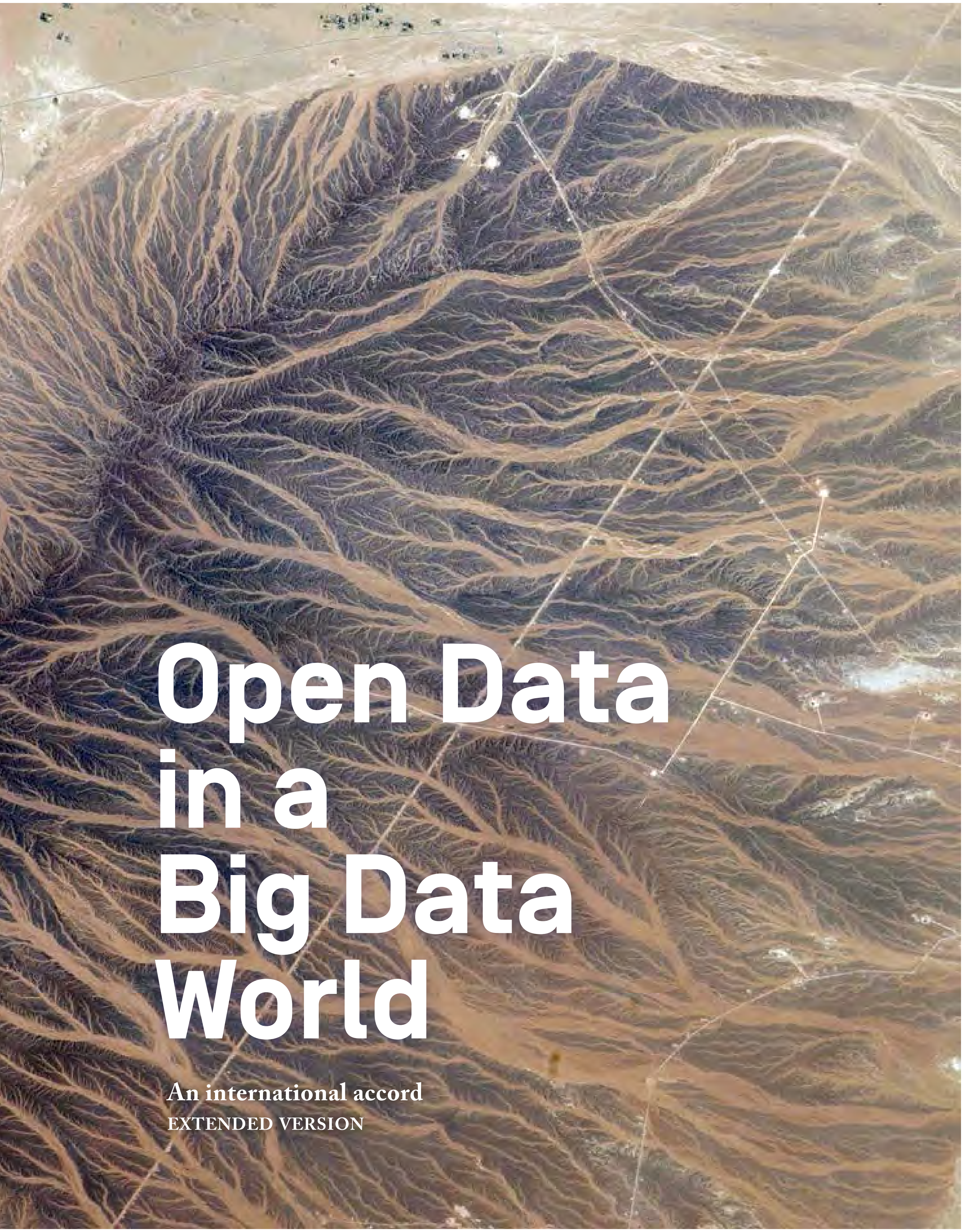
資料５－別添４

提案２２



Open Data in a Big Data World

An international accord

EXTENDED VERSION

ICSU International Council for Science

iap the interacademy partnership

ISSC international social science council

twas

Preface

Four major organisations representing global science, the International Council for Science (ICSU), the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC), are collaborating in a series of action-oriented annual meetings, dubbed "Science International". These meetings are designed to articulate the views of the global scientific community on international matters of policy for science and to promote appropriate actions.

The following accord is the product of the first Science International meeting. The accord identifies the opportunities and challenges of the data revolution as one of today's predominant issues of global science policy. It sets out principles that are consistent with ones being carried out in practice in some national research systems and in some disciplinary fields. It adds the distinctive voice of the scientific community to those of governments and inter-governmental bodies that have made the case for open data as a fundamental pre-requisite in maintaining the rigour of scientific inquiry and maximising public benefit from the data revolution. It builds on ICSU's 2014 statement on open access by endorsing the need for an international framework of open data principles.

In the months ahead, Science International partners will promote discussion and adoption of these principles by their respective members and by other representative bodies of science at national and international levels. We will ask that these organizations review the accord and endorse it, and thereby provide further support in global policy venues for these constructive and vitally important principles.

An abbreviated version of this accord summarises the issues in section A of this document and presents the open data principles that it advocates.

A. Opportunities in the Big Data World

A world-historical event

1. The digital revolution of recent decades is a world-historical event as profound and more pervasive than the introduction of the printing press. It has created an unprecedented explosion in the capacity to acquire, store, manipulate and instantaneously transmit vast and complex data volumes[1]. The rate of change is formidable. In 2003 scientists declared the mapping of the human genome complete. It took over 10 years and cost $1billion – today it takes mere days and a fraction of the cost ($1000)[2]. Although this revolution has not yet run its course, it has already produced fundamental changes in economic and social behaviour and has profound implications for science[3], permitting patterns in phenomena to be identified that have hitherto lain beyond our horizon and to demonstrate hitherto unsuspected relationships. Researchers were amongst the first users of digital networks such that many areas of research across the humanities, natural and social sciences are being transformed, or have the potential to be transformed, by access to and analysis of such data.

2. The worldwide increase in digital connectivity, the global scale of highly personalized communications services, the use of the World Wide Web as a platform for numerous human transactions, the "internet of things" that permits any device with a power source to collect data from its environment together with advances in data analytics have coalesced to create a powerful platform for change. In this networked world, people, objects and connections are producing data at unprecedented rates, both actively and passively. This not only creates large data volumes, but also distinctive data streams that have been termed "big data", characterised by the four Vs[4]:

- the **volume** that systems must ingest, process and disseminate;
- the **variety** and complexity of datasets, originating from both individuals and institutions at multiple points in the data value chain;
- the **velocity** of data streaming in and out of systems in real time;
- the **veracity** of data (referring to the uncertainty due to bias, noise or abnormality in data), which is often included. This is a desirable characteristic, not an intrinsic feature of Big Data. The veracity and the peer review of results based on big data, however, pose severe problems for effective scrutiny, with a clear need to establish a "reproducibility standard."

3. A second pillar of the data revolution is formed by "linked data." Separate datasets that relate to a particular phenomenon and that are logically connected can be semantically linked in ways that permit a computer to identify deeper relationships between them. Semantic search links similar ideas together, permitting the World Wide Web to evolve from a web of documents into a Semantic Web in which meaning can be more readily deduced from linked data, connecting related data that were not necessarily designed for mutual integration. Such processes offer profound ways of understanding the structure and dynamics of systems where very diverse elements are coupled together to produce complex behaviour. They have the potential to yield an enormous dividend of understanding by breaking down the barriers that tend to separate disciplinary silos, although only if the data is openly available and free to be linked.

4. The great achievements of science in recent centuries lie primarily in understanding relatively simple, uncoupled or weakly coupled systems. Access to increasing computational power has permitted researchers to simulate the dynamic behaviour of highly coupled complex systems. But the advent and analysis of big and linked data now add to this the complementary capacity to characterise and describe complexity in great detail. Coupling these two approaches to the analysis of complexity has the potential to usher in a new era of scientific understanding of the complexity that underlies many of the major issues of current human concern. "Global challenges" such as infectious disease, energy depletion, migration, inequality, environmental change, sustainability and the operation of the global economy are highly coupled systems, inherently complex, and beyond the reach of the reductionist approaches and the individual efforts

1 We use the term ***data*** to refer to "representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship". C.L. Borgman, 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World.* The MIT Press, p. 28.

2 Illumina announces landmark $1,000 human genome sequencing. *Science* 15 January 2014

3 The word ***science*** is used to mean the systematic organisation of knowledge that can be rationally explained and reliably applied. It is used, as in most languages other than English, to include all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines.

