecular | Materials upon request | Protocols upon request | Conditions of publication | % of papers |
| New England Journal of Medicine | 52.589 | ■ | | | | | | | 0 |
| Cell | 29.887 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 1 |
| Nature | 28.751 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Lancet | 28.638 | ■ | | | | | | | 0 |
| Nature Medicine | 26.382 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Science | 26.372 | ■ | ■ | ■ | ■ | ■ | ■ | | 1 |
| Nature Immunology | 26.218 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 9 |
| Nature Genetics | 25.556 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| JAMA | 25.547 | | ■ | ■ | | | | | 1 |
| Nature Biotechnology | 22.848 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 5 |
| Nature Materials | 19.782 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Immunity | 19.266 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Nature Cell Biology | 17.623 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Journal of Clinical Investigation | 16.915 | ■ | ■ | ■ | | ■ | ■ | ■ | 0 |
| Archives of General Psychiatry | 15.976 | | ■ | ■ | | | | | 0 |
| Journal of the National Cancer Institute | 15.678 | ■ | ■ | ■ | | ■ | ■ | | 0 |
| Nature Neuroscience | 15.664 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 1 |
| Journal of Experimental Medicine | 15.612 | ■ | ■ | ■ | ■ | ■ | ■ | | 0 |
| Annals of Internet Medicine | 15.516 | | | | | | | | 0 |
| Journal of Clinical Oncology | 15.484 | | | | | | | | 0 |
| Nature Methods | 15.478 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 6 |
| Genes and Development | 14.795 | | ■ | ■ | ■ | ■ | | | 3 |
| Nature Physics | 14.677 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| PLoS Biology | 13.501 | ■ | ■ | ■ | ■ | | | | 2 |
| Neuron | 13.41 | ■ | ■ | ■ | ■ | ■ | | | 0 |
| Molecular Cell | 13.156 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Circulation | 12.755 | ■ | ■ | ■ | | | | | 0 |
| PLoS Medicine | 12.601 | ■ | ■ | ■ | ■ | | | | 0 |
| Development Cell | 12.436 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Gastroenterology | 11.673 | ■ | ■ | ■ | | ■ | | | 0 |
| Genome Research | 11.224 | ■ | ■ | ■ | ■ | ■ | | ■ | 6 |
| American Journal of Human Genetics | 11.092 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 3 |
| Nature Structural and Molecular Biology | 11.085 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Journal of the American College of Cardiology | 11.054 | | | | | | | | 0 |
| Blood | 10.896 | ■ | ■ | | ■ | | | | 0 |
| Hepatology | 10.374 | | ■ | | | | | | 0 |
| Current Biology | 10.539 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| Gut | 10.015 | | | | | | | | 0 |
| British Medical Journal | 9.723 | | | | | | | | 0 |
| Circulation Research | 9.721 | ■ | ■ | | ■ | ■ | ■ | | 1 |
| Plant Cell | 9.653 | | | | ■ | | | ■ | 0 |
| Nano Letters | 9.627 | | | | | | | | 0 |
| Journal of Cell Biology | 9.598 | ■ | ■ | ■ | ■ | | | | 0 |
| PNAS | 9.598 | ■ | ■ | ■ | ■ | ■ | | | 1 |
| Molecular and Cellular Proteomics | 9.425 | | ■ | ■ | | | | | 7 |
| PLoS Pathogens | 9.336 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |
| American Journal of Psychiatry | 9.127 | | | | | | | | 0 |
| American Journal of Respiratory and Critical Care Medicine | 9.074 | ■ | ■ | | | | | | 0 |
| Annals of Neurology | 8.813 | | | | | | | | 0 |
| PLoS Genetics | 8.721 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | 0 |

Publishers should require datasets that relate to published papers to be lodged in an appropriate, electronically accessible format that is identified in the published paper. Alternatively, the publication should indicate when, and under what conditions, the data will be available for others to access. The Royal Society has recently updated its journal data policy in line with these requirements (see Box 2.2). Appropriate standards for data and metadata provision need to be applied, and the funders of research need to incorporate the costs of data and metadata compilation as part of the cost of the research process. Suggestions and recommendations for ways in which this can be done are presented in chapters 4 and 5.

58 Alsheikh–Ali A A, Qureshi W, Al-Mallah M H & Ioannidis J P A (2011). *Public Availability of Published Research Data in High-Impact Journals*. PLoS ONE, 6, e24357.

### Box 2.2 Royal Society Publishing data and material sharing policy

To allow others to verify and build on the work published in Royal Society journals it is a condition of publication that authors make the available the data and research materials supporting the results in the article.

Datasets should be deposited in an appropriate, recognised repository and the associated accession number, link or DOI to the datasets must be included in the methods section of the article. Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available). Where no discipline-specific data repository exists authors should deposit their datasets in a general repository such as Dryad (http://datadryad.org/).

Where possible any other relevant research materials (such as statistical tools, protocols, software etc) should also be made available and details of how they may be obtained should be included in the methods section of the article.

Authors must disclose upon submission of the manuscript any restrictions on the availability of research materials or data.

### 2.1.2 Making information accessible: diverse data and diverse demands

This report is unequivocal that there is an imperative to publish intelligently open data when that data underlies the argument of a scientific paper. But many more data are produced from scientific projects than are used to support resulting publications. Combining data in structured datasets offers considerable opportunities for discovery and considerable loss of potential if they are not. There is a trend towards this kind of collaborative data management, often through public and accessible databases. But it is by no means universal across all the areas of science that could benefit.

Success in understanding and unravelling biological problems is more and more reliant on the routine ability to analyse data that is collated in major databases. Since the 1996 Bermuda Principles[59], genome sequencing data must be immediately released into the public domain. The release of this data is relatively straightforward and new sequencing techniques have recently taken sequence data storage into the petabyte range (see Box 2.3). Keeping this growing data accessible is a continual challenge. In 2012, 200 terabytes of data from the international 1000 Genomes Project will be uploaded onto the Amazon Web Services Cloud in an attempt to overcome current problems of access. Developing storage, access and analysis tools alongside this have led to a large and growing bioinformatics community - melding molecular biology with informatics and producing curated, globally accessible resources. These tools are an essential part of recent progress in using genomic information to understand human diseases and in the identification of new molecular targets for drug discovery.

Sophisticated modelling tools have been built using open data resources. The international *In Silico Oncology* collaboration developed a new mathematical model for cancer growth[60] that uses recent advances in medical modelling and real patient data to model the likely tumour response to different therapeutic regimes. The 'oncosimulator' changes its parameters according to the clinical conditions of a new patient. It is anticipated that this model will support doctors and patients making decisions about treatment by providing a personalised scheme for treatment. This bespoke treatment only came about because researchers had access to databases of cancer patient records.

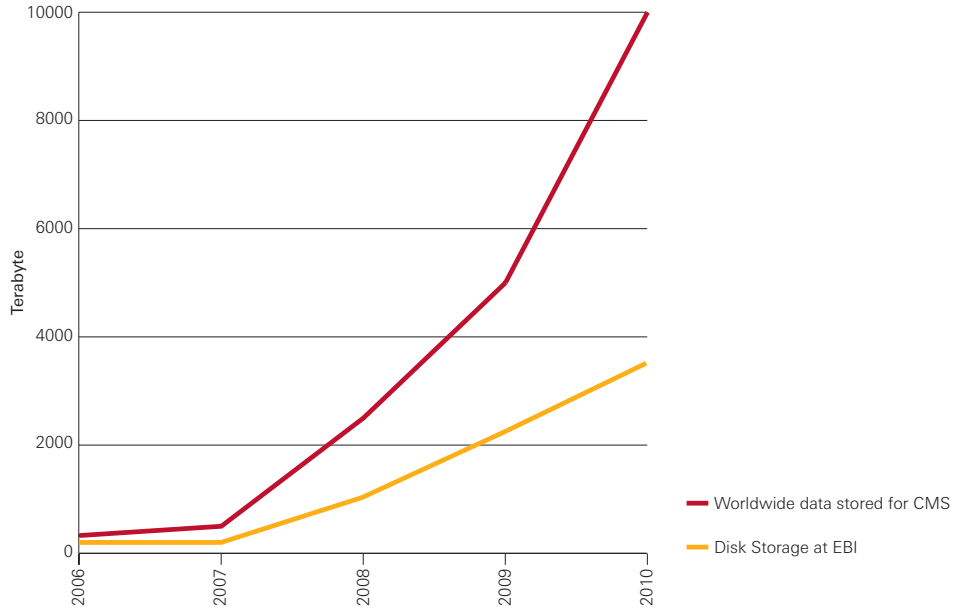
59 In 1996, during the First International Strategy Meeting on Human Genome Sequencing in Bermuda, it was agreed that primary genomic sequence should be rapidly released into the public domain: "all human genomic sequence information, generated by centres funded for large scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society". Human Genome Project (2003). Available at: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1

60 In Silico Oncology Group (2011). Available at: http://in-silico-oncology.iccs.ntua.gr/english/index.php

### Box 2.3 Growing scientific data – the European Bioinformatics Institute (EBI) and the Compact Muon Solenoid (CMS) detector at CERN

The figure below shows the increase in data storage capacity at EBI and worldwide data storage for the CMS experiment at CERN between 2006 and 2010. See appendix 1 for more details of each case.

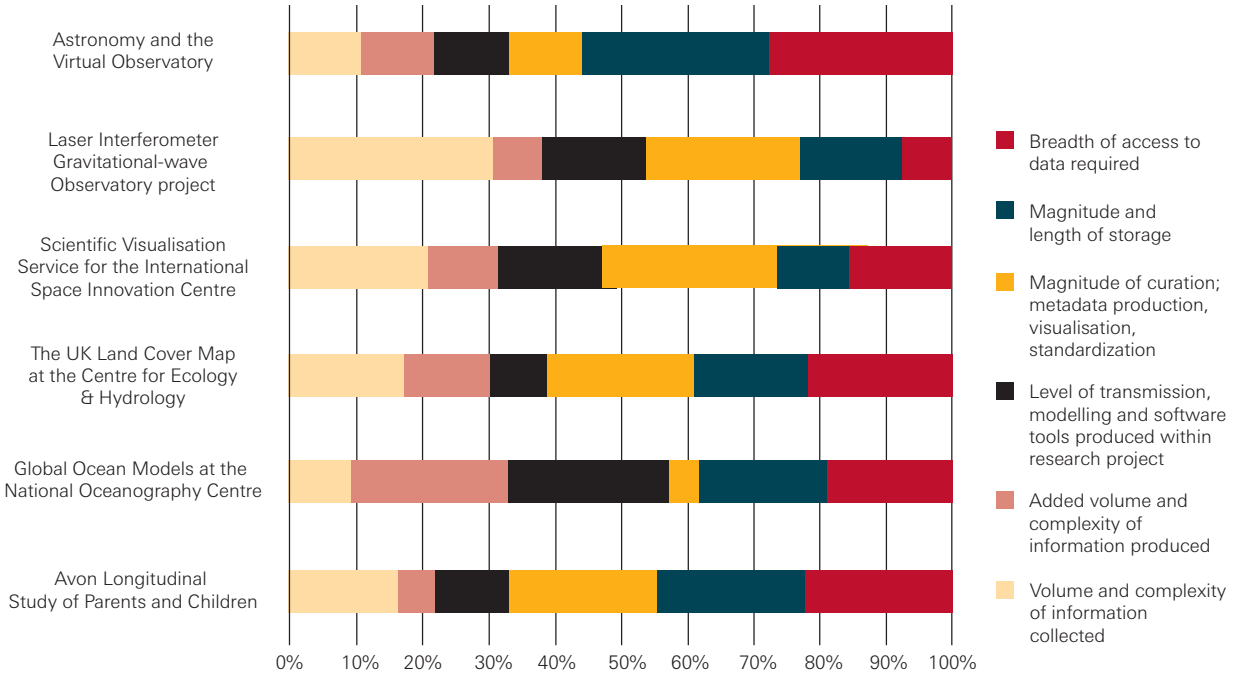
Data volumes at EBI will keep increasing as the institute hosts data on behalf of an ever widening research community including basic biology researchers, clinical and environmental life scientists. But they are unlikely to catch-up with the volumes of particle physics data produced by CMS. This project keeps raw data collected by the detector and derived data, which is used in analysis and simulation data. There are multiple copies of data: raw data is always replicated at two out of seven large computing facilities for redundancy. Derived data is stored at multiple sites in a distributed computing grid of approximately 50 sites to allow the data to be analysed efficiently.




There are now many scientific fields where the collection and integration of data in major databases is seen as a community good in itself, for testing theories as widely as possible and as a source of new hypotheses. Appendix 1 gives examples of the different ways researchers share data. Figure 2.2 illustrates how these diverge according to the type of data and demands for access and reuse.

**Figure 2.2 Changing data demands in the research lifecycle**

An illustration of the diverse attributes of data production in a range of scientific studies. The type of open data is tailored to the nature of the data, the curation and storage effort, and requirements for data access. Full description of examples in appendix 1. To illustrate the attributes of each project, the relative percentage below roughly equate to work done on that attribute.




In contrast to the examples in appendix 1, that show formal data sharing in fields of science that have recognised its potential for benefit, there are many areas of science where data sharing individuals or small groups is neither common nor expected. It tends, almost exclusively, not to occur in an open mode but between close, trusted collaborators. However, it has been argued that in the longer term this "small science" research could create at least as valuable data as big, formal data-rich science.[61] This is partly due to the increase in readily available research tools that produce datasets that are very large compared to the size of a laboratory.[62] Scientists in these areas are increasingly turning to their university libraries and institutional repositories for support for their data, where universal curation models - typical of major databases - are inappropriate because of the complexities of production and communication in diverse small science disciplines. An international community initiative based at Stanford University (Lots of Copies Keep Stuff Safe – LOCKSS[63]) provides tools and support for institutions to collect and preserve their own e-content. Support at an institutional level is important given the tendency for traditional but successful small-science activities to evolve into medium- or large-science collaborative science as they forge novel science opportunities. It is important to stimulate awareness of the potential for data curation and of modern data-intensive science and to ensure that the diversity of skills and tools needed to facilitate sustainable curation and data-intensive science are readily available. These issues of data and skills management are addressed in chapter 4, which also addresses the timing of data release and the attribution of credit for data compilation.

61 Carlson S (2006). *Lost in a sea of science data*. The Chronicle of Higher Education, June.
62 Cragin M H, Palmer C L, Carlson J R and Witt M (2010). *Data sharing, small science and institutional repositories*. Philosophical Transactions Royal Society A, 368, 4023-2038.
63 Library of Congress (2012). Lockss program. Available at: http://www.digitalpreservation.gov/partners/lockss.html

### 2.1.3 A fourth paradigm of science?

A staple part of scientific inquiry has been the observation of patterns in nature followed by testable theories of their causes. The power of modern computers permits highly complex and hitherto unperceived relationships to be identified, and has become the central thrust of the e-science effort in the UK[64] and elsewhere. It recognises that informatics need not merely support traditional ways of conducting inquiry in a particular discipline, but can fundamentally change the development of a discipline.

Some have argued[65,66] that this represents a fourth paradigm of scientific research. The classic duo of experiment and theory were joined by a third paradigm of science, that of simulation after the advent of the modern computer. The data collections summarised in section 2.1.2 create the potential through the immense data-sorting, analysis and manipulative abilities of computers to infer information and relationships on a scale that is so much greater than hitherto possible that it represents a fourth paradigm of science. Rather than hypotheses being tested and developed from data collected for that purpose, hypotheses are constructed after identifying relationships in the dataset. In this data-led approach the data comes first, embedded in a sequence of data capture, curation and analysis.

An example of the application of this approach is represented by the UK Biobank, which contains blood, urine and saliva samples from 500,000 people who have provided personal data and agreed to have their health status followed. This database will offer a new resource for understanding the prevalence and development of cancer, strokes, diabetes, arthritis, depression and forms of dementia. It exemplifies the power of a large cohort of data and how databases need not be restricted to digital information.

### 2.1.4 Data linked to publication and the promise of linked data technologies

There is an important trend towards more effective harvesting of data from published literature and towards linking publication through live links to data sources, together with dynamic up-dating of data and metadata. This is not only an important means whereby the gap that has emerged between published ideas and underlying data can be closed, but also a way of making the publication-to-data relationship more dynamic. A powerful recent enhancement of this trend has been the creation of PubChem: a free, open database of chemical structures of small organic molecules and information on their biological activities; and PubMed[67], an open access database that comprises more than 21 million citations to the biomedical literature from MEDLINE[68], from life science journals, and from online books. Citations may include links to full-text content from PubMed Central and publishers' websites.

An international group of academics, librarians, publishers and funders, Force 11, submitted evidence to this report's study arguing that the journal article is one among many forms of knowledge exchange. Knowledge exchange relies on a "research object, a container for a number of related digital objects - for example a paper with associated datasets, workflows, software packages, etc, that are all the products of a research investigation and that together encapsulate some new understanding". Live links between an article and the data that underlie it, permitting the reader to manipulate the data while reading the article, are means of realising the Force 11 vision. Figure 2.3 illustrates a scheme called the *Collage Authoring Environment* for adding interactive elements to standard online journal articles and producing 'executable' papers.[69]

64 Walker D W, Atkinson M P, Brooke J M and Watson P (2011). *Special theme: E-science novel research, new science and enduring impact*. Philosophical Transactions of the Royal Society A, 369, 1949.

65 Gray J (2009). *E-Science: a transformed scientific method*. In: The Fourth Paradigm: data-intensive scientific discovery. Hey T, Tansley S and Tolle K (eds.). Microsoft Research: Washington.

66 Shadbolt N, Berners-Lee T, and Hall W (2006). *The Semantic Web Revisited*. IEEE Intelligent Systems, 21, 3, 96-101. ISSN 1541-1672. Available at: http://eprints.soton.ac.uk/262614/

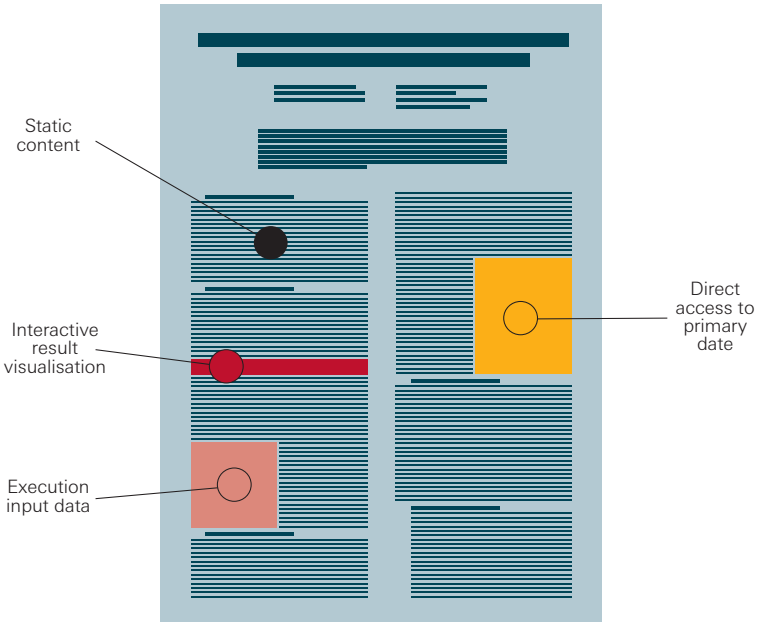
67 NCBI (2012). *PubMed*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/

68 U.S. National Library of Medicine (2012). *MEDLINE®/PubMed® Resources Guide*. Available at: http://www.nlm.nih.gov/bsd/pmresources.html

69 Elsevier (2011). *Executable Paper Grand Challenge*. Available at: http://www.executablepapers.com

**Figure 2.3 An Executable Paper in the Collage Authoring Environment**[70]

Conceptual view of the executable paper. Static content (the body of the publication) is extended by interactive elements. readers can access primary data and reenact computations in order to validate the presented conclusions or navigate result spaces. Subject to the authors' approval, readers can also obtain access to the underlying code of the experiments presented in the publication. It is a web based infrastructure, which can be integrated with the publisher's portal.




In addition to linking data to articles, it is now possible to link databases directly to other related databases. This is not just by mutual citation; linked semantic data technologies promise a much deeper integration. This is because semantic data is data that are tagged with particular metadata - metadata that can be used to derive relationships between data. In a simple example, imagine if a computer can understand not just that the title deed to a house and the name of the homeowner are related, but how they are related. This could automate the process of updating deeds when a house is sold.[71] Including machine readable information that describes as well as identifies data creates opportunities for computerised data comparison – using the type of relationship between data and not just the fact that they are related.

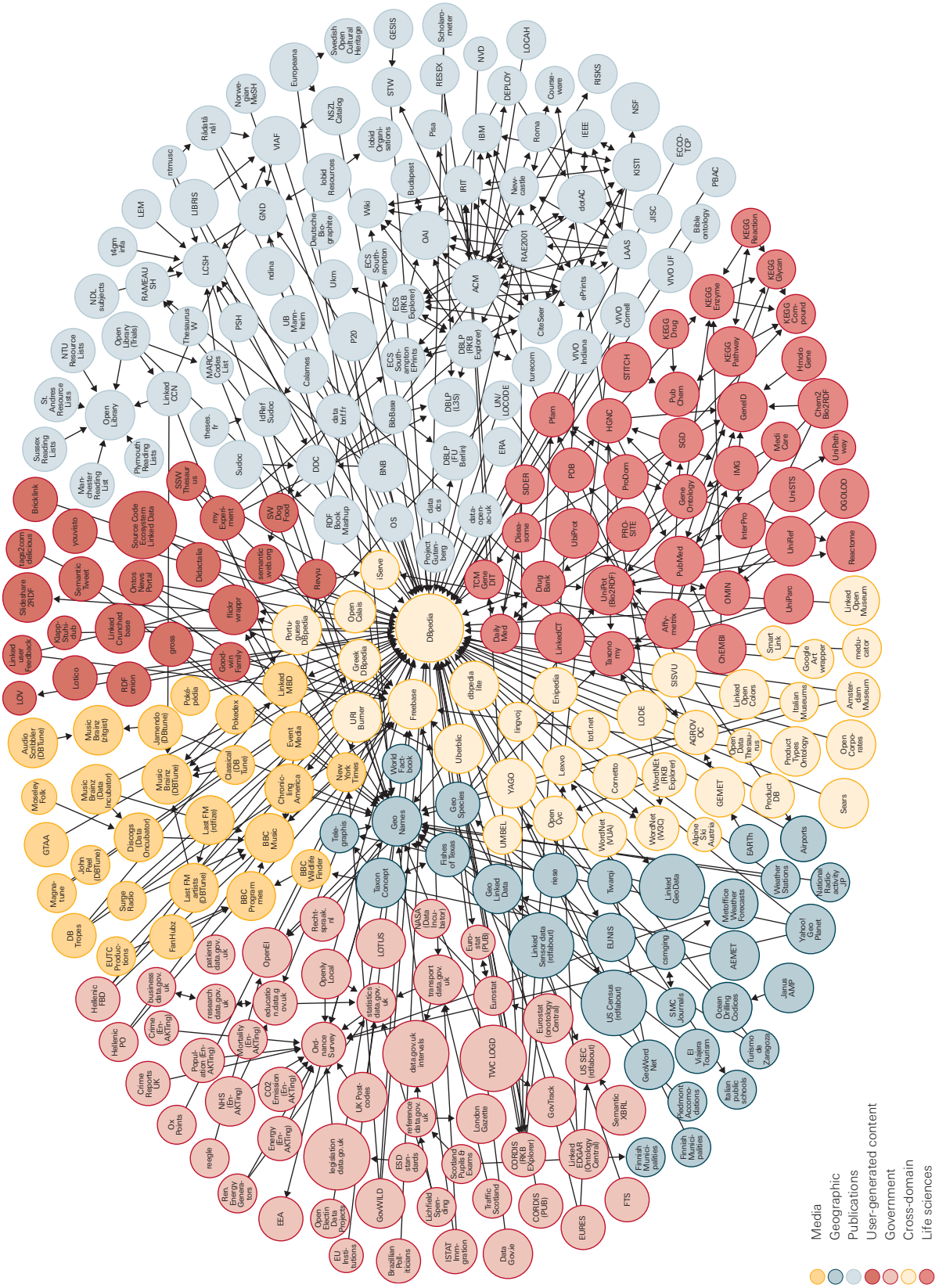
By *standardising* identifiers and descriptions, a rich global web of linked datasets has been developed. Using the same Uniform Resource Identifiers (URIs) and ascribed metadata via Resource Description Frameworks (RDFs), organisations such as the BBC, Thomson Reuters and the Library of Congress now link together 295 datasets. An organisation to champion this system, W3C SWEO Linking Open Data Community Project, was launched in 2007 and the number of RDF links between datasets has since increased from 120,000 in 2007 to 504 million in 2011 (Figure 2.4).

70 Nowakowski *et al* (2011). *The Collage Authoring Environment*. Procedia Computer Science, 4, 608–617. Available at: http://www.sciencedirect.com/science/article/pii/S1877050911001220

71 This example was first used by Berners-Lee T (1994). *The first World Wide Web Conference in 1994.* Plenary at world wide web Geneva conference. Available at: http://www.w3.org/Talks/WWW94Tim/. The more general overview is Shadbolt N, Berners-Lee T, and Hall W (2006). *The Semantic Web Revisited*. IEEE Intelligent Systems, 21, 3, 96-101. ISSN 1541-1672

72 W3C (2012). Available at: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

## Figure 2.4 A map of interlinked data from the Linking Open Data Community Project[72]



If search engines represent a first generation of tools to sort networked knowledge, semantic analysis represents a second generation which not just identifies lists of documents (or databases) but also the relationships between them. This is an exciting prospect not just because the ability to mash data together can produce new knowledge – a global extension of the fourth paradigm described in 2.1.3 - but also promises as yet unpredictable changes to scholarly practice. From the first generation of tools, Google Scholar has become the primary form of dissemination for some researchers,[73] with journal publication providing no more than the official stamp of quality on their work. What changes might a mature semantic web of data make to the way researchers share data?

There are, however, problems in linking datasets to produce deeper and better integrated understanding. The vocabulary used in the semantic description of data – ie in the metadata – can so greatly vary between heterogeneous linked datasets that the whole lacks a shared vocabulary capable of revealing the underlying meaning. The consequence is to produce siloed sections of the web of linked data. The datasets are linked but are not truly interoperable. Some argue that the RDF system is not well suited to address this problem and that other ways of linking data are needed.[74] One way to achieve this would be to improve systems for searching metadata; leaning on the free-text indexing that sits behind today's sophisticated search engines.

Although many of the datasets in Figure 2.4 are periodically updated, the RDF derivative is not usually updated in parallel with the database, leaving stale data in the web of data. There is much to do before the semantic web can offer the curatorial functions that make some bespoke databases trustworthy.

73 Academics often keep working papers and have well-kept publication pages because these appear on Google Scholar alongside online journal pages.

74 Freitas A, Curry E, Gabriel Oliveira J, O'Riain S (2012). *Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends*. Internet Computing, IEEE, 16, 1, 24-33. Available at: http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=4236

### 2.1.5 The advent of complex computational simulation

Mathematical modelling of phenomena has long been a tool of science. A mathematical model is a formal quantitative articulation of a scientific theory. It enables scientists to approximate an exact, but non-quantitative theoretical formulation of a problem to yield quantitative predictions and insights. The manipulation of raw data in ways that make it usable frequently involves modelling. The exponential rise in available computing power has taken mathematical modelling to new levels of sophistication. For the first time this permits scientists to undertake simulations able to explore the behaviour of truly complex systems, such as the evolution of the climate, the actions of whole organisms or the structure and dynamics of cities.[75] Today's simulation techniques so pervade scientific practice that they have added a third basic tool to those of theory and experiment.[76] They have moved on from assisting scientists in doing science, to transforming both how science is done and what science is done. They are fundamental to many open questions in science, in domains that span from Climate and Earth System science (eg Slingo *et al*[77]) to Epidemiology (especially pandemic models, eg Ferguson[78]), from Species Distribution modelling (eg Benton[79]) to Immunology (eg Coen[80]).

A computer simulation is analogous to a physical experiment, but an experiment conducted with mathematical equations rather than with physical entities. Their output could, therefore, be regarded as data that is analogous to the data produced by a physical experiment. This is not to confuse it with the original measurement data on which many models rely, either as input or as a means of correcting results of intermediate simulations before performing further iterations. In principle, therefore, there is no reason why data from simulations should be regarded as different from other forms of scientific data. Although they need to be accompanied by a full description of the algorithms on which the computation is built, and, possibly, by details of the code and computer characteristics. In some areas of the biological sciences it has become routine for researchers to make computational models available, for example EBI run a biomodels database of peer reviewed, published computational biological models.[81]

The complexities of many simulations pose problems of intelligibility. The complexity of the computations within the simulation is often such that it is difficult to make a simple statement about the relationship between cause and effect, a relationship that reductive science has historically attempted to enunciate. For example, quantum chemistry produces *ab initio* simulations that predict chemical reactions from first principles of quantum mechanics, producing predictions without a clear causal relationship between input and output. In some areas, this has led to scepticism about the validity of such simulations. "It's only a model" has been a perennially dismissive comment, for example about climate simulations, which are, however, essential for understanding a highly complex system.

