INTERNATIONAL ACCORD ON OPEN DATA FOR OPEN SCIENCE

Preface

*The International Council for Science (ICSU), the Inter-Academy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC) have created a joint enterprise**, Science International**, to be the global science community's voice of policy for science. This accord is its first foray in that domain. The accord identifies the challenges and opportunities of the global data revolution as the predominant issue of current policy for science. It wishes to add 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 if science is to maximise its 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 principles of open data as set out in the following document.*

Science International partners will promote discussion and adoption of these principles by their respective members and by other representative bodies of science at national, regional and international levels. The realisation of these principles in practice is being promoted strategically within several national research systems, in some disciplinary fields and by international bodies such as the ICSU Committee on Data for Science and Technology (CODATA) on practice and policy in national research systems the ICSU World Data System (WDS) on database services and the Research Data Alliance (RDA) on data interoperability.

A. The Scientific Challenges of a Data-Intensive (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]. 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

[1] We use the term **data** to refer to a set of values of qualitative or quantitative variables in which each value represents a piece of information. The vast and growing corpus of data are preponderantly in digital form and represented or coded in ways that permit them to be processed in order to reveal higher order patterns of information. Digital data can be "born digital" or be a product of digitization of data from some other format (e.g. printed text, painted images, three-dimensional objects). The arts and humanities increasingly use digital data, but the highly interpretive nature of much of their work is at odds with the ethos of data as "given". Many such phenomena are not discrete or observer-independent, so that the term data is considered by some to be inappropriate. The term *capta* (from the Latin *capere*, "to take") has been suggested in such cases, which emphasizes the act of observation as the essential element.

science[2], permitting patterns in phenomena to be identified that have hitherto lain beyond our horizon and to demonstrate hitherto unsuspected relationships.

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 world of "big data" is now the driving force of the data revolution. It can be characterised by the four Vs[3]: 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 that streams in and out of systems in real time; and veracity (referring to the uncertainty due to bias, noise or abnormality in data). The peer review of results based on big data poses severe problems for effective scrutiny, with a clear need to establish a "reproducibility standard."

3. Such data collecting capacity, when coupled with great processing power, permits machines 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. Scientists were amongst the first and most pervasive users of digital networks such that many areas of research across the natural and social sciences are being transformed, or have the potential to be transformed, by access to and analysis of such data.

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 now, the analysis of big data adds to this the capacity to characterise and describe complexity in great detail. Coupling these two approaches to the understanding 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 that nonetheless remain powerful tools in the armoury of science.

5. 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 broad-data lie far beyond the analytical power of these methods, such that we now need to move on from them in

---

[2] 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.

[3] www.ibmbigdatahub.com/infographic/four-vs-big-data

adapting topological and related methods to their analysis in ensuring that inferences drawn from big 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 profound implications for discovery, but also for the world of work and for what it means to be a researcher in the 21st century. It poses profound questions about the potential disconnect between machine analysis and human cognition.

B. Responding to the Challenge: the Open Data Imperative

Maintaining "self-correction"

6. Openness has been the bedrock on which the progress of science in the modern era has been based. It has permitted the logic connecting evidence (the data) and the claims derived from it to be scrutinised, and the replicability of observations or experiments to be tested, thereby supporting or invalidating those claims – a principle that has been termed "self-correction". Big data, and data-intensive science, challenge this vital principle through the sheer complexity of making data available in a form that is readily subject to rigorous scrutiny. Open 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 complex datasets, 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 their editorial boards 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 is 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. It must be "**intelligently open**"[4], meaning that it can be thoroughly scrutinised and appropriately re-used. The following criteria should be satisfied: data must be ***discoverable*** - a web search can readily reveal its existence; ***accessible*** **–** the data can be electronically imported into 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 useable), 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,

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

authentic, and of scientific relevance. For longitudinal datasets, the metadata must 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.

Openness: the default position for publicly funded research

9. 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, that the fundamental role of the publicly-funded scientific enterprise is to add to the stock of knowledge and understanding, and therefore 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.

10. If this is social revolution in science is to be achieved, it is not only a matter of making data that underpins a scientific claim intelligently open, but one of making all publicly funded data open. In some disciplinary communities data is released into the public domain immediately it has been created, such as in the case of genome sequencing data and based on the 1996 Bermuda Principles[5]. The circumstance and timescale of release are important. Data that are collected during the period of a research grant should not be expected to be released until the termination of the grant, and even then the grant holders should have an opportunity to have a first bite of the publication cherry before data release. Although it is tempting to suggest a witholding period, perhaps of the order of a year, it would be better for individual disciplines to develop their procedures that are sympathetic to disciplinary exigencies, but without involving excessive delay.