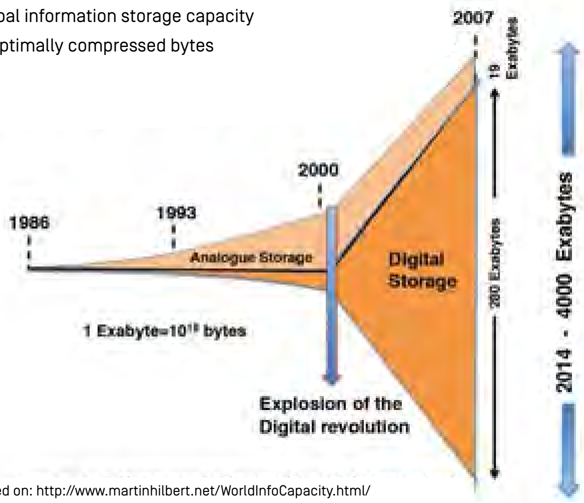
4 www.ibmbigdatahub.com/infographic/four-vs-big-data

BOX 1

The beginning of the digital revolution

Global information storage capacity

In optimally compressed bytes



Based on: http://www.martinhilbert.net/WorldInfoCapacity.html/

BOX 2

Linked Data and the Semantic Web

Linked Data use the techniques and concepts of the World Wide Web to describe the real world. They use web identifiers (Uniform Resource Identifier or URIs, often in the form of an http location) to identify facts, concepts, people, places and phenomena as well as documents that have common attributes. This allows connections to be discovered between different datasets, thereby increasing the value of each through the Network Effect, permitting a researcher to discover data important to their work. Programmes such as Resource Discovery for Extreme Scale Collaboration (http://rdesc.org) use these approaches to search for and discover data resources relevant to a particular scientific purpose. The approach is being increasingly applied in environmental fields. Operational examples relevant to business include OpenPHACTS, which uses the technology to provide easy access to more than 14 million facts about chemical and pharmacological data; the European Environment Agency's provision of reference datasets for species; and the Slovenian Supervizor portal which matches public spending to contracts to businesses, providing a powerful tool against corruption.

Linked Data is a subset of the wider Semantic Web, in which queries do not retrieve documents as in the standard web, but semantic responses that harvest information from datasets that are connected by logical links. This approach is being much exploited in genomics, one example being through a resource description framework (RDF) platform implemented through the European Molecular Biology Laboratory Elixir programme (see Box 6).

that nonetheless remain powerful tools in the armoury of science. The potential of big data in such cases is to permit analysis of complex system whilst still producing general explanations.

5. A further consequence of the increasing capacity to acquire data at relatively low cost, when coupled with great processing power, is to permit machines that sense data from their immediate environment to learn complex, adaptive behaviours by trial and error, with the disruptive potential to undertake what have hitherto been regarded as highly skilled, and necessarily human, tasks.

B. Exploiting the Opportunities: the Open Data Imperative

Maintaining "self-correction"

6. Openness and transparency have formed the bedrock on which the progress of science in the modern era has been based. They have permitted the logic connecting evidence (the data) and the claims derived from it to be scrutinised, and the reproducibility of observations or experiments to be tested, thereby supporting or invalidating those claims. This principle of "self-correction" has steered science away from the perpetuation of error. However, the current storm of data challenges this vital principle through the sheer complexity of making data available in a form that is readily subject to rigorous scrutiny. Ensuring that data are open, whether or not they are big data, is a vital priority if the integrity and credibility of science and its utility as a reliable means of acquiring knowledge are to be maintained.

7. It is therefore essential that data that provide the evidence for published claims, the related metadata that permit their re-analysis and the codes used in essential computer manipulation of datasets, not matter how complex, are made concurrently open to scrutiny if the vital process of self-correction is to be maintained. The onus not only lies on researchers but also on scientific publishers, the researchers who make up the editorial boards of scientific journals and those managing the diverse publication venues in the developing area of open access publishing, to ensure that the data (including the meta-data) on which a published scientific claim are based are concurrently available for scrutiny. To do otherwise should come to be regarded as scientific malpractice.

The definition of open data

8. Simply making data accessible is not enough. Data must be **"intelligently open"**[5], meaning that they can be thoroughly scrutinised and appropriately re-used. The following criteria should be satisfied for open data, that it should be:

- **discoverable** – a web search can readily reveal their existence;
- **accessible** – the data can be electronically imported into or accessed by a computer;
- **intelligible** – there must be enough background information to make clear the relevance of the data to the specific issue under investigation;
- **assessable** – users must be able to assess issues such as the competence of the data producers or the extent to which they may have a pecuniary interest in a particular outcome;
- **usable** – there must be adequate metadata (the data about data that makes the data usable), and where computation has been used to create derived data, the relevant code, sometimes together with the characteristics of the computer, needs to be accessible.

Data should be of high quality wherever possible, reliable, authentic, and of scientific relevance. For longitudinal datasets, the metadata must be sufficient for users to be able to make a comparative analysis between timelines, and the sources must be valid and verifiable. It is important to be aware that the quality of some scientifically important datasets, such as those derived from unique experiments, may not be high in conventional terms, and may require very careful treatment and analysis.

Non-Replicability

9. The replication of observations and experiments has a central role in science. It is the justification for the statement made by Galileo in Brecht's play[6] that "the aim of science is not to open the door to infinite wisdom, but to set a limit to infinite error." Recent attempts to replicate systematically the results of series of highly regarded published papers in, for example, pre-clinical oncology (53 papers)[7], social psychology (100 papers)[8] and economics (67 papers)[9], were successful in only 11 %, 39 % and 33 % of cases respectively. The reasons adduced for these failures included falsification of data, invalid statistical reasoning and absent or incompleteness of the data or metadata. Such failures were highlighted in *The Economist*[10] magazine under the headline: "Scientists like to think of science as self-correcting. To an alarming degree it is not." These failures will threaten the credibility of the scientific enterprise unless corrective action is taken. If data, meta-data and the code used in any manipulations are not available for scrutiny, published work, whether right or wrong, cannot be subject to an adequate test of replication.

10. An implication of the above results is that pre-publication peer review has failed in these cases in its primary purpose of checking whether the research has been performed to reasonable standards, and whether the findings and conclusions drawn from them are valid. Given the depth of analysis required to establish replicability, and the increasing pressure on reviewers because of the dramatic rise in the rate of publication[11], it is unsurprising that peer review fails in this regard. Under these circumstances, it is crucial that data and metadata are concurrently published in an intelligently open form so that it is also accessible to "post-publication" peer review, whereby the world decides the importance and place of a piece of research[12].

Open data and "self correction"

11. The reputational and other rewards for scientific discovery can be considerable, with an inevitable temptation for misconduct involving the invention of data or intentional bias in their selection. In general we would expect open data to deter fraud, on the principle that "sunlight is the best disinfectant". In contrast, there are cases where the integration of datasets derived from different open sources could enable fraud by effectively hiding fraudulent components because of the difficulty of disentangling datasets. Without a standard of openness that permits, even in these cases, others to subject the related scientific claim to the test of reproducibility,

5 Science as an Open Enterprise. 2012. The Royal Society Policy Centre Report, 02/12. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

6 Bertolt Brecht, 1945. *The Life of Galileo.*

7 Begley, C.G. and Ellis, L.M. 2012. Nature, 483, p. 531 – 533.

8 Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716.

9 Chang, A. and Li, P. 2015. Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System

10 The Economist. 2013, October 19 – 25. pp. 21 – 23.

11 The term publishing and publisher simply refer to the act of making written or spoken work publically and permanently available. It is not restricted to conventional printed publication.

12 Smith, R. 2010. Classical peer review: an empty gun. *Breast Cancer Research* 2010, 12 (Suppl 4): S13 http://breast-cancer-research.com/supplements/12/S4/S13.

such claims may prove to be an irreducible barrier to scientific progress. The integrity of data is often of greater significance than the claim based on them. To quote Charles Darwin[13]: " false facts are highly injurious to the progress of science, for they often long endure; but false views, if supported by some evidence, do little harm, as everyone takes a salutary pleasure in proving their falseness"; leading to an outcome described by Arthur Koestle[14] as one in which "the progress of science is strewn, like an ancient desert trail, with the bleached skeletons of discarded theories that once seemed to possess eternal life".

Valid reasoning

12. A major priority for data-intensive science must be greater analytical rigour and the establishment, discipline by discipline, of acceptable standards of replicability. Regression-based, classical statistics have long been the basic tools for establishing relationships in data. Many of the complex relationships that we now seek to capture through big or linked data lie far beyond the analytical power of these methods, such that we now need to supplement them in adapting topological and related methods to data analysis to ensure that inferences drawn from big or linked data are valid. Data-intensive machine-analysis and machine-learning are becoming ubiquitous, creating the possibility of improved, evidence-informed decision making in many fields. The creative potential of big data, of linking data from diverse sources and of machine learning not only have implications for discovery, but also for the world of work and for what it means to be a researcher in the 21st century. The potential disconnect between machines that learn from data and human cognitive processes poses profound issues for how we understand machine-analysed phenomena and their accessibility to human reasoning.

Openness: the default for publicly funded research

13. We regard it as axiomatic that knowledge and understanding have been and will continue to be essential to human judgements, innovation and social and personal wellbeing. The fundamental role of the publicly funded scientific enterprise is to add to the stock of knowledge and understanding, such that high priority should be given to processes that most efficiently and creatively advance knowledge. The productivity of open knowledge, of having ideas and data made open by their originators, is illustrated by a comment attributed to the playwright George Bernard Shaw: *"if you have an apple and I have an apple and we exchange these apples, then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas".* The technologies and processes of the digital revolution as described above provide a powerful medium through which such multiplication of productivity and creativity can be achieved through rapid interchange and development of ideas by the networked interaction of many minds.

14. If this social revolution in science is to be achieved, it is not only a matter of making data that underpin a scientific claim intelligently open, but also of having a default position of openness for publicly funded data in general. In some disciplinary communities data are released into the public domain immediately after they have been produced, such as in the case of genome sequencing data since the agreement of the 1996 Bermuda Principles and the 2003 Fort Lauderdale Principles[15]. The circumstance and timescale of release are important. It many disciplines it is reasonable to expect that data need only be released upon the termination of the grant that funded their collection. Even then it may be appropriate for grant holders to have the first bite of the publication cherry before data release. Although it is tempting to suggest a embargo period, perhaps of the order of a year, it would be better for individual disciplines to develop procedures that are sympathetic to disciplinary exigencies, but without involving excessive delay.