The predictive power of a simulation or its capacity to represent reality depend on its accuracy, the closeness of the approximations that are made to an exact theoretical formulation of the problem, and its precision or repeatability, determined by the numerical algorithms used to solve them for the processes in the simulation and the numerical precision of the computer. The predictions of many simulations are subject to very large errors due to the uncertainty in assumptions that they need to make about poorly understood properties or relationships and which are important to their operation. A frequent problem arises from the sensitivity of the systems of non-linear equations that many simulations manipulate. They can create large changes in output from small changes in data, model code or in the computing environment. Part

75 Wilson A (2012). *The science of cities and regions: lectures on mathematical model design*. Springer: Heidelberg.
76 Bell G, Hey T & Szalay A (2009). *Computer Science. Beyond the data deluge*. Science: New York, 323, 5919, 1297-8. Available at: http://www.sciencemag.org/content/323/5919/1297.short.
77 Slingo *et al* (2009). *Developing the next-generation climate system models: challenges and achievements*. Philosophical Transactions of the Royal Society A – Mathematical, Physical and Engineering Sciences, 367, 1890, 815-831.
78 Ferguson *et al* (2007).*The role of mathematical modelling in pandemic preparedness*. International Journal of Antimicrobial Agent, 29, 2, S16-S16.
79 Benton M J (1997). *Models for the diversification of life*. Trends in Ecology & Evolution, 12, 12, 490-495.
80 Coen P G (2007). *How mathematical models have helped to improve understanding the epidemiology of infection*. Early Human Development, 83, 3, 141-148.
81 EMBL-EBI (2012). *BioModels Database – A Database of Annotated Published Models*. Available at: http://www.ebi.ac.uk/biomodels-main/

of the problem arises from chaotic behaviour, so that validation needs to be statistically based rather than based on exact replication.

As computer simulation matures as a basic tool of science, high commercial standards of coding must increasingly become the norm in research laboratories. This is particularly difficult in a research environment where simulations are the main tool of inquiry and where code is continually developing and incorporating new insights.[82] An important aim, however, must be to ensure that the behaviour of a simulation is determined only by the algorithms on which it is based and not by variations in the style of coding or differences between computers or compilers.

Increasing commoditisation of computing is creating increasingly powerful software tools such as the BLAST[83] sequence matching tool, and process-based algorithms and languages such as SpiM[84] that enable complex dynamical biological and ecological (ie natural) processes to be more effectively represented in a form that allows them to be tested as models and 'executed' in practice. In the light of these developments, the need for better ways for the scientific community to share and communicate models is growing, and likely to become a central requirement of publishing over the next few years. The scientific community has begun to develop common standards to address some of the basic issues in model sharing and model communication, for example in biological modelling, where formats such as SBML (Systems Biology Mark-up Language) and CellML (Cell Mark-up Language) already exist, EBI runs a biomodels database of peer reviewed and published computational biological models.[85]

Such standards set out minimum requirements for specifying models so that they can be compared and tested against data. Standards alone will not fulfil the requirements for sharing and communicating models. This will need to be achieved through the requirements to make model software codes both open and accessible to the scientific community and to peer reviewers. This movement may well be largely driven by research funding agencies. Although the issue of the publication of models is outside the scope of this report, it is appropriate to include a pointer here to this critical issue which is on the horizon for the open publishing debate.

A significant part of the training of scientists and the practice of established scientists who create computer simulations should be the discipline that has been intrinsic to experimental science - that of recording details of the computational experiment with fidelity. By the same token, it is important that the details of simulations are exposed to scrutiny by other competent experts to analyse its operation and replicate its findings. The details required to permit replication will vary according to the nature of the simulation. For example, the British Atmospheric Data Centre provides excellent guidance on when and how to curate information and data for climate simulation.[86]

In principle, simulations whose conclusions have major implications for society should be assessable by citizens. However, simulations such as those that evaluate the operation of cities as a basis for planning or that simulate the operation of climate and environmental systems and forecast their futures, are often highly complex, and only truly assessable by experts. Nonetheless, it is important that the operation and output of simulations where major public interest is at stake should be set out in ways that expose problematic issues and identify levels of uncertainty both in their forecasts and in their implications for policy. Efforts of bodies such as the Climate Code Foundation in improving the transparency and public communication of the software used in climate science are important in this regard.[87]

---

82 As was addressed in the roundtable on computer modelling (see appendix 4), large and complex models in addition frequently represent many man-years of work, and there is understandable reluctance to publish models that may either have commercial value or could be used to create high impact publications. Issues of commercial value, and of incentives for sharing data and models are addressed elsewhere in the report..

83 Basic Local Alignment Search Tool (2012). Available at: http://blast.ncbi.nlm.nih.gov/Blast.cgi

84 Larus J (2011). *SpiM: a MIPS32 Simulator*. Available at: http://spimsimulator.sourceforge.net/

85 EMBL-EBI (2012). BioModels Database – *A Database of Annotated Published Models*. Available at: http://www.ebi.ac.uk/biomodels-main/

86 NERC (2012). Archiving of Simulations within the NERC Data Management Framework: BADC Policy and Guidelines. Available at: http://badc.nerc.ac.uk/data/BADC_Model_Data_Policy.pdf

87 Climate Code Foundation (2012). Student Internships. Available at: http://climatecode.org/

### 2.1.6 Technology-enabled networking and collaboration

The geography and sociology of research is changing. Since the 1940s there has been a conscious recognition of the role which science plays in supporting national economies. This has ushered in an era of research funding directed towards national priorities.[88] John Ziman argued[89] that science has reorganised itself around these priorities and that research working methods are no longer centred on the individual or small group but on large collectives. These formal collaborations are now a well-established part of the scientific landscape, from nations time-sharing large telescopes to pooled climate data analysis through the UNFCCC.[90,91] Over 35% of articles published in journals are based on international collaboration compared with 25% 15 years ago. A great many have been enabled by computational and communications technologies that have stimulated new ways of conducting science, many of which depend upon on enhanced levels of collaboration through virtual global networks[92] and professional communities of shared interest, motivated by the exchange of scientific insight, knowledge and skills that are changing the focus of science from the national to the global.

Science is increasingly interdisciplinary:[93] the boundaries between previously distinct fields are blurring as ideas and tools are exported from one discipline to another. These shifts challenge the way that science is funded, conducted, communicated, evaluated and taught. Effective access to data resources are important in this transition, but more proactive data sharing is necessary if new opportunities are to be seized.

Novel communication technologies permit modes of interaction that change the social dynamics of science and exploit the collective intelligence of the scientific community. Free online resources and search engines have become integral to science in ways that have replaced the library as a source of information, searches and cataloguing. New tools, for example, myExperiment,[94] offer much more enhanced abilities to share and execute scientific workflows. Live and open debate played out via wikis and blogs have changed the dynamic of academic discussion – sometimes in extreme ways. In January 2009 Tim Gowers, an eminent mathematician and recipient of the Fields Medal, launched the Polymath Project, a blog serving as an open forum for contributors to work on a complex unsolved mathematical problem. He posed the question: "Is massively collaborative mathematics possible?" He then set out the problem, his ideas about it and an invitation for others to contribute to its solution. 27 people made more than 800 comments, rapidly developing or discarding emerging ideas. In just over a month, the problem was solved. Together they not only solved the core problem, but a harder generalisation of it. In describing this, Gowers said, "It felt like the difference between driving a car and pushing it".[95]

Since Gowers' successful demonstration, the concept of massive collaboration is spreading in mathematics. At the last count there were ten similar projects under way, some of which address problems even more ambitious than the original. Mathematicians have also adapted collaborative tools from the software community. MathOverflow follows the StackOverflow model, crediting contributors with bronze, silver and gold badges for responding to each other's queries. These platforms allow anyone to join in. But this access does not mean they are intelligible discussions outside a very specialised community – this openness, that is in practice very closed, has been described as "almost an anti-social network".[96]

---

88 Bush V (1945). Science: *The Endless Frontier*. United State Government Printing Office: Washington.
89 Ziman J (1983). *The Collectivization of Science*. Proceedings of the Royal Society B, 219, 1-19.
90 Genuth J, Chompalov I & Shrum W (2007). *Structures of Scientific Collaboration*. Available at: http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11233.
91 Wuchty S, Jones B, Uzzi B (2007). *The increasing dominance of teams in the production of knowledge*. Science, 316, 5827, 1036-1039..
92 Nielsen M (2012). *Reinventing discover: the new era of networked science*. Princeton University Press: Princeton.
93 Nowotny H, Scott P, Gibbons M (2001). Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty. Polity Press: London, UK.
94 University of Manchester and University of Southampton (2011). *Myexperiment*. Available at: http://www.myexperiment.org/
95 Gowers T, Nielsen M (2009). *Massively Collaborative Mathematics.* Nature, 461, 879-881.
96 Keller J (2010). *Beyond Facebook: How the World's Mathematicians Organize Online.* The Atlantic. September 28. Available at: http://www.theatlantic.com/technology/archive/2010/09/beyond-facebook-how-the-worlds-mathematicians-organize-online/63422/

## 2.2 Open science and citizens

### 2.2.1 Transparency, communication and trust

Public communication of scientific knowledge should not simply disclose conclusions but also communicate the reasoning and evidence that underlie them. Good communication is assessable communication, which allows those who follow it not only to understand what is claimed, but also to assess the reasoning and evidence behind the claim. If scientific communication is not assessable by its audiences they will be unable to judge when or why its claims are trustworthy. For example, participants in the public dialogue on open data led by Research Councils UK were worried that multiple interpretations of the same data could cause widespread and unnecessary confusion.

Many advocates of transparency simply assume that what is disseminated will be intelligible, relevant, accurate and honest. However, by emphasising dissemination rather than basic epistemic and ethical standards (such as those contained within the UK Government's Universal Ethical Code for Scientists[97]) they lose sight of the real demands of good communication.[98]

These issues expose a difficult dilemma for science. Whereas science eschews claims based on authority, as embodied in the Royal Society's motto, *nullius in verba* ("on the word of no one" or "take nobody's word for it"), understanding scientific analysis of issues of public importance or concern can require very high levels of expertise. If informed, democratic consent is to be gained for public policies that depend on difficult or uncertain science, it may be in the public interest to provide deep access to the data underlying that science, but this must be supported by relevant evidence to enable the intelligent placing – or refusal – of trust in scientific claims.

The potential loss of trust in the scientific enterprise through failure to recognise the legitimate public interest in scientific information was painfully exemplified in the furore surrounding the improper release of emails from the University of East Anglia.[99] These emails suggested systematic attempts to prevent access to data about one of the great global issues of the day - climate change. The researchers had failed to respond to repeated requests for sight of the data underpinning their publications, so that those seeking data had no recourse other than to use the Freedom of Information Act (FoIA) to request that the data be released.

The need to invoke FoIA reflects a failure to observe what this report regards as should be a crucial tenet for science, that of openness in providing data on which published claims are based. Greater accessibility of datasets would reduce the need for claims for access to data under FoIA and would also make data available in a more intelligible and reusable form. FoIA provides only one of a range of approaches to data sharing, and its approach is often unsatisfactory both to those who request data and to those who are required to release it. FoIA requests can be anonymous, and the procedure for responding to them leaves little room for direct communication between requester and data holders.

As it stands, the FoIA is not a satisfactory way of opening up access to research information, either for the applicant who requests the data, or for the scientists who are required to supply it.[100] What the applicant obtains through a FoIA request is usually decontextualised and not in a format that makes for easy analysis or reuse. However, the 2012 Protection of Freedoms Act requires that in response to FoIA requests there will be a duty to "provide the information, so far as reasonably practicable in an electronic form which is capable of reuse",[101] and a duty to provide a license for reuse. This demand may seriously underestimate what is "reasonably practicable" or affordable. Some scientific data can

97 Government Office of Science (2007). *Rigour, Respect, Responsibility: a universal ethical code for scientists*. Government Office of Science: London.

98 O'Neill O (2006). Transparency and the Ethics of Communication. In: *Transparency: The Key to Better Governance?* Heald D & Hood C (eds.). Proceedings of the British Academy 135. Oxford University Press: Oxford.

99 Reviews of these events include the "Muir-Russell" Independent Climate Change E-mails Review, which the chair of this report contributed to, as well as reviews by the House of Commons Science and Technology Committee and the Science Assessment Panel.

100 It is also important that all the costs borne by individuals or institutions in responding to FoI requests scientists are capped at a reasonable level to prevent such requests becoming a significant unfunded liability.

101 Parliament (2012). *Protection of Freedoms Act 2010-12*, Clause 102. Available at: http://services.parliament.uk/bills/2010-12/protectionoffreedoms.html

be readily rendered reusable by others who are unfamiliar with a research project, but it would be an unreasonably large task for many other data.

One way of dealing with this problem might be for scientists to prepare affordable sets of 'public interest data' for release within the boundaries described in chapter 3, together with the metadata required to make them intelligible and usable. However, if costs are to be contained it would have to be agreed that the availability of such public interest data would preempt further FoIA requests about the same datasets (by making this explicit in the exemption for "information available by other means" in the current FoIA). Part of the future work of international and national data centres, as well as the UK's Research Councils' Gateway to Research initiative, could be to identify these datasets and begin to build mechanisms for broader access.

Modern communication technologies can be used just as effectively to disseminate material that is neither intelligible nor reasonably accurate as it can to disseminate information that is intelligible to its audiences and reasonably accurate. The British Science Association, Nuffield Foundation and Wellcome Trust among others have taken on facilitating roles developing resources for well-informed public debate. Learned societies have a complementary role to play here in providing technically correct, scientifically honest and readable accounts of what this report terms public interest science. This could be the basis for ensuring that data-backed information, together with the necessary meta-data is available for public access, together with an intelligible summary of current hypotheses and uncertainties.

### 2.2.2 Citizens' involvement in science

Access to data that fulfils the requirements of accessibility, intelligibility, assessability and usability is important for citizens' involvement in science and their pursuit of scholarship through data which, after all, for publicly funded science they have paid for through their taxes. The level of metadata required will vary according to the extent to which the data can be readily translated into information and knowledge and to the level of knowledge of the user. It may also require a significant description of the scientific background if the data is to be a source of knowledge for the inquirer. Given the effort required from data originators or data scientists to make data available to users who may range from the highly expert in the field to the non-specialist, how should the necessary choices be made about those data that should be prioritised for wider public use? This follows the earlier comment that the concept of public interest science should be used in prioritising those datasets that should be made accessible, intelligible, assessable and usable for others, while recognising that assessability must be tailored to the differing needs and capacities of expert and lay audiences.

Openness to the public must be audience-sensitive. It must recognise a diversity of demands from citizens. Many wish to interrogate scientific understanding about a particular issue, often related to problems, such as illness, that affect them personally and where, for example, bodies such as patient groups or disease-specific charities can play an important role. At the other extreme, there is a small, but increasingly numerous body of engaged "citizen scientists" that wish to dig deeply into the scientific data relating to a particular issue. They are developing an increasingly powerful "digital voice," though many lack formal training in their area of interest. Some have been highly sceptical about research findings on issues such as GM crops, nanotechnology, HIV/AIDS, anthropogenic climate change, etc. Some ask

tough and illuminating questions, exposing important errors and elisions.[102] Others have effectively become members of particular scientific communities by dint of their rigorous and valuable observations and measurements, and become formally involved in scientific projects (see Box 2.4). The increased availability of major open databases is increasingly proving to be an asset to this citizen science community, and a way to increase their contribution to scientific progress. The growth of the citizen science movement could turn out to be a major shift in the social dynamics of science, in blurring the professional/amateur divide and changing the nature of the public engagement with science. Free or affordable access to scientific journals and data would provide important encouragement to the movement.

### Box 2.4 Examples of citizen science projects

#### Fold.it

The fold.it website offers participants a game in which players solve the intricate puzzles figuring out the ways in which amino acids fold to create different protein molecules. Knowing the structure of a protein is key to understanding how it can be targeted with drugs. Some human proteins are extremely complex with up to 1000 amino acids and there are myriad possibilities of how a protein can fold. Figuring out which of the many possible structures is the best one is regarded as one of the hardest problems in biology today and current methods take a lot of money and time, even for computers. Fold.it attempts to predict the structure of a protein by taking advantage of human puzzle-solving intuitions and having people play competitively to fold the best proteins. Some players, despite no previous experience in biology have become so good at folding proteins that scientists have invited them to study the intuitive principles they employ while solving puzzles during the fold.it game.

#### Galaxy Zoo

Galaxy Zoo enables users to participate in the analysis of the imagery of hundreds of thousands of galaxies drawn from NASA's Hubble Space Telescope archive and the Sloane Digital Sky Survey. It was started in 2007 by the Oxford doctoral student Kevin Schawinski, who decided to involve the community of amateur astronomers by using crowdsourcing. To understand how these galaxies formed, astronomers classify them according to their shapes. Humans are much better at classifying shapes than even the most advanced computer. More than 320,000 people have taken part in Galaxy Zoo. Over 120 million classifications have been completed and there are now more than 25 peer-reviewed publications based on data from Galaxy Zoo. Galaxy Zoo has led to four similar participatory projects: Planet Hunters, The Milky Way Project, Old Weather and Solar Stormwatch, which have already produced a further six papers.

#### BOINC

Numerous Citizen Science projects employ so-called volunteer computing, where individuals provide the resources of their home computers to contribute to big science research. Today there are over 50 active projects based on the BOINC platform developed at the University of California Berkeley. The most well known volunteer computing project is SETI@home, launched in 1997. Individuals can provide their computing resources to help in data analysis for the search for extra-terrestrial intelligence – so far no contact has been made. Many crowdsourcing projects have developed in the area of climate change.

102 McIntyre S (2012). *Climate Audit.* Available at: www.climateaudit.org/

## 2.3 System integrity: exposing bad practice and fraud

The rewards for attracting resources for research and making important scientific discoveries are considerable. These can create temptations for some scientists, whether in the public or private sector, to indulge in poor practices that range from blatant fraud, where data are invented, to selective reporting of findings in order to promote a particular hypothesis.

This is not a new phenomenon. The Piltdown hominid fraud of 1912 was exposed some 40 years after the original find as the contrived association of skull bones became increasingly incompatible with subsequent finds of fossil hominids from strata of the same age. Such frauds or bad practice tend to be exposed as the corpus of scientific evidence grows, although the short term effects can be costly in financial and personal terms. This is particularly damaging where there are immediate consequences for public policy or for areas of research that may have public safety implications such as those related to medical interventions.

Good science , simply put, "looks at all the evidence (rather than cherry picking only favourable evidence), uses controls for variables so we can identify what is actually working, uses blind observations so as to minimise the effects of bias, and uses internally consistent logic."[103] To ignore this kind of rigour is at the least poor practice.

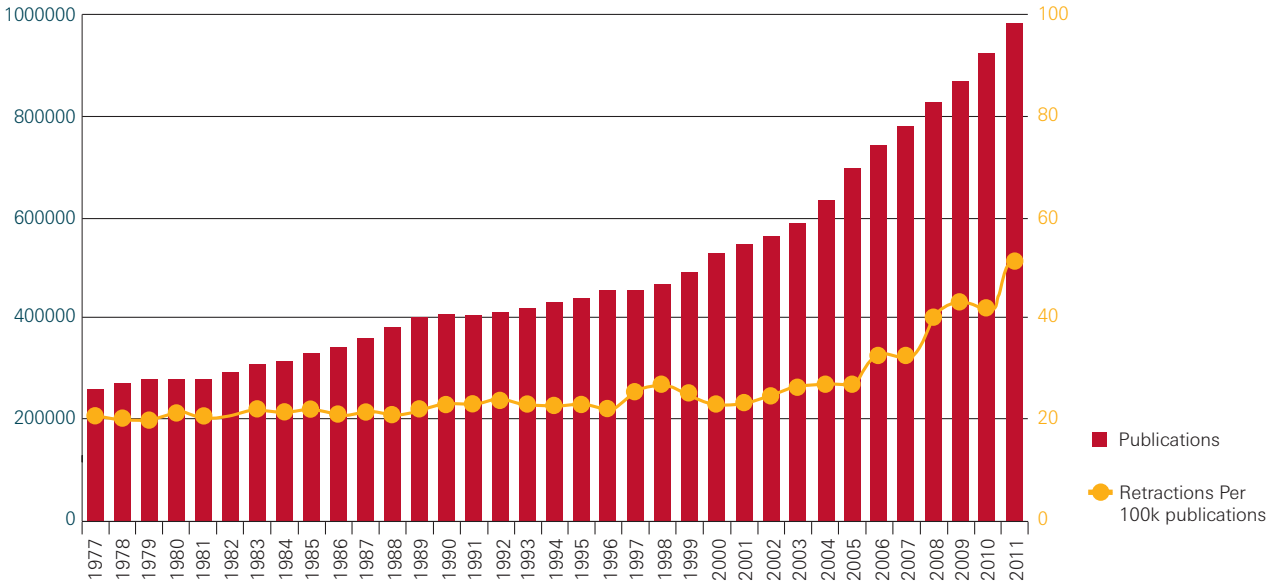
Retracting papers from scientific literature removes misleading information with the potential to distort scientific knowledge. It is critical that journals are prepared to publish retractions. 742 of the hundreds of thousands of English language papers submitted to PubMed between 2000 and 2010 were retracted.[104] Of those with a formal retraction notice, about one quarter were fraudulent.

Serious cases of scientific fraud have occurred in medicine[105] and physics[106] resulting in the retraction of a considerable number of papers. In 2002, Jan Hendrick Schön, a physicist at the Bell Laboratories in New Jersey, was found guilty of falsifying or fabricating results in at least 17 papers he had published within the previous two years. Schön's results, if proved to have been correct, would have revolutionised solid state physics. Woo Suk Hwang, a South Korean scientist; hit the headlines in 2006 when two of his papers[107,108] detailing the creation of the first cloned human embryo and derivation of 'patient-specific' stem cell lines were retracted from the journal *Science*[109]. The results had been fabricated. The affair was very damaging to the already controversial field of stem-cell research. It had raised the hopes of many patients suffering from incurable diseases that new, tailored stem-cell cures were on the horizon.

Between 2000 and 2010, an estimated 80,000 patients underwent clinical trials associated with research that was later retracted.[110] The numbers of retractions of journal articles relating to clinical trials are growing much more quickly than numbers of journal articles (Figure 2.5), although it is unclear whether the increase in retractions reflects improved detection or increasingly poor quality and problematic research. New initiatives like CrossMark[111] can run an article against live journal databases, telling a reader whether the version they have is up-to-date and whether the article has been redacted. These services make the vital step from an increasing willingness to redact to making sure amendments reach other researchers.

103 Novella S (2011). A Skeptic in Oz. Available at: http://www.sciencebasedmedicine.org/index.php/a-skeptic-in-oz/
104 Steen R G (2011). *Retractions in the scientific literature: is the incidence of research fraud increasing?* Journal of Medical Ethics, 37, 249-53.
105 Marcus A (2009). *Fraud case rocks anesthesiology community*. Anesthesiology News, 35, 3.
106 Service R F (2002). *Bell labs fires star physicist found guilty of forging data*. Science, 298, 30e1.
107 Hwang W S, *et al*. (2004). Evidence of a Pluripotent Human Embryonic Stem Cell Line Derived from a Cloned Blastocyst. Science 303, 1669–1674.
108 Hwang WS, *et al*. (2005). *Patient-Specific Embryonic Stem Cells Derived from Human SCNT Blastocysts*. Science, 308, 1777–1783.
109 Cyranoski D (2006). *Rise and fall.* Nature News. Available at: http://www.nature.com/news/2006/060111/full/news060109-8.html
110 Steen R G (2011). *Misinformation in the medical literature: what role do error and fraud play?* Journal of Medical Ethics, 37 (8), 498-503.
111 Crossmark (2012). *Helping Researchers Decide What Scholarly Content to Trust.* Available at: http://www.crossref.org/crossmark/

**Figure 2.5 Number of publications (columns) and number of retractions (line) relating to clinical trials: 1977 - 2011[112]**




Scientific papers that are merely wrong rather than fraudulent should not, however, be retracted. Understanding how errors have arisen is part of the process through which science self-corrects and through which discoveries are made. Nobel Laureate, Richard Feynman, put this very clearly: "if you've made up your mind to test a theory, or if you want to explain some idea, you should always publish it whichever way it comes out. If we only publish results of a certain kind, we can make the argument look good. We must publish *both* kinds of result."[113]

Bias in understanding can arise from poor experimental design, data collection, data analysis, or data presentation, and from earnest error or statistical naïvety. A particular form of bias that can seriously distort understanding has been highlighted in medical science through the failure to publish negative results - where no beneficial result is observed - from clinical trials.[114]

A recent series of articles published in the British Medical Journal[115] (see Box 2.5) examined the extent, causes and consequences of unpublished evidence. Unpublished Food and Drug Administration (FDA) trial outcome data in the US affects the reported outcome of drug efficacy.[116] For instance, 94% of antidepressant trials in the US were positive according to the published literature. In contrast, an FDA analysis of all trials showed that 51% were positive. The apparent increase in effect when only published results were taken into consideration ranged from 11-69% for individual drugs.[117] It has been argued that under-reporting of research results in this way is scientific malpractice.[118]

112 Reprinted with permission of Neil Saunders, the creator of PMRetract, a web application for monitoring and analysing retraction notices in PubMed (2012). *Retractions – by Year*. Available at: http://pmretract.heroku.com/byyear
113 Feynman R P (1974). Cargo cult science. Available at: http://calteches.library.caltech.edu/3043/1/CargoCult.pdf
114 Boulton, G., Rawlins, M., Vallance, P. and Walport, M. (2011). *Science as a public enterprise: the case for open data*. The Lancet, 377, May 14.
115 Lehman R and Loder E (2012). *Missing clinical trial data: A threat to the integrity of evidence based medicine.* British Medical Journal, 344, d8158. Available at: http://www.bmj.com/node/554663?tab=related
116 Hart B, Lundh A and Bero L (2012). *Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses*. British Medical Journal, 344, d7202.
117 Turner E H, Matthews AM, Linardatos E, Tell R A & Rosenthal R (2008). *Selective Publication of Antidepressant Trials and its influence on apparent efficacy*. The New England Journal of Medicine, 358, 252-260. Available at: http://www.nejm.org/doi/full/10.1056/NEJMsa065779
118 Chalmers I (1990). *Under-reporting research is scientific misconduct*. Journal of the American Medical Association, 263, 1405-1408.

### Box 2.5 Clinical Trial Registries

Failure to report the results of clinical trials can lead to publication bias (the selective reporting of positive results at the expense of inconclusive or negative results) and has been identified as a problematic issue for a number of years. Various policies have been implemented since 2004 to try and tackle these issues. Such policies include the registration of clinical trials in a public registry, including the US National Institute of Health (NIH) database ClinicalTrials.gov or the European Medicines Agency (EMA) EU Clinical Trials Register[119],[120], before the onset of patient enrolment as a condition of consideration for publication.[121] There should be compulsory registration of a minimal dataset consisting of 20 items on the registry at the onset of a study, including the trial's title, the condition studied in the trial, the trial design and the primary and secondary outcome[122]; and the posting of basic results, including primary and secondary efficacy endpoints, on ClinicalTrials.gov and other registers after completion of the trial.

It is hoped that these policies will ensure an open source of online information for all registered trials and promote transparency and accountability to help resolve the issue of publication failure or bias.

However, recent studies published in the British Medical Journal show that there needs to be a continued focus on compliance with these requirements and that publication failure and bias are still a concern today. Whilst there has been progress in the timely publication of clinical trials, fewer than half of the clinical trials funded by the NIH were published in a peer reviewed journal within 30 months of trial completion and a third remained unpublished after 51 months.[123] Additionally, only 22% of trials had released mandatory trial summary results on ClinicalTrials.gov within one year of completion of the trial. However, compared to other funders, industry-funded trials were increasingly likely to report their results (40%).[124] A further study found that in some cases it was necessary to combine results published in clinical trial registries, in peer reviewed journals and in internally produced study reports to obtain more comprehensive information about trial outcomes and even this did not ensure completeness (defined as providing adequate data for meta-analysis).[125] Policies mandating the registration of all clinical trials and summary results on public registries are a step in the right direction – compliance with and the enforcement of these policies is, however, crucial.

The editors of the British Medical Journal concluded that effective concealment of data is "a serious ethical breach" and that "clinical researchers who fail to disclose data should be subject to disciplinary action by professional organisations".[126] This report strongly supports the need to implement a mandatory system of open reporting.

Although every effort should be made to encourage high levels of personal and professional integrity as set out in the Universal Ethical Code[127], system integrity (in particular revealing faults in a published paper by securing open access to the underlying data) is arguably the most efficient way both to deter and to identify fraudulent or poor practice.

119 ClinicalTrials.gov (2012). Available at: http://clinicaltrials.gov/
120 EU Clinical Trials Register (2012). Available at: https://www.clinicaltrialsregister.eu/
121 De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC and Van Der Weyden MB (2004). *Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors*. The New England Journal of Medicine, 351, 1250-1251.
122 World Health Organisation (2005). *WHO Technical Consultation on Clinical Trial Registration Standards.* Available at: http://www.who.int/ictrp/news/ictrp_meeting_april2005_conclusions.pdf
123 Ross J S, Tse T, Zarin D A, Xu H, Zhou H, Krumholz H M (2012) *Publication of NIH-funded trials registered in ClinicalTrials.gov: cross sectional analysis*. British Medical Journal, 344, d7292.
124 Prayle A P, Hurley M N and Smyth A R (2012) *Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study*. British Medical Journal, 344, d7373.
125 Wieseler B, Kerekes MF, Vervoelgyi V, McGauran N and Kaiser T (2012) *Impact of document type on reporting quality of clinical drug trials: a comparison of registry reports, clinical study reports, and journal publications*. British Medical Journal, 344, d8141.
126 Lehman R and Loder E (2012). *Missing clinical trial data: A threat to the integrity of evidence based medicine*. British Medical Journal, 344, d8158. Available at: http://www.bmj.com/node/554663?tab=related
127 Government Office of Science (2007). *Rigour, Respect, Responsibility: a universal ethical code for scientists*. Government Office of Science: London.