Boundaries of openness

11. Open data should be the default position for publicly funded research data. However, not all data can 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 should constrain the way that data systems operate and data are used. 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 openness are also possible in

---

. [5] Human Genome Project (2003). Available at: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1

these cases. These categories should not however be used as the basis for blanket exceptions, which should be applied on a case by case basis, with the onus on a proponent to demonstrate a specific reason for an exception to the default.

Changing the dynamic

12. 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.

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

14. The rationale for a national open data policy lies in ensuring the rigour of its own science based on its reproducibility and the accessibility of its results, in capturing the value of open data[6] for national benefit and as an efficient collaborator in international science in a data-rich era. 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. 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 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 by both research institutions and researchers. By contrast, the physical infrastructure required to sustain data storage, analysis, broadband transmission and long-term preservation is not separable from that required to support a strong science base.

15. Although many well-funded national science systems are adapting rapidly to seize the open data challenge, the costs of adaptation referred to above pose particular problems for science systems in low- and middle-income countries. It is important that the "knowledge divide" between them and better-funded

---

[6] 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 (www.mckinsey.com/.../open_data_unlocking_innovation_and_performance).

systems do not widen. It is particularly crucial in relation to global challenges, where global solutions, almost inevitably based on data-intensive science, will only be achieved if there is global participation. In order to minimise such a knowledge divide, CODATA in collaboration with the RDA has organised relevant training workshops, 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. A major 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. However, the issue of "fair data" is a fundamental one for the international science community.

16. The ways in which big data can be used for data-driven development and be leveraged to positively impact the lives of the most vulnerable are becoming clearer[7]. 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. However, there is the possibility of a dystopic future dominated by "digital extractive industries" that override local public interests. It is vital that fair data 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.

17. Responsibilities also fall on international bodies, such as the International Council for Science's (ICSU) Committee on Data for Science and Technology (CODATA) and World Data System (WDS), and the Research Data Alliance (RDA), 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 as far as possible compatible with national procedures. 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. For example, is a single-author article with a fixed publication date in a "high impact" journal, which plays such a key role in criteria for researcher promotion and advancement, a barrier to more creative ways of communicating science? How do we recognise and reward, and therefore incentivise, the importance of data management, preservation, curation?

18. 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 scientific communities that will determine the adoption, extent and impact of these processes. Such take-up is happening, with well-developed

[7] World Economic Forum 2015. Data Driven Development: Pathways for Progress.

processes of open data sharing in areas such as linguistics[8], bioinformatics[9] and chemical crystallography[10]. These developments are sensitive to the needs of the disciplines involved, they provide 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, and that can be adapted to local possibilities and resources.

Open Science

19. Current moves towards "Open Science" reflect a dynamic in which scientific practice is emerging more and more from behind closed laboratory doors to engage widely as a necessary public enterprise in a networked era when reliable knowledge and its effective communication are vital if global sustainability and equity are to be achieved. Open data 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 science training in serious research programmes, 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 eroding the boundary between professional and amateur scientists. 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?

20. Open scientific knowledge has generally been regarded as a public good and a fundamental basis for human judgement, innovation and the wellbeing of society. Many governments now recognise the potential power of being open with their own data holdings in order to enhance financial gain through 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 in evaluating social and economic trends and in medical sciences in evaluating optimal public health strategies from population health records. Care needs to be taken to avoid privatisation of a public resource or uncontrolled and unconsented access to personal information. There are inter-governmental initiatives to promote openness, such as the *Open Government Partnership*, which now involves 66 participating countries worldwide, the *G8 Open Data Charter* and the report to the UN Secretary-General from his Independent Advisory Group on *the Data Revolution for Sustainable Development.*

---

[8] http://www.linguistic-lod.org/llod-cloud

[9] https://www.**elixir**-europe.org/

[10] http://www.crystallography.net/

21. It is tempting to think that the boundary of open data is the boundary between the publically-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. Conversely there is great research potential if the massive data volumes that are daily captured by retail and service industries could be made available to social science researchers.

22. 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 seem sensible not to discriminate between not-for-profit and for-profit users.