Boundaries of openness

15. Although open data should be the default position for publicly funded research data, not all data can or should be made available to all people in all circumstances. There are legitimate exceptions to openness on matters of personal privacy, safety and security, whilst further ethical concerns ought to constrain the way that data systems operate and data are used, as discussed in the next section. Given the increasing incidence of joint public/private funding for research, and with the premise that commercial exploitation of publicly funded research data can be in the broader public interest, legitimate exceptions to the default position for openness are also possible in these cases. These categories – which are largely discipline-dependent – should not however be used as the basis for blanket exceptions. Exceptions to the default should be made on a case-by-case basis, with the onus on a proponent to demonstrate specific reasons for an exception.

Ethical issues

16. Open data and data sharing have important ethical dimensions that relate to researchers' responsibilities to the public, to those who provide personal data, and to fellow researchers. Although we advocate a normative view that publicly funded researchers have an obligation to make data that they have collected openly available as a public good in the interests of science and society, we recognise that this creates further dilemmas that require attention:

- Datasets containing personal information have the potential to infringe the right to privacy of data subjects and require governance practices that protect personal privacy.
- A substantial body of work in computer science has demonstrated that conventional anonymisation procedures cannot guarantee the

13 Darwin, C. 1871. *The Descent of Man.* John Murray, London. 2 vols.

14 Koestler, A. 1967. *The ghost in the machine.* Macmillan, London, 384 pp.

15 Human Genome Project (2003). Available at: http://www.genome.gov/10506376 and http://www.wellcome.ac.uk/About-us/Publications/Reports/Biomedical-science/WTD003208.htm

BOX 3

Managing Ethical Risk

The Administrative Data Research Centre for England (ADRC-E), has examined attitudes to data handling of administrative data and developed a model for managing ethical risks that attempts to address public concerns. An IPSOS-Mori poll on behalf of the UK Economic and Social Research Council (ESRC) found that the public held few objections over the use of administrative data for research purposes, subject to certain important caveats: that there should be strong governance, effective de-identification and de-linkage, and a clear public — not commercial — benefit in using the data. The ADRC network model recognises three critical levels of scrutiny: of researchers, of project aims, and of the role of the ADRC itself. ADRC will accredit researchers, then, when an accredited researcher requests access to data, a panel will evaluate whether or not the proposed project will deliver a clear public benefit, is drawn on data that is essential to their research and is not available elsewhere. Once a request is approved, the ADRC assembles the requested data sets and takes responsibility for linkage and de-identification. Importantly, in the language of data protection, the ADRC acts only as a 'data processor', not as a 'data controller'. This approach accommodates public concerns, and creates an acceptable synergy between researchers, the nature of data supplied and where the data are located. The model has proved financially sustainable following significant start-up funding.

In, Science Europe Social Sciences Committee (September 2015), 'Workshop Report: Ethical Protocols and Standards for Research in Social Sciences Today': D/2015/13.324/7

security of personal records[16], such that stronger, more secure practices may be required[17].
- Researchers have a moral obligation to honour relationships that they have developed with those who have entrusted them with personal information. Data sharing threatens these relationships because it entails a loss of control over future users and future usage of data. In the humanities and social sciences, data are often co-constructed by researchers and respondents, and also contain much sensitive information relating to both respondents and researchers.
- Open data can override the individual interests of the researchers who generate the data, such that novel ways of recognizing and rewarding their contribution must be developed (see Section D). Junior researchers, PhD students and/or technicians may be particularly vulnerable to lack of recognition, and with limited say in data reuse.
- In international projects, data sharing may become a form of scientific neo-colonialism, as researchers from well-funded research systems may stand to gain more than those from poorly funded systems. This could happen because of differences in infrastructure investment, or different levels of granularity.

Open Global Participation

17. The ways in which big data, linked data and open data can be used for data-driven development and can be leveraged to positively impact the lives of the most vulnerable are becoming clearer[18]. There is great potential for data-driven development because of its detail, timeliness, ability to be utilized for multiple purposes at scale and in making large portions of low-income populations visible. Although many well-funded national science systems are adapting rapidly to seize the data challenge, the great promise of big data remains remote for many less affluent countries, and especially for the least developed countries (LDCs), where the costs of adaptation referred to in the next section pose particular problems.

18. LDCs typically have poorly resourced national systems. If they cannot participate in research based on big and open data, the gap could grow exponentially in coming years. They will be unable to collect, store and share data, unable to participate in the global research enterprise, unable to contribute as full partners to global efforts on climate change, health care, and resource protection, and unable fully to benefit from such efforts, where global solutions will only be achieved if there is global participation. Thus, both emerging and developed nations have a clear, direct interest in helping to fully mobilize LDC science potential and thereby to contribute to achievement of the UN Sustainable Development Goals. It is vital that processes that deliver local benefit are developed based on effective governance frameworks and the legal, cultural, technological and economic infrastructures necessary to balance competing interests.[19]

Changing the dynamic

19. Creative and productive exploitation of this technologically-enabled revolution will also depend upon the creation of supporting "soft" and "hard" infrastructure and changes in the social dynamics of science, involving not only a willingness to share and to release data for re-use and re-purposing by others but the recognition of a responsibility to do so.

16 For example: Denning D (1980). *A fast procedure for finding a tracker in a statistical database.* ACM Transactions on Database Systems (TODS), 5, 1. *Differential Privacy*. International Colloquium on Automata, Languages and Programming (ICALP), 1–12; Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007).

17 For example: Thomas R & Walport M (2008). Data Sharing Review. Available at: http://www.justice.gov.uk/reviews/docs/data-sharing-review-report.pdf

18 http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html#

19 Linnet Taylor & Ralph Schroeder (2015) 'Is bigger better? The emergence of big data as a tool for international development policy, *GeoJournal* 80(4), pp. 503–518. 10.1007/s10708-014-9603-5

20. Although science is an international enterprise, it is largely done within national and disciplinary systems that are organised, funded and motivated by national and disciplinary norms and practices. Effective open data in a data-intensive age can only be realised if there is systemic action at disciplinary, national and international levels. At the national level there is need for government to recognise the value to be gained from open data, for national science agencies to adopt a coordinating role, for science policy makers to set incentives for openness from universities and research institutes, for these institutions to support open data processes by their researchers and for the learned societies that articulate the priorities and practices of their disciplines to advocate and facilitate open data processes as important priorities.

21. The rationale for a national open data policy lies in ensuring the rigour of national science based on its reproducibility and the accessibility of its results, in capturing the value of open data[20] for national benefit and as the basis for efficient collaboration in international science. New partnerships, infrastructures and resources are needed to ensure that researchers and research institutions work with government and private-sector big data companies and programmes to maximize data availability for research and for its effective exploitation both for public policy and direct economic benefit.

22. Soft and hard enabling infrastructures are required to support open data systems. Soft infrastructure comprises the principles that establish behavioural norms, incentives that encourage their widespread adoption

20 The economic value of open data has been estimated as $3–5 trillion per annum across seven commercial sectors. McKinsey Global Institute: Open Data, 2013.

BOX 4

Open research data in South America

The Latin American region is one with a strong tradition of cooperation in building regional information and publishing systems. Today, an estimated 80% of active journals are open access, complemented by repositories (regional subject repositories, and more recently institutional repositories) which are gaining momentum promoted by national open access legislation approved in Peru, Argentina, Mexico, and in discussion in Brazil and Venezuela. These require publicly-funded research results to be deposited in open access repositories, in some cases explicitly including research data.
The issue of open research data is starting to take off in the region, with activities to build awareness and consensus on good practices, sponsored by national research agencies (e.g. national systems for data–climate, biological, sea, genomics-coordinated by the Ministry of Science, Technology and Innovation of Argentina);
the initiative datoscientíficos.cl promoted by the National Commission of Scientific and Technological Research in Chile to seek opinions for a proposed policy for open research data; and a national meeting of open data organized by the Brazilian Institute of Information in Science and Technology. These national actions provide context and guidance for new institutional and national open research data initiatives within the region, which also look at other existing open research data programmes (e. g. UN Economic Commission for Latin America and open research data at the National Autonomous University of Mexico-UNAM).

In parallel, there is a movement in Latin America towards open government data, open knowledge and open data in general, as part of international movements and initiatives. Governments and civil society organize open data events and projects, open data schools, unconferences and data hackathons that build awareness about the need and opportunities to open government data, which also benefits research. To facilitate regional research cooperation and exchange of big research data, the National Research and Education Networks (NREN) are members of the Latin American Cooperation of Advanced Networks (RedCLARA) which provides advanced Internet networking facilities to countries and institutions of the region.

and practices that ensure efficient operation of a national open data system that is also consistent with international standards. This part of the soft infrastructure is not financially costly, but depends upon effective management of the relationships summarised in the preceding paragraph and effective international links. The costly component is the need for time-intensive data management both by research institutions and researchers. By contrast, the physical or hard infrastructure required to sustain data storage, analysis, broadband transmission and long-term preservation is not separable from that required to support a strong national science base. Both soft and hard infrastructures are essential enabling elements for producing and using scientific data, though, as commented above, they pose especially difficult challenges for doing research in low- and middle-income countries.