# The boundaries of openness

This report advocates a default position in favour of open data, but recognises that openness is not an unqualified good. There are four areas where there are warranted restrictions on openness, which relate to: commercial interests, personal information, safety and national security. The following outlines ways that these areas can be managed to optimise the benefits of openness without transgressing warranted limitations.

## 3.1 Commercial interests and economic benefits

Creating marketable products from a scientific idea is normally costly. People are prepared to bear this cost if they can protect innovations from immediate mimicry and unfair competition. Patents play a key role by giving inventors a privileged position so that they can exclusively exploit their work whilst making the underlying knowledge available.

There is a balance to be struck between creating incentives for individuals or groups to exploit new scientific knowledge for financial gain and societal benefits through the products and services that are developed and the macroeconomic benefits that accrue when knowledge is broadly available and can be exploited creatively in a wide variety of ways. In recent decades such tensions have been exemplified in the explosive growth of knowledge about the human genome, which promises major biomedical therapeutic opportunities. An international publicly funded consortium, which ultimately delivered the full sequence of the human genome, was challenged by a parallel commercial effort.[128] The public-consortium made all draft sequences of genes openly available. If the commercial effort had dominated, data could have been shared only after the stakes were claimed, and could theoretically have provoked an international genome gold rush. These tensions contributed to the adoption of the UN Declaration on the Human Genome and Human Rights in 1997. However, it is far from clear what effect, if any, this had on practice, as national patent offices have granted thousands of patents which include human DNA sequences. The search for balance continues in ongoing debates about whether genetic discoveries meet the legal requirements for patentability and whether it is ethical to patent what many see as the common heritage of humanity.[129]

As data are increasingly seen as commercial assets in themselves, the pressure to hoard rather than share could rise. There is a growing consensus that "data is the greatest raw material of business, on a par with capital and labour".[130] Google has more data than European Bioinformatics Instituteand the Large Hadron Collider put together, and has spawned a data analysis industry. Firms specialising in data management and analytics are estimated to be worth more than $100 billion and growing at almost 10% a year, roughly twice as fast as the software business as a whole. These services allow companies to understand the habits and priorities of potential customers and identify the most effective ways of selling services.[131] But the value a company finds in trawling customer data is not the same as the value in most scientific datasets. Of the data-related potential £216 billion gain to the UK economy between 2012 and 2017 (section 1.5.2), six times more comes from customer intelligence, supply chain management and other business efficiency gains than data-driven R&D.[132]

It might seem that the natural boundary of openness should coincide with the boundary between publicly funded and privately funded research: with private business maintaining confidentiality of their data and publicly funded researchers opening their data. However, effective commercial exploitation of some publicly funded research is in the public interest and may require limitations on openness whilst some commercial business models thrive on openness. The following characterises the current boundary of openness as it is shaped by commercial interests, suggesting how policy might adapt to encourage beneficial reuse of open research data.

---

128 Shreeve J (2005). *The Genome War: How Craig Venter Tried to Capture the Code of Life and Save the World*. Ballantine Books: New York.
129 Human Genetics Commission (2010). *Intellectual Property and DNA Diagnostics.* Available at: http://www.hgc.gov.uk/UploadDocs/DocPub/Document/IP%20and%20DNA%20Diagnostics%202010%20final.pdf.
130 The Economist (February 2010). *Data,data everywhere.* Special Edition on Managing Information. Available at: http://www.economist.com/node/15557443?story_id=15557443
131 The Economist (February 2010). *Data,data everywhere.* Special Edition on Managing Information. Available at: http://www.economist.com/node/15557443?story_id=15557443
132 CEBR (2012). *Data equity: unlocking the value of big data.* Available at: http://www.cebr.com/wp-content/uploads/1733_Cebr_Value-of-Data-Equity_report.pdf

### 3.1.1 Data ownership and the exercise of Intellectual Property rights

Governments worldwide recognise the potential commercial value of reusing publicly funded research data. Government policies are not inimical to data reuse by others and does not necessarily require the funders of research to assert ownership of resultant data, indeed most encourage commercial exploitation of research results by those they have funded. At the same time it is particularly important that newly acquired datasets that add significantly to an existing body of data, for example, time series data on environmental change or social behaviour, are made available for reuse. The UK Economic and Social Research Council has a clearly articulated policy to maximise reuse of data whose collection it has funded, whilst retaining the right to assert ownership of the intellectual property so that it can be exploited for national benefit.[133] Ownership is asserted through Intellectual Property (IP) rights. They are a means of control but need not be used to restrict access. Developments since the Royal Society's 2003 study into IP in academic research show how rapidly changes in technology and legislation are occurring.[134] Copying of digital material has become so easy that it has undermined the effectiveness of traditional copyright IP, and has created new categories of ownership. The GNU General Public Licence and Creative Commons ShareALike licences have the explicit aim of maintaining the free flow of information.[135]

**Box 3.1: Intellectual Property rights**

Intellectual Property (IP) rights refer to a variety of different ways in which the use of ideas can be restricted under the law, including copyright, patents and database rights.

*Copyright* confers a right on the creators of original works to prevent others from copying the expression of ideas in a work. Copyright is vital for protecting the integrity of works. At present most journals require assignment of copyright. The Copyright term is 70 years after the author's death in the EU, and must be at least 50 years after the author's death for all members of the World Trade Organisation.[136] But control over copying in a digital environment is almost meaningless. It can limit the reuse of the work – because it prevents further analysis of the data by, for example, text mining. The UK Hargreaves Review recommended a series of changes to copyright law to allow the text mining of articles for non-commercial research purposes.[137] The Wellcome Trust and the Research Councils are moving towards a grant condition that stipulates that, when an open access fee is paid, the article must be licensed under a Creative Commons licence (CC-BY) that allows full reuse (including commercial). Open access articles in Royal Society journals are published as CC-BY, as is this report.

*Patents* protect inventions such as products, processes or apparatus. To be patentable, an invention must be new, industrially applicable and involve an inventive step. Patents last for 20 years, but can be longer in some circumstances. Patents offer a vital incentive for innovation in many spheres, but can contribute significantly to the costs of innovation in fields such as software, where the presence of multiple interlocking patents can create significant uncertainty and expense.

A *database right* exists in Europe that offers protection to the maker of a database who has put substantial investment into obtaining, verifying or preserving the contents of the database. This right exists over and above any copyright in the material content of the database. It affords protection against unauthorised extraction from or reuse of all or a substantial part of the database.

133 ESRC (2010). *Research Data Policy*. Available at: http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

134 Royal Society (2003) *Keeping science open: the effects of intellectual property policy on the conduct of science.* Royal Society: London. Available at: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2003/9845.pdf

135 The GNU General Public Licence is explicit on this point: "The licences for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public Licence is intended to guarantee your freedom to share and change all versions of a program - to make sure it remains free software for all its users." Such licences are sometimes known as "copyleft", because they use copyright law to ensure that information remains freely shareable.

136 World Trade Organisation (1994). *Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS)*. Available at: http://www.wto.org/english/tratop_e/trips_e/t_agm0_e.htm

137 Hargreaves I (2011). *Digital Opportunity: A Review of Intellectual Property and Growth*. Available at: http://www.ipo.gov.uk/ipreview-finalreport.pdf

Patents are often cited as a barrier to openness, albeit that a core purpose of the patent system is to enable information to be shared that would otherwise be treated as a trade secret. Disclosure of an invention and the means to work it – ie to reproduce it - is a condition of proceeding with a patent application. Patent databases have thus become a vast reserve of information about current technology and the commercial potential of scientific disciplines. They contain pending patents, patents in force and lapsed patents. For patents in force, the right to exploit the invention described in the patent is controlled by the patent holder - which is typically done through licensing - but the information *per se* is freely available and can form the basis for further innovation or research. A patent can last up to twenty years (and sometimes longer) and decisions to grant or not to grant licences can have a major impact on access to scientific information and the possible trajectories of scientific inquiry.

Most countries provide research exemptions in their IP laws, although there is no harmony of approach in Europe, and the US courts have so narrowly defined the patent exemption as to render it impotent. The recent UK Hargreaves Review of Intellectual Property's recommendation to exempt non-commercial text or data mining will help UK researchers make the most of these new tools: whether searching journals for papers on particular compounds or for secondary analysis of census data. The UK Government has accepted this recommendation in principle and this report looks forward to its realisation as a change to the fair dealing exemptions in UK Law. There has never been an assessment of the effects of the EU Database Directive (96/9/EC) on research. The impact of the database right on the scientific community should be an explicit objective of the next review of the Database Directive.[138] If necessary, consideration should be given to the introduction of a compulsory licensing scheme.

It is important to recognise that the denial or removal of intellectual property protection is unlikely to make more scientific data available. A more likely consequence is that the work would not be carried out because of the uncertainty of a financial return on investment, or investors and researchers would fall back on trade secrets which are designed to keep information out of the public domain. Instead of concentrating on the *existence* of intellectual property, there needs to be fair and equitable approaches to the *exercise* of intellectual property rights. The UK and 21 other countries have signed the Anti-Counterfeiting Trade Agreement (ACTA), debated by the European Union in June 2012.[139] There are widespread fears and protests against this act; there are legitimate concerns about the potential for adverse restriction of online content. In a similar vein, the Stop Online Piracy Act (SOPA) and the Protect Intellectual Property Act (PIPA) are being considered by the Senate and Congress in the US[140]; both extend the powers of prosecution of online entities in order to protect intellectual property rights. There are many international intellectual property laws to which the UK and other countries are subject, and the new legislation could extend and greatly affect the online community.

There is still much that the scientific community can do through collective action to promote co-existence of intellectual property and openness. Database rights holders can publish their willingness to grant non-exclusive licences and terms of use. Patent pools can be set up to allow patent owners to agree coordinated licensing action and can help to avoid the problem of patent thickets - when a scientific domain is so clogged with IP that it is impossible to navigate. Patent clearing-houses could operate to administer patents in a particular field and levy returns for IP owners while facilitating access by others. The Hargreaves Review did not make specific recommendations for changes to Intellectual Property (other than for text mining) in relation to the research community. The Review failed to unearth evidence that harm is being done that cannot be reversed through better local practices.[141] The problems associated with intellectual property rights are not primarily due to its format, nor to ideas about how best to deploy it. The problems lie with those who use it.

138 Periodic review is a legal requirement under Article 16(3) of the Directive.
139 United States Government (2012). ACTA. Available at: http://www.ustr.gov/acta
140 United States House of Representative (2012). Committee on the Judiciary. Available at: http://judiciary.house.gov/issues/issues_RogueWebsites.html
141 Intellectual Property Office (2011). Supporting Document U: Universities, Research and Access to IP. Available at: http://www.ipo.gov.uk/ipreview-documents.htm

A recent report on intellectual property and DNA diagnostics by the Human Genetics Commission (HGC) reflects typical tensions.[142] There is evidence that some clinical laboratories have stopped using tests as a result of aggressive attempts to enforce patents, although this is currently a North American rather than a European problem. The HGC recommended that funders of biomedical research should review their guidelines on licensing. Similar considerations by research institutions and other funders would be beneficial. There is, however, a broader pattern of the tightening of control over IPs by universities that is more worrying.

### 3.1.2 The exercise of Intellectual Property rights in university research

In the UK, universities are the principal recipients of government funding for research. Following the 2003 Lambert review, there have been consistent efforts by the UK government to increase the economic value of this research to the wider society.[143] UK universities are now more aware of business needs than at perhaps any other time in their history, and some have created technology transfer offices which support commercialisation of research through mechanisms such as the protection and licensing of IP and the creation of spin-out companies.

Evidence is emerging that this change in attitudes may, in some cases, have gone too far, straining delicate relationships. A survey of Engineering and Physical Sciences Research Council (EPSRC) collaborations found a growing proportion of firms reporting barriers to business-university collaboration between 2004 and 2008. Potential conflicts over IP were cited as a barrier by 32.4% in 2004, increasing to 55.6% in 2008.[144]

The economic rationale for tighter control of intellectual property by universities is dubious. In the seven years from 2003/04 to 2009/10[145], UK universities income rose by 35% (from around £2,200 million to £3,086 million). An average of 2.6% of this was derived from IP, including the sales of shares,[146] and showed no significant increase over the whole period. The low return on formal technology transfer activities is also apparent in the USA[147], where the average technology transfer income is about 2.2% of research income (and where Harvard and MIT make less than the average). This suggests that the value of research is not primarily in its ownership, and does not warrant the strict control over IP in some technology transfer offices. The commercial value of some intellectual property may be overestimated and rights exercised too early in the process of knowledge generation.[148] It is important that the search for short term benefit to the finances of universities does not work against the longer term benefit to the national economy.

A more discriminating approach may be needed in identifying and supporting technologies that have the potential to deliver long term economic value (Box 3.2), as well as strengthening the collaborative and contract research that make up the majority of universities' income (see section 3.1.3). The Intellectual Property Office's May 2011 updated guide to IP strategy for universities is a welcome addition, recommending that universities adopt a more flexible, bespoke approach to IP management.[149] A promising development is the Easy Access Innovation Partnership, in which the University of Glasgow, King's College London and the University of Bristol have agreed not to enforce some patents, allowing businesses to use them for commercial purposes.[150]

142 Human Genetics Commission (2010). *Intellectual Property and DNA Diagnostics.* Available at: http://www.hgc.gov.uk/UploadDocs/DocPub/Document/IP%20and%20DNA%20Diagnostics%202010%20final.pdf
143 HM Treasury (2003). *Lambert Review of Business—University Collaboration.* Final Report. Stationery Office: London, UK.
144 Bruneel J, d'Este P, Neely A, Salter A (2009). *The Search for Talent and Technology: Examining the Attitudes of EPSRC Industrial Collaborators Towards Universities.* Advanced Institute of Management Research: London, UK
145 Higher Education Funding Council for England (2010). Higher Education - Business and Community Interaction Survey 2009-10. Available at: http://www.hefce.ac.uk/pubs/hefce/2011/11_25/
146 In addition, the costs associated with defending IP are high. Higher Education Institutes spent £29 million on defending IP in 2009-10, out of an income of £84 million.
147 British Consulate-General of San Francisco (2003), *Key lessons for technology transfer offices: Viewpoints from Silicon Valley*, Note produced by the Science and Technology Section.
148 Evidence from economic roundtable (see appendix 4).
149 Intellectual Property Office (2011). *Intellectual asset management for universities.* Available at: http://www.ipo.gov.uk/ipasset-management.pdf
150 University of Glasgow (2012). *Easy Access IP deals.* Available at: http://www.gla.ac.uk/businessandindustry/technology/easyaccessipdeals/

### Box 3.2 Balancing openness and commercial incentives – the Medical Research Council's handling of Monoclonal Antibodies

The development and commercialisation of monoclonal antibodies exemplifies the value of an approach that combines data sharing with steps to retain appropriate commercial protection. Publically funded research by the Medical Research Council (MRC) led to the development of monoclonal antibodies in 1975.[151] Today, these make up a third of new drug treatments for a variety of major diseases. At key points in this process, scientific findings were shared which generated interest and fuelled further research and development. Openly sharing the early technology in this area was important in stimulating the field and driving the science to the point at which commercialisation was possible.[152] Protecting certain parts of the science with patents and exclusivity led to a start-up company, Cambridge Antibody Technology. In 2005, the US pharmaceutical company Abbott paid the MRC over £100 million in lieu of future licensing royalties. The MRC have also benefitted from the sale of Cambridge Antibody Technology and the antibody company Domantis Ltd to the pharmaceutical firms Astra Zeneca and GSK respectively. All the income the MRC receives from these commercial activities is ploughed back into further research to improve human health. An optimal combination of data-sharing and securing intellectual property ensured the health and wealth benefits in this area were delivered to the UK.

151 Kohler & Milstein (1975). *Continuous cultures of fused cells secreting antibody of predefined specificity*. Nature, 256, 495
152 MRC (2012). *Therapeutic Antibodies and the LMB*. Available at: http://www2.mrc-lmb.cam.ac.uk/antibody/

### 3.1.3 Public-private partnerships

Many companies increasingly adopt a process of open innovation, bringing external ideas to bear on product and service development.[153] Academic research is a common source for these external ideas, and new kinds of relationship are being built between firms and researchers. Some of these are short term and specific to well defined problems, as in Box 3.3a. Others are more formal, longer term partnerships. They include company funded university centres in technology (Box 3.3c), more complex, multi-partner relationships (Box 3.3b) and innovative partnerships in new areas of science and technology (Box 3.3d). In all cases, there must be clear definition of the boundary between open and restricted material, including data, as well as arrangements for allocating IP. Often a shared space, such as in Box 3.3c or in the UK Government's new Catapult Centres,[154] is needed to build informal connections before businesses and researchers can enter into more formal arrangements. Rapid resort to formal arrangements can have a chilling effect on the relationship.

A recent twist to this fragile relationship is the threat of Freedom of Information (FoI) requests for data from public-private partnerships. Under the 2012 Protection of Freedoms Act, universities will have to share information assets with any requester in a reusable format and must allow the requester to republish that information. The liability to be subject to these requests could undermine the confidence of businesses in partnering with UK universities. Arrangements will need to specify who owns the data, and perhaps who owns the locations in which they are stored.

Data are not usually the major currency in these partnerships. For example, Syngenta are contributing to the development of the open source Ondex data visualisation software (Box 3.2e), but whilst the software is open, much of the data it processes will not be. Policies that encourage collaboration based on data produced by universities and business are in their infancy in the UK, and they may remain so if current legislation leaves a legacy of uncertainty. Sir Tim Wilson's recent review of Business-University collaboration[155] concentrated on the human networks needed for collaborative working without looking closely at what these networks share or at the way in which data is controlled.

As research becomes increasingly data-rich, there is a potential niche for entrepreneurial businesses to set up as knowledge brokers and to point others to research data that is publicly available and to repackage such data in ways that are usable by others. Projects such as Imanova (Box 3.2d), show how a collaboration can leverage existing data. Given the maturity of data analytics in the private sector, there is a clear opportunity for partnerships to leverage the skills of the new cohort of data scientists (see section 4.1.4) that is needed both inside and outside the research community.

This report recommends that funding be provided by the Department for Business, Innovation and Skills and the Technology Strategy Board to enhance business take-up of openly accessible scientific outputs. In the same way that Catapult Centres provide physical infrastructure for university-business collaboration, there needs to be enhanced (but possibly dispersed) digital infrastructure to enable data-based knowledge brokers. But such developments will be inhibited unless there is clarity about the status of data produced through public-private partnerships under the UK FoI Act. Legislative ambiguity could undermine collaborations of the kind that data-intensive research could facilitate.

153 Chesborough H (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology.* Harvard Business Press: Harvard.
154 Technology Strategy Board (2012). *Catapult Centres*. Available at: http://www.innovateuk.org/deliveringinnovation/catapults.ashx
155 Wilson T (2012). *A Review of Business-University Collaboration*. Department of Business, Innovation and Skills: London.

### Box: 3.3: Public-Private Partnerships

#### a) InnoCentive

InnoCentive is a service for problem solving through crowdsourcing. Companies post challenges or scientific research problems on InnoCentive's website, along with a prize for their solution. More than 140,000 people from 175 countries have registered to take part in the challenges, and prizes for more than 100 Challenges have been awarded. Institutions that have posed challenges include Eli Lilly, NASA, nature.com, Procter & Gamble, Roche and the Rockefeller Foundation.

#### b) The Structural Genomics Consortium

The Structural Genomics Consortium (SGC) is a not-for-profit public-private partnership to conduct basic science.[156] Its main goal is to determine 3D protein structures which may be targets for drug discovery. Once such targets are discovered, they are placed in the public domain. By collaborating with the SGC, pharmaceutical companies save money by designing medicines that they know will 'fit' the target. The SGC was initiated through funding from the Wellcome Trust, the Canadian Institute of Health Research, Ontario Ministry of Research and Innovation and GlaxoSmithKline. More recently other companies (Novartis, Pfizer and Eli Lilly) have joined this public-private partnership. The group of funders recently committed over US$50 million to fund the SGC for another four years.

#### c) Rolls-Royce University Technology Centres

In the late 1980s Rolls-Royce created University Technology Centres (UTCs), of which there are now 19 in 14 UK universities. The network was later extended to the USA, Norway, Sweden, Italy, Germany and South Korea. Each UTC addresses a key technology collectively tackled through a range of engineering disciplines. Research projects are supported by company sponsorship, by Research Councils and international government agencies. Ownership of IP for emerging technologies depends on the company's arrangements with the universities. Alternatively they may be governed by national regulation, but Rolls-Royce retains access to the IP while simultaneously providing for its use by the academic bodies, including for research and teaching purposes. Where IP remains with the universities, it is licensed back for use by its sponsor.

#### d) Imanova

Imanova is an innovative alliance between the UK's Medical Research Council and three London Universities: Imperial College, King's College and University College. It was established in April 2011 as a jointly-owned company, aiming to enhance rapid development of novel methodologies for biomedical research using imaging data. Incorporating expert staff and facilities from GSK's Clinical Imaging Centre at Hammersmith Hospital, Imanova is a state-of-the-art research centre for imaging methodologies. It trains scientists and physicians, and hopes to become an international partner for pharmaceutical and biotechnology companies.

#### e) Syngenta

The Technology Strategy Board and Syngenta are building a system that helps scientists visualise the similarities between the molecules in ChEMBL, an openly accessible drug discovery database of over one million drug-like small compounds, and those in their own research.[157] They are developing the open source Ondex[158] software, allowing users to visualise and analyse company data when it is integrated with data from ChEMBL. This can lead, for instance, to the discovery of a new protein target for a Syngenta molecule because it is similar to a ChEMBL molecule which affects that protein. This could tell the company something new about the kinds of pesticides that work on a particular weed that contains that protein.

156 Structural Genomics Consortium (2012). Available at: http://www.thesgc.org/

157 The database also includes details of over 8,000 biological targets, mainly proteins, which are activated or inhibited by particular molecules. The data is manually extracted from the scientific literature, and curated to enhance usability. Extraction needs to be done manually because compound structures are often presented only as images. Machine extractable structures, and a legal regime that made text mining easier, as recommended by the Hargreaves Review, would make ChEMBL's work much easier. See Gaulton A *et al* (2012). *ChEMBL: a large-scale bioactivity database for drug discovery*. Nucleic Acids Research, 40. Available at: https://www.ebi.ac.uk/chembldb/.

158 BBSRC (2012). *Ondex: Digital Integration and Visualisation*. Available at: http://www.ondex.org/

### 3.1.4 Opening up commercial information in the public interest

There is a strong case for greater openness of data and information from privately funded research that has the potential to impact the public, while respecting the boundaries described in this chapter. In many of these areas regulatory bodies such as the Medicines and Healthcare Product Regulatory Agency make decisions on behalf of the public on the availability (and continued availability) of products. These regulators play a key role in making information available to the public while protecting legitimate commercial interests, personal privacy, safety and national security.

Managing legitimate commercial interests in the public disclosure of data and information from privately funded research that is of public interest warrants careful consideration. For example, information could be made public and data made available after Intellectual Property has been secured or after a particular product or service is made available to the public (Box 3.1). It should be recognised, however, that trade secrets are an important component of intellectual property and that certain types of research (eg research relating to a manufacturing process) are of limited public interest. Where the research relates to a particular and immediate safety issue, the need to make information and data available in an expeditious manner should take priority over immediate commercial considerations.

Clinical Trials Registries have the potential to balance the public interest that accrue from commercial endeavour against those interests served by access to research data for purposes of safety assessment or for the scrutiny of public decisions. (Box 2.5) Clinical trial registries, such as ClinicalTrials.gov, require that summary results (but not the raw data) from industry-sponsored clinical trials are publicly disclosed on their database at a time when this disclosure does not undermine the ability of the sponsor to obtain a patent, thus allowing information of public interest to be disclosed without unduly encumbering the ability to draw commercial advantage from the research data.

## 3.2 Privacy

The use of datasets containing personal information is vital for a lot of research in the medical and social sciences, but poses considerable challenges for information governance because of the potential to compromise individual privacy. Citizens have a legitimate interest in safeguarding their privacy by avoiding personal data being used to exploit, stigmatise or discriminate against them or to infringe on their personal autonomy (see box 3.4).[159] The legal framework for the "right to respect for private and family life" is based on article 8 of the European Convention on Human Rights (ECHR)[160] for member states of the Council of Europe. Some aspects of privacy rights are codified by the EU Data Protection Directive (95/46/EC) and implemented in the UK by the Data Protection Act 1998 (DPA).

159 Laurie G (2002). *Genetic Privacy: A Challenge to Medico-legal Norms.* Cambridge University Press: Cambridge.

160 Korff D (2004). The Legal Framework, In: *Privacy & Law Enforcement, study for the UK information Commissioner.* Douwe K & Brown I (eds.). Available at: http://www.ico.gov.uk/upload/documents/library/corporate/research_and_reports/legal_framework.pdf

### Box 3.4 Attitudes to use of personal data in health research

#### a. Huntington's disease

Huntington's disease (HD) is a progressive neurodegenerative condition that usually presents in mid-life. It is inherited as an autosomal dominant trait and each child of a parent with HD therefore has a 50% chance of developing the condition. It is possible for such "at risk" people to undergo a genetic test to determine whether they will, or will not, develop HD. For a variety of reasons, however, less than 15% of those "at risk" accept the offer of genetic testing. Data from HD families that include information about whether or not members have the HD mutation are also likely to include details such as the age and gender of the family members. As HD is a relatively rare condition, with sufferers and their families often forming close-knit communities, it would be easy for the identities of each member – and hence their HD risk status – to be discerned. The consequences for both individuals and their families could be devastating.[161]

#### b. UK National Cancer Registration

Data is collected on UK cancer patients via hospitals, cancer centres, hospices, screening programmes and GPs. It is then curated through 11 national registries. This data is identifiable for several reasons, including for research purposes. Names of patients are disclosed to researchers forprojects investigating the causes of, or outcomes from, specific cancers. To release the information, cancer registries require that such studies are approved by the appropriate research ethics committees and their Ethics and Confidentiality Committee. Data for geographical studies (eg studies of cancer risk in people living near landfill sites) can only be undertaken if the full postcode is available. Support for registration and the use of identifiable data from cancer sufferers and their families is as high as among other groups. 85% of cancer sufferers support cancer registration, compared to 81% of non-sufferers. 72% in each group are not concerned about breaches of privacy in this system.[162]

#### c. UK surveys on access to patient records

A 2006 Ipsos MORI survey for the Medical Research Council[163] reported that approximately 69% of the UK public are favourable - including 14% who say they are certain - to allowing their personal health information to be used for medical research purposes. This is compared with around a quarter who feel they would not be likely to do so, including 7% would certainly not do so. There was less enthusiasm among 16-24 yr olds where only 27% were in favour of sharing their records for research purposes. When asked if they would want to be asked about sharing personal health information for medical research purposes after diagnosis of a serious disease, only 17% refused.[164]

Amongst those unhappy to share their information, concerns about privacy was the most common factor (cited by 28%). 4% cited worries about whether the information was anonymised. 62% would be more likely to consent to their health information being used if they knew how the confidentiality of their records would be maintained. But 89% did not trust public sector researchers handling their medical research records, and 96% did not trust researchers in the private sector.

In comparison, GPs were trusted to access records by 87% of respondents, and 59% trusted other health professionals – such as consultants or hospital doctors. The Scottish emergency care summary extracts data from GP records and hospital notes, contains records for nearly five million patients, of which less than 500 opted out when it was created in 2007. Similarly shared

161 Almqvist *et al* (1999). *A Worldwide Assessment of the Frequency of Suicide, Suicide Attempts or Psychiatric Hospitalization after Predictive Testing for Huntington Disease.* American Journal of Human Genetics, 64, 1293. This study suggests that suicide rates among people given a positive test result for Huntington's Disease was ten times the national average.

162 National Cancer Registry (2006). *National survey of British public's views on use of identifiable medical data.* British Medical Journal, 332.

163 Ipsos MORI (2006). *The Use of Personal Health Information in Medical Research: general public consultation*. Available at: http://www.ipsos-mori.com/Assets/Docs/Archive/Polls/mrc.pdf.

164 Ipsos MORI and Association of Medical Research Charities (AMRC) (2012). *Public support for research in the NHS*. Available at: http://www.ipsos-mori.com/researchpublications/researcharchive/2811/Public-support-for-research-in-the-NHS.aspx

summary care records are rolling out across England. 37 million people have been contacted and 1.27% opted out. Almost 13m records have now been created.

This disparity between trust in GPs and researchers[165] may prove crucial as the UK Government moves forward with its Life Science Strategy. The change in the NHS Constitution that opens up medical records to researchers will need to be accompanied by more public information on how confidentiality will be maintained.