23. 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. It is 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. The Principles of Open Data

24. The following principles are advocated by Science International to national science bodies and international unions. They are based on review and synthesis of the many previous policy statements on open research data and which are referenced in section D.

i. Responsibility of scientists and their institutions

Publicly funded scientists and scientific institutions 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 have a responsibility to make such data openly available to others in ways that permit them to be re-used and re-purposed.

ii. Testing Scientific Claims

The data that provide evidence for published scientific claims must be concurrently made publicly available in an intelligently open form 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.

iii. Responsibility of publishers of scientific results

Publishers of research papers that present scientific concepts should require the evidential data to be concurrently made intelligently open, if possible in a reliable data repository. Data need to be available to reviewers during the review stage.

iv. Boundaries
Open data should be the default position for publicly funded research, although there should be proportional exceptions for cases of legitimate commercial exploitation, privacy, confidentiality, safety and security.

v. Timeliness
Data should be released into the public domain as soon as possible after its creation, and no later than upon the publication of results based on the data.

vi. Legally enabling reuse (Better title?)
Research data should be made open in the public domain by means of an international agreement or national legislation or policy, or by the application of private waiver of rights or non-restrictive licences applied voluntarily by the rights holder. Users of any of these legal mechanisms should make it clear that the data may be reused with no more arduous requirement than that of acknowledging the creator.

vii. Citation and provenance
When, in scholarly communications, researchers use data created by others, they must be cited with reference to their originator, their provenance and a permanent digital identifier.

viii. Text and data mining
The historical record of scientific discovery and analysis published in scientific journals should be accessible to text and data mining (TDM) at no additional cost by scientists from journals to which their institution already subscribes.

ix. Interoperability
Research data, and the metadata which allows it to be assessed and reused, should be able to be interoperable to the greatest degree possible.

x. Sustainability
To the extent possible, research data should be deposited in managed repositories of databases that maintain data in an intelligently open form that have low access barriers and are sustainable in the long term.

xi. Incentives
Research funders and institutions should provide incentives for appropriate open data practices. Metrics of research contribution can include citation metrics, funders' research assessment analyses and other impact assessments to recognise the considerable contribution to research of making data available for reuse.

D. The Practice of Open Data

Note – this section is being substantially expanded and added to, together with references to accessible work that provides practical guidance on important processes. Note that the references referred to at the beginning of paragraph 20 have yet to be added.

25. This section expands on the rationale for above principles and consequential issues of practice that need to be addressed. They are related to the individual principles by Roman numerals.

Responsibilities (i)

National

26. The assertion that researchers and their institutions have a responsibility to be open with their data frequently conflicts with contrary pressures on both. There are two principal issues for individual researchers:

- preparing data and metadata in a way that would satisfy the criteria of "intelligent openness" is costly in time and effort;
- data is 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.

There are two principal issues for institutions:

- The costs of managing research data, of giving support to data-intensive research and of minimising the burden of metadata preparation on their researchers.
- How to motivate their researchers to make data open by giving credit for open data deposition.

27. Personal 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. The capacities required to efficiently implement and to maximise benefit from the application of the principles set out above and the responsibility to do so are not exclusively those of researchers and their institutions. They are systemic responsibilities and capacities that need to be embedded at every level of a national science system that operates as an interactive ecology as follows:

- ***Government:*** in expressing broad national policies and objectives which provide a frame for system priorities without prescribing how they should be delivered, and which, in part would be reflected by acceptance of the concordat.
- ***Funders of Research and related Strategic Bodies:*** in setting thematic priorities and creating incentives for research performing institutions.
- ***National Academies and Learned Societies***: in expressing the principles and priorities of for research in its varied fields.
- ***Universities and Research Institutes:*** in providing the immediate environment of support and management for open data/big data, in training researchers, devising incentive structures for their staff and exercising responsibility for the knowledge that they create.
- ***Researchers:*** in recognising that the essential contribution to society of publicly funded research is to generate and communicate knowledge, and that open data is essential to its credibility and utility.

28. 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. As underlined

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. 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 (ICSU/IAP/TWAS; CODATA/WDS/RDA etc)

29. International science organisations can 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, but there are others.

30. 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.

Skills and education

31. 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 skills is widely recognised. 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.

Data that is used as evidence for a scientific claim (ii)

32. 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 electronically available through a hot link in the published article or in an accessible data trusted data repository (principle x).

33. 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. Open data is a fundamental part of the process of doing science properly, and cannot be separated from it.

34. 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, and by publishers of scientific work, who should require, as a condition of publication, deposition of open data that provides the evidence for a concept that is submitted for publication.

Scientific publishers (iii)

35. Publishers of research papers that present scientific concepts 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[3.8]. A model for good practice can be found in the Joint Data Archiving Policy that underpins the role of the Dryad Data Repository[4]. 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[4.2]. The international science community should promote worldwide capability in these areas. Furthermore, journal publishers and editors have increasingly realised that providing direct access to the data, sometimes with visualisation, increases the appeal of the journal[4.4]. 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 supports research articles should not lie behind a paywall, although monetising data products that integrate or present reference data for researchers can offer useful and value-added resources. However, it is not

---

[3.8] 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.