23. Responsibilities also fall on international bodies, such as the International Council for Science's (ICSU) Committee on Data for Science and Technology (CODATA)[21] and World Data System (WDS)[22], and the Research Data Alliance (RDA)[23], to promote and support developments of the systems and procedures that will ensure international data access, interoperability and sustainability. Members of these bodies represent a wide range of countries, and both through them and through other national contacts, international norms should aim to be compatible with national procedures as far as possible. In establishing where change is required, it is important to distinguish between those habits that have arisen because they were well adapted to a passing technology but which may now be inimical to realisation of the benefits of a new one, and those habits that reflect essential, technology-independent priorities and values. In this regard, it is a priority to establish new ways of recognising, rewarding and therefore incentivising efforts in data management, preservation and curation. It involves questioning ingrained assumptions about the primacy of "high-impact" publications as a measure of scientific excellence, and finding ways to acknowledge communication of science, such as the development and dissemination of "open software", and participation in international programmes of data donation and curation.

21 http://www.codata.org/

22 https://www.icsu-wds.org/

23 https://rd-alliance.org/node

24. Although the articulation by international representative bodies of the ethical and practical benefits of open data processes is important, it is the actions of practising scientists and scientific communities that will determine the adoption, extent and impact of these processes. These are fundamental issues for science, society and the economy and depend on the willingness of scientists to open up their data for sharing, re-use and re-purposing, even if there are personal, technical, organizational and political barriers to doing so. New solutions for making data open are required that demand collective efforts from all stakeholders involved in the production of knowledge, including individual researchers, the institutions in which they work, and the myriad organizations which influence their work. It is of course recognised that the gap between aspiration and practical implementation is a large one, both in terms of the willingness of individuals and institutions to change mindsets, and the capacity to adapt behaviour because of the availability of tools, management systems and hard infrastructure.

25. Major bottom-up changes are however happening at the level of disciplinary and multi-disciplinary communities. Strong processes of open data sharing have developed in areas such as linguistics[24], bioinformatics[25] and chemical crystallography[26]. In human palaeogenetics, it appears that open data sharing is almost universal (> 97 %), not as a consequence of top-down requirements, but because of awareness of its value by the relevant research community.[27] Moreover, a growing number of researchers share their data from the start of their research projects, both to receive comments from peers and to engage in open collaboration. These developments are sensitive to the needs of the disciplines involved, they provide

24 http://www.linguistic-lod.org/llod-cloud

25 https://www.elixir-europe.org/

26 http://www.crystallography.net/

27 Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D. and Destro Bisol, G. 2015. *When Data Sharing Gets Close to 100 %: What Human Paleogenetics Can Teach the Open Science Movement.* PLOS ONE · March 2015, DOI: 10.1371/journal.pone.0121409

BOX 5

Open Data Initiatives in Africa

Many African countries are energetically developing their own capacities to exploit the data revolution for the benefit of their public policies and economies.

United Nations perspective

The UN report, A World that Counts (2015 – www.undatarevolution.org), sets out the public policy imperative to improve data gathering and to make data open for maximum impact and reuse: "Data are the lifeblood of decision-making. Without data, we cannot know how many people are born and at what age they die; how many men, women and children still live in poverty; how many children need educating; how many doctors to train or schools to build; how public money is being spent and to what effect; whether greenhouse gas emissions are increasing or the fish stocks in the ocean are dangerously low; how many people are in what kinds of work, what companies are trading and whether economic activity is expanding".

Open Data for Africa portal (www.opendataforafrica.org)

This includes such data on food prices, GDP per capita, energy statistics, demographics, water, energy and energy forecasts, food, education, government debt, healthcare infrastructure, malaria, migration, mortality, urbanization etc.

National initiatives

The Kenyan Data Forum (http://www.dataforum.or.ke/) emphasizes the need for the domestication of the data revolution as a key step in accelerating implementation of the national development agenda, which is aligned with regional and global goals. It convenes stakeholder communities from government, private sector, academia, civil society, local communities and development partners who engage on the informational aspects of development decision-making.

Agriculture: Agriculture accounts for 65 % of Africa's workforce and 32 % of the continent's GDP. In some of Africa's poorest countries, including Chad and Sierra Leone, it accounts for more than 50 % of GDP. The Global Open Data for Agriculture and Nutrition initiative (http://www.godan.info) recently published a report which asks 'How can we improve agriculture, food and nutrition with Open Data'. The report presents numerous case studies of precisely how Open Data can advance research and practice in these areas with numerous positive outcomes.

AgTrials is an example of improving crop varieties with open data about breeding trials. Scientists have used 250 open AgTrials datasets to build crop models specific to the West Africa region. The models are used to project the local impacts of climate change, addressing issues such as drought tolerance, heat stress, and soil management and defining breeding programmes for adaptation.

Mobilising Science Capacity

To accompany this open data accord, Science International will promote a collaborative initiative involving the South African Government's Department of Science and Technology, other national science bodies in sub-Saharan Africa and CODATA and its international partners (RDA and WDS) in mobilising the African research community in developing big data/open data capacities.

BOX 6

Open Data Platforms

A national data platform

The National Science and Technology Infrastructure (NSTI) of the Peoples Republic of China is the networked, ITC based system that provides shared service for technology innovation and economic and social development. The NSTI programme supports 10 scientific data centres and 3 scientific data sharing networks. It integrates more than 50,000 science and technology databases in 32 categories and 10 technical fields, including agriculture, meteorology, seismicity, population health, materials, energy, geology, etc. It has established a managed service for scientific and technology data and information sharing, based on a series of standard specifications. Under this programme, a number of high-profile data and information sharing services have been set up. In 2011, NSTI supported the creation of 6 scientific data platforms, which facilitate standardised management of data resources and offer a quality-controlled service. By 2014, the NSTI platform website had received more than 50 million visits and provided 60 terabytes of information. The platform currently provides a service for nearly 3000 national key science and technology projects and plays an important role in innovation and public service. The NSTI is demand-driven: in specific instances it responds with comprehensive, systematic, special services, and creates scientific data products.

A disciplinary platform: ELIXIR — an integrated data support system for the life sciences.

ELIXIR is the European life-science infrastructure for biological information. It is a unique and unprecedented initiative that consolidates Europe's national centres, services, and core bioinformatics resources into a single, coordinated infrastructure. It brings together Europe's major life-science data archives and, for the first time, connects them with national bioinformatics infrastructures throughout ELIXIR's member states. By coordinating local, national and international resources the ELIXIR infrastructure is designed to serve the data-related needs of Europe's 500,000 life-scientists. Open access to bioinformatics resources provides a valuable path to discovery. National nodes develop national strategies and are the sources of support for national communities and the route through which ELIXIR resources, including data, analytic software and other tools are accessed. There is a strong ethos of data sharing in many life science communities, but even here practices vary. In structural biology and genomics it is established practice to deposit sequence data as soon as it is acquired. In many fields it is a requirement to deposit data for publishing. In other areas, such as biomedical research, practice is varied, though there is strong pressure from funders for openness.

an open corpus of information for their communities that is far greater than any single researcher could acquire, offer support and advice, and animate creative collaboration between their members. It is important that top-down processes do not prescribe mechanisms that inhibit the development of such initiatives, but are able to learn from their success and be supportive of and adaptive to their needs through the provision of appropriate soft and hard infrastructures that are sensitive to local possibilities and resources.

Open Science and Open Data

26. The idea of "open science" has developed in recognition of the need for stronger dialogue and engagement of the science community with wider society in addressing many current problems through reciprocal framing of issues and the collaborative design, execution and application of research. "Open data" (as a set of practices and a resource) is an essential part of that process. In an era of diminished deference and ubiquitous communication it is no longer adequate to announce scientific conclusions on matters of public interest and concern without providing the evidence (the data) that supports them, and which can therefore be subject to intense and rigorous scrutiny. The growth of citizen science, which involves many participants without formal research training, and the increasing participation of social actors other than scholars in co-creation of knowledge, are enriching local and global conversations on issues that affect us all and are eroding the boundary between professional and amateur scientists. At the same time, the apparent increase in fraudulent behaviour, much of which includes invention or spurious manipulation of data, risks undermining public trust in science, for which openness to scrutiny must be an important part of the necessary corrective action.

Public Knowledge or Private Knowledge?

27. Open scientific data and the resulting knowledge have generally been regarded as public goods and a fundamental basis for human judgement, innovation and the wellbeing of society. Many governments now recognise the benefits of being open with their own data holdings in order to provide opportunities for creative commercial re-use of a public resource, to achieve specific public policy objectives, to increase government accountability and to be more responsive to citizens' needs. Access to such data can also be of considerable scientific value, particularly in the social sciences for evaluating social and economic trends, and in the medical sciences for evaluating optimal public health strategies from population health records. There are inter-governmental initiatives to promote openness, such as the *Open Government Partnership*[28], which now involves 66 participating countries worldwide, the G8 Open Data Charter[29] and the report to the UN Secretary-General from his Independent Advisory Group on *the Data Revolution for Sustainable Development*[30].

28. It is tempting to think that the boundary of open data is the boundary between the publicly funded and the commercially held, but this is not necessarily the case. Different business sectors take different approaches, with some benefitting from openness. For example, it is in the interests of manufacturers of environmental data acquisition systems for the data to be open in ways that stimulate new businesses based on novel ways of using them, thereby increasing demand for the hardware. The massive data volumes that are daily captured by retail and service industries offer great research potential if made available to social science researchers. Thus, policy makers have a responsibility to consider new ways of incentivising

BOX 7

Opening up government data: the Indian strategy

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 recognises the need for open data in order to: maximise use, avoid duplication, maximise integration, spread ownership of information, and 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. 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.