The DPA regulates the collection, storage and processing of personal data, defined as data from which an individual is identifiable or potentially identifiable. Data protection requires that the processing of personal data is fair and lawful, which means that it must be done on one or more legitimate grounds contained in the DPA and in conformity with other legal requirements such as the common law duty of confidentiality and ECHR article 8. Personal data may be processed where informed consent has been given and where the data subject has a clear understanding of the facts, implications and potential consequences of processing. This does not on the one hand absolve the data controller from the need to comply with standards of notification and security. On the other hand data can be processed without explicit consent on grounds such as a significant public interest.

It had been assumed in the past that the privacy of data subjects could be protected by processes of anonymisation, such as the removal of names and precise addresses of data subjects. However, a substantial body of work in computer science has now demonstrated that the security of personal records in databases cannot be guaranteed through anonymisation procedures where identities are actively sought.[166] All datasets that contain information about individuals, even if they are anonymised, will provide support for inferences about the probability of other information about subjects.[167] The most that can be achieved by so-called anonymisation procedures is to ensure that the risk of revealing private information about individuals is not substantially increased by the compilation of the database and by specific requests to it.

*Ad hoc* approaches to anonymisation of public records for example have failed when researchers have managed to identify personal information by linking two or more separate and apparently innocuous databases. This was powerfully demonstrated by Latanya Sweeney[168] in the case of the Group Insurance Commission (GIC) in Massachusetts. GIC is responsible for purchasing health insurance for approximately 135,000 state employees and their families. In the mid-1990s, detailed patient data from GIC, ostensibly anonymised by the removal of explicit identifiers such as name, address, and Social Security number, were made available to researchers and industry. The released data contained the ZIP code, birth date and gender of each person, as well as details of diagnoses and prescriptions. Sweeney was able to

165 UK Life Sciences Strategy, Ibid.

166 See for example, Denning D (1980). *A fast procedure for finding a tracker in a statistical database*. ACM Transactions on Database Systems (TODS), 5, 1; Sweeney L (2002). *k-anonymity: a Model for Protecting Privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10, (5), 557-70; Dwork C (2006). *Differential Privacy*. International Colloquium on Automata, Languages and Programming (ICALP), 1–12; Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007). *L-diversity: Privacy beyond k-anonymity*. ACM Transactions on Knowledge Discovery from Data (TKDD), 1, 1.

167 Dwork C (2006). *Differential Privacy*. International Colloquium on Automata, Languages and Programming (ICALP), 1–12.

168 Data Privacy Lab (2005). *Recommendations to Identify and Combat Privacy Problems in the Commonwealth, Sweeney's Testimony before the Pennsylvania House Select Committee on Information Security* (House Resolution 351), Pittsburgh, PA, October 5. Available at: http://dataprivacylab.org/dataprivacy/talks/Flick-05-10.html#testimony

purchase the voter registration list for Cambridge Massachusetts for $20, which contained the name, address, ZIP code, birth date and gender of each voter. The two datasets were linked by their common fields (ZIP code, birth date and gender), allowing the diagnoses, procedures, and medications in the published GIC data to be matched to particular individuals. From this, for example, it was possible to identify the medical records of the state governor. Six people had the governor's birth date. Only three of these were men, and he was the only one with the matching 5-digit ZIP code. While this example may be extreme - releasing Zip codes has an obvious risk of reidentification - there are other less inevitable routes to stripping off anonymity.

Faced with the limitations of anonymisation, it is difficult to make judgements about the balance between personal privacy and the potential benefit to the broader public of collating information that might, for example, confer public health benefits.[169] Two extreme positions were exemplified in the evidence submitted to this enquiry and expressed in almost identical, but conflicting terms, namely that the risks to privacy posed by data release should either trump or be trumped by broader public benefits to be gained from data release. Contrasting positions are exemplified by the Joseph Rowntree Trust[170] *Database State* report, which argues that the public are neither served nor protected by the increasingly complex holdings of personal information that extend government oversight of every aspect of our lives, and by the recent changes to the NHS constitution that extend access to patient records from GPs to medical researchers. The recent public dialogue on open data by RCUK found that participants were generally relaxed about data confidentiality, so long as appropriate governance provisions were in place, but a significant minority were very concerned about the issue.[171] These are issues of public values that are not easily amenable to technical solutions and deserve a broader public debate.

In the context of privacy interests, it would be inappropriate to have a commitment to openness in science that includes putting data that could result in the identification of individuals directly into the public domain. This report advocates a proportionate approach to sharing, compilation and linkage of datasets containing personal data for research purposes. Public benefit and risks to confidentiality need to be assessed and balanced in individual cases, recognising that no processing of data can entirely preclude these risks.[172] A variety of governance mechanisms has been developed to minimise risks to privacy while facilitating access to data for research and other purposes. Examples include informed consent and the use of safe havens that limit data access to researchers with a legitimate interest in the data but who are subject to penalties if they breach confidentiality. Some procedures involve oversight by an independent body to advise on the ways that specific forms of data may be used, although the consequent patchwork of legal, ethical and practical considerations can often be difficult for researchers to navigate.

Consent to use a person's data is often thought to be a gold standard in information governance. However, it is neither a necessary nor a sufficient step in protecting the range of interests at stake either from an ethical or legal perspective. Consent functions both legally and ethically as a way of waiving benefits or protections to which an individual would otherwise be entitled. Where there is no underlying entitlement, there is no requirement to

169 The policy implications of these results are explored in Ohm P (2010). *The Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.* UCLA Law Review, 57, 1701), O'Hara K (2011). *Transparent Government, Not Transparent Citizens.* A Report on Privacy and Transparency for the Cabinet Office. Available at: http://www.cabinetoffice.gov.uk/resource-library/independent-transparency-and-privacy-review

170 Anderson R, Brown I, Dowty T, Inglesant P, Heath W and Sasse A (2009). *Database State*. Joseph Rowntree Reform Trust, 67.

171 TNS BMRB (2012). *Public dialogue on data openness, data re-use and data management Final Report*. Research Councils UK: London. Available at: http://www.sciencewise-erc.org.uk/cms/public-dialogue-on-data-openness-data-re-use-and-data-management/

172 Decisions to respond to FoI requests based on concerns about personal data are based on the method of anonymisation: see (1) *Common Services Agency v Scottish Information Commissioner* [2008] UKHL 47 in which privacy concerns arose about the release of very small numbers of incidences of childhood leukemia as a result of a freedom of information request. The data had been 'barnadised' – randomly adding 0, +1 or -1 to cells of data - but still the data custodian was concerned that children could be identified from this table and other information in the public domain; (2) *Department of Health v Information Commissioner* [2011] EWHC 1430 (Admin). An FOI request for release of full aggregated abortion statistics was upheld on the grounds that this was not a release of 'personal data'. The court confirmed that release of data under FOIA is equivalent to release to the public as a whole, without conditions. The risks must be assessed accordingly. (England and Wales High Court (Administrative Court) Decisions (2011).)

seek consent.[173] Consent also struggles to adequately defend interests that go beyond the individual.[174] Much research using datasets involves research questions that were not anticipated when consent was given for the original data collection. In such cases, it is often prohibitively difficult and expensive to re-contact all the original data subjects. Broad prospective consent (of the type used by UK Biobank) raises questions about the extent to which data subjects fully understand what this might imply about possible research uses of their personal data.[175] This form of consent can be complemented through oversight by an independent body that advises on the appropriate uses of the data and determines whether re-consent is required. The UK Biobank and its Ethics & Governance Council is an example of this practice.[176] In other contexts an authorising body can be used as an alternative to seeking consent from data subjects, and designed to operate in the public interest.

Safe havens are created as secure sites for databases containing sensitive personal data that can only be accessed by authorised researchers. They were an important recommendation of the 2008 Data Sharing Review.[177] Their use was advocated for population-based research in order to minimise the risk of identifying individuals. The Review recommended that approved researchers be bound by strict rules, with criminal sanctions and the prospect of a custodial sentence up to a maximum of two years if confidentiality is breached. Safe havens transfer the information governance problem from one of protecting privacy through anonymisation to requiring trusted and accredited researchers to maintain confidences. They have proven successful for large scale studies such as the Scottish Longitudinal Study (see box 3.6), although because of the cost and manpower overheads associated with them, it is recommended that they are only used for large high-value datasets. The emphasis on confidential relationships is, however, a universal consideration and consent should be the norm for safe haven data. Confidentiality agreements guard against abuses of the trust that is put in researchers by granting them access to sensitive datasets. All researchers accessing datasets containing personal data should be required to sign confidentiality agreements, with clear sanctions for breaching them, which are both contractual and professional.

173 Brownsword R (2004). *The Cult of Consent: Fixation and Fantasy*. King's College Law Journal, 15, 223-251. ["it is a mistake to view consent as a free-standing or detached principle (on the same level as privacy, confidentiality and non-discrimination); rather consent is implicated in the right to privacy, the right to confidentiality, and the right against discrimination – in each case the right-holder may consent to waive the benefit of the right in question."]

174 O'Neill O (2003). *Some limits of informed consent.* Journal of Medical Ethics, 29, 4-7.

175 These factors mean, for example, that in the given examples consent is unlikely to be a lawful basis for processing personal data under the new Data Protection Regulation which would require consent to be "…freely given specific, informed and explicit…" – draft Article 4(8).

176 UK Biobank (2012) *Ethics and Governance Council.* Available at: http://www.egcukbiobank.org.uk/

177 Thomas R & Walport M (2008). Data Sharing Review. Available at: http://www.justice.gov.uk/reviews/docs/data-sharing-review-report.pdf The idea of a location where records are kept, and made available to only bona fide researchers is not new. Innovation over the past few years has centred around how such arrangements can be reinvented for a world of electronic records.

### Box 3.6 Safe Havens: The Scottish Longitudinal Study

The Scottish Longitudinal Study (SLS) is a good example of how safe havens can be used to share complex and sensitive data.[178] Data from routine administrative and statistical sources, including Census data (1991 and 2001), vital events data (births, marriages, deaths), NHS Central Register (migration in or out of Scotland) and NHS data (cancer registration and hospital discharges), are collated to examine migration patterns, inequalities in health, family reconstitution and other demographic, epidemiological and socio-economic issues.[179] Such data are an invaluable source of information for social policymaking. To protect people's privacy a series of safeguarding measures exist. First, the SLS sources its data from individuals with one of 20 predetermined birthdates. Only a small group of researchers actually know these dates. Second, the dataset is anonymous – individuals included in the survey are assigned an SLS number ensuring that no names or addresses are retained on the database and that anonymity is maintained. Third, the actual data are stored on an independent network that is password protected and access to the data can only occur in 2 rooms that are protected by a keypad. Fourth, a Steering Committee oversees the maintenance and use of the SLS and a Research Board, which reviews every research proposal, will not authorise any studies to be undertaken in which individuals may be identified. Fifth, data are not made publicly available. In addition, access to the data for authorised projects is rigorously controlled. A data subset strictly tailored to the research needs is created – no unnecessary data are included and no data are sent off site. If researchers want to analyse the data remotely, a statistical program can be run on their behalf by the SLS Centre. Only the results are returned to the researcher following checks to ensure that no personal information is included. Alternatively, the researcher can analyse data in one of the two 'safe rooms' alongside a member of the SLS Support Team.

The regulatory landscape is changing rapidly. The Data Protection Directive is under review and the European Commission has published a proposal to substitute it by a Data Protection Regulation[180] that will now be considered by the European Parliament and Council of Ministers. Unlike a Directive, a regulation removes discretion from member states on the implementation of the requirements in national law. In the UK, the majority of the functions of the National Information Governance Board are likely to be taken over by the Care Quality Commission.[181] The Government has indicated that it intends to transfer the functions of the Ethics and Confidentiality Committee to a new Health Research Authority.[182] These changes are an opportunity to revisit good governance in this area, with an eye on feasibility and securing public benefit.

Future governance practices need to reflect the speed that data analysis technologies are changing. Protecting privacy will only get harder as techniques for recombining data improve. Governance processes need to weigh the potential public benefit of research against the very latest technical risks.

This report recommends 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.

178 Longitudinal Studies Centre – Scotland (2012). *Scottish Longitudinal Survey*. Available at: http://www.lscs.ac.uk/sls/

179 Hattersley L & Boyle P (2007) *The Scottish Longitudinal Study: An Introduction*. Available at: http://www.lscs.ac.uk/sls/LSCS%20WP%201.0.pdf

180 European Commission (2012). *Regulation of the European Parliament and of the Council*. Available at: http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf

181 Other important bodies and organisations are the National Research Ethics Service and the Research Information Network.

182 Paragraph 7.5 of The National Archives (2011). *The Health Research Authority (Establishment and Constitution) Order 2011*. Available at: http://www.legislation.gov.uk/uksi/2011/2323/memorandum/contents

### 3.3 Security and safety

Open scientific information poses challenges to both systems and hardware engineers to develop ways of sharing confidential, sensitive or proprietary data in ways that are both safe (protected against unintended incidents) and secure (protected against deliberate attack). A more data-intensive future is likely to increase concerns about information security. Data leakage has become essentially irreversible, increasing the stakes for the owners of sensitive data. The situation is most severe for those who hold personal data; a leak is not only hard to manage, but de-anonymisation techniques (in section 3.2) can add further details to the data even when it is in a redacted format. There are already signs that digital security is falling behind; less than a third of digital information can be said to have at least minimal security or protection and only half of the information that should be protected is protected.[183]

Keeping the source code and architecture of systems secret is an untrustworthy method of ensuring information security. Modern cryptography starts from the presumption that what is required is a system that remains secure even when would-be attackers know its internal workings.[184] Whilst revealing the source code and architecture of systems allows attackers to analyse systems for weaknesses, it also permits systems to be tested more thoroughly.[185] This attitude towards security – that openness ultimately breeds better security - can be applied to scientific data as well.

Scientific discoveries often have potential dual uses - for benefit or for harm.[186] Cases where national security concerns are sufficient to warrant a wholesale refusal to publish datasets are rare. Nature Publishing Group received 74,000 biology submissions between 2005 and 2008, of which only 28 were flagged as having potential for dual use. None were rejected for publication. Although this report is not aware of any cases where harmful consequences have arisen, it would be foolish to assume that this could not happen. Restrictions on the export of sensitive information are in place in the UK through the Export Control Organisation.[187]

183 IDC (2011). *Digital Universe study: Extracting Value from Chaos*. Available at: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

184 Shannon C (1949). *Communication Theory of Secrecy Systems*. Bell System Technical Journal, 28 , 4, 656-715. This was first argued in: Kerckhoffs A (1883). *La Cryptographie Militaire*. Journal des Sciences Militaires, 9, 5, 38. Available at: http://www.cl.cam.ac.uk/users/fapp2/kerckhoffs/

185 In real world circumstances, open systems may be either more secure or less secure than closed ones, depending on a number of factors such as the willingness of users to report bugs, the resources available for patching reported bugs, and the willingness of vendors to release updated versions of software. See Anderson R (2002). *Security in Open versus Closed Systems – The Dance of Blotzmann, Coase and Moore*. Open Source Software Economics. Available at: http://www.ftp.cl.cam.ac.uk/ftp/users/rja14/toulouse.pdf

186 See Parliamentary Office of Science and Technology (July 2009). *POSTNote 340, The Dual-use Dilemma*. Available at: http://www.parliament.uk/documents/post/postpn340.pdf

187 See BIS (2010). *Guidance on Export Control Legislation for academics and researchers in the UK*. Available at: http://www.bis.gov.uk/assets/biscore/eco/docs/guidance-academics.pdf. According to the Export Control Act (2002) s8, the Secretary of State may make a control order which has the effect of prohibiting or regulating the communication of information in the ordinary course of scientific research, or the making of information generally available to the public only if "the interference by the order in the freedom to carry on the activity in question is necessary (and no more than is necessary)".

### Box 3.7 Dual use: publishing two new avian flu papers

The H5N1 avian flu virus rarely infects humans and does not spread easily from person to person. There is ongoing concern that the virus could evolve into a form transmissible among humans, thereby creating a serious global public health threat. Research on factors that can affect the transmissibility of the H5N1 virus is vital in understanding this possible threat. But information that helps understand a threat and help with prevention can also be misused for harmful purposes.

Two manuscripts were submitted for publication that describe laboratory experiments which concluded that the H5N1 virus has greater potential than previously believed to be transmitted among mammals, including humans. The US National Science Advisory Board for Biosecurity (NSABB) recommended that the authors and the editors of the journals should publish the general conclusions, with significant potential benefit to the global influenza surveillance and research communities, but not the details that could enable replication of the experiments by those who would seek to do harm.[188]

The journals Science and Nature were sent the manuscripts and initially supported the NSABB verdict. But both also emphasised[189] the need for researchers to access the work in order to maintain proper scrutiny of scientific results. After researchers met at the World Health Organisation (WHO) in February 2012, there was consensus that redacting parts of the papers was not a practical option and that the best solution was to publish the full text,[190] although this decision is subject to a series of further WHO meetings. Nature has now published their paper.

#### An earlier dilemma

In 2005 a team of US scientists sequenced the 1918 flu virus.[191] Another group then published a paper describing how they had used this sequence to reconstruct the complete virus.[192] Although understanding the genetic makeup of this strain was extremely helpful given both its virulence and its high mortality, the contents of the articles would also make it easier for terrorist groups to make use of this pandemic strain. Some voices therefore urged caution. The NSABB examined both articles before publication, and concluded that the benefits of publishing outweighed the potential harms resulting in the publication of the research.

188 Science: Journal Editors and Authors Group (2003). *Statement on Scientific Publication and Security*. Science, 299, 1149. Available at: http://www.sciencemag.org/site/feature/data/security/statement.pdf

189 Butler D (2011). *Fears grow over lab-bred flu*. Nature, 480, 421–422. Available at: http://www.nature.com/news/fears-grow-over-lab-bred-flu-1.9692

190 World Health Organisation (2012). *Report on technical consultation on H5N1 research issues: Geneva, 16–17 February 2012.* Available at: http://www.who.int/influenza/human_animal_interface/mtg_report_h5n1.pdf

191 Taubenberger, J K, Reid A H, Lourens R M, Wang R, Jin G & Fanning T G (2005). *Characterization of the 1918 influenza virus polymerase genes*. Nature, 437, 889-893.

192 Tumpey T M *et al* (2005). *Characterization of the Reconstructed 1918 Spanish Influenza Pandemic Virus.* Science 310, 5745, 77–80.

A joint report by the Royal Society, the Inter-Academy Panel and International Council for Science in 2006 concluded that "restricting the free flow of information about new scientific and technical advances is highly unlikely to prevent potential misuse and might even encourage misuse".[193] The sequence of the polio virus was published in 1980. Live polio virus was recreated from cloned DNA in 1981. Between 1981 and 2001 the ability to recreate polio virus and other picornaviruses from cloned DNA created medical benefit, enhancing understanding of the viruses and permitting a more stable vaccine to be derived which also reduced the threat of misuse.

Funders currently screen research for potential dual uses and there is a common sense acceptance of responsibility by publishers and the wider scientific community. But the unusual series of events around avian flu research (box 3.7) raises a more general concern about the safety and security of scientific materials. In this case, the National Science Advisory Board for Biosecurity has no power over versions of the paper accidentally in the public domain, and which may be stored, for example, on a university server. This opens the wider security issue of the 'cyberhygiene' of research data. It is important that there are clear rules for access and copying information and that they evolve as the nature of data evolves. There are good examples of best practice at an institutional level, such as JISC's development of the US Shibboleth single sign-in system that removes the need for content providers to maintain user names and passwords, and allows institutions to restrict access to information at the same time as securing it for remote access for approved users.

Historically, confidentiality of data has been almost synonymous with security. Keeping personal data safe was the main motivation behind creating secure systems. More recent developments indicate that ensuring data integrity and provenance are also significant motivations for creating secure systems, as well as the need to keep data available to those that created it. There are a number of accepted standards for such practices in commercial settings. These should be regarded as a minimum in standard protocols for secure scientific data management.[194] These should be development and sharing of good practice and common security and information sharing protocols.

Codes of conduct for professional scientists also have a part to play in encouraging individual responsibility. All scientists sign a contract with their employer which states that they agree to comply with all local and national safety legislation. Despite this security concerns should continue to be addressed separately, as they are in the UK's Universal Ethical Code for Scientists.[195] Tertiary scientific education should include an understanding of the process for identifying and reporting risks resulting from research and should be accompanied by a clear description of the benefits. Any guidelines should reflect the point that greater security can come from greater openness as well as from secrecy.

193 Royal Society (2006). *Report of the RS-ISP-ICSU international workshop on science and technology developments relevant to the Biological and Toxin Weapons Convention*.

194 Examples of sources, both extant and in the process of generation, for such standards include: The ISO 27000 Directory (2009). *An Introduction to ISO 27001, ISO 27002....ISO 27008*. Available at: http://www.27000.org/ ; aiim (2012). *The Global Community of Information Professionals.* Available at: http://www.aiim.org/Resources/Standards/Catalog ; The National Archives (2012). *Standards*. Available at: http://webarchive.nationalarchives.gov.uk/+/http://www.dti.gov.uk/sectors/infosec/infosecadvice/legislationpolicystandards/securitystandards/page33369.html ; W3C (2011). *Connection Task Force Informal Report.* Available at: http://www.w3.org/2011/prov/wiki/Connection_Task_Force_Informal_Report; Moreau L & Foster I (2006). *Provenance and Annotation of Data Springer*. Springer: Heidelberg. Available at: http://www.w3.org/2011/prov/wiki/Connection_Task_Force_Informal_Report

195 Government Office of Science (2007). *Rigour, Respect, Responsibility: a universal ethical code for scientists.* Government Office of Science: London.

# Realising an open data culture: management, responsibilities, tools and costs

Chapter 1 argues that the default position for data should be intelligent openness rather than closed data – for the sake of better science and to be an asset in society. Chapter 2 examines how emerging communication technologies create opportunities for this openness and Chapter 3 looks at the boundaries to this openness in order to protect competing values.

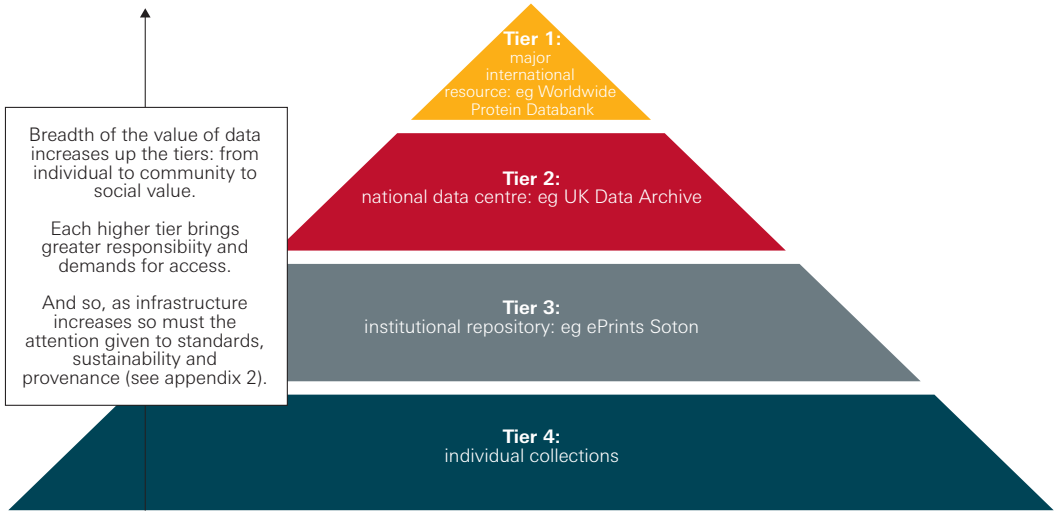
This chapter focuses more on the practical qualifications on openness. Sharing research data can be complex and costly and needs to be tempered by realistic estimates of demand for those data. 4.1 sets out a tiered taxonomy of ways research data are managed, and the demands on that data that lead to these differing levels of curation. In 4.2, there is a set of principles for how custodians of data should operate within this tiered approach and where changes need to be made. The limitations of tools of data management (4.3) and operational costs (4.4) must be taken into account.

## 4.1 A hierarchy of data management

In understanding patterns of the current management of scientific data it is helpful consider it as being done within four tiers of activity that reflect the scale, cost and international reach of the managed data and, to some degree, their perceived importance.[196] This formalises the discussion of diverse data needs in 2.1.2. Each Tier requires different financial and infrastructural support (Box 4.1)

### Box 4.1 The Data Pyramid – a hierarchy of rising value and permanence

Details of examples given in appendix 3.




196 There are various ways of dividing these levels. Our version owes much to: National Science Foundation (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Available at: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf

**Tier 1** comprises of major international programmes that generate their own data, such as those clustered at the Large Hadron Collider, or that curate data from a large number of sources, such as the Worldwide Protein Data Bank (see appendix 3). The former rely on a complex network to distribute their data, whilst the latter rely on distributing tools for the submission, curation and distribution of data.

**Tier 2** includes the data centres and resources managed by national bodies such as UK Research Councils or charitable funders of research such as the Wellcome Trust. Major funding is earmarked for these centres for the specific purpose of supporting their curation activities. Many are managed by agents on behalf of national bodies. For example HepData[197] is a database for high energy experimental physics funded by the UK Science and Technology Facilities Council, which has been managed by Durham University for 25 years.

**Tier 3** is the curation undertaken by individual institutions, universities and research institutes for the data generated by their programmes of research. Their approaches vary greatly. Although many have policies for research data, this tends to take the form of broad advice to researchers rather than stronger, targeted support or active curation based on comprehensive oversight of the full range of data generated by the research efforts that they house. This issue is addressed below.

**Tier 4** is that of the individual researcher or research group. Outside the fields where the disciplinary norm is to make data available to international databases or where there are national databases designated as repositories for researchers' data by their funders, research groups collate and store their own data. They tend only to make it available to trusted collaborators, but may also make it publicly accessible via their own or institutional websites, which are also an important means of data storage. The curation effort is generally supported using the funding of researchers' projects and programmes. They are often dependent for data filing on a small range of conventional off-the-shelf tools, such as Excel or MATLAB, but which lack functionality for many of the needs of efficient curation, data use and sustainability.

Even for small groups of bench or field scientists, data is increasingly digital and increasingly measured in mega-, giga-, tera- and petabytes. They need powerful and easy-to-use data management tools if they are to exploit their data to maximum effect and if they are to respond to the challenge of the data-gap and the opportunities of new tools for data analysis (Box 4.2 shows how natural history scientists have started to build this kind of distributed infrastructure using linked data technologies described in 2.1.4). The needs of Tier 4 are poorly served. A meeting with a group of the Royal Society's University Research Fellows involved in data intensive science confirmed problems of data management and that the skills required for efficient curation are often more sophisticated than can be expected of scientists from non-informatics disciplines.[198]

There are new tools that facilitate data sharing at this level. Figshare[199] allows immediate pre-publication data sharing among scientists through a web-based portal, focusing on sharing negative results or results that would not otherwise be published. These data are valuable, as illustrated by the fact that many major databases have arisen from compilations at this level. It is important therefore that the mechanisms developed are able to identify and support data that deserve sustainable curation.

197 Durham University (2012). *HEPDATA – the Durham-HEP database.* Available at: http://www.ippp.dur.ac.uk/Research/ProjectsHEPDATA.html

198 Roundtable with URFs, 15 January 2012. See appendix 4 for participants.

199 Figshare (2012). Available at: http://figshare.com/

### Box 4.2 DIY data curation - Scratchpad Virtual Research Environment[200]

The Scratchpad Virtual Research Environment is an online system for people creating research networks to support natural history science. Since it was set up in March 2007, 340 websites had been created, with 6000 registered users from more than 60 countries. Over 400,000 pages of content are currently hosted by the Natural History Museum. One of the initial sites - devoted to fungus gnats - was discovered on Google by a group in Finland, who then contacted the author and asked to contribute to the site.

Anyone can apply for a Scratchpad. The only requirement for a statement of scope for the site that has some bearing on natural history. The original objective was to support taxonomy of plant and animal species, but this has broadened to include sites for local activities and hobby groups.

Scratchpad is piloting a mechanism that links content from a webpage to produce an XML file that is submitted for publication in a journal. The publisher renders the XML into a PDF and automatically forwards the manuscript to reviewers. If the reviewers comments received are positive, then the paper is published as: a paper version, as an open-access PDF, as HTML and as XML. The XML format is then available for text-mining and certain data are automatically extracted to be fed to the many international aggregators for taxonomies, such as Catalogue of Life and Globalnames. All new taxon names are registered prior to publication in international registries, such as International Plant Name Index and ZooBank. Scratchpad and its partner publishers are committed to the transition from the publishing of texts to the publishing of data, including marked-up text to be mined by machines.

The above hierarchy does not cover databases created by private companies for commercial purposes. These include databases designed to support a company's business practices, including data on customers and suppliers. Other databases are collected and disseminated by companies for profit. When these databases are related to research activities, there are examples (3.1.3) of profitable collaborations from sharing data at each Tier. Some of the most promising private enterprises are concerned with developing tools for data access and reuse rather than hoarding data. In a simple example, search engines have a vital function in the public data domain as intermediaries between data sources and data users. Although some private data is rightly proprietary, industry demand and supply of data should be a factor in the design of data management in every Tier.