28 www.opengovpartnership.org

29 https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex; see also the related G8 Science Ministers Statement, London, 12 June 2013: https://www.gov.uk/government/publications/g8-science-ministers-statement-london-12-june-2013

30 www.undatarevolution.org

private companies to make their data open. New forms of university-industry engagement around public and private data could generate important insights and benefits for science, society and the economy.

29. There is currently an important international debate about whether to make public data freely available and usable by everyone, or just the not-for-profit sector. Should the private, for-profit sector pay for access and use of publicly funded data? This is a complex issue, but as long as the original data remain openly available on the same terms to all, it does not seem sensible, appropriate or productive to discriminate between not-for-profit and for-profit users. Robust evidence is accumulating of the diverse benefits and broader economic and societal value derived from the open sharing of research data.[31]

30. It is however important to recognise that there is a countervailing trend to openness, of business models built on the capture and privatisation of socially produced knowledge through the monopoly and protection of data. Such trends towards privatisation of a public resource or uncontrolled and unconsented access to personal information are at odds with the ethos of scientific inquiry and the basic need of humanity to use ideas freely. If the scientific enterprise is not to founder under such pressures, an assertive commitment to open data, open information and open knowledge is required from the scientific community.

C. Principles of Open Data

31. Such is the importance and magnitude of the challenges to the practice of science from the data revolution that Science International believes it appropriate to promote the following statement of principles of responsibility and of enabling practice for data-intensive science. Science International partners will advocate them for adoption by scientific unions, national representative science bodies and others that influence the operation of national and international science systems. The principles are an evolution of – but consistent with – priorities and recommendations set out in earlier reports on data-intensive science by Science International partners, by governmental and Inter-governmental bodies and by academic groups.[32] These principles recognise not only the benefits of open data and open science, but also the complexity of the international research landscape, with sometimes overlapping and sometimes competing needs and interests between different stakeholders. Section D sets out further rationale for the principles and practical options for their implementation.

Responsibilities

Scientists

i. Publicly funded scientists have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They should make such data openly available to others as soon as possible after their production in ways that permit them to be re-used and re-purposed.

ii. The data that provide evidence for published scientific claims should be made concurrently and publicly available in an intelligently open form. This should permit the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To the extent possible, data should be deposited in well-managed and trusted repositories with low access barriers.

iii. Research institutions and universities
have a responsibility to create a supportive environment for open data. This includes the provision of training in data management, preservation and analysis and of relevant technical support, library and data management services. Institutions that employ scientists and bodies that fund them should develop incentives and criteria for career advancement for those involved in open data processes. Consensus on such criteria is necessary nationally, and ideally internationally, to facilitate desirable patterns of researcher mobility. In the current spirit of internationalisation, universities and other science institutions in developed countries should collaborate with their counterparts in developing countries to mobilise data-intensive capacities.

iv. Publishers
have a responsibility to make data available to reviewers during the review process, to require intelligently open access to the data concurrently with the publication which uses them and to require the full referencing and citation of these data. Publishers also have a responsibility to make the scientific record available for subsequent analysis through the open provision of metadata and open access for text and data mining.

vi. Funding agencies
should regard the costs of open data processes in a research project to be an intrinsic part of the cost of doing the research, and should provide adequate resources and policies for long term sustainability of infrastructure and repositories. Assessment of research impact, particularly any involving citation metrics, should take due account of the contribution of data creators.

vii. Professional associations, scholarly societies and academies
should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.

viii. Libraries, archives and repositories
have a responsibility for the development and provision of services and technical standards for data to ensure that data are available to those who wish to use them and that data are accessible over the long term.

Boundaries of openness

viii. Open data should be the default position for publicly funded science. Exceptions should be limited to issues of privacy, safety, security and to commercial use that is in the public interest. Exceptions should be justified on a case-by-case and not blanket basis.

Enabling practices

ix. Citation and provenance
When, in scholarly publications, researchers use data created by others, those data should be cited with reference to their originator, their provenance and to a permanent digital identifier.

x. Interoperability
Both research data, and the metadata which allows them to be assessed and reused, should be interoperable to the greatest degree possible.

xi. Non-restrictive reuse
If research data are not already in the public domain, they should be labelled as reusable by means of a rights waiver or non-restrictive licence that makes it clear that the data may be reused with no more arduous requirement than that of acknowledging the prior producer(s).

31 An brief yet comprehensive survey of current evidence is provided in Paul Uhlir for CODATA (2015) *The Value of Open Data Sharing: A White Paper for the Group on Earth Observations* http://dx.doi.org/10.5281/zenodo.33830

32 Reports by Science International partners include: ICSU-CODATA 2000; IAP 2003; CODATA 2014. Governmental or inter-governmental statements include: Bromley 1991; WMO 1995; OECD 2007 and 2008; and G8 2013. Academic statements include: the Bermuda Principles 1996; Berlin Declaration 2003; The Royal Society 2012; Bouchout Declaration 2014; Hague Declaration 2014; and RECODE Project 2015. A compendium of many national and international policy documents for Open Data may be found at: Sunlight Foundation 2015 or Open Access Directory 2015. Further statements are referenced in appendix 2.

xii. Linkability
Open data should, as often as possible, be linked with other data based on their content and context in order to maximise their semantic value.

D. The Practice of Open Data

32. This section expands on the rationale for the above principles and consequential issues of practice that should be addressed.

Responsibilities

Normative values
33. The accord makes the normative assertion that publicly funded research should be undertaken in a way that creates maximum public benefit. It argues that the open release of data is the optimal route by which this is achieved.

34. The argument that such openness should be openness to the world and not merely contained within national boundaries is part of both the utilitarian and normative arguments for open publication:

- that no one country dominates the international scientific effort and that maximum national benefit is gained if all openly publish their results and all are able to utilise them;
- that the acquisition of knowledge is an essential human enterprise and should be open to all.

Statements and reports that emphasise these priorities are referenced in appendix 2.

Data used as evidence for a scientific claim
35. The data that provide evidence for a published scientific claim must be concurrently published in a way that permits the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To do otherwise should be regarded as scientific malpractice. The intelligent openness criteria of principle ii should be applied to the data. It is generally impracticable for large data volumes to be included in a conventional scientific publication, but such data should be referenced by means of a citation including a permanent digital identifier and should be curated in and accessible from a trusted repository.

36. The main responsibility for upholding this important principle of science lies with researchers themselves. However, given the onerous nature of this task in areas of data-intensive science, it is important that institutions create support processes that minimise the burden on individual scientists. It is a false dichotomy to argue that there is a choice to be made between funding provision for open data and funding more research. The practice of open data is a fundamental part of the process of doing science properly, and cannot be separated from it.
37. Responsibilities for ensuring that this principle is upheld also lie with the funders of research, who should mandate open data by researchers that they fund,[33] and by publishers of scientific work, who should require, as a condition of publication, deposition of open data that provides the evidence for a claim that is submitted for publication. Funders should also accept that the cost of curation of open data is part of the cost of doing research and should expect to fund it.[34]

National responsibilities
38. The capacities required to efficiently implement and to maximise benefit from the application of the principles set out in this accord and the responsibility to do so are not exclusively those of researchers and their institutions. They depend upon mutually supporting, systemic responsibilities and relationships that need to be embedded at every level of both national and international science systems, operating as parts of a dynamic ecology. It is also important to recognise that individual and institutional interests are not necessarily identical to the interests of the scientific process or to national interests in stimulating and benefiting from open data. These issues of motivation need to be identified and addressed. Box 9 shows relationships between the two key elements of national infrastructure for open data, the hard technologies and the soft relationships and responsibilities (based on Deetjen, U., E. T. Meyer and R. Schroeder (2015), "Big Data for Advancing Dementia Research: An Evaluation of Data Sharing Practices in Research on Age-related Neurodegenerative Diseases", OECD Digital Economy Papers, No. 246, OECD Publishing. http://dx.doi.org/10.1787/5js4sbddf7jk-en).

39. We characterise responsibilities and relationships as follows:

Publicly funded scientists should recognise that the essential contribution to society of publicly funded research is to generate and communicate knowledge, and that open data practices are essential to its credibility and utility. This latter requirement poses two problems of motivation:

- preparing data and metadata in a way that would satisfy the criteria of "intelligent openness" is costly in time and effort;
- data are regarded by many as "their" data, and as a resource which they are able to draw on for successive publications that are conventional indices of personal productivity, sources of recognition and grist for promotion.

Universities and Research Institutes have a responsibility to address the above motivational issues by:

- providing support that minimises the burden of compliance for individual researchers and allows them to focus less on process and more on research;
- developing processes of advancement and recognition that recognise and reward open data activities, with the need to ensure broad commonality at international level so as not to inhibit researcher mobility.

They also need to provide a managed environment to train researchers in big data and linked data analytics and in open data management, to provide expert support in these areas, and to manage open data processes.

Institutional Libraries have a continuing role to collect, to organize, to preserve knowledge, and to make it accessible. Many are now adapting to the technological change from paper to digital formats and to the open data management issues highlighted by this accord, but it is a major and difficult transition that requires sustained effort.
Funders of Research and Research Institutions have a responsibility to promote and enable open data processes by funding relevant hard and soft infrastructure; by stimulating research on fundamentals of data science; and by creating incentives for research performing institutions that help them to exercise their responsibilities and accepting that the cost of open data is an inseparable cost of doing research.

Governments hold data that are of great value to the scientific enterprise if made open, particularly in the social sciences, in addition to the broader societal value that they may create. Governments should also express broad national policies and objectives that are important in providing a frame for national efforts in developing an open data environment and system priorities, though they should not prescribe how they should be delivered.

National Academies and Learned Societies are distinctive in speaking to scientists directly without institutional intermediaries and influencing "bottom-up" initiatives by expressing the principles and priorities of research in their specific fields. They should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.

33 See the comprehensive survey of funder data policies: Hodson and Molloy (2014) *Current Best Practice for Research Data Management Policies* http://dx.doi.org/10.5281/zenodo.27872

34 See for example the RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/; see also the discussion of policy positions on the costs of RDM in Hodson and Molloy (2014) http://dx.doi.org/10.5281/zenodo.27872 pp. 11 – 12.

40. Ensuring a sustainable data infrastructure (including the management systems, standards, procedures and analysis tools for what is often called 'live' or 'active' data and the infrastructure of 'Trusted Digital Repositories' – TDRs – for long term curation of valuable data) is a core responsibility of research funders and research performing organisations (see below, para. 62–64). As emphasised above, it is a false dichotomy to argue that there is a choice to be made between funding provision for open data and funding more research. The practice of open data is a fundamental part of the process of doing science properly, and cannot be separated from it. Data infrastructure forms an essential tool for science, as necessary as networked and high performance computers, access to high quality scientific literature, in vitro labs and organic or inorganic samples.

International responsibilities

41. International science organisations play an important role in establishing principles and encouraging practices to ensure the worldwide adoption of "open data" and "open science" regimes to maintain the rigour of scientific processes and take advantage of the data revolution. Many have already developed their own data principles or protocols, as noted above. They can also help ensure that some of the most influential stakeholders are mobilised. The most effective examples of open data transformations have occurred when individual research communities, including funders, learned societies or international scientific unions, journals and major research performing organisations have endorsed community principles for open data sharing. Those established for the international genomics community are the most well known and successful, but there are others.[35]

42. It is a responsibility of the international science community to ensure that as far as possible, the capacities and the means to take up the big data and open data challenges are developed in all countries, irrespective of national income. It is for this reason that Science International and its parent bodies collaborate with low- and middle-income countries in capacity building programmes. In order to minimise such a knowledge divide, and resulting fragmentation, CODATA in collaboration with the RDA has organised relevant training workshops,[36] and Science International is currently discussing the possibility of launching a major big data/open data capacity mobilisation exercise for low- and middle-income countries, starting with an initiative in Africa. The rationale for this initiative is the danger that if a low income country has little capacity in modern data handling, its own data resources are likely either to be kept behind closed doors to protect it from foreign exploitation or, if open, to be exploited by such groups without reciprocal benefit to the host. If national capacities are mobilised, not only is a country able to exploit its own national data resources but also those that are available internationally.

43. Transformative initiatives, however resoundingly endorsed in principle, will be ineffective without investment in education and skills. The need to inculcate the ethos of Open Science outlined above and to develop data science and data handling skills for researchers is widely recognised.[37] Additionally, there are well-documented calls to develop skills and career paths for the various data-related professions that are essential to research institutions in a data-intensive age: these include data analysts, data managers, data curators and data librarians.[38]

Scientific publishers

44. Publishers of research papers that present scientific claims should require the evidential data to be concurrently made intelligently open in a trusted data repository. It is a fundamental principle of transparency and reproducibility in research that the data underlying a claim should be accessible for testing[39]. A model for good practice can be found in the Joint Data Archiving Policy that underpins the role of the Dryad Data Repository[40]. Journal editors, editorial boards, learned societies and journal publishers share responsibility to ensure such principles are adopted and implemented. Data infrastructure, comprising specialist, generic data archives and institutional data repositories which support these practices are now emerging in national jurisdictions and some international programmes[41]. The international science community should promote worldwide capability in these areas. Furthermore, journal publishers and

35 See the summary of genomics data sharing agreements at http://www.genome.gov/page.cfm?pageID=10506537; there is longstanding but far from comprehensive data sharing in the astronomical and geophysical sciences as well as in the social sciences; crystallographers successfully publish final, 'science ready' data using the CIF standard http://www.iucr.org/resources/cif

36 See the CODATA-RDA Research Data Science 'Summer Schools' or short courses http://www.codata.org/working-groups/research-data-science-summer-schools

37 The CODATA-RDA Research Data Science courses start from the premise that 'Contemporary research — particularly when addressing the most significant, transdisciplinary research challenges — cannot effectively be done without a range of skills relating to data. This includes the principles and practice of Open Science and research data management and curation, the use of a range of data platforms and infrastructures, large scale analysis, statistics, visualisation and modelling techniques, software development and annotation, etc, etc. The ensemble of these skills, we define as 'Research Data Science'.'

38 See for example the ANDS page on 'Data Librarians' http://ands.org.au/guides/dmframework/dmskills-information.html and the Harvard 'Data Science Training for Librarians' http://altbibl.io/dst4l/

39 The Royal Society's 'Science as an Open Enterprise' report stated: '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.' Royal Society 2012, p. 7.

40 Joint Data Archiving Policy (JDAP): 'This journal requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as GenBank, TreeBASE, Dryad, or the Knowledge Network for Biocomplexity.' http://datadryad.org/pages/jdaphttp://datadryad.org/pages/jdap

41 For example, the Pangaea data archive has bidirectional linking between datasets and articles in Elsevier journals. Dryad, FigShare and now Mendeley provide repositories for data underlying journal articles. In addition to specialist, discipline specific repositories, the generic repositories like FigShare and Zenodo provides places where researchers can deposit datasets. An increasing number of research institutions are providing repositories for data outputs of research conducted in the institution.

BOX 8

The infrastructure requirements for an efficient open data environment. Technology is only a part. The vital, submerged elements relate to processes, organisation and personal skills, motivation and ethos.

editors have increasingly realised that providing direct access to the data, sometimes with visualisation, increases the appeal of the journal[42]. It is not however sufficient for data to be accessible only as poorly described 'supplementary materials' provided in formats that hamper reuse. Data that directly support research articles should not lie behind a paywall. As the OECD Principles and Guidelines on Access to Research Data from Public Funding make clear, it is not legitimate for purely commercial reasons to close access to those data which have been gathered with the support of public funds and those which support published research findings.[43] However, it can be legitimate for repositories to monetise data products for which there has been considerable value-adding investment in order, for example, to present useful and reliable reference data for researchers.

The boundaries of openness

45. Openness as defined above should be the default position for scientific data although there are proportional exceptions for cases of legitimate commercial exploitation, privacy and confidentiality, and safety and security. Not all data should be made available and there are well-recognised reasons when this is the case. However, it should be recognised that open release of data is the default, such that the exceptions listed must not be used to justify blanket exceptions to openness. Rather, as it is difficult to draw sharp, general boundaries for each of these cases, they should be applied with discrimination on a case-by-case basis. Important considerations at these boundaries include:

Commercial interests

46. There can be a public interest in the commercialisation of scientific discovery where that is the route to the greatest public benefit in the national jurisdiction in which the discovery is made. The case for long-term suppression of data release on commercial grounds is weak however. Patenting is a means of protecting intellectual property whilst permitting release of important scientific data. Demands for confidentiality from commercial partners may exercise a chilling effect on swathes of research activity and the openness that should characterise it. There have been many major discoveries where suppression of data release or the privatisation of knowledge would have been highly retrograde, such as the discovery of electricity, the human genetic code, the internet etc. Difficult and potentially contentious issues include: where there has been a public/private partnership in investing in a scientific discovery; where the contribution of a private contributor should not be automatically assumed to negate openness; where commercial activities carry externalities that influence societal individual wellbeing; and where the data supporting a risk analysis should be made public.

Privacy and confidentiality

47. The sharing of datasets containing personal information is of critical importance for research in many areas of the medical and social sciences, but poses challenges for information governance and the protection of confidentiality. There can be a strong public interest in managed openness in many such cases provided it is performed under an appropriate governance framework. This framework must adapt to the fact that other than in cases where the range of data is very limited, complete anonymisation of personal records in databases is impossible. In some cases, consent for data release can be appropriate. Where this is not possible, an effective way of dealing with such issues is through what are sometimes called "safe havens", where data are kept physically secure, and only made available to bona fide researchers, with legal sanctions against unauthorised release.[44]

Safety and security

48. Careful scrutiny of the boundaries of openness is important where research could in principle be misused to threaten security, public safety or health. It is important in such cases to take a balanced and proportionate approach rather than a blanket prohibition. Scientific discoveries often have potential dual uses — for benefit or for harm. However, cases where national security concerns are sufficient to warrant wholesale refusal to publish datasets are rare.[45] and cultural choice whether to encourage or obstruct its pursuit.

Enabling practices

Timeliness of data release

49. Data should be released into the public domain as soon as possible after their creation. Data that underpin a scientific claim should be released into the public domain concurrently with the publication of the claim. Where research projects have created datasets with significant reuse value, and particularly when such projects are publicly funded, the data outputs should also be released as soon as possible.[46] Recognising the effort involved in data creation and the intellectual capital invested, the policies of some funders allow public release to be delayed for precisely limited periods, allowing data creators privileged access to exploit the asset. In contrast, however, the genomics community has demonstrated the benefits of immediate data release.[47] It is important to evaluate the benefits of immediate release in other research domains.