## 4.2 Responsibilities

An effective data ecology must adapt to changing research needs, behaviours and technologies. Data collections often migrate between the tiers particularly when data from a small community of users become of wider significance and move up the hierarchy or are absorbed by other databases.[201] The catch-all life sciences database, Dryad, acts as a repository for data that needs to be accessible as part of journals' data policies. This has led to large collections in research areas that have no previous repository for data. When a critical mass is reached, groups of researchers sometimes decide to spin out a more highly curated specialised database. Most recently, population geneticists interested in interbreeding data created their own database hosted by the University of British Columbia.

200 Scratchpad (2012). Available at: http://scratchpads.eu/

201 National Science Board (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* Available at: http://www.nsf.gov/pubs/2005/nsb0540/.

This dynamic system requires a set of principles of stewardship that the custodians of scientific work should share:

- Openness as standard
  The default position is openness and responsible sharing, with only justifiable and reasonable restrictions on access and use.

- Clear policies
  All custodians should have transparent policies for custodianship, data quality and access.

- Clear routes to access
  Positive action is taken to facilitate openness through the creation of clearly signposted data registers.

- Pre-specification of data release
  When and how newly acquired data will be released for reuse must be specified in advance.

- Respect for values that bound openness
  Governance mechanisms for protection of privacy and commercial interests.

- Rules of sharing
  Explicit terms and conditions for data sharing.

There are four areas where changes are needed in order for these responsibilities to be executed.

### 4.2.1 Institutional strategies

There is a particular concern about Tier 3 in the citation hierarchy, the institutions, and particularly universities, which employ scientists. Notwithstanding the financial pressures under which these bodies currently find themselves, the massive and diverse data resources that they generate from their research suggests two timely questions: what responsibility should they have in supporting the data curation needs of their researchers and what responsibilities should they have for curating the data they produce?

In the UK, JISC and the Digital Curation Centre have developed support tools ranging from onsite development of data management for 17 large research institutions in the UK[202] to training for individual researchers.[203] The pervasive data deluge in science means that familiarity with such tools and principles of data management should be an integral part of the training of scientists in the future. However, it is unreasonable to expect a scientist in a non-informatics field to be as adept as data scientists. Ideally, they should be supported by those who are.

Data created by research is routinely discarded as it has little long term value. Much important data with considerable reuse potential, is also lost, particularly when researchers retire or move institution. This report suggests that institutions should create and implement strategies designed to avoid this loss. Ideally data that has been selected as having potential value, but for which there is no Tier 1 or Tier 2 database available, and which can no longer be maintained by scientists actively using the data, should be curated at institutional (Tier 3) level. Commercial companies now offer services that claim to organise and archive laboratory data, digital laboratory notebooks, spreadsheets and images that can be readily retrieved, shared with collaborators and made public if desired.[203] Such functions could and should be developed by or for institutions.

A particular dilemma for universities is to determine the role of their science libraries in a digital age. In the majority of cases (86%), libraries have led responsibility for the university repository.[204] The traditional role of the library has been as a repository of data, information and knowledge and a source of expertise in helping scholars access them. That role remains, but in a digital age, the processes and the skills that are required to fulfil the same function are fundamentally different. They should be those for a world in which science literature is online, all the data is online, where the two interoperate, and where scholars and researchers are supported to work efficiently in it.

201 Through the JISC (2012). *Managing Research Data Programme 2011-13*. Available at: http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx
202 Data Curation Centre (2012). *Data management courses and training*. Available at: http://www.dcc.ac.uk/training/data-management-courses-and-training.
203 Labarchives (2012). Available at: http://www.labarchives.com
204 Repositories Support Project (2011). *Repositories Support Project Survey 2011*. Available at: http://www.rsp.ac.uk/pmwiki/index.php?n=Institutions.Summary

### 4.2.2 Triggering data release

The timing of data release is an important issue for an open data culture. In some areas such as genomics, immediate release has been the rule and scientists who wish to publish on data that they have contributed have adapted to that rule. It is understandable, however, that researchers should be reluctant to publish datasets that they have collected for fear of being "scooped" to publication, and it is quite appropriate for researchers to have a short and well defined period of exclusive access to their data to give them time to analyse and publish their results. Given the diversity of established practice (see appendix 1), a 'one size fits all' approach to release would be unhelpful. In some cases data is released immediately without restriction, in others subject to conditions that reflect demand for use of data, commercial value and the need to ensure that standards are in place for data and metadata.

This report suggests that the timing of release should be pre-specified in ways that are consistent for an institution and for disciplinary practices. For grant-funded research this pre-specification should be part of the data management plan that most research funders require at the time of application for funding. As the volume of information produced by research increases, further practical barriers to data sharing will no doubt occur.[205] This is not an excuse for outright refusal to provide data. It is important to develop a broad consensus that regards the hoarding of data as inimical to scientific progress, in which independent validation of datasets, replication of experiments, testing of theories and reuse of data in novel ways by others are essential elements. Despite this, universal instant access to data is neither a realistic nor a desirable goal for research.

### 4.2.3 The need for skilled data scientists

Data science is a fast increasing discipline in its own right (as reflected in the decade of growth of papers with the data as the topic in Figure 4.3a). The skills of data scientists are crucial in supporting the data management needs of researchers and of institutions. They are mathematically adept, and are often informatics trained scientists expert in the tools and processes of data management and the principles that underlie them. A US National Science Foundation (NSF) report describes them as combining the skills of informaticians and librarians.[206] As part of new funding announced in 2012, the NSF will issue a $2 million award for undergraduate training in complex data, whilst also encouraging research universities to develop interdisciplinary graduate programs in Big Data.[207] There are now courses designed to train such highly skilled professionals at some UK universities (eg Southampton, Edinburgh). The private sector is hungry for informaticians and data scientists (Figure 4.3b), and has serious worries about the supply of suitably trained candidates. A well designed career path will be required if universities and research institutes are also to attract and retain them.

## 4.3 Tools for data management

Given the volume and complexity of modern digital databases, filing, sorting and extracting data have become demanding technical tasks that require sophisticated software tools if they are to be carried out efficiently and effectively. Although this report is able to specify the functions that it recommends such tools to fulfil, and although many new tools have been developed, a lot of research and development is still needed from computer scientists if the full potential of the digital revolution is to be realised. For both the data used for scientific purposes and

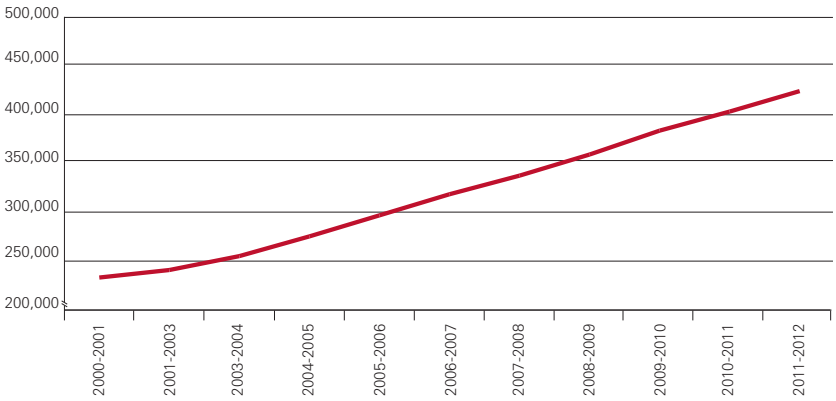
205 In a 2011 survey, 1295 authors of articles in the Royal Society journals were asked: "should all data relating to an article be deposited in a public domain database?" 57% responded "yes" and 43% "no". Reasons for responding negatively included the practical difficulties of sharing data.

206 National Science Board (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Available at: http://www.nsf.gov/pubs/2005/nsb0540/

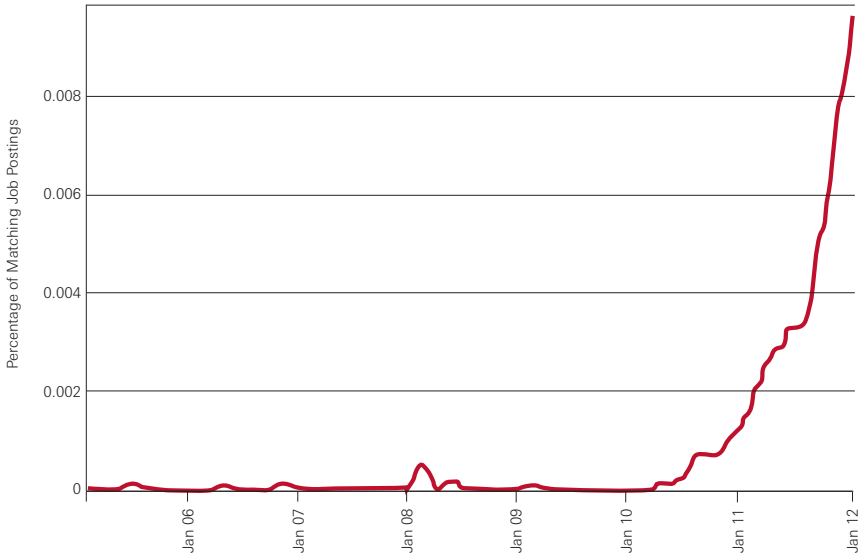
207 White House Press Release (2012). *Obama Administration unveils 'Big Data' initiative. Announces $200 million in new R&D investments*: *29 March* Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

**Figure 4.3**

**a** Number of articles with 'data' as topic in the Thomson Reuter's Web of Knowledge




**b** Job adverts using the phrase 'data scientist' on Indeed.com




those used for commercial benefit. Moreover the sustainability of current levels of data archiving is seriously in question.

Appendix 2 summarises some of the software tools that are needed and some of the related issues that must be addressed in supporting the changes in scientific practice that are advocated in this report. Much data is *dynamic*, changing as new, better data is acquired or data treatment procedures are improved. There needs to be methods to ensure that linked databases can be readily updated rather

than becoming "stale". Better tools for *indexing* data are needed so that the precise data that is required can be readily extracted from larger data resources. There is an urgent need for new tools to support the whole *data cycle* from data capture by an instrument or simulation through: processes of selection, processing, curation, analysis and visualisation with the purpose of making it easy for a bench or field scientist who collects large or complex datasets to undertake these tasks.

Tracking the *provenance* of data from its source is vital for its assessment and for the attribution of data to their originators. The *citation* of data is crucial in evaluating the contributions of individual scientists, and giving them credit on a par with the citation of scientific articles. This is an essential part of giving scientists incentives to make data available in open sources. If this report's criteria for intelligent openness (accessibility, intelligibility, assessability and usability) are to be observed, *standards* need to be set for them. Common standards and structures are also needed to allow reusers not only to manipulate data but also to integrate them with other datasets.

*Financial sustainability* is needed in order to maintain databases with long term value. It must also include *energy sustainability*; within a decade a significant proportion of global electricity supply will be needed to maintain database servers at the present rates of database growth.

## 4.4 Costs

The costs of maintaining a repository for research data have been estimated to be in "an order of magnitude greater than that for a typical repository focused on e-publications".[208] These figures are consistent with the evidence in appendix 3. arXiv.org, which plays an increasingly prominent role in physics, mathematics and computer science; and currently stores only articles and not data. It requires the equivalent of six full time staff. Two indicative world-leading repositories for data are the Worldwide Protein Data Bank and the UK Data Archive. Each requires a multi-million pound budget and upwards of 65 full time staff.

Except for massive datasets, such as the data produced by the Large Hadron Collider or European Bioinformatics Institute, the costs of data storage and backup are small compared with the total costs of running a repository. For example, the Worldwide Protein Data Bank (wwPDB) archive, which is the worldwide repository of information about the 3D structures of large biological molecules, holds around 80,000 structures, but the data it holds occupies no more than 150GB in total (less than the hard disk storage of an inexpensive laptop). A rule of thumb estimate of cost used in some universities is that the provision of storage and backup of research data is approximately £1/gigabyte/5yrs total cost, excluding extended data curation.[209] This would translate into £150 over five years for the costs of storage for the Worldwide Protein Data Bank. The total cost of €6.5 million per year is dominated by staff costs. Nonetheless, curating all known protein structures determined worldwide since the 1960s via wwPDB costs less than 1% of the cost of re-generating the data.

Similarly at Tier 2, large scale digital curation is typically between 1-10% of the cost of research. Box 4.3 shows the balance of funding for Earth Science research currently funded in the UK through the Research Councils.

208 Beagrie N, Chruszcz J and Lavoie B (2008). *Keeping Research Data Safe 1*. JISC. Available at: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

209 Evidence provided by RCUK. In addition, in the costing model produced for the University of London National Digital Archive of Datasets, the costs of physically storing each megabyte of information were found to be only one percent of the total costs of curating a megabyte. (Beagrie N, Lavoie B and Woollard M (2010). *Keeping Research Data Safe 2*. JISC. Available at: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf)

At Tier 3, the 2011 Repositories Support Project survey received responses from 75 UK universities, which showed that the average university repository employed a total 1.36 FTE, combining managerial, administrative and technical roles. At present, only 40% of these repositories accept research data.[210] Larger budgets and highly skilled staff will be required if the roles that are suggested are to be fulfilled by institutions such as universities. Larger university repositories, such as ePrints Soton (3.2FTE) and the Oxford Research Archive (2.5FTE, and expanding), give an idea of levels of service that could be provided and the benefits these could bring, and are presented in appendix 3. DSpace@ MIT represents a mature version of this scheme. It may once have been a stand-alone project, but today it is one aspect of an enterprise of integrated digital content management and delivery systems operated by the MIT Libraries. It is this report's view that funding should be provided as part of the costs of doing research, but with efficiencies of scale being sought through local or regional consortia.

The services required to host a useful and efficient database include: platform provision; maintenance and development; multiple formats or versions (e.g. PDF, html, postscript, latex); web-access pages; registration and access control (if appropriate); input quality control (consistency with format standards, adequate technical standards); the tools for provenance and management of dynamic data mentioned in 4.3 and detailed in appendix 2; hosting or linking to relevant analysis tools (e.g. visualisation, statistics); measuring and documenting impact through downloads, and data citations. Each of these services adds value to the data curated, but those that are labour intensive add significant costs. Because this report regards digital curation to be part of the research process, the basis for judging the efficacy of investment is not to concentrate on the absolute cost, but the return on investment that enhanced scientific productivity represents.[211]

A recent *Nature* article compared the cost of producing an academic publication through a conventional grant-aided inquiry with the academic publications produced from material from a data repository. It examined the use that had been made by authors other than those who had created the datasets in the Gene Expression Omnibus (GEO). They showed that over 2,700 GEO datasets uploaded in 2000 led to 1,150 published articles by people who had not been involved in the original data acquisition. On a financial basis this would translate into over 1000 additional papers from an investment equivalent to $400,000, which compares very favourably with the 16 papers that would be expected from the same amount of money invested in original research.[212] Similarly, the ESRC's policy of both funding the UK Data Archive, and requiring researchers to demonstrate, before funding the collection of new data, that no suitable data are available for reuse aims to maximise the benefits of sharing.[213]

---

210 A summary of the results:Repositories Support Project (2011). *Repositories Support Project Survey 2011*. Available at: http://www.rsp.ac.uk/pmwiki/index.php?n=Institutions.Summary. A more detailed breakdown by institution is available from: Repositories Support Project Wiki (2012). Institutional Repositories. Available at: http://www.rsp.ac.uk/pmwiki/index.php?n=Institutions.HomePage

211 JISC (2010). *Keeping Research Data Safe 2*. Available at: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf. Also see Research Information Network (2011), *Data centres: their use, value and impact*.Available at: http://www.rin.ac.uk/our-work/data-management-and-curation/benefits-research-data-centres.

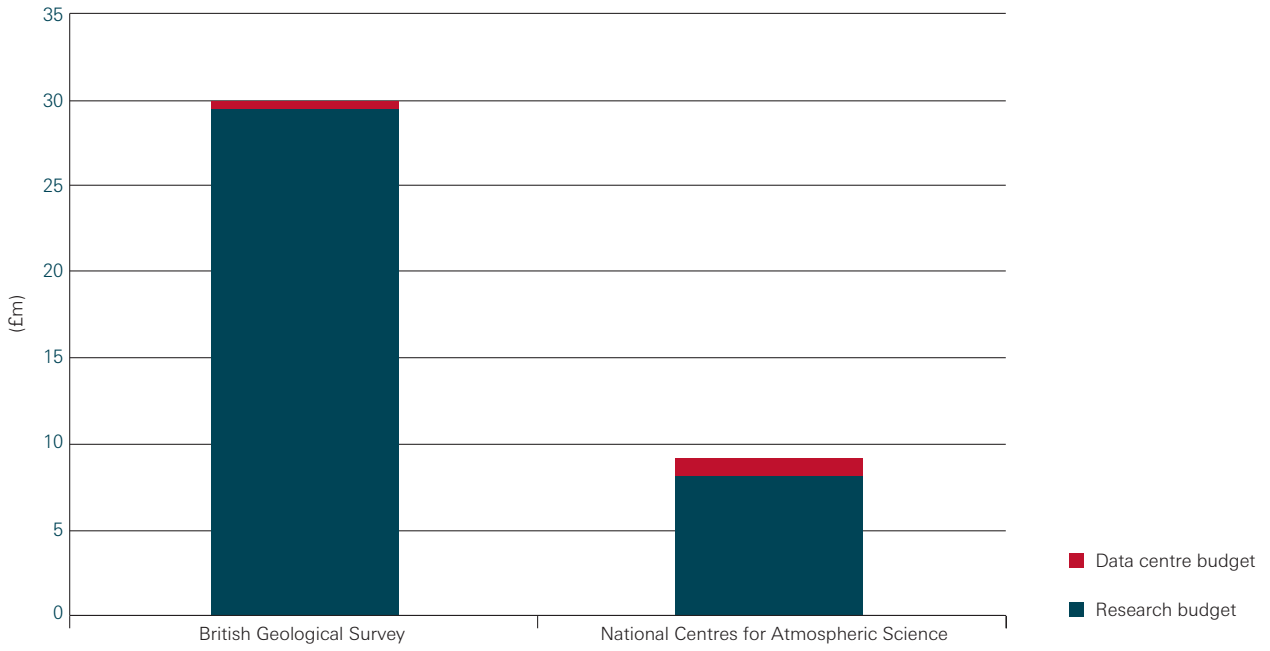
212 Piwowar H A, Vision T J, Whitlock M C (2011). *Data archiving is a good investment.* Nature, 473, 285. Available at: http://dx.doi.orgdoi:10.1038/473285a The indicative repository costs used in the article were those for Dryad (appendix 2).

213 ESRC (2010). *Research Data Policy*. Available at: http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

### Box 4.3 Earth Science data centres

The British Geological Survey has an annual budget of approximately £30 million. It spends £350,000 of this on the National Geosciences Data Centre, which holds over 9 million items dating back over 200 years and includes information from oil companies as well as publicly funded research. The UK's National Centre for Atmospheric Science receive £9 million a year from the science budget. Approximately £1 million of this is spent on curating 228 datasets through the British Atmospheric Data Centre. For the UK's Geological Research, roughly 1% is spent on data curation; for atmospheric research this figure is 10%.

Without this curation, not only would historical data be lost, but much of the rest of the research in this area could not be captured for reuse. Over 70% of research data in these centres is reused by other research projects.



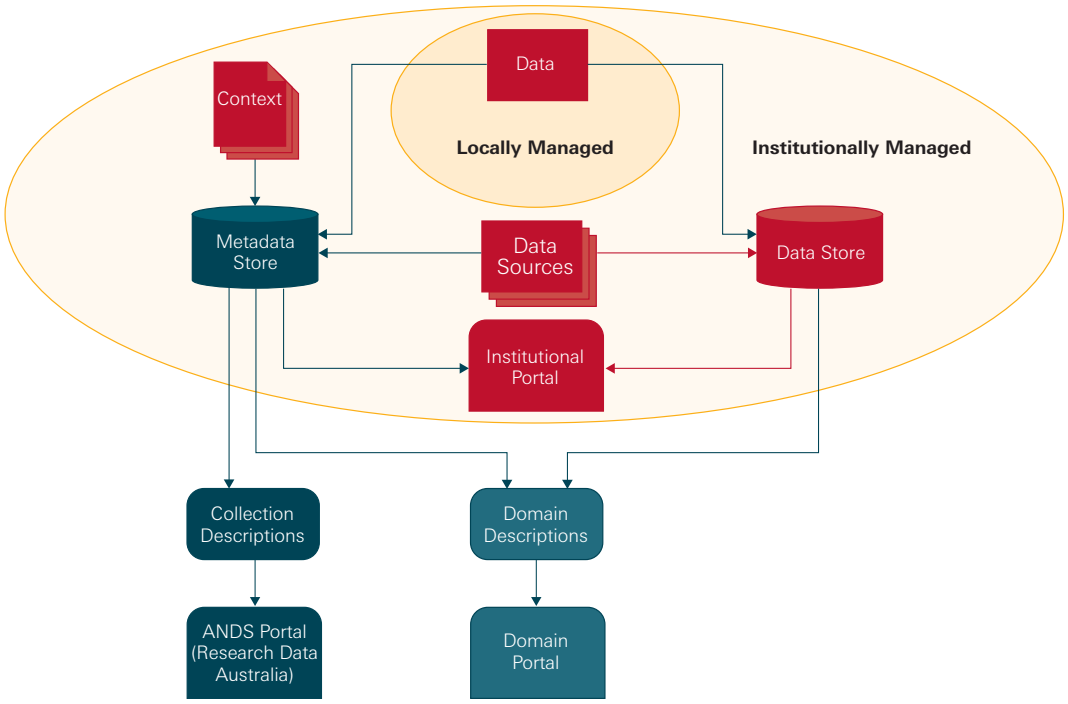

In the UK there is support for Tier 3 and 4 curation from the Joint Information Systems Committee (JISC), which provides leadership in the use of ICTs in post-16 learning and research. This Committee had a 2010-2011core budget of £89.2 million and £27.6 million in capital funding. A joint HEFCE and JISC £10 million project *Shared Services and the Cloud Programme* is establishing shared cloud infrastructure to offer discounted data management and storage services to Higher Education institutions.[214]

The Australian National Data Service[215] was established in 2008 to create and develop an Australian Research Data Commons (ARDC) infrastructure over two years. One of their major activities is Data Capture (AUS$11.6 million), creating infrastructure within institutions to collect and manage data, and to improve the way metadata is managed. Much of this initiative is concerned with managing data in universities (the "institutionally managed" bubble in figure 4.4). The model poses an interesting question that relates to the UK's dual support system: how is responsibility for research data management split between research projects and institutions? The Australia system anticipates the long term systematic change that is necessary

214 JISC (2012). *UMF Shared Services and the Cloud Programme*. Available at: http://www.jisc.ac.uk/whatwedo/programmes/umf.aspx
215 Australian National Data Service (2011). Available at: http://www.ands.org.au/

**Figure 4.4 Australian Research Data Commons**[216]

The management of data in the Australian Research Data Commons. The bubble shows the parts of the process that are institutionally managed, highlighting the division of responsibility between ("local") projects, institutions and the national Research Data Australia infrstructure. Domain descriptions and portals are discipline-specific databases.




to capture the wide range of data produced by the majority of scientists not working in partnership with a data centre.

The Australian model also provides funding for the development of metadata tools through its Seeding the Commons initiative (AUS $4.55 million). The discontinuation of the UK's decade long eScience programme has removed a central focus for similar and vital tool development in the UK since 2009. At the announcement of a recent boost in funding for tools for data management in the US, the deputy director of the White House Office of Science and Technology Policy said "the future of computing is not just big iron. It's big data".[217] The UK Government has recently found an extra £158 million for UK e-infrastructure to be spent on: software development, computer power, data storage, wide bandwidth networks, cyber security and skills. Most of this investment is in the physical infrastructure needed to underpin data-heavy science. The value of e-Infrastructure can only be harnessed by the right tools and skilled professionals. The UK's eScience programme was world leading. A formal review concluded that the programme had developed a skilled pool of expertise, but that these achievements were at a project level rather than by generating infrastructure or transforming disciplines. The changes were not self-sustainable without further investment. The legacy of the eScience programme – in the work of JISC, the national Digital Curation Centre and others - should not be lost.

216 Australian Research Data Commons (2012). Available at: http://www.ands.org.au/about/approach.html#ardc

217 White House Press Release (2012). *Obama Administration unveils 'Big Data' initiative. Announces $200 million in new R&D investments 29 March*. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

# Conclusions and recommendations

The opportunities of intelligently open research data are exemplified in a number of areas of science. With these experiences as a guide, this report argues that it is timely to accelerate and coordinate change, but in ways that are adapted to the diversity of the scientific enterprise and the interests of: scientists, their institutions, those that fund, publish and use their work and the public. This report's recommendations are designed to enhance the impact of researchers work in a new era for science that will transform its conduct and the ways in which the results of research are communicated. These changes will improve the conduct of science and respond to changing public expectations of openness. But not all data are of equal interest and importance. Some are rightly confidential for commercial, privacy, safety or security reasons, and there are both costs as well as opportunities in effective communication of data and metadata. The recommendations set out the key principles, but these must be applied with careful judgement.

Strategies to promote better exploitation of the data universe must respect the contrast between top-down planning, of the type needed to design and build an aeroplane, and the dynamics of bottom-up, emergent behaviour exemplified by the development of the internet and the uses made of it. Planning infrastructure in ways that prescribe or makes assumptions about patterns of use is likely to be misconceived and expensive. It must stimulate rather than crush creativity.

The priority is to ensure that actors in the science community – scientists, their institutions, funders, publishers and government – agree on six broad 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 recommendations artciulate what these changes mean for each actor.

## 5.1 Roles for national academies

The intrinsically international character of open science makes it vital that any recommendations are reinforced by the development of international standards and capacities. The Royal Society will work with other national academies and international scientific unions to encourage the international scientific community to implement intelligently open data policies. It will also support attempts to create global standards for scientific data and metadata. International scientific professional bodies must take the lead in adapting this ideal in the way that data are managed in their communities. In April 2012, the member academies of the ALL European Academies (ALLEA), including the Royal Society, signed a commitment to promoting open science principles for publications, research data and software.[218]

The Royal Society supports efforts in the global scientific community to ensure that countries with a relatively limited national research capacity are able to benefit equitably from efforts to expand global access to research data. Research funding in low and middle income countries should encompass efforts to enhance national capacity for data management and analysis in a sustainable manner.

It is vital to share data in a way that balances the rights and responsibilities of those who generate and those who use data, and which recognises the contributions and expectations of the individuals and communities who have participated in the research. This report argues that the UK can create value from research data while sharing it internationally. The country has the absorptive research capacity to make the most of data as an open resource. Whereas those in poorer countries might be disadvantaged by others taking the data they have generated.[219] While open access publishing models will help to ensure that research publications are freely accessible to researchers the world over, it is crucial to ensure that

218 ALLEA (2012). *Open Science for the 21st century: A declaration of ALL European Academies*. Available at: http://www.allea.org/Content/ALLEA/General%20Assemblies/General%20Assembly%202012/OpenScience%20Rome%20Declaration%20final_web.pdf

219 Walport M and Brest P (8.1.2011). *Sharing research data to improve public health*. The Lancet, 377,9765, 537-539.

the transition to author-pays fees does not limit the ability of scientists in developing countries to publish their work.

## 5.2 Scientists and their institutions

### 5.2.1 Scientists

Scientists aim to seek new knowledge by the most effective routes. This report offers a vision of open science that exploits the potential for achieving this, and is encouraged by the strong trends in some disciplines in this direction. Pathfinder disciplines have committed themselves to and are benefitting from an open data culture. Some researchers are exploring crowd sourcing mechanisms, and some are increasingly reaching beyond the professional boundary of the disciplines.

In contrast to these signs of increased openness, there remains an unhelpful tendency for some scientists to hoard their data. It is understandable that those who have worked hard to collect data are reluctant to release it until they have had an opportunity to publish its most significant implications. This report regards hoarding data as a serious impediment to the scientific process. It inhibits independent validation of datasets, replication of experiments, testing of theories and reuse of data in novel ways by others. Except in those areas where immediate release has become the norm, researchers should have a well defined period of exclusive access to give them time to analyse and publish their results, including negative results. A research grant should pre-specify the timing and conditions of data release.

**Recommendation 1**

**Scientists should communicate the data they collect and the models they creat, via methods that 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 the data justifies 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.**

**Detailed action**

The Royal Society will work with the learned societies and professional bodies that represent the diverse parts of the scientific community to press for the adoption of this recommendation (see also recommendation 4).

### 5.2.2 Institutions (universities and research institutes)

Universities and research institutes have continually adapted to new opportunities for creating and disseminating knowledge. The modern data-rich environment for research and learning and the open culture that is needed to exploit it presents new challenges. These are twofold: creating a setting that will encourage their researchers to adapt their ways of working and developing, and implementing strategies to manage the knowledge that they create.

If the benefits of open science are to spread to new areas of research, the systems of reward and promotion in universities and institutes needs to do more to recognise those who develop and curate datasets. Institutions need to make information and knowledge management part of their organisational strategy. This should include schedules for data and article publication. This is also an opportunity to update IP strategy, helping institutions to take a more diverse approach to the exercise of IP rights.

### 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.**

### Detailed actions

#### Recognition

a. Develop more sophisticated systems of attributing credit for researchers' development and dissemination of data resources – and crucially the use of such resources by others.

b. Monitor these activities amongst their staff, ensuring that these are used as criteria for career progression, promotion and reward.

c. Develop and adopt common and persistent unique individual researcher identifiers through an open system, such as that in development by Open Researcher and Contributor ID (ORCID).