Non-restrictive re-use

50. Research data should be dedicated to the public domain by legal means that provide certainty to the users of the right of their re-use, re-dissemination and, for cases where research is conducted over multiple datasets, their "legal interoperability".[48] This can be accomplished by a variety of means, either broadly, as a governmental agreement, statute or policy, or as a narrow waiver of rights or a non-restrictive license that applies to a specific database or data product on a voluntary basis. The RDA-CODATA Interest Group on Legal Interoperability of Research Data has produced Principles and Implementation Guidelines that are currently in review.[49]

42 Both FigShare http://figshare.com/blog/figshare_partners_with_Open_Access_mega_journal_publisher_PLOS/68 and Dryad now provide 'widgets' which allow simple visualisations of data associated with a given article. Nevertheless, the so-called 'article of the future' is taking quite a long time to become a reality in the present … (e. g. see http://scholarlykitchen.sspnet.org/2009/07/21/the-article-of-the-future-lipstick-on-a-pig/)

43 See OECD Principles and Guidelines for Access to Research Data from Public Funding http://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm and other statements of principle like the RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/; Uhlir, Paul for CODATA (2015) marshals evidence to demonstrate that greater economic benefits and return on public investment are achieved through open data that through charging regimes designed to recover costs of data distribution.

44 See, e.g. the workshop and report on data safe havens from the Academy of Medical Sciences http://www.acmedsci.ac.uk/policy/policy-projects/data-in-safe-havens/; see also the UKDA Secure Data Service https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab; and the restricted use data held by ICPSR http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/

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

46 These categories of research data to be shared are identified, for example, in the EC's Horizon 2020 Data Policy, see Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p. 10; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot- guide_en.pdf

47 See the summary of data release policies in the genomics field at http://www.genome.gov/page.cfm?pageID=10506537 and the more general discussion and summary of period of privileged access in Hodson and Molloy (2014) Current Best Practice for Research Data Management Policies, p. 18 http://dx.doi.org/10.5281/zenodo.27872

48 "Legal interoperability occurs among multiple datasets when: i) use conditions are clearly and readily determinable for each of the datasets; ii) the legal use conditions imposed on each dataset allow creation and use of combined or derivative products; and, iii) users may legally access and use each dataset without seeking authorization from data rights holders on a case-by-case basis, assuming that the accumulated conditions of use for each and all of the datasets are met." Definition provided in GEO (2014) *White Paper: Mechanisms to Share Data as Part of the GEOSS Data-CORE. Data Sharing Working Group.* Available at: https://www.earthobservations.org/documents/dswg/Annex%20VI%20-%20%20Mechanisms%20to%20share%20data%20as%20part%20of%20GEOSS%20Data_CORE.pdf

49 The final Implementation Guidelines for the Principles on the Legal Interoperability of Research Data developed by the CODATA-RDA Group will be released in March 2016 following community review.

51. The broadest approach to placing research data in the public domain is to develop and use a convention or executive agreement at the international level, or legislation or executive policies at the national level. For example, the U.S. federal government excludes all information produced within its ambit from copyright protection under the 1977 Copyright Law. Different ministries or research agencies may adopt a policy that allows research data produced through their funding to be placed in the public domain. Because it is more difficult to agree to such far-reaching exemptions from intellectual property protection, the rights holder also may expressly state on a voluntary basis that the data are in the public domain.

52. In the absence of a broad law that enables the re-use, re-dissemination and legal interoperability of data, a voluntary rights waiver or a non-restrictive, "common-use" licence can be used by the rights holder (see: www.creativecommons.org). If a non-restrictive license is used, it should make it clear that the data may be reused with no more arduous requirement than that of acknowledging the original producer of the data. It is good practice to use a public domain waiver of rights (e.g. CC0) or non-restrictive licence (such as CC-BY). The license requires nothing more than that the producer of the data is acknowledged. Imposing further restrictions against commercial use defeats the objectives of open data and the dedication of those data to the public.[50]

53. Although the use of an attribution-only (CC-BY) license may be appropriate in some circumstances, the challenges associated with providing recognition to the generators of datasets integrated into complex data products, a phenomenon of data-intensive research, means that many authorities argue that licences such as CC-BY that require attribution are not sustainable or appropriate in a Big Data age.[51]

Citation and provenance

54. When used in scholarly communication, research data must be cited with reference to specific information and a permanent digital identifier[52]. The information attached to the citation and the identifier must allow the provenance of the data to be assessed. The practice of citing data in scholarly discourse is important for two reasons. First, citing sources is essential to the practice of evidence-based reasoning and distinguishes scientific texts from other writing. Second, 'citations' are one of the metrics by which research contributions are assessed. Although not without flaws and subject to possible gaming, article-level citation metrics are the "least bad" means of measuring research contribution and are without doubt an improvement on journal level impact factors.[53]

55. It would be naïve to pretend that citation is not an important component of the system of academic recognition and reward. Therefore, integrating the practice of data citation must be seen as an important step in providing incentives for 'data sharing'.

56. Citations also provide essential information – metadata – that allow the data to be retrieved. A permanent digital identifier (for example, a Digital Object Identifier issued by the DataCite organisation)[54] allows other researchers to determine without ambiguity that the data in question were indeed those which underpin the scientific claim at issue. This is particularly important when dynamically created subsets or specific versions of time-series datasets may be at issue.[55]

57. Additional metadata is necessary to determine the provenance of the data and to understand the circumstances in which they were created and in what way they may be reused. Standards exist in most research disciplines for the way in which data should be described and the circumstances of their creation reported.[56]

Text and data mining

58. The historical record of scientific discovery and analysis published in scientific journals should be accessible to text and data mining (TDM). At the very least, this should be at no additional cost by scientists from journals to which their institution already subscribes, though there is a case for broader access to the corpus of scientific literature for TDM. The importance for science lies in the unprecedented capacity offered by text and data mining to harvest the cumulative scientific knowledge of a phenomenon from already published work. TDM has the potential to greatly enhance innovation. It can lead to an exponential increase in the progress of the rate of discovery, such as when facilitating the discovery of cures for serious diseases.

59. The Hague Declaration on Knowledge Discovery in the Digital Age[57], lays out the scientific and ethical rationale for the untrammelled freedom to deploy TDM in order to analyse scientific literature at scale. The Hague Declaration asserts that 'Intellectual property was not designed to regulate the free flow of facts, data and ideas, but has as a key objective the promotion of research activity'. In the digital age, the benefits of TDM are vast and necessary in order to support systematic review of the literature through machine analysis. Publisher resistance to TDM on the grounds of defending intellectual property are weak in the light of a skewed business model in which scientists sign copyright transfer agreements, make up journals' editorial boards and reviewer cohorts at no cost to the publisher, whilst scientists then pay to publish, and institutions pay for electronic copies of journals. There has been strong academic criticism of commercial publishers of research for claimed restrictive business practices and excessive profits[58].

Interoperability

60. Research data, and the metadata which allow them to be assessed and reused, should be interoperable to the greatest degree possible. Interoperability may be defined as the 'property of a product or system … to work with other products or systems, present or future, without any restricted access or implementation.'[59] Interoperability is an attribute that greatly facilitates usability of research data. For example, semantic interoperability depends on shared and unambiguous properties and vocabulary, to which data refer, allowing comparison or integration at scale.

61. In relation to data, interoperability implies a number of attributes. These include the following:

- The encodings should be open and non-proprietary and there should be ready sources of reference, of a high quality, that allow the data to be ingested to other systems.
- The values which the data represent should use units describing properties for which there are standardised definitions.
- Standardised ontologies that are a key to interoperability.
- Metadata, particularly those reporting how the data were created

50 The DCC Guide 'How to Licence Research Data' is a very useful resource on this issue http://www.dcc.ac.uk/resources/how-guides/license-research-data

51 Carroll MW (2015) Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol 13(8): e1002235. doi:10.1371/journal.pbio.1002235

52 See the Joint Declaration of Data Citation Principles https://www.force11.org/group/joint-declaration-data-citation-principles-final.

53 For example: Arnold, Douglas N. and Kristine K. Fowler. "Nefarious Numbers." Notices of the American Mathematical Society v. 58, no. 3 (March 2011): 434 – 437. http://www.ams.org/notices/201103/rtx110300434p.pdf
Carlton M. Caves, "High-impact-factor Syndrome", APSNEWS November 2014 · Vol. 23, No. 10, http://aps.org/publications/apsnews/201411/backpage.cfm

54 https://www.datacite.org/

55 Research Data Alliance Working Group on Data Citation: https://rd-alliance.org/filedepot/folder/262?fid=667

56 See the RDA Metadata Standards Directory http://rd-alliance.github.io/metadata-directory/ building on work by the UK's Digital Curation Centre http://www.dcc.ac.uk/resources/metadata-standards; and the BioSharing catalogue of standards https://www.biosharing.org/standards/

57 The Hague Declaration on Knowledge Discovery in the Digital Age http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/

58 Harvie, D., Lightfoot, G., Lilley, S. and Weir, K. 2014. Publisher be damned! From price gouging to the open road. Prometheus: Critical Studies in Innovation. Vol. 31, No. 3, 229–239, http://dx.doi.org/10.1080/08109028.2014.891710

59 See http://interoperability-definition.info/en and the characteristics of the properties should use, where possible, accepted standards.

Sustainable data deposition

62. To ensure long-term stewardship in a sustainable data infrastructure, research data should be deposited in trusted digital repositories (TDR).[60] A TDR has the following attributes:

- an explicit mission to provide access to data and to preserve them in a defined area of competency;
- expertise and practices that conform to the principles laid out above;
- responsibility for long-term preservation and management of this function in a planned and documented way;
- an appropriate business model and funding streams to ensure sustainability in foreseeable circumstances;
- a continuity plan to ensure ongoing access to and preservation of its holdings in the case of wind-down.

63. Most trusted digital repositories cater for well-defined research disciplines, providing an appropriate and efficient focus of effort. However, the scale of the challenges and opportunities are such that multi-disciplinary repositories are emerging and research-performing institutions need also to provide TDRs to manage their research data outputs.

64. Research funders and national infrastructure providers have an obligation to ensure that an ecology of TDRs functions on a sustainable footing. This involves some serious rethinking of business and funding models for these essential but often undervalued elements of the research infrastructure.

Incentives

65. Actions that encourage appropriate open data practices fall into three categories – those that encourage researchers to make data open, those that encourage the use of open data, and those that discourage closed data practices. The potential roles of four key actors need to be considered – research funders, institutions, publishers and researchers themselves. These actors are the key elements of the research community. They need to work together to ensure that data are considered legitimate, citable products of research; with data citations being accorded the same importance in the scholarly record as citations of other research objects, such as publications[61].

66. A developing method for researchers to gain credit for their data activities is through the formal publication and then citation of datasets, often via the route of a peer-reviewed data paper. There are a growing number of journals which either focus on publishing data papers, or have data papers as one of the article types within the journal.[62] These published datasets can then be formally cited within a research paper that makes use of the data, allowing the use and impact of the datasets to be tracked and rewarded in the same way as research papers. Many specialised data repositories – as well as the new multi-disciplinary data repository infrastructures, such as Dryad,[63] Figshare[64] and Zenodo,[65] which place particular emphasis on this feature – provide digital object identifiers (DOIs) for datasets they hold, which can then be referenced when the data are reused, providing credit for the data provider.

67. Institutions, especially funders, can reward data sharing by refining their research assessment analyses and other impact assessments, including those related to tenure and promotion, to include recognition of the considerable contribution to research of making data available for reuse.

68. By providing dedicated funding lines to support the reuse of open data, funders can start to encourage researchers to begin to unlock the value within open data. For example, the UK's Economic and Social Research Council is supporting a Secondary Data Analysis Initiative[66] which aims to deliver high-quality, high-impact research through the deeper exploitation of major data resources created by the ESRC and other agencies. Such dedicated funding can help facilitate the development of a re-use culture within research communities.

69. Journals have a key role in ensuring that researchers make their data open, by requiring that the data that underpin the research are openly available for others, and that research papers include statements on access to the underlying research materials. Major publishers, such as PLoS and Nature now have formal data policies in place, and many publishers are actively considering how to ensure that data availability becomes a mandatory part of the publication workflow. [67]

70. It is now common for research funders to have policies that require data arising from the research they fund to be made openly available where practical.[68] What is currently less common is for funders to monitor the adherence to their policies and to sanction researchers who do not comply. However, some funders are now starting to address this issue.[69]

60 See the foundational work done by OCLC on 'Attributes of Trusted Digital Repositories' http://www.oclc.org/research/activities/trustedrep.html. The Data Seal of Approval http://datasealofapproval.org/en/ and the ICSU World Data System's certification procedure https://www.icsu-wds.org/services/certification each offer lightweight and basic approaches to assessment of trusted digital repositories. More in-depth accreditation is offered by DIN 31644 — Criteria for trustworthy digital archives http://www.din.de/en/getting-involved/standards-committees/nabd/standards/wdc-beuth:din21:147058907 and ISO 16363 — Audit and certification of trustworthy digital repositories http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

61 See the Joint Declaration of Data Citation Principles (ref Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/datacitation]).

62 Examples include: Nature Scientific Data, CODATA Data Science Journal, Wiley — Geoscience Data Journal, Ubiquity Press Metajournals like the Journal of Open Archaeology Data http://openarchaeologydata.metajnl.com/ and the Journal of Open Research Software http://openresearchsoftware.metajnl.com/

63 http://datadryad.org/

64 http://figshare.com/

65 https://zenodo.org/

66 http://www.esrc.ac.uk/research/our-research/secondary-data-analysis-initiative/

67 See: http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/ and http://www.nature.com/authors/policies/availability.html

68 For example, in the UK see http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

69 For example EPSRC dipstick testing — https://www.jisc.ac.uk/guides/meeting-the-requirements-of-the-EPSRC-research-data-policy

Appendix 1: Working Group members

Geoffrey Boulton, Regius Professor of Geology Emeritus in the University of Edinburgh and President of the Committee on Data for Science and Technology. Working Group Chair.

Dr. Dominique Babini, Coordinator of the Latin American Council of Social Sciences Open Access Program (ISSC representative).

Dr. Simon Hodson, Executive Director of the Committee on Data for Science and Technology (ICSU representative).

Dr. Jianhui LI, Assistant Director General of the Computer Network Information Centre, Chinese Academy of Sciences (IAP representative).

Professor Tshilidzi Marwala, Deputy Vice Chancellor for Research, University of Johannesburg (TWAS representative).

Professor Maria G. N. Musoke, University Librarian of Makerere University, Uganda, and Professor of Information Sciences (IAP representative).

Dr. Paul F. Uhlir, Scholar, US National Academy of Sciences, and Consultant, Data Policy and Management (IAP representative).

Professor Sally Wyatt, Professor of Digital Cultures in Development, Maastricht University, & Programme Leader of the eHumanities Group, Royal Netherlands Academy of Arts and Sciences (ISSC representative).

Appendix 2: Statements and reports on open data

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003).
Available at: http://openaccess.mpg.de/Berlin-Declaration.

Bermuda Principles. *Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing.* Human Genome Organization (1996). Available at: http://www.casimir.org.uk/storyfiles/64.0.summary_of_bermuda_principles.pdf.

Bouchout Declaration. *Bouchout Declaration for Open Biodiversity Knowledge Management.* Plazi (2014).
Available at: http://bouchoutdeclaration.org/.

Bromley, D. Allen. Principles on Full and Open Access to "Global Change" Data, Policy Statements on Data Management for Global Change Research. Office of Science and Technology Policy (1991).

Carroll, MW. Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol 13(8) (2015.

CODATA. *Nairobi Principles on Data Sharing for Science and Development in Developing Countries.* CODATA (2014). Available at: https://rd-alliance.org/sites/default/files/attachment/NairobiDataSharingPrinciples.pdf.

G8. *Open Data Charter* (2013). Available at:
https://www.gov.uk/government/publications/open-data-charter.

Group on Earth Observations. *Implementation Guidelines for the GEOSS Data Sharing Principles.* GEO VI, Document 7, Rev. 2 (17 – 18 November 2009). Available at: http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf
GEOSS Data Sharing Principles Post-2015. Data Sharing Working Group (2014). Available at: https://www.earthobservations.org/documents/dswg/Annex%20III%20-%20GEOSS%20Data%20Sharing%20Principles%20Post-2015.pdf.

The Hague Declaration on Knowledge Discovery in the Digital Age. LIBER (2014). Available at: http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/.

Hodson, Simon and Molloy, Laura, for CODATA (2014) *Current Best Practice for Research Data Management Policies* (commissioned from CODATA by the Danish e-Infrastructure Cooperation and the Danish Digital Library) http://dx.doi.org/10.5281/zenodo.27872

ICSU-CODATA Ad Hoc Group on Data and Information, Access to databases: A set of principles for science in the internet era (June 2000). Available at: http://www.icsu.org/publications/icsu-position-statements/access-to-databases/.

Interacademies Panel, IAP Statement on Access to Scientific Information (2002). Available at: http://www.interacademies.net/10878/13916.aspx.

Organisation for Economic Co-operation and Development (OECD). *Principles and Guidelines for Access to Research Data from Public Funding.* OECD (2007). Available at:
http://www.oecd-ilibrary.org/content/book/9789264034020-en-fr.

Recommendation on Public Sector Information. OECD (2008).
Available at: http://www.oecd.org/sti/44384673.pdf.

Open Access Directory (2015). Available at:
http://oad.simmons.edu/oadwiki/Declarations_in_support_of_OA.

RECODE Project. *Policy Guidelines for Open Access and Data Dissemination and Preservation.* European Commission (2015). Available at: http://recodeproject.eu/wp-content/uploads/2015/02/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf.

The Royal Society (2012). *Science as an Open Enterprise.* The Royal Society Policy Centre Report, 02/12. https://royalsociety.org/topics.../science...enterprise/report/

Uhlir, Paul for CODATA (2015) *The Value of Open Data Sharing: A White Paper for the Group on Earth Observations*
http://dx.doi.org/10.5281/zenodo.33830

A summary version of this accord can be found at
http://www.science-international.org

Imprint

Suggested citation:
Science International (2015): Open Data in a Big Data World.
Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP)

Science International
www.science-international.org

Cover photo: NASA

Design: Curie Kure, Hamburg
www.curiekure.de

International Council for Science

the interacademy partnership

international social science council

www.icsu.org
www.interacademies.net
www.worldsocialscience.org
www.twas.org