#### Data and information strategies

a. Ensure the datasets that have the scope for wider research use or hold significant long term value are either made available in a recognised subject repository or curated by the university in an accessible form with a link to appropriate repositories.

b. Support the development of local databases with the potential for wider use.

c. Provide education and training in the principles and practice of the management of scientific datasets.

d. Develop and publish a register of data assets that specifies the estimated timetable for data release for funded but as yet unfinished projects.

e. Support a career structure for data scientists and information managers as part of the core business of the organisation. This should include individuals charged with creating and implementing institutional data strategies, as well as those directly supporting researchers in data curation and others critically involved in the development, construction and maintenance of research data infrastructures.

#### UK specific

a. Good practice needs to be shared between institutions. Developing institutional strategies has the coordinated support of JISC and HEFCE. But implementing this will require university board-level scrutiny of knowledge management within an institution.

b. The Intellectual Property Office's recent calls for universities to adopt a more flexible approach to IP management should be taken seriously.[220] The Hargreaves Review of IP in the UK concluded that there is little evidence that current IP legislation harms universities in ways that cannot be reversed by better practices.[221] While Government revisit the issues in the review over the next five years (see Recommendation 8), universities should work to provide more conclusive evidence of IP legislation as a barrier to innovation.

220 Intellectual Property Office (2011). *Intellectual asset management for universities*. Available at: http://www.ipo.gov.uk/ipasset-management.pdf

221 Intellectual Property Office (2011). Supporting Document U: Universities, Research and Access to IP. Available at: http://www.ipo.gov.uk/ipreview-documents.htm

### 5.3 Evaluating University Research

There is an increasing trend towards national assessments of the excellence and impacts of university research. The way in which this is done depends upon the channels through which public funding reaches universities and the way in which accountability for the use of those funds in research is exercised, either directly through government departments or through intermediary bodies. The way in which outputs are assessed is crucial in influencing the behaviour of universities, in particular, in relation to data through the criteria they use in evaluating the research contributions of staff and how they are rewarded and promoted. This report argues that the skill and creativity required to successfully acquire data represents a high level of scientific excellence and should be rewarded as such. At its best, creative, inspired individuals work in a network of intellectual interaction where data and data sharing are key goals in their own right as well as routes to research publications. Citations for open data should therefore be treated as on a par with conventional research publication.

The operation of the equipment and facilities that generate large datasets require the work of teams of people. Similarly the informatics requirements for the manipulation, storage, curation and presentation of large datasets and the underlying metadata require team working. It is important that more effective methods of communal working are not undermined by incentives that exclusively reward conventional modes of working by individuals and small groups.

**Recommendation 3**

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

#### Detailed actions

**Dataset metrics should:**

a. Ensure the default approach is that datasets which underpin submitted scientific articles are accessible and usable, at a minimum by scientists in the same discipline.

b. Give credit by using internationally recognised standards for data citation.

c. Provide standards for the assessment of datasets, metadata and software that combines appropriate expert review with quantitative measures of citation and reuse.

d. Offer clear rules on the delineation of what counts as a dataset for the purposes of review, and when datasets of extended scale and scope should be given increased weight.

e. Seek ways of recognising and rewarding creative and novel ways of communal working, by using appropriately validated social metrics.[222]

**UK-specific**

a. These principles should be adopted by the UK Higher Education Funding Councils as part of their Research Excellence Framework (REF). The REF is a powerful driver for how universities evaluate and reward their researchers. Use in the REF of metrics that record citable open data deposition would be a powerful motivation for data release.

b. JISC's Managing Research Data programme, or a similar initiative, should be expanded beyond the pilot 17 institutions within the next five years. The aim of any initiative should be to support a coordinated national move towards institutional data management policies.

222 Altmetrics (2012). *altmetrics: a manifesto*. Available at: http://www.altmetrics.org/manifesto/

### 5.4 Learned societies, academies and professional bodies

Scientists tend to have dual allegiances, to their disciplines and to the institutions that employ them. Their disciplinary allegiance is strongest in relation to the traditions and habits of research, reflecting the standards, values and priorities of the learned societies, academies and professional bodies that represent their disciplines. The learned societies are well placed to play an important role in promoting a culture of open data as the norm in their disciplinary area, in articulating how it will operate and in seizing the new opportunities that follow from a more open culture.

The Royal Society has hosted several discussion meetings related to data sharing, most recently 'Web Science: a new frontier'.[223] As part of this study, a roundtable on open data was held in January 2012 with University Research Fellows who work with large datasets. In September 2012, the Royal Society and the ICSU Committee on Freedom and Responsibility in the conduct of Science will hold a joint residential discussion on the value of scientific output in the digital age. The conclusions of this meeting will feed into discussions about the provisions the Society makes for its own grant holders.

#### 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 research habits might need to change to achieve this.**

#### Detailed actions

Learned societies and academies should:

a. Define good practice in digital curation for their constituency.

b. Promote collaboration to exploit the opportunities provided by more effective data sharing.

c. Promote the benefits of new data-sharing tools, including providing training opportunities for members.

### 5.5 Funders of Research: Research Councils and Charities

Funders increasingly ask for greater access to data produced from the research they fund. Since 2006, UK Research Councils have had policies[224] for open access to research outputs. Research Councils established its Common Principles on Data Policy .[225] This baseline borrows from the stronger policies some Councils have held for many years.[226] Many other non-commercial funders have now brought into place policies that require the sharing of data within a limited time period after the generation of completed datasets,[227] and most require applicants to submit data management and sharing plans at the grant proposal stage. In 2011, the US National Science Foundation (NSF) went a step further. They require proposals for funding to include a data management plan showing how the proposal will conform to NSF policy: "to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants".[228]

223 Royal Society (2010). *Web Science: a new frontier*. Available at: http://royalsociety.org/Event.aspx?id=1743
224 Research Councils UK (2006). *Research Councils UK updated position statement on access to research outputs*. Available at: http://www.rcuk.ac.uk/documents/documents/2006statement.pdf
225 Research Councils UK (2012). *Excellence with Impact: RCUK Common Principles on Data Policy*. Available at: http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
226 Natural Environment Research Council (2012). *NERC data policy*. Available at: http://www.nerc.ac.uk/research/sites/data/policy2011.asp
227 Digital Curation Centre (2012). *Overview of funder's data policies*. Available at: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
228 National Science Foundation (2011).Dissemination and Sharing of Research Results. Available at: http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Monitoring of the rates compliance with funders' policies is often sporadic, and no funders yet release statistics on these rates. As research funders progress from common principles to more detailed data management requirements, levels of compliance must be addressed. National funding bodies - the Research Councils in the UK – have a leadership role to play in progressive moves towards refusing applications for funding for those that do not share data. This should progress must be sensitive to the norms for data management in different disciplines. Costly and complex data sharing must not be enforced where there is no demand for access to the data or without suitable infrastructure to support its curation.

The Royal Society has begun to include data management policies for larger grants, which cover some costs of research as well as wages.[229] In the next review of fellowship grant policy, the Society will explore the possibility of extending this across its portfolio of grants. It will relaunch its Research Fellows database in 2012, providing access to information about the researchers the Royal Society fund and the research they do.

### 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.**

### Detailed actions

#### UK-specific

a. Research Councils UK has announced a *Gateway to Research*[230] initiative to provide expanded access to information about Research Council grants. This initiative will provide a database detailing successful funding applications, aimed at making research more accessible for small businesses. This scheme will not be running until 2013. In the mean time, more work needs to be done to understand interdisciplinary and public interest in using the Gateway, so it can be quickly expanded to meet these needs as well. This process should consider demands for research data and not just data about research.

b. The common data policies for Research Councils in the UK need to be updated within the next year to require data management plans. Data management planning is not part of the assessment of a grant application, but a way to join up research and the support that is already available through UK data centres and JISC. A plan could be as simple as demonstrating compliance with an institutional information management strategy.

229 Royal Society (2012). *Sir Henry Dale Fellowships*. Available at: http://royalsociety.org/grants/schemes/henry-dale/
230 RCUK (2011). *Gateway to Research Initiative*. Available at: http://www.rcuk.ac.uk/research/Pages/gtr.aspx

### 5.6 Publishers of Scientific Journals

Most scientific research finds its way into the public domain via academic journals. Ideally, all the data that underlie the research or argument presented in an article, but which is not included for reasons of space, should be accessible electronically via a link in the article. Alternatively the publication should indicate when and how the data will be available for others to access. In unusual cases in which there are compelling reasons for not releasing data, researchers should explain in a publicly accessible manner why the data are being withheld from release. An increasing number of journals have explicit policies that require data to be made available, but the rates of compliance are low.

**Recommendation 6**

**As a condition of publication, scientific journals should progressively enforce requirements for traceable and usable data available through an article, when they are intrinsic to the arguments in that article. This should be in line with the practical limits for that field of research. Materials should be uploaded to a repository before publication of the article, though their release may be subject to a temporary embargo. The publication should indicate when, and the conditions under which data will be available for others to access.**

**Detailed actions**

**Publishers should**:

a. Actively encourage the development of standards and protocols for accessing data.

b. Encourage and support incentives for the citation of datasets.

c. Continue moving towards the development of journals devoted to data publication and support the development of wider best practice and common standards.

d. Support and engage with global, open and persistent researcher identification initiatives such as ORCID to ensure connectivity and accurate attribution of researchers and data.

### 5.7 Business funders of research

An easy flow and exchange of ideas, expertise and people between the public and private sectors is key in delivering value from research. This report describes how greater openness can enhance and deliver commercial value. Greater openness can also provide opportunities to develop commercial products and services utilising data, information and knowledge that are freely available. There are striking examples where opening up research and government data has provided opportunities for innovation and new businesses. This means that closed processes are sometimes necessary – either temporarily in order to attract further investment or permanently to protect trade secrets.

Greater openness is also desirable when commercial research data - such as data from clinical trials - has the potential for public impact. This includes negative data as well as the data that underlies positive published results. In particular, where there is a safety issue related to a particular technology (such as a medicine or medical device), a need to make information available in an expedited manner via the regulator or private funder should take priority over commercial considerations.

**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. Any release of data should be clearly signposted and effectively communicated.**

Data management is part of good 21st century business practice, whether that is customer data used to modify services or external data harvested to provide new services. The next generation of data scientists and analytic tools will serve these needs as well. Articulating industry demand for these skills and for IP arrangements will be vital to the success of government policies detailed below.

### 5.8 Government

Many governments regard their national science base as a crucial contributor to the future wellbeing of the nation. But is business *effectively* exploiting

the opportunities opened up by data-intensive science? Do businesses have the scientific capacity that is required to support this? And how should the free release of government data be balanced with charging for access to sophisticated information products?

These are issues that need to be addressed by many governments around the world in ways that are consistent with national priorities and processes. The US government has recently recognised this priority by investing millions of dollars annually in infrastructure and personnel.[231] The effective exploitation of data from the UK Government's Open Data initiative will depend on similar support for data scientists and analytic tools. Intelligent openness for research data requires a tiered infrastructure that is sensitive to the breadth of data types and demands for access.

**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.**

## Detailed actions

### UK-specific

a. The UK Government, through the Department of Business, Innovation and Skills, should revisit the work behind its roadmap for e-infrastructure. The urgent need for software tools and data scientists identified in this report need to be given due prominence alongside the physical infrastructure, skills and tools needed to exploit the data revolution. It should consider a major investment in these areas ensuring that the UK is able to exploit the data deluge.

b. BIS and the Technology Strategy Board should use its funding to enhance business take up of openly accessible scientific information and outputs. In the same way that Catapult Centres provide physical infrastructure for university-business collaboration, there needs to be enhanced digital infrastructure to enable data-based knowledge brokers. The Gateway to Research initiative provides some of this access, but does not go far enough.

c. Work undertaken by BIS in the wake of the Hargreaves Review of Intellectual Property[232] should continue. Recent attempts to remove the copyright barrier to data and text mining are welcome. Concerns from the research community voiced during Hargreaves' Review need to be revisited within the next five years.[233]

d. Government should ensure that the Freedom of Information regime does not reduce the potential of public-private partnerships to exploit scientific data by requiring FoI-based release of commercially-sensitive data.

e. Government should continue to explore whether an open data policy for some of its own agencies and institutions would be more economically productive than a policy of selling their data products. A two tier solution – providing some data for free and more detailed data or information products under a license – has been useful for meteorological and geological data. This has allowed for the continuation of the sophisticated, and commercially valuable, data interpretation done inside the Met Office and the British Geological Survey.

231 White House Press Release (29 March 2012). *Obama Administration unveils 'Big Data' initiative. Announces $200 million in new R&D investments*. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

232 Department for Business, Innovation and Skills (2011). *Innovation and research strategy for growth*. Available at: http://www.bis.gov.uk/assets/biscore/innovation/docs/i/11-1387-innovation-and-research-strategy-for-growth.pdf

233 Intellectual Property Office (2011). *Supporting Document U: Universities, Research and Access to IP*. Available at: http://www.ipo.gov.uk/ipreview-doc-u.pdf

### 5.9 Regulators of privacy, safety and security

Future governance practices need to reflect the speed that data analysis technologies are changing. Protecting privacy and security will only get harder as techniques for recombining data improve. Governance processes need to weigh the potential public benefit of research against the very latest technical risks.

For every decision made about whether to share personal data for research purposes, there must be scrutiny of the balance between public benefits that can flow from the sharing of research, the protection of individual privacy and the management of other risks, such as reputational risks. It is equally important that guidance for researchers is clear and consistent on this topic.

**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.**

#### Detailed actions

All regulatory and governance bodies, as well as data custodians, should adopt a risk-based approach to the promotion of open data policies and the protection of privacy interests, deploying the most appropriate governance mechanisms to achieve greater openness while protecting privacy and confidentiality.

**UK-specific**

a. The Ministry of Justice should promote clarity about the commitment to public interest research and advocate mechanisms of proportionate governance in negotiating the new EU Data Protection Regulation.

b. The new Health Research Authority, in consultation with other regulatory bodies, should produce guidance for researchers and ethics committees on the coverage of such legislation and its interpretation. This should aim to avoid the current variation in guidance.

**EU-Specific**

a. The EU Commission should be more explicit in the Data Protection Regulation about its commitment to research in the public interest and clearer about the relative roles of consent, anonymisation and authorisation in research governance. In doing so, the Commission should recognise that anonymisation cannot currently be achieved.

b. An assessment of the impact of the Database Right on the scientific community should be an explicit objective of the next review of the EU Database Directive.

Whilst security concerns are real, they should not be used as a blanket excuse to avoid opening up data. There has been very little dual use of new scientific findings in comparison with the documented public benefit of opening up research.

**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. Any guidelines should reflect that security can come from greater openness as well as from secrecy.**

# Glossary

| Term | Definition |
|---|---|
| ACTA | Anti-Counterfeiting Trade Agreement |
| ALSPAC | Avon Longitudinal Study of Parents and Children |
| ALLEA | ALL European Academies |
| Amazon Web Services Cloud | A set of services that together form a scalable and inexpensive computing platform. |
| Anonymisation | The process of removing identifying features of data from datasets in an effort to protect privacy and increase security. |
| ARDC | Australian Research Data Consortium |
| AUD | Australian Dollars |
| BADC | British Atmospheric Data Centre |
| BGI | British Genomics Institute |
| BIS | The Department for Business, Innovation and Skills |
| BLAST | Basic Local Alignment Search Tool |
| BOINC platforms | open source software platforms for volunteer computing and grid computing |
| BADC | British Atmospheric Data Centre |
| CCF | Climate Code Foundation |
| CEDA | Centre for Environmental Data Archival |
| CEH | Centre for Ecology and Hydrology |
| CellML | Cell Mark-up Language |
| CERN | European Organisation for Nuclear Research |
| ChEMBL | database of bioactive drug-like small molecules |
| Copyright | Confers a right on creators of original works to prevent others from copying the expression of ideas in a work. |
| Creative Commons | An organistaion enabling authors and creators to voluntarily share their work, by providing free copyright licences and tools. |
| CrossMark | A new initiative that can run an article against live journal databases, telling a reader whether the version they have is up-to-date and whether the article has been redacted. |
| crowdsourcing | The process of opening up science to public input in order to solve problems. |
| cryptography | The art of writing or solving codes. |
| cyberhygiene | Having clear rules for access and copying information and that they evolve as the nature of data evolves. |
| LOCKSS | Lots Of Copies Keep Stuff Safe, global archive that preserves content for libraries and scholars. |
| DaMaRO | Data Management Rollout project at the University of Oxford |
| DDI | Data Documentation Initiative |
| DMPOnline | Data Management Planning online tool |
| DNA | Deoxyribonucleic acid |
| DOI | Digital Object Identifier |
| DPA | Data Protection Act 1998 |
| dual-use | Something has a use other than it's intended one. |
| DVD data distribution | The recording of scientific data on DVDs, so that developing countries can access the data despite less access to advanced technology. |
| EBI | European Bioinformatics Institute |

| | |
|---|---|
| ECHR | European Commission on Human Rights |
| e-content | Content that is online. |
| ELIXIR network | A future pan-European research infrastructure for biological information |
| EMA | European Medicines Agency |
| EPSRC | Engineering and Physical Sciences Research Council |
| ESFRI | European Strategy Forum on Research Infrastructures |
| ex cathedra | With the full authority of office |
| FDA | Food & Drug Administration |
| FITIS | Flexible Image Transport System |
| FoIA | Freedom of Information Act |
| FTE | Full-time equivalent |
| GDP | Gross Domestic Product |
| GEO | Gene Expression Omnibus |
| GIC | Group Insurance Commission |
| Gigabyte | $10^9$ bytes of information |
| GitHub | A web-based hosting service for software development projects that use Git. Revision control system. |
| GM crops | Genetically Modified crops |
| GNU | Unix-like operating system that is free software |
| GPS | Global Positioning System |
| GPs | General Practitioners |
| GSK | GlaxoSmithKline |
| H5N1 | Avian Influenza |
| HD | Huntington's Disease |
| HE Institutions | Higher Education Institutions |
| HEFCE | Higher Education Funding Council for England |
| HGC | Human Genetics Commission |
| HIV/AIDS | Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome |
| HTML | Hypertext Mark-up Language |
| IBM | International Business Machines; a US company |
| ICSU | International Council of Science |
| informaticians | Someone who practices informatics |
| IP | Intellectual Property |
| IPNI | International Plant Name Index |
| IPRs | Intellectual Property Rights |
| Ipsos MORI | Research company in the UK. |
| ISIC | International Space Innovation Centre |

| | |
|---|---|
| ISP address | Internet Service Provider address |
| IVOA | International Virtual Observatory Alliance |
| JISC | Joint Information Systems Committee |
| LCM | UK Land Cover Map |
| LIGO | Laser Interferometer Gravitational-wave Observatory |
| LSE | London School of Economics |
| MATLAB | high-level technical computing language and interactive environment for algorithm development |
| Megabyte | $10^6$ bytes of information |
| Meta-analysis | Analysing metadata |
| MetOffice | Meteorological Office |
| MHRA | Medicines and Healthcare products Regulatory Agency |
| MIT | Massachusetts Institute of Technology |
| MRC | Medical Research Council |
| NASA | National Aeronautics and Space Administration |
| NCEO | National Centre for Earth Observation |
| NCSU | North Carolina State University |
| NERC | Natural Environment Research Council |
| NGO | Non-Governmental Organisation |
| NHS | National Health Service |
| NIH | National Institutes of Health |
| NSABB | National Science Advisory Board for Biosecurity |
| NSF | US National Science Foundation |
| OECD | Organisation for Economic Co-operation and Development |
| OGC | Open Geospatial Consortium |
| Ondex | A website that enables data from diverse biological sets to be linked, integrated and visualised through graph analysis techniques |
| Open Access journal | A journal that can be accessed without payment or restriction |
| Open Source | Of or relating to or being computer software for which the source code is freely available. |
| OPERA | Oscillation Project with Emulsion-Racking Apparatus |
| ORA | Oxford University Research Archive |
| ORCID | Open Researcher and contributor Identification |
| Patent | A legal contract that protects inventions such as products, processes or apparatus. To be patentable, an invention must be new, industrially applicable and involve an inventive step. |
| PDF | Portable Document Format |
| Per se | By or in itself or themselves; intrinsically |
| Petabyte | 1000 terabytes or $10^{15}$ bytes of information |
| PIPA | Protect Intellectual Property Act |
| Piwik | Open source website analytics software |
| PLoS | Public Library of Science |

| | |
|---|---|
| Public interest science | Scientific subjects or areas that are either by demand or potential impact are deemed to be of interest to the public |
| PURL | Persistent Uniform Resource Locator |
| RCUK | Research Councils UK |
| RDFs | Resource Description Frameworks |
| REF | Research Excellence Framework |
| Pre-print archive | A repository for journal articles, accessible before the journal has published them – often six months in advance. |
| RSS | Really Simple Syndication |
| Safe haven | Secure sites for databases containing sensitive personal data that can only be accessed by authorised researchers |
| SGC | Structural Genomics Consortium |
| Shibboleth system | Removes the need for content providers to maintain user names and passwords, and allows institutions to restrict access to information at the same time as securing it for remote access for approved users |
| SLS | Scottish Longitudinal Study |
| SMBL | Systems Biology Mark-up Language |
| SOPA | Stop Online Piracy Act |
| StackOverFlow | A language-independent collaboratively edited question and answer site for programmers |
| STFC | United Kingdom Science and Technology Facilities Council |
| SVSeo | Science Visualisation Service for Earth Observation |
| STFC | United Kingdom Science and Technology Facilities Council |
| SVSeo | Science Visualisation Service for Earth Observation |
| Text-mining | Using software to search databases of text |
| Terabyte | $10^{12}$ bytes of information |
| UCL | University College London |
| UKDA | United Kingdom Data Archive |
| UKOLN | United Kingdom Office for Library and Information Networking |
| UNFCCC | United Nations Framework Convention on Climate Change |
| URI | Uniform Resource Identifier |
| USNAS | United States National Academy of Science |
| USNRC | United States National Research Council |
| UTC | University Technology Centre |
| UUID | Universally Unique Identifier |
| VO | Virtual Observatory |
| WHO | World Health Organisation |
| WMS | Web Map Service |
| wwPDB | worldwide Protein Data Bank |
| XML | Extensible Markup Language |

# Appendix 1: Diverse databases

## CASE STUDIES USED IN BOX 2.3

### Discipline-wide openness – major international bioinformatics databases

The European Bioinformatics Institute (EBI) is at the forefront of the transition to data-heavy research in some areas of the life science, working with volumes of information that were barely comprehensible 60 years ago. Scientists around Europe deposit their biomolecular data into one of the EBI's data resources. These data are collected, curated, archived, and exchanged with partners in international consortia to maintain a shared global archive. The data are then made freely available to all through the internet. Despite the falling costs of data storage, data volumes are so large that data storage alone at EBI now requires an annual budget of almost £6 million. All of EBI's data resources are growing exponentially, especially the nucleotide sequences which are doubling every nine months.

The Wellcome Trust Sanger Institute, the biomedical research facility that provided the UK's contribution to the Human Genome Project, has about the same storage capacity as EBI and this has grown at a similar exponential rate. This is used to analyse and process the raw data before depositing the abstracted sequences at the EBI. Similar approaches for abstracting and storing data are used in almost all scientific disciplines.

These UK databases are part of an international network of coordinated biomedical data resources. For example the International Nucleotide Sequence Database Collaboration, consisting of the European Nucleotide Archive (Box 2.3), GenBank in the US and the DNA Database of Japan[234] coordinate the collection and distribution of all publically available DNA sequences. Over the last five years, a cross-Europe body have planned a new distributed data infrastructure for biological science - ELIXIR. ELIXIR will develop a sustainable, distributed yet coordinated infrastructure for biological information. The UK government recently committed £75 million to establish the central ELIXIR hub at EBI. Five other European countries have also already committed funds to establish national nodes.

### Processing huge data volumes for networked particle physics

This experiment maintains its own database and also illustrates how raw data often need to be progressively condensed into a usable form as derived data. The experiment produces about 25MB of data per event. There are 23 events per beam crossing, and 40 million beam crossings per second, which in total produced 23 petabytes per second of raw data; almost as much as all the data stored by the European Bioinformatics Institute in 2008. This raw data is trawled for the most interesting events, keeping data related to only a few hundred out of the 920 million events per second. Once compressed, this data requires 100MB of disk space per second - a few petabytes each year. Grid computing is used to reconstruct the events as researchers try to work out the kinds of particles that could produce each event. The experiment exemplifies the major, highly specialist effort often required to translate raw data into information and knowledge.

Data exchanged over the CERN Grid is proprietary to scientific groups working on the project. 3000 people work on another CERN project – the Compact Muon Solenoid (CMS). These researchers develop their own software and models in laboratories across the world. The large number of independent groups involved in the project ensures that the conclusions of any one group will be subject to expert critique by another. The CMS is working with CERN to make data publicly available: under the new proposals, data which is over two years old will be made publicly accessible via a virtual machine. This virtual simplified detector allows users to model collisions without the large infrastructure necessary to manipulate full datasets. The Virtual Observatory (below) in Astronomy performs a similar function for a large international community that shares a few telescopes. Data in particle physics is continually reprocessed and reanalysed in order to better understand it and correct it: after two years, the data has stabilised and whilst not definitive, it is in a much better shape than when it was collected. Old processed data are not usually kept, as they have been superseded.

234 More details can be found here: EMBL-EBI (2012). *EMBL Nucleotide Sequence Database*. Available at: http://www.ebi.ac.uk/embl/

## POLICY-RELATED DATA MANAGEMENT

### Epidemiology and the problems of data heterogeneity

There are cases where such large scale coordination presents severe problems. Epidemiologists studying infectious diseases rely on health data collected by national governments or agencies, often curated through the World Health Organisation. But datasets are often heterogeneous, with different information collected at irregular intervals and with poor data collection in developing or unstable regions. For access to some datasets, researchers still rely on special relationships with private companies or particular national statistics agencies.[235] The Vaccine Modelling Initiative has gone some way to creating an epidemiological repository for vaccine research, as well as digitising historical vaccine datasets.

### Improving standards and supporting regulation in nanotechnology

Nanotechnology is a large interdisciplinary field concerned with understanding phenomena and materials at atomic, molecular and macromolecular scales, where properties differ significantly from those at larger scales. There are significant concerns, particularly in Europe, over the inclusion of nanomaterials in everyday products. The drive to share nanomaterial data therefore has two motives, to increase the efficiency of an interdisciplinary field and to help regulators charged with licensing products that include nanomaterials. It has led to moves by researchers[236] to improve and standardise the way they describe nanomaterials.

## CASE STUDIES USED IN FIGURE 2.2

### The Avon Longitudinal Study of Parents and Children (ALSPAC)

This aims to investigate genetic and environmental factors that affect health and development. Researchers have been collecting large amounts of data from mothers and their children at 55 time points since 1991 in the form of biological samples, questionnaires, information from medical notes, and in some cases genome analysis. The nearly 40,000 variables, 55 time points and 94 data collection events of this study can be explored through a prototype online gateway developed by the MRC, the MRC Research Gateway[237]. Researchers from approved collaborations can view 'deep metadata' of variables of studies and export these to support data sharing requests. Researchers then liaise with a 'data buddy' who releases the required data according to the degree of risk of breaching participant anonymity. If there is a risk that study participants may be identified, data are made available via a two-stage process: first potentially identifying but unmatched data are provided to collaborators, the study team later matches these with the dataset. Data with a low risk of disclosure are more readily accessible and subject to a less stringent release process. Genotype data are only made available via data transfer agreements. The MRC Research Gateway is striving to enhance data sharing within defined limits to protect participant anonymity.

### Global Ocean Models at the UK National Oceanography Centre

Researchers at the National Oceanography Centre in Southampton[238] run high resolution global ocean models to study the physics of ocean circulation and the bio-geochemical consequences of changes in this circulation over timescales spanning multiple decades. Data on the ocean properties, sea-ice cover, ocean currents and biological tracers are recorded and a typical 50 year run produces between 10 and 50 terabytes of data. To analyse the data, researchers' cycle through the time series of output using software specifically developed in-house. Standard packages can be used to visualise the data although in-house packages are also developed for specific needs. The data are stored locally and at data centres for up to 10 years or until superseded and are made freely available to the academic community.

235 Samet J M (2009) *Data: To Share or Not to Share?* Epidemiology, 20, 172-174.
236 For instance, International Council for Science (2012). ICSU-CODATA *workshop on the description of nanomaterials*. Available at: http://www.codata.org/Nanomaterials/Index-agenda-Nanomaterial.html
237 Medical Research Council (2011) *MRC Data Support Service*. Available at: www.datagateway.mrc.ac.uk
238 National Oceanography Centre, University of Southampton (2010). Available at: http://www.noc.soton.ac.uk/

### The UK Land Cover Map at the Centre for Ecology and Hydrology

The UK Land Cover Map (LCM2007) has classified approximately 10 million land parcels into the UK Biodiversity Action plan Broad Habitats by combining satellite imagery and national cartography. It is the first land cover map to provide continuous vector coverage of 23 of the UK Broad Habitats derived from satellite data. To process and classify the 2 terabytes of data involved, researchers have developed novel techniques and automated production tools. The data are curated by the Natural Environment Research Council (NERC) Centre for Ecology and Hydrology (CEH) so it can be reused for further research. Metadata, technical descriptions, visualisation services and download of summary datasets are available through the CEH Information Gateway[239] The national product is available in a range of formats from 1 km summary to 25 m resolution for the UK for all 23 habitat types.

### Scientific Visualisation Service for the International Space Innovation Centre

The Science Visualisation Service for Earth Observation (SVSeo)[240], developed by CEDA as part of the development of the International Space Innovation Centre (ISIC), is a web-based application that allows users to visualise and reuse Earth Observation data and climate model simulations. Users can visually explore large and complex environmental datasets from observations and models, view, step through and zoom in to gridded datasets on a map view, overlay different parameters, export images as figures and create animations for viewing and manipulation on the ISIC videowall, on Google Earth or other similar software. Datasets from the National Centre for Earth Observation (NCEO) in the CEDA archives have been included in the visualisation service and provide satellite derived products relating to clouds, plankton, air-sea gas exchange and fire, and model output. The visualisation service will be updated as additional datasets are produced and provided to CEDA for long term archival. The service is also capable of including any remote data which are exposed via a Web Map Service (WMS) interface. CEDA data are made available for visualisation through the CEDA Open Geospatial Consortium (OGC) Web Services framework (COWS[241]. *(Interactive visualisation software developed by partners in STFC e-Science and the University of Reading can also be used at the ISIC facility to create animations on a virtual globe or multiple, synchronised virtual globes displayed on a large videowall.)*

### Laser Interferometer Gravitational-wave Observatory project

The Laser Interferometer Gravitational-wave Observatory (LIGO) project[242] is an 800-person international open collaboration, involving approximately 50 institutions. It aims to detect gravitational waves, tiny ripples in the structure of spacetime caused by astrophysical events like supernovae, neutron stars or black holes. They were first predicted by Albert Einstein in 1916 as part of his theory of general relativity but remain to be directly observed. The UK is involved in this collaboration via the UK-German GEO600 project[243], a 600m laser interferometer infrastructure built near Hannover.

The collaboration has generated in the order of 1 petabyte of data so far, a volume which is expected to increase to a rate of around 1 petabyte per year by 2015. These data are stored at the US LIGO sites, some or all of which is also maintained at various European sites. Despite the core dataset being relatively straightforward, it also includes important but complex auxiliary channels, such as seismic activity and environmental factors, and several layers of highly-reduced data products, mostly specific to custom software suites. Such data require careful curation. The management of the data and the processing software has so far been designed to support an ongoing research project. A long term data preservation plan has also recently been agreed, including an algorithm for data release. Data collected remain proprietary to the collaboration until its release is triggered by a significant event such as an announced detection of a gravitational wave, or a certain volume of spacetime being explored by the detector.

239 Centre for Ecology and Hydrology (2011). Information Gateway. Available at: www.gateway.ceh.ac.uk
240 International Space Innovation Centre (2011). Available at: http://isicvis.badc.rl.ac.uk/viewdata/
241 CEDA OGC Web Services Framework (2011). Available at: http://proj.badc.rl.ac.uk/cows
242 Laser Interferometer Gravitational-Wave Observatory (2012). Available at: http://www.ligo.caltech.edu/
243 GEO600 The German-British Gravitational Wave Detector (2012). Available at: http://www.geo600.org/

### Astronomy and the Virtual Observatory

In the field of astronomy, scientists have for some time already recognised the importance of greater openness in science. Astronomers from around the globe have initiated the Virtual Observatory (VO) project to allow scientists to discover, access, analyse and combine astronomical data archives and make use of novel software tools. The International Virtual Observatory Alliance (IVOA)[244] coordinates various national VO organisations and establishes technical and astronomical standards. The establishment of such standards is vital so that datasets and analysis tools from around the world are interoperable. Metadata are also standardised using the Flexible Image Transport System (FITS) standard and the more recent XML-based IVOA format, IVOTable. It is also an IVOA standard to register datasets in a registry, a sort of web-based Yellow Pages for astronomy databases. These are important to document the existence and location of datasets so that they can be easily found and accessed. IVOA itself collates a registry of registries. In Europe, the main VO organisations have come together to form Euro-VO[245]. Euro-VO is responsible for maintaining an operational VO in Europe by supporting the utilisation of its tools and services by the scientific community, ensuring the technology take up and compliance with international standards and assisting the build of the technical infrastructure. Deposition of data in data centres is common practice in astronomy, especially since it is a condition of access to large facilities. Access to data may be embargoed for up to a year to allow the scientists who carried out the research to have a first chance to analyse their data; data are however made publically available at the end of this period.

244 International Virtual Observatory Alliance (2012). Available at: http://www.ivoa.net/
245 The European Virtual Observatory EURO-VO (2011). Available at: http://www.euro-vo.org/pub/

# Appendix 2: Technical considerations for open data

## Dynamic data

Databases tend to be dynamic rather than static. On the face of it, scientific data should not be expected to change. In practice, a great deal of scientific data does evolve rapidly because of refinements in methods of acquisition or improvements in data treatment. Many curated databases, including several cited in this report, are constantly updated with new data, and up to 10% of existing data is overwritten each year.[246]

This poses problems for systems of linked databases, such as the web of data shown in Figure 2.3, where the many of the individual databases evolve as new data is added. Unless metadata in the web of data is updated too, then the data quickly becomes out of date. For example, the World Factbook[247] is probably the most widely used source of demographic information and although the information is frequently updated, the host institution does not expose the history of these updates.

## Indexing and searching for data

Google and Wikipedia are important tools for many researchers. Scientists increasingly publish on their own or their organisation's website and let search tools ensure dissemination to interested scientists.[248] Journal publication provides no more than the official stamp of quality on their work. Authors' credentials can be checked using an online bibliography,[249] rather than using the digital object identifiers and special-purpose archives that have evolved from traditional scholarship.

Data management may in the future benefit from similar systems. Large volumes of data will always require the special mechanisms of compilation, preservation and dissemination that characterise major databases, but these formal systems must be seen, not as the entirety of future systems for indexing data but as one amongst several mechanisms. As systems for searching metadata improve, the free-text indexing in today's search engines will also work for data. Some data resources are already highly accessible through generic internet search. For example a Google search on"Metabotropic glutamate receptors" yields one of the leading curated databases on the topic[250]. It contains descriptive text, substantial tabular data and links to external sources. In addition each section has the authority of named contributors and indicates how they should be cited.

## Servicing and managing the data lifecycle

Data first need to be appraised for whether they are to be retained, for how long,[251] how they need to be treated and the audience for which they are intended, whether a research group, users of an international database or non-specialists[252], whether they are to be protected as intellectual property and whether the cost of curation is proportional to their value. Experimental replication by others requires precise specification of the processes of initial data acquisition, manipulation and storage. For simulation output, specification of the exact computing environment may be necessary, with replication achieved by the use of a downloadable virtual machine (eg the Virtual Observatory – Case X in appendix 1). The UK Met Office, for example, preserves all measured weather data, but only a subset of the data generated by simulations. In areas such as genomics, the cost of sequencing genetic information is falling more quickly than the cost of storing it, suggesting that it may soon be cheaper to re-sequence samples as required than to store the data (see Figure A).

---

246 Buneman P, Khanna S, Tajima K and Tan W-C (2002). *Archiving Scientific Data*. Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, 29, 1, 2-42. Available at: http://repository.upenn.edu/cis_papers/116/

247 CIA (2012). *The World Factbook*. Available at: https://www.cia.gov/library/publications/the-world-factbook/

248 Eg the Wayback machine or Internet Archive (2012). Available at: http://archive.org

249 Computer scientists make substantial use of DBLP (2012). *The DBLP Computer Science Bibliography.* Available at: http://www.informatik.uni-trier.de/~ley/db/, which extracts citations from journals and conference proceedings.

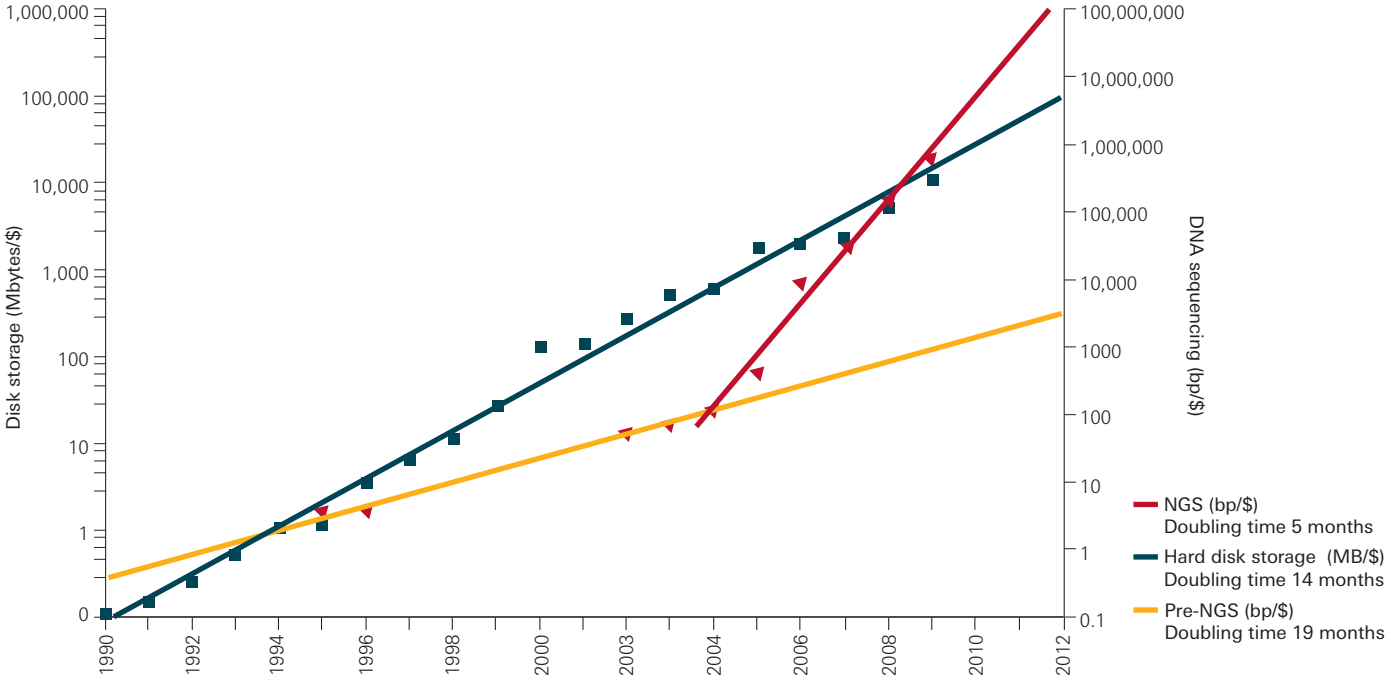
250 IUPHAR Database (2012). *Committee on Receptor Nomenclature and Drug Classification*. Available at: http://www.iuphar-db.org/

251 Ellis J (1993). (ed.). *Keeping Archives 2nd ed*. Autralian Society of Archivists: Melbourne. See further the Digital Curation Centre (2010). How to Appraise and Select Research Data For Curation. Available at: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data. Selection policies need to take account of the value of datasets over time. Some data are valuable now and will be expected to continue to be valuable in the future. The value of others will decay quickly over time, whilst others, for example longitudinal studies, will increase in value as time passes. On this point, see Borgman (2011). *The conundrum of sharing research data.* Journal of the American Society for Information Science and Technology.

252 See also the following points from the evidence submitted by Research Councils UK: "It is important that openness (data sharing) is pursued not as an end in itself, but to maximise the value of the data and the ultimate benefits to the public. This requires custodians of data and those who wish to have access to understand the data lifecycle, when in that lifecycle sharing best adds value, and the risks associated with inappropriate access (eg to confidential information)."

**Figure A The data sequenced per dollar in next generation sequencing has increased faster than the data that can be stored per dollar.[253]**

The blue squares describe the historic cost of disk prices in megabytes per US dollar. The long term trend (blue line, which is a straight line here because the plot is logarithmic) shows exponential growth in storage per dollar with a doubling time of roughly 1.5 years. The cost of DNA sequencing, expressed in base pairs per dollar (bp/$), is shown by the red triangles. It follows an exponential curve (yellow line) with a doubling time slightly slower than disk storage until 2004, when next generation sequencing (NGS) causes an inflection in the curve to a doubling time of less than six months (red line). These curves are not corrected for inflation or for the 'fully loaded' cost of sequencing and disk storage, which would include personnel costs, depreciation and overhead.




There is an urgent need for new tools to support the whole data cycle from data capture from an instrument or simulation through processes of selection, processing, curation, analysis and visualisation with the purpose of making it easy for a bench or field scientist who collects large or complex datasets to undertake these tasks. The cases in appendix 1 illustrate how different projects have varying data management needs throughout their lifetime.

Commercial Laboratory Information Systems exist[254], but they tend to be specific to a particular task and are costly. Research Council support for building generic tools for researchers is vital, and was specifically supported in the UK through the decade of support for eScience.[255] This programme was discontinued in 2009. As funding is dispersed to various agencies, a coordinating body is much needed. By contrast, in March 2012, US Government agencies announced $200 million of new funding specifically to improve the tools and techniques needed to makes discoveries from large volumes of digital data[256] as part of the US cyber-infrastructure programme.

253 Stein, Lincoln D (2010). *The case for cloud computing in genome informatics*. Genome Biology, 11, 207. Available at: http://genomebiology.com/2010/11/5/207

254 Labvantage (2012). LIMS (Laboratory Information Management). Available at: http://www.labvantage.com. Or Starlims (2012). Available at: http://www.starlims.com/

255 For details of current initiatives under the eScience umbrella see Research Councils UK (2012) e-Science. Available at: http://www.rcuk.ac.uk/research/xrcprogrammes/prevprogs/Pages/e-Science.aspx

256 White House Press Release (29 March 2012). *Obama Administration unveils 'Big Data' initiative. Announces $200 million in new R&D investments*. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

After selecting that data that is to be curated, it is calibrated, cleaned and gridded, and may be modelled or smoothed. Most usable data that is curated in a database is not raw data from an instrument, but processed data, in contrast to data which is born digital such as that from a computer simulation.

Active research soon accumulates the equivalent of millions of files, from which researchers may wish to extract all data with a particular property; for example deep-sea sediment samples that contain a particular kind of fauna. Automated tools are needed for this kind of query. When data is added to, re-calibrated or amended in some other way, the provenance trail of these changes needs to be recorded and personal identifiers are required to ensure that credit can be given to data originators. Such tools are vital to ensure the efficient use of data and to reduce costs, and are as relevant in Tiers 3 and 4 as they are in Tier 1.

Data are not often static digital objects. Measurement data, for instance, is regularly updated. And there can be but may have a rich relationships between data that need to be retained through these changes. It is these feature that have created problems urgently require solutions and tools to cope with them. Annotation must not be lost when data are moved from one place to another, *archiving* must efficiently preserve the history of a data collection that evolves over time - of particular important for longitudinal studies - and *provenance* needs to be recorded.

## Provenance

Tracking the provenance of data from its source[257,258] is vital for its assessment and for attribution to its originators. First, permanent data identifiers need to be assigned, giving each datum a unique, unalterable digital identification. Second, links to other relevant data sources need to be included to allow researchers to explore related datasets; and third, metadata need to be provided alongside the data to enable researchers to understand the linkage methods and to assess the quality of the data and its context. Linked semantic data (see section 2.1.4) are meant to provide some of this function, but there is still work to be done to provide a trustworthy system that preserves all the properties that are vital for accessing, assessing and reusing data.

Identification of scientific data can sometimes be done through a Digital Object Identifier (DOI). They are already a common way of referring to academic journal articles and can be displayed as linkable, permanent URLs without changing over time, even if the location and information about the data do. Over 40 million DOI names have already been assigned internationally. DataCite, established in 2009 and based at the British Library[259] develops DOIs for research datasets. There are similarly well established systems for identifying researchers that could be modified to identify producers of datasets. The ORCID (Open Researcher and contributor ID)[260] system is one that unambiguously establishes the identity of the data creator. There are also moves to create systems for reusing research data. The Data Documentation Initiative (DDI) is designed for the exchange of complex data and metadata between scientific groups. It was developed explicitly to track the processes through which data are reshaped and reanalysed in a data lifecycle.[261]

Capturing provenance has recently been recognised as an important challenge for informatics, but there is very little understanding of the full needs of researchers, let alone solutions that go beyond static identifiers or tracking data through a predictable lifecycle. Research data needs its own form of version-control, tracking changes in a way that is linked to the metadata description and that should move with the data.

## Citation

Citation plays an important role in modern science. It is a locator of published articles and tracks the provenance of information. It is important in evaluating and the contributions of individual scientists and influences their reputation and career progression. The modes of citation currently in

257 Buneman P, Khanna S and Tan W-C (2000). *Data Provenance: Some Basic Issues*. In Foundations of Software Technology and Theoretical Computer Science. Available at: http://db.cis.upenn.edu/DL/fsttcs.pdf

258 Simmhan Y L, Plale B and Gannon D (2005). *A Survey of Data Provenance Techniques*. Technical Report IUB-CS-TR618, Computer Science Department, Indiana University: Bloomington. 47405. Available at: ftp://ftp.cs.indiana.edu/pub/techreports/TR618.pdf

259 DataCite (2012). Available at: http://datacite.org/

260 Open Researcher & Contributor ID (2012). Available at: http://about.orcid.org/

261 Data Documentation Initiative (2009). *Technical Specification, Part I: Overview, Version 3.1*. Available at: http://www.ddialliance.org

general use have two major drawbacks. First they fail to recognise the contribution of novel collaborative processes or open sources. Second, although there is widespread recognition of the need for data citation through persistent identifiers, it is not yet clear how to put data citation on a par with article citation. Tools and standards for data citation exist but need to be improved particularly for contributions to evolving databases.

Recognition for collaborative ways of working is most developed among software programmers, and built around their open source practices. GitHub[262] allows members to collaborate in writing software in a way that retains provenance for all changes and allows members to see the number of members "watching" their published projects for updates. There is prestige associated with the most-watched projects.

### Standards and interoperability

Curation should be done to format standards that observe this report's criteria of accessibility, intelligibility, assessability and usability. Common structures allow reusers not only to manipulate data but also to integrate it with other datasets. This is the thrust behind the simple set of standards developed for the web of linked data (see Chapter 2.1.3). There are attempts to create global standards for the curation of scientific data. The International Council for Science (ICSU) hope to develop a World Data System[263] to provide long term provision of quality-assessed data and data services to the international science community. The World Bank Microdata Library (Box A) illustrates how quickly these global standards can spread.

A drive for broad standards should not, however, override the specific needs of disciplinary communities. The microarray community established 'MIAME' standards; the crystallographers created the CIF standard; and the Statistical Data and Metadata Exchange is designed to facilitate the sharing of official statistics typically generated by governments for the purpose of monitoring social, economic, demographic and environmental conditions. Each standard helps that community share the data accompanied by the kind of descriptive metadata that makes sense for research purposes.

As with the web of linked data, creating interdisciplinary standards for scientific data is made difficult by the distinctive vocabularies in a particular field.[264] The same terms can describe wholly different data properties, and different terms can describe the same properties. Integrating datasets in the future requires a leap forward in the systems that can create this interoperability.

### Box A The World Bank's Microdata Management Toolkit[265]

The World Bank's Microdata Library[266] holds over 700 national survey datasets, so that anyone anywhere can access the 1997 Moldovan Reproductive Health survey or Bangladesh's 2009 survey of citizens' experience of the legal system. But the Microdata Library has done more than collect survey results. They have implemented standards, including the DDI, for metadata and data formats, as well as providing financial support for implementing these in 60 developing nations. It is not just the World Bank and its data users that benefit from this effort; national statistics around the globe are now prepared and preserved to higher standards, making them easier to find, compare and reuse.

The World Bank Data Group has also created a tool to automate this standardisation of survey data. In collaboration with the International Household Survey Network[267], they have developed a 'Microdata Management Toolkit'. This open source app checks and formats data on behalf of the user. It also allows users to export data in various common formats for reuse in different contexts.

262 GitHub (2012). Available at: https://github.com/
263 ICSU World Data System (2012). Available at: http://www.icsu-wds.org/
264 Freitas A, Curry E, Gabriel Oliveira J, O'Riain S (). *Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends*. Internet Computing, IEEE, 16, 1, 24-33.
265 International Household Survey Network (2011). *Microdata Management Toolkit*. Available at: http://ihsn.org/home/index.php?q=tools/toolkit
266 The World Bank (2012). *Microdata Library*. Available at: http://microdata.worldbank.org
267 International Household Survey Network (2011). Available at: www.ihsn.org

## Sustainable data

Written records which are hundreds of years old are regularly unearthed and are still readable. Digital records only ten or fifteen years' old can become unreadable. The BBC Domesday laserdiscs, released in 1986, 900 years after the original Domesday Book, were intended to provide a durable record of the country in that year, just as the original had. However, the hardware necessary to read the discs quickly became so rare that there were serious concerns that the ability to read their contents would be lost.[268] Tools such as the Internet Archive's Wayback machine offer access to previous webpages, but curating digital objects requires far more than this kind of automated storage. Active curation is vital, and costly, involving cleaning the data, backing up, ensuring that the data are updated to keep pace with format or technological changes, reprocessing to maintain usability, and development and maintenance of an accessible, well signposted guide to the data and the metadata that makes it usable.

Financial sustainability is guaranteed for most major databases that are funded through international agreements and databases funded by Research Councils. But such arrangements do not shore up Tier 3 and 4 collections, and it is far from clear who takes responsibility for databases as they become increasingly valuable and need to move between tiers.

Energy use is a further issue. Data centres currently consume 1% of the world's electricity generating capacity.[269] If the nine month doubling rate suggested by EBI in Box 2.3 is a universal phenomenon, and assuming that energy use increases in proportion to data produced, then data centres will need more than the today's total electricity capacity within the decade. Recognition of this problem has led to searches for more power-efficient ways of running database systems:[270] Google claim to use less than 1% of global data centre electricity in 2010 because of highly optimized servers.[271]

Simply replicating data in order to store it can no longer be the norm - as is it for some of the most promising initiatives like LOCKSS.There must be more emphasis on distributed data, accessible in the cloud, rather than replication of data for local storage and analysis. Although cloud services currently account for less than 2% of IT spending, it is estimated that by 2015 nearly 20% of online information will be "touched" by cloud computing service providers.[272] As data moves to cloud repositories, signposting rather than replication should become the norm. Ease of digital copying has improved resilience to the loss of data, but multiple copies are not a sustainable solution for long term data storage.

Ubiquitous replication cannot be part of a sustainable data storage solution, nor will it be necessary for future large scale data analysis and modelling. There are now algorithms that can operate simultaneously on data on multiple servers. Hadoop MapReduce is a software framework that creates ways of rapidly processing vast amounts of data in parallel on large clusters of servers.[273] The Map Reduce algorithm can run operations on servers where the data is located. Rather than copy the data over the network in order to run an analysis, the programme is exported to the machines and the results are merged.

---

268 The content has now been successfully transferred into a more durably reusable form and is available from: BBC (2012). Doomsday reloaded. Available at: www.bbc.co.uk/history/domesday

269 Koomey J (2010). *Analytics Press. Growth in data center electricity use 2005 to 2010*. Available at: http://www.koomey.com/post/8323374335

270 Xu Z, Tu Y-C. and Wang X. (2009). *Exploring Power-Performance Tradeoffs in Database Systems*. Available at: http://web.eecs.utk.edu/~xwang/papers/icde10.pdf

271 Google Data Centers (2012). *Data Center Efficiency*. Available at: http://www.google.com/about/datacenters/inside/efficiency/servers.html

272 IDC (2011). *Digital Universe study: Extracting Value from Chaos*. Available at: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

273 Hadoop (2012) *Hadoop MapReduce*. Available at: http://hadoop.apache.org/mapreduce/

# Appendix 3: Examples of costs of digital repositories

Chapter 4 distinguishes between four tiers of digital repositories. Tier 1 comprises the major international data initiatives that have well defined protocols for the selection and incorporation of new data and access to them. Tier 2 includes the data centres and resources managed by national bodies such as UK Research Councils or prominent research funders such as the Wellcome Trust. Tier 3 refers to curation at the level of individual universities and research institutes, or groupings of them. Tier 4 is that of the individual researcher or research group that collates and stores its own data, often making it available via a website to collaborators or for public access.

This appendix presents costings and capabilities for a representative sample of Tier 1, Tier 2 and Tier 3 repositories, gathered by a standardised survey instrument. The data presented below were gathered in January-February 2012, and are accurate as of this time. As the figures particularly for the universities repositories indicate, this is a fast moving field. Data on some additional repositories was provided, but has not been reproduced here for reasons of space.[274]

## International and Large National Repositories (Tier 1 and 2)

### 1. Worldwide Protein Data Bank (wwPDB)

The Worldwide Protein Data Bank (wwPDB) archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. It was founded in 1971, and is managed by the Worldwide PDB organisation (wwpdb.org). As of January 2012, it held 78477 structures. 8120 were added in 2011, at a rate of 677 per month. In 2011, an average of 31.6 million data files were downloaded per month. The total storage requirement for the repository was 135GB for the archive.

The total cost for the project is approximately $11-12 million per year (total costs, including overhead), spread out over the four member sites. It employs 69 FTE staff. wwPDB estimate that $6-7 million is for "data in" expenses relating to the deposition and curation of data.

| wwPDB – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | Yes |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Yes |
| 'Front end' - web-access to pages? | Yes |
| Registration and access control? | No |
| Input quality control: consistency with format standards, adequate technical standards? | Yes |
| Input quantity control: ensure community coverage? | Yes |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | Yes |
| Provide accession number to log deposition? | Yes |
| Alert registrants to new additions? | Yes |
| Provide means by which the data can be cited and credited to originators? | Yes |
| Host or link to relevant analysis tools (eg visualisation, statistics)? | Yes |
| Measure and document impact: downloads, data citations? | Yes |

274 We are grateful for the returns from PANGAEA, the Tier 1 georeferenced earth system data repository, and Tier 3 repositories at the University of St Andrews, University of Edinburgh and University of Portsmouth.

## 2. UK Data Archive

The UK Data Archive, founded 1967, is curator of the largest collection of digital data in the social sciences in the United Kingdom. It contains several thousand datasets relating to society, both historical and contemporary. The UK Data Archive provides services to the ESRC and JISC: including the Economic and Social Data Service, the Secure Data Service, the Census Registration Service, the Census Portal. The Archive maintains the History Data Service (unfunded) and undertakes a variety of research and development projects in all areas of the digital life cycle. UKDA is funded mainly by Economic and Social Research Council, University of Essex and JISC, and is hosted at University of Essex.

The main storage 'repository' holds multiple versions of approx 1.26 million files (ie digital objects), other 'repositories' hold a little under than 1 files (in a primary version.) UKDA tends to work on the basis of core data collections, of which there are currently 6,400. Of the 6,400 data collections, there were 53,432 direct downloads in 2011 (approx 4,500 per month). This does not include downloads of freely-available material which are estimated to be over 1 million.

This also does not include online tabulations through Nesstar, nor images browsed through websites hosted at the UK Data Archive (eg, www.histpop.org).

On average around 2,600 (new or revised) files are uploaded to the repository monthly. (This includes file packages, so the absolute number of files is higher.) The baseline size of the main storage repository is <1Tb, though with multiple versions and files outside this system, a total capacity of c.10Tb is required.

The UKDA currently (26/1/2012) employs 64.5
The UKDA currently (26/1/2012) employs 64.5
people. The physical storage systems and related security infrastructure is staffed by 2.5 FTE. The total expenditure of the UK Data Archive (2010-11) was approx £3.43 million. This includes additional infrastructural costs eg lighting, heat, estates etc. Total staff costs (2010-11) across the whole organisation: £2.43 million. Total non-staff costs (2010-11) across the whole organisation: £360,000, but these can fluctuate by more than 100% across given years. Non-staff costs in 2009-10 were approx £580,000, but will be much higher in 2011-12, ie almost £3 million due to additional investment.

| UKDA – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | Yes |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Yes |
| 'Front end' - web-access to pages? | Yes |
| Registration and access control? | Yes |
| Input quality control: consistency with format standards, adequate technical standards? | Yes |
| Input quantity control: ensure community coverage? | Yes |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | Yes, including metadata creation |
| Provide accession number to log deposition? | Yes |
| Alert registrants to new additions? | Optional |
| Provide means by which the data can be cited and credited to originators? | Yes |
| Host or link to relevant analysis tools (eg visualisation, statistics)? | Yes – not all data |
| Measure and document impact: downloads, data citations? | Yes – not all data |
| Other: UKDA also provides a range of other services including Content creation, Content hosting, Content hosting (secure), Content licensing, Content selection, Data curation, Data preservation, Data curation (secure), Licence negotiation, Documentation creation, Resource discovery, Content Development, Ingest (QA/Validation), Access Control (liaison with data owners), Consultancy, Creating & maintaining expertise, Developing advice & guidance (eg on data management), Requirements expertise, Solutions expertise, Training, Thesaurus/controlled vocabulary development, Horizon scanning, Trend analysis, General helpdesk support, Online help (FAQ, help manuals), Specialist helpdesk support, Event organisation & management, Funding engagement, Funding application, Market research, Promotion and PR, Impact promotion, Vendor engagement, Project & programme management. | |

### 3. arXiv.org

arXiv.org is internationally acknowledged as a pioneering and successful digital archive and open-access distribution service for research articles. The e-print repository has transformed the scholarly communication infrastructure of multiple fields of physics and plays an increasingly prominent role in a unified set of global resources for physics, mathematics, computer science, and related disciplines. It is very firmly embedded in the research workflows of these subject domains and has changed the way in which material is shared, making science more democratic and allowing for the rapid dissemination of scientific findings. It has been running since 1991, and is hosted by Cornel University Library, and is funded by Cornell University Library and contributing institutions.

As of January 2012, it held over 750,000 articles. Around 7,300 are added per month. The size of the repository is currently 263GB. arXiv.org employs just over six people. Its projected running costs for 2012 (including indirect costs) are in the region of $810,000 per year, of which roughly $670,000 are staff costs. Storage and computing infrastructure accounts for around $45,000 per year.[275]

| arXiv.org – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | Yes |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Yes |
| 'Front end' - web-access to pages? | Yes |
| Registration and access control? | Allows user registration, but all papers are open access. |
| Input quality control: consistency with format standards, adequate technical standards? | Yes, please see the policies at http://arxiv.org/help |
| Input quantity control: ensure community coverage? | See http://arxiv.org/help/moderation |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | We rely on metadata provided during submission and are in the process of considering ORCID or other similar initiatives for author name disambiguation |
| Provide accession number to log deposition? | Yes |
| Alert registrants to new additions? | Yes |
| Provide means by which the data can be cited and credited to originators? | Yes (for arXiv documents – see http://arxiv.org/help/faq/references |
| Host or link to relevant analysis tools (eg visualisation, statistics)? | We have some R&D natured tools such as http://arxiv.culturomics.org/ |
| Measure and document impact: downloads, data citations? | none |
| Other: Provides support for ancillary files: http://arxiv.org/help/ancillary_files. Support for datasets as a R&D project, not a streamlined operation: http://arxiv.org/help/datasets. | |

275 http://arxiv.org/help/support/2012_budget and https://confluence.cornell.edu/display/culpublic/arXiv+Sustainability+Initiative. There is also a 5-year budget projection included in the "membership program" document on the sustainability website.

## 4. Dryad

Dryad (datadryad.org) is a repository of data underlying peer reviewed articles in the basic and applied biosciences. Dryad closely coordinates with journals to integrate article and data submission. The repository is community driven, governed and sustained by a consortium of scientific societies, publishers, and other stakeholder organisations. Dryad currently hosts data from over 100 journals, from many different publishers, institutions, and countries of origin. It was founded in 2008.

As of 24 January 2012, Dryad contained 1280 data packages and 3095 data files, associated with articles in 108 journals. It received 7518 downloads per month in December 2011, and 79 new data packages in December, 2011, with approximately 2.3 files per data package. Its current size is 0.05 TB.

Dryad has 4-6 FTE, with 50% devoted to operational core and 50% to R&D. Its total budget is around $350,000 per year, with staff costs of approximately $300,000, and $5,000-$10,000, of infrastructure costs including subscription services (eg DataCite, LOCKSS, etc.). It has received R&D funding from NSF and IMLS in the US, and JISC in the UK. Dryad's sustainability plan and business model ensure that long term, revenues from payments for the submission of new data deposits cover the repository's operating costs (including curation, storage, and software maintenance). The primary production server is maintained by the North Carolina State University Digital Library Program. The Dryad is currently applying to the State of North Carolina and the US IRS to be recognised as an independent not-for-profit organisation.

| Dryad – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | Usage is primarily through the centrally managed web platform at NCSU and its mirrors. The Dryad is responsible for provision, maintenance and development of this service. Since Dryad is built using open-source software, in large part DSpace, it can also be locally deployed for institutional use. |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Both multiple data machine and content formats and multiple versions are supported. Dryad does not generally host the articles themselves, but rather the datafiles associated with them. |
| 'Front end' - web-access to pages? | Yes, see http://datadryad.org |
| Registration and access control? | Yes, but not required for viewing/download, only for submission. Data and metadata can be embargoed from public access until article acceptance, publication, or beyond, depending on journal policy. |
| Input quality control: consistency with format standards, adequate technical standards? | • Quality control of bibliographic and subject metadata, including author name control<br>• Validation of the integrity of uploaded files, including screening for copyrighted and sensitive content<br>• Migrating files to new or more preservation-robust formats<br>• Providing user help.<br><br>The formatting of file contents varies with discipline and is controlled by journal policy, not by Dryad. In fields with mature data standards, journals frequently specify that users use a specialised repository. Dryad is designed to provide a home for the "long tail" of data, where such formats and repositories do not (yet) exist. At the same time, Dryad is developing means to coordinate the submission process with specialised repositories in order to ensure each data file is appropriately managed. |
| Input quantity control: ensure community coverage? | Dryad is interdisciplinary and spans multiple scientific communities; annotation functions are under discussion. |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | The repository controls the quality and completeness of bibliographic metadata (title, authors, DOI, etc), including subject keywords to enable search. Provenance and other context provided is always provided at least partially by the associated article. Authors may supplement this upon deposit (eg with a ReadMe file) or include such information within a metadata-rich data file (eg XML) |
| Provide accession number to log deposition? | Yes, DataCite DOIs. |
| Alert registrants to new additions? | Yes, eg through an RSS feed |

| | |
|---|---|
| Provide means by which the data can be cited and credited to originators? | Yes, Dryad is frequently noted as an exemplar of data citation policy best practice. http://datadryad.org/using#howCite |
| Host or link to relevant analysis tools (eg visualisation, statistics)? | No. |
| Measure and document impact: downloads, data citations? | Views and downloads are reported on a per file and per data package basis (eg see http://datadryad.org/depositing#viewStats). Tracking data citations is a long-range objective, but not currently feasible technically. |
| Other: Dryad is governed by a diverse set of stakeholder organisations. The Dryad is itself a service to its membership in providing a forum for the discussion of data policies and the promotion of best practice in data archiving.<br><br>Dryad has an open access business model in which curation and preservation costs are paid upfront to ensure that the data can be provided at no cost to those who wish to use it. Nearly all content in the repository is made available for reuse through a Creative Commons Zero waiver, and so can be built upon both by academic researchers and third party value-added services (eg more specialised data repositories that provide additional curation). Dryad also enables partner journals to integrate manuscript and data submission through automated exchange of metadata emails. This ensures that data records are prepopulated with bibliographic information in order to reduce the submission burden on authors, and partner journals are notified of all completed submissions, including DOIs. Partner journals may allow or disallow authors to set a one year embargo on access to a datafile, and editors may specify custom embargo lengths. Partner journals may offer editors and peer reviewers anonymous and secure access to data from manuscripts prior to their acceptance. | |

### Institutional Repositories (Tier 3)

Most university repositories in the UK have small amounts of staff time. The Repositories Support Project survey in 2011 received responses from 75 UK universities. It found that the average university repository employed a total 1.36 FTE – combined into Managerial, Administrative and Technical roles. 40% of these repositories accept research data. In the vast majority of cases (86%), the library has lead responsibility for the repository.[276]

### 5. ePrints Soton

ePrints Soton, founded in 2003, is the institutional repository for the University of Southampton. It holds publications including journal articles, books and chapters, reports and working papers, higher theses, and some art and design items. It is looking to expand its holdings of datasets.

It currently has metadata on 65653 items. The majority of these lead to an access request facility or point to open access material held elsewhere. It holds 8830 open access items. There are 46,758 downloads per month, and an average of 826 new uploads every month. The total size of the repository is 0.25TB.
It has a staff of 3.2 FTE (1FTE technical, 0.9 senior editor, 1.2 editors, 0.1 senior manager). Total costs of the repository are of £116, 318, comprised of staff costs of £111,318, and infrastructure costs of £5,000. (These figures do not include a separate repository for electronics and computer science, which will be merged into the main repository later in 2012.) It is funded and hosted by the University of Southampton, and uses the ePrints server, which was developed by the University of Southampton School of Electronics and Computer Science.

276 A summary of the results of this Respositories Support Project survey is available at http://www.rsp.ac.uk/pmwiki/index.php?n=Institutions.Summary. A more detailed breakdown by institution is available at http://www.rsp.ac.uk/pmwiki/index.php?n=Institutions.HomePage

| ePrints Soton – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | Yes |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Yes |
| 'Front end' - web-access to pages? | Yes |
| Registration and access control? | Yes |
| Input quality control: consistency with format standards, adequate technical standards? | Yes, although stronger on metadata. Starting to do more on recommended storage formats for objects for preservation purposes but more to do in this complex area |
| Input quantity control: ensure community coverage? | Yes |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | Yes, a new data project means the repository will be working more on provenance and contextual information for data. Up to now mostly publications rather than data. |
| Provide accession number to log deposition? | Yes |
| Alert registrants to new additions? | Yes- users can set up alerts |
| Provide means by which the data can be cited and credited to originators? | Yes |
| Host or link to relevant analysis tools (eg visualisation, statistics)? | Stats visualisation |
| Measure and document impact: downloads, data citations? | Yes, downloads, harvest ISI citation counts |
| Other: Integration with other systems – eg user/project profile pages, reporting for internal and external stakeholders, import/export in various formats including open data RDF format. | |

### 6. DSpace@MIT

DSpace@MIT is MIT's institutional repository built to save, share, and search MIT's digital research materials including an increasing number of conference papers, images, peer reviewed scholarly articles, preprints, technical reports, theses, working papers, and more. It was founded in 2002.

As of December 2011 DSpace@MIT held 53,365 total items, comprising 661,530 bitstreams. The scope of its holdings of research data is unknown, as whilst submitters have the ability to designate new items as being of a 'research dataset' content type, this information is not required.[277] It receives around one million browser-based file download per month, and an additional 1.3 million crawler-based file downloads per month. It receives around 700 uploads of new items per month. The total size of the repository is currently 1.1TB. Growth is anticipated at ~250GB/yr with current service scope.

The repository has 1.25 FTE dedicated to overall program administration technical support.[278] Additional capacity of 1.5 FTE supports the identification, acquisition, ingest, and curation of MIT's database of Open Access Faculty Articles http://libraries.mit.edu/sites/scholarly/mit-open-access/open-access-at-mit/mit-open-access-policy/ . While there are additional staff costs associated with identifying and managing the collections which are curated by the MIT Libraries and disseminated via the DSpace platform, e.g., theses, technical report series, working papers, etc., these costs are independent of DSpace@MIT and are borne in other Libraries' units independent of the service platform. The total cost of the repository itself is approximately $260,000 per year, of which around $76,500 are infrastructure costs, and around $183,500 direct or indirect staff costs.

277 They report that, There are 14 items in our repository with this designation but we know that there are dozens more without it.
278 0.3 FTE Development; .25 FTE SysAdm; 0.6 FTE Program Manager.; and 0.1-Operations.

| DSpace@MIT – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | DSpace@MIT is run as a single repository instance for all contributing communities at MIT. Provision, maintenance and development are done in house for this library-run service. |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Multiple formats are supported but are not automatically generated by the system upon ingest. Versioning is supported through creation of multiple items with cross-reference links and descriptive text. |
| 'Front end' - web-access to pages? | Yes |
| Registration and access control? | Yes |
| Input quality control: consistency with format standards, adequate technical standards? | Support from within the Libraries varies here depending upon the source community and target collection. The MIT Open Access Articles collection is heavily curated, as are other collections mediated by the Libraries. However, the DSpace@MIT service is open to the faculty and research community at large and aside from specific collections is largely unmediated – i.e., there is no specific review of pre-ingested content to determine the quality and completeness of entry. |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | The ability to input this metadata is supported within the system. They provide a best practices guide that aids submitters with respect to describing research datasets. The guide includes recommendations for describing the hardware, software and conditions that created the dataset, file format descriptions, and requirements for reuse of the data. |
| Provide accession number to log deposition? | Internally, the system creates an identifier for the submitted items that are directly referenceable to back-end database queries. Additionally, each item receives a handle URI (similar to a DOI) that is a permanent, persistent and citable URI. It does not yet support DataCite or other file-level identifiers (DSpace items can contain multiple files). |
| Alert registrants to new additions? | Yes. Users can set up e-mail notification of new content or via RSS. |
| Provide means by which the data can be cited and credited to originators? | Yes. Permanent, persistent handle URI for citation at the item level. |
| Host or link to relevant analysis tools (eg visualization, statistics)? | Supported if these links to relevant tools are added by the submitter. It does not have embedded 'dissemination' services that would produce such visualizations, analytics or derivatives on the fly from submitted bitstreams. |
| Measure and document impact: downloads, data citations? | The repository captures internal usage statistics but these are not publically displayed or redistributed to authors/creators/submitters. They do not attempt to track subsequent citation of their content. |
| Measure and document impact: downloads, data citations? | Yes, downloads, harvest ISI citation counts |
| Other: Most of the comments above have not directly referenced research data specifically. As an institutional repository, DSpace@MIT serves as the single repository for the breadth of research and teaching output of the Institute. As such, DSpace was designed to support submission of all formats, but without description, dissemination and search facilities that were specialized for various format types. Moreover, DSpace@MIT has historically been modelled as an unmediated service open to the faculty and research community at MIT.<br><br>The data model and metadata schema allows for the notation of related items, either held within the repository or externally. This allows for linking a locally-held dataset to an externally published article or to denote relationships among items. Also, DSpace@MIT supports the application of a Creative Commons license for submitted research data. | |

### 7. Oxford University Research Archive and DataBank

The Oxford University Research Archive (ORA) and DataBank services are being developed as part of the Bodleian Libraries' digital collections provision for Oxford. ORA is a publications repository, which holds a mixture of 'text-like' items. The data repository, DataBank, is well developed and is being developed further as part of the JISC-funded Damaro project. It will form one service within a broader research data infrastructure for Oxford. The Bodleian Libraries are also developing a research data catalogue (DataFinder) to record metadata about Oxford research data for discovery of datasets. ORA was founded in 2007, DataBank in 2008: both are still in development.

ORA currently hold 14,500 items, and DataBank 12 datasets. There are 1100 downloads from ORA per month; figures for DataBank are not available. ORA has around 100 uploads per month, excluding bulk ingests. DataBank currently has no deposit interface (one is in development), and requires assisted deposit. The service is not yet broadly advertised.

ORA has a staff of 2.5 FTE (0.5 manager; 1.0 developer; 1.0 assistant). Staffing that will be required for DataBank is not yet clear, but these staff will overlap with ORA. Total running costs not available. The service is hosted by the Bodleian Libraries. Funding for DataBank is under discussion within the University. ORA use Fedora, whilst DataBank uses Oxford DAMS (Digital Asset Management System).

| The Oxford University Research Archive (ORA) and DataBank – Services Provided | |
|---|---|
| Platform provision, maintenance and development? | Yes. |
| Multiple formats or versions (eg PDF, html, postscript, latex; multiple revisions of datasets)? | Format agnostic. Advise open formats if possible. All datasets in DataBank should be 'publishable' and can therefore be assigned a DOI. Updated versions can be accepted (DOI can indicate version number). |
| 'Front end' - web-access to pages? | Yes |
| Registration and access control? | Open access if permitted. Embargo facility if not. |
| Input quality control: consistency with format standards, adequate technical standards? | On the repository side, yes. |
| Add metadata and references to describe authorship, provenance and experimental or simulation context? | Working towards mandating DataCite kernal for data but may mandate additional fields (eg rights) [To be discussed]. |
| Provide accession number to log deposition? | Every item assigned a UUID as well as internal system PID |
| Alert registrants to new additions? | ORA: [Feature on home page]; RSS feed; Twitter |
| Provide means by which the data can be cited and credited to originators? | DOI for datasets; UUID for every item in both repositories; PURL resolver currently being deployed; DataFinder will provide a record including location for Oxford data even if not stored at Oxford. |
| Host or link to relevant analysis tools (eg visualisation, statistics)? | ORA: Statistics analytics (Piwik) |
| Measure and document impact: downloads, data citations? | ORA: Record accesses and downloads |
| Other services (please add additional rows as appropriate) | DOI assignment for datasets (DataCite) |
| Other: DataBank is not yet fully functioning (deposit and search features under development and also user interface design). The handful of datasets in the repository can be freely accessed by using the accurate URL. The Damaro project will see development of DataFinder. Policies and sustainability and training will be also be developed as part of Damaro. A colleague is working on the Oxford DMPOnline project (data management planning tool) which runs parallel to Damaro. We are expecting the basic service to be launched during 2013. ORA is small as yet and still in development. We see increasing numbers of doctoral theses (institutional mandate). We are currently starting promotion of easy deposit into ORA using Symplectic. We are aiming to run more bulk uploads where possible. | |

# Appendix 4: Acknowledgements, evidence, workshops and consultation

The Royal Society gratefully acknowledges the financial support received from The David and Lucile Packard Foundation USA, the Kohn Foundation and Sir Tom McKillop FRS.

## Evidence Submissions

### Email

- Prof Sheila M Bird, Royal Statistical Society (RSS)
- Prof Mark Blamire, Professor of Device Materials, University of Cambridge
- Jonathan Brüün, Director of Communications & Business Development, British Pharmacological Society
- David Carr, Policy Adviser, Wellcome Trust
- Dr Lee-Ann Coleman, Head of Science, Technology and Medicine, The British Library
- Stephanie Dyke, Policy Adviser, Sanger Institute
- Joshua Gans, Skoll Chair of Technical Innovation and Entrepreneurship and Professor of Strategic Management, University of Toronto
- William Hardie, Consultations Officer, The Royal Society of Edinburgh
- Prof Yuecel Kanoplat, President of Turkish Academy, Turkish Academy of Sciences
- Prof Michael J Kelly FRS, Prince Philip Professor of Technology, University of Cambridge
- Alan Palmer, Senior Policy Adviser, the British Academy
- Dr Anjana Patel, Independent
- Dr Rachel Quinn, Policy Adviser, Academy of Medical Sciences
- Dr Leonor Sierra, International Science Policy Manager, Sense About Science
- Helen Wallace, Director, GeneWatch UK
- Office of Vice Provost (Research), UCL

### Online

- Dr Helen Anthony, Programme Manager, National Physical Laboratory
- Mr Nicholas Barnes, Director, Climate Code Foundation
- Prof Sheila M Bird, Chair of Royal Statistical Society's working party on Data Capture - for the Public Good, and formerly a Vice-President Royal Statistical Society
- Dr Chas Bountra, SGC, Sage, Said Business School, School of Medicine - UCSF
- Prof Ian Boyd, Director
- Asa Calow, Director, Madlab (http://madlab.org.uk)
- Sir Iain Chalmers, Coordinator, James Lind Initiative
- David De Roure, ESRC National Strategic Director of e-Social Science
- Emmanuel
- Dr Sameh Garas, Senior Supervisor at Accredo health
- Prof Erol Gelenbe, UKCRC Executive Committee
- Prof Carole Goble
- Mr Bernard Godding
- Ms Ann Grand
- Dr Ivo Grigorov, Project Officer
- Dr Trish Groves, Deputy Editor, BMJ and Editor-in-chief, BMJ Open
- Professor Stevan Harnad
- Dr Tony Hirst, Lecturer
- Prof Tessa Holyoake, University of Glasgow
- Dr Ralph G. Jonasson, Independent
- Mr Andrew Lewis, Simul Systems Ltd
- Dr Philip Lord, Lecturer
- Mr Edgar R. McCarvill
- Miss Jenny Molloy, Coordinator, Working Group on Open Data in Science, Open Knowledge Foundation
- Mr Peter Mulderry
- Dr Cameron Neylon
- Mr J.D. Pawson
- Dr Pawel Sobkowicz
- Chloe Somers, Policy Manager for Research, Research Councils UK
- Dr Elizabeth Wager, Chair, Committee on Publication Ethics (COPE)
- Emeritus Prof A.C. Wardlow

- Steve Wood, Head of Policy Delivery, Information Commissioner's Office
- Lady Kennet (Elizabeth Young)

### Evidence Gathering Meetings Evidence session 1, 3 May 2011

- Simon Bell, Head of Strategic Partnerships & Licensing, British Library
- Kevin Fraser, Head of EU and International Data Protection, Ministry of Justice
- Dr Audrey McCulloch, Executive Director for the UK, The Association of Learned and Professional Society Publishers
- Simon Tanner, Director, King's Digital Consultancy Services, King's College London
- Dr Max Wilkinson, Datasets programme, British Library

### Evidence session 2A &2B, 7 June 2011

- Professor Ross Anderson FRS FREng, Professor of Security Engineering, University of Cambridge
- Debi Ashenden, Senior Lecturer, Dept of Informatics & Systems Engineering, Cranfield University
- Andy Clark, Director, Primary Key Associates Limited
- Professor Douwe Korff, Professor of International Law, London Metropolitan University.
- Toby Stevens, Director, Enterprise Privacy Group

### Southbank Centre public meeting, 9 June 2011

Speakers included

- David Dobbs, freelance science writer
- William Dutton, Oxford Internet Institute, University of Oxford
- Stephen Emmott, Head of Computational Science, Microsoft Research
- Timo Hannay, Managing Director, Digital Science
- Cameron Neylon, Senior Scientist, Science and Technology Facilities Council
- Sir Paul Nurse, President, Royal Society
- Charlotte Waelde, Professor of Intellectual Property Law, University of Exeter

### Big Datasets and Data Intensive Science Evidence Session, 5 August 2011

- Phil Butcher, Head of IT, The Wellcome Trust Sanger institute
- Dr David Colling, High Energy Physics Group, Imperial College London
- Prof Ian Diamond FBA FRSE AcSS, Vice Chancellor, University of Aberdeen
- Dr Anthony Holloway, Head of Computing, Jodrell Bank Centre for Astrophysics and Jodrell Bank Observatory, University of Manchester.
- Dr Sarah Jackson, Chief Advisor to Government, Met Office
- Prof Anne Trefethen, Director, Oxford e-Research Centre, University of Oxford

### Digital Curation Evidence Session, 5 August 2011

- Dr Kevin Ashley, Director, Digital Curation Centre
- Dr Michael Jubb, Director, Research Information Network
- Angela McKane, Information Capability Manager, BP
- Dr Stephen Pinfield, Chief Information Officer, Information Services, University of Nottingham
- Dr David Shotton, Head, Image Bioinformatics Research Group, University of Oxford

### Policy Lab on Reinventing Discovery, 1 September 2011

Presentation from

- Michael Nielsen, author and previously Perimeter Institute

### Future of Libraries Evidence Sessions, 21 October 2011

- Chris Banks, Head Librarian and Director of Library Special Colletions and Museums, University of Aberdeen
- Rachel Bruce, Innovation director for digital infrastructure, JISC
- Ellen Collins, Research Officer, Research Information Network
- Liz Lyon, Director, UKOLN, University of Bath
- Dr Stephen Pinfield, Chief Information Officer, Information Services, University of Nottingham
- Phil Sykes, University Librarian, University of Liverpool
- Simon Tanner, Director, King's Digital Consultancy Services, King's College London

### Roundtable with Vice Chancellors, 2 December 2011

- Prof Nigel Brown, Senior Vice-Principal Planning, Resources and Research Policy, University of Edinburgh
- Prof Ian Diamond FBA FRSE AcSS, Vice Chancellor, University of Aberdeen
- Christopher Hale, Deputy Director of Policy, Universities UK
- Prof Christopher Higgins, Durham University
- Prof Sir Rick Trainor KBE, Principal and President of Social History, King's College London

### Roundtable on Open Data and Economic Competitiveness, 10 November 2011

- Dr Sam Beale, Head of Technology Strategy, Rolls Royce
- Hadley Beeman, Technology Strategy Board
- Prof Sir Alasdair Breckenridge CBE, Chair, Medicines and Healthcare products Regulatory Agency
- Ellen Collins, Research Officer, Research Innovation Network
- Dr Andy Cosh, Assistant Director, Enterprise and Innovation Programme, Centre for Business Research, Cambridge University
- Prof Patrick Dunleavy, Professor of Political Science and Public Policy, LSE
- Tony Hickson, Managing Director, Technology Transfer, Imperial Innovations
- Prof Sir Peter Knight FRS, President, Institute of Physics
- Dr Brian Marsden, Principal Investigator, Structural Genomics Consortium, University of Oxford
- Dr Tony Raven, Chief Executive, Cambridge Enterprise, University of Cambridge

### Science and the public good: a workshop with evidence from the social sciences, 21 November 2011

Jointly organised by the Royal Society, the Royal Society of Edinburgh, and the ESRC Genomics Policy and Research Forum.

Presentations by:

- Prof Geoffrey Boulton OBE FRSE FRS, Regius Professor of Geology Emeritus, University of Edinburgh
- Dr Iain Gillespie, Innogen Visiting Professor, ESRC Genomics Network
- Dr Jack Stilgoe, Senior Research Fellow, Business School, University of Exeter
- Prof Andrew Stirling, Professor of Science & Technology Policy, SPRU
- Prof Steve Yearley, Director, ESRC Genomics Policy & Research Forum

### Open Public & Panel Debate: Why and How should Science be Open? 21 November 2011

Jointly organised by the Royal Society and Royal Society of Edinburgh.

Panellists:

- Prof Geoffrey Boulton OBE FRSE FRS, Regius Professor of Geology Emeritus, University of Edinburgh
- Sir Kenneth Calman KCB, DL, FRSE, Chancellor, University of Glasgow
- Prof Graeme Laurie, Professor of Medical Jurisprudence, University of Edinburgh
- Prof Wilson Sibbett OBE FRS, Wardlaw Professor of Physics, University of St Andrews
- Prof Steve Yearley, Director, ESRC Genomics Policy & Research Forum

### A Seminar: Opening up scientific data, 30 November 2011

Jointly organised by the Norwegian Academy of Science and Letters, the Royal Society and the University of Bergen.

Seminar leaders:

- Prof Geoffrey Boulton OBE FRSE FRS, Regius Professor of Geology Emeritus, University of Edinburgh
- Prof Ole Laerum, Professor of Experimental Pathology and Oncology, The Gade Institute
- Prof Truls Norby, Department of Chemistry, University of Oslo
- Prof Inger Sandlie, Professor, University of Oslo

### Seminar at The Center, Brussels, 14 December 2011

Speakers included

- Dr Christoph Best, Senior Software Engineer, Google UK Ltd
- Dr Donatella Castelli, D4Science project, Scientific Coordinator at CNR
- Prof Sir Roger Elliott FRS, ALLEA Standing Group on IPRs
- Dr Konstantinos Glinos, Head of Unit GEANT and e-Infrastructures European Commission, DG INFSO
- Prof Wouter Los, Project Leader, LifeWatch
- Prof Laurent Romary, former Director, Max-Planck Digital Library, Chairman of the Council of the international Text Encoding Initiative
- Prof Dr Joseph Straus, Max Planck Institute for Intellectual Property and Competition Law

### URFS roundtable, 27 January 2012

- Dr Sebastian Ahnert, Theory of Condensed Matter (TCM) group, Cavendish Laboratory, University of Cambridge
- Professor Nicholas Grassly, Department for Infectious Disease Epidemiology, Imperial College London
- Dr Francis Jiggins, Department of Genetics, University of Cambridge
- Dr Karen Lipkow, Department of Biochemistry, University of Cambridge
- Dr Christopher Martin, Oxford Neuroscience, University of Oxford
- Dr Jessica Metcalf, Institute for Emerging Infections, University of Oxford
- Dr Emily Nurse, Department of Physics & Astronomy, University College London
- Dr David Payne, Department of Materials, Imperial College London
- Dr Colin Russell, Department of Zoology, University of Cambridge
- Dr Paul Williams, Meteorology Department, University of Reading

## Roundtable on Computer Modelling, 15 February 2012

- Mr Nick Barnes, Board Member, Climate Code Foundation
- Prof Neil Ferguson FMedSci, Professor of Mathematical Biology, Imperial College London
- Prof Tim Palmer FRS, Royal Society Research Professor in Climate Physics and Professorial Fellow, Jesus College, Oxford University
- Prof John Shepherd CBE FRS, National Oceanography Centre, University of Southampton
- Prof Adrian Sutton FRS, Department of Physics, Imperial College London
- Prof Simon Tavare FRS, Dept of Applied Mathematics and Theoretical Physics, University of Cambridge
- Prof Simon Tett FRS, Professor of Earth System Dynamics, University of Edinburgh
- Prof Sir Alan Wilson, Professor of Urban and Regional Systems, UCL

Valuable discussions about the issues raised in this report have been held with several national academies, including the Chinese Academy of Sciences, the US National Academy of Sciences and the Norwegian Academy of Science and Letters.

## Further Consultation

We would also like to acknowledge the vital contribution the following individuals made to this report's scoping work or for their comments on drafts of the report.

- Professor Peter Buneman FRS, School of Informatics, University of Edinburgh
- Professor Tim Clarke, Director of Bioinformatics, MassGeneral Institute for Neurodegenerative Disease & Instructor in Neurology, Harvard Medical School
- Professor Geoff Smith FRS, Department of Pathology, University of Cambridge

**The Royal Society**

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- Promoting science and its benefits
- Recognising excellence in science
- Supporting outstanding science
- Providing scientific advice for policy
- Fostering international and global cooperation
- Education and public engagement

**For further information**

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

Founded in 1660, the Royal Society is the independent scientific academy of the UK, dedicated to promoting excellence in science

Registered Charity No 207043

ISBN: 978-0-85403-962-3
Issued: June 2012 Report 02/12 DES2482

ISBN 978-0-85403-962-3
9 780854 039623

Price £39