[4] 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

[4.2] For example, the Pangaea data archive has bidirectional linking between data sets 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 data sets. An increasing number of research institutions are providing repositories for data outputs of research conducted in the institution.

[4.4] 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/)

legitimate to close access to data that has been gathered with the support of public funds and which supports published research findings[4.5].

The boundaries of openness (iv)

36. 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

37. There is 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 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, where the data supporting a risk analysis should be made public.

Privacy and confidentiality

38. 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 so-called “safe havens”, where data are kept physically secure, and only made available to *bona fide* researchers, with legal sanctions against unauthorised release.

[4.5] 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/

Safety and security

39. 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 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 (ref).

Timeliness of data release (v)

40. Data should be released into the public domain as soon as possible after its creation. This should be a matter of course for data that underpin a scientific claim and should be released into the public domain concurrently with the publication of the claim. Where research projects have created datasets with significant reuse value, particularly when such projects are publicly funded, then the data outputs should also be released (ref. e.g. H2020 data policy). 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 (ref. genomics agreements). It is important to evaluate whether such benefits of immediate release could be realised in other research domains.

Legally enabling reuse (vi)

41. 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. 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.

42. **[This paragraph will describe and reference the international and national laws.]**

43. A voluntary rights waiver or a non-restrictive, “common-use” licence can be used by the rights holder. 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 generator of the data. An inhibiting factor in data reuse can be uncertainty around the public domain status of the data and restrictions imposed. It is good practice, therefore, to use a public domain waiver (e.g. CC0) or non-restrictive licence (such as CC-BY) which requires nothing more than that the generator of the data is acknowledged. Imposing further restrictions against commercial use defeats the objectives of open data and a dedication of those data to the public.

44. Although individual data points, as facts, are not subject to copyright, ‘Anglo-Saxon’ IP law and EC legislation governing IP in databases means that in these legal regimes it is appropriate and necessary to dedicate most research datasets to the public domain through a waiver or non-restrictive licence. (Ref

Principles of CODATA-RDA Group on Legal Interoperability). 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.

Citation and provenance (vii)

45. When used in scholarly communication, research data must be cited with reference to specific information and a permanent digital identifier[5]. 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 fiction. 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[16].

46. 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 citing data must be seen as an important step in providing incentives for 'data sharing'.

47. 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)[7] allows other researchers to determine without ambiguity that the data in question was indeed that which underpins the scientific claim at issue. This is particularly important when dynamically-created subsets or specific versions of time-series data sets may be at issue[8].

48. 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[9].

Text and data mining (viii)

---

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

[6] [1] See the 'San Francisco' Declaration on Research Assessment (DORA) http://www.ascb.org/dora/

[7] https://www.datacite.org/

[8] See Ball, A. & Duke, M. (2015). 'How to Cite Datasets and Link to Publications'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides - See more at: http://www.dcc.ac.uk/resources/how-guides/cite-datasets; and Recommendations from Research Data Alliance Working Group on Data Citation: https://rd-alliance.org/filedepot/folder/262?fid=667

[9] 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/

49. The historical record of scientific discovery and analysis published in scientific journals should be accessible to text and data mining (TDM) at no additional cost by scientists from journals to which their institution already subscribes. 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.

50. The *Hague Declaration on Knowledge Discovery in the Digital Age*[10], 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 and 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.

Interoperability (ix)

51. Research data, and the metadata which allows it 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.'[11] 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.

52. 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 was created and the characteristics of the properties should use, where possible, accepted standards.

Sustainable data deposition (x)

---

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

[11] See http://interoperability-definition.info/en

53. To ensure long-term stewardship in a sustainable data infrastructure, research data should be deposited in trusted digital repositories (TDR)[12]. A TDR has the following attributes:

- an explicit mission to provide access to and preserve data in a defined area of competency;
- expertise and practices that conform to the principles laid out above;
- responsibility for long-term preservation and manages 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.

54. 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.

55. 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 undervalued elements of the research infrastructure.

Incentives (xi)

56. 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, and 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[13].

57. A developing method for researchers to gain credit for their data activities is through the formal publication and then citation of data sets, 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

---

[12] 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

[13] 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]).

the article types within the journal.[14] These published data sets can then be formally cited within a research paper that makes use of the data, allowing the use and impact of the data sets to be tracked and rewarded in the same way as research papers. Data repository infrastructures, such as Figshare.com, provide digital object identifiers (DOIs) for data sets they hold, which can then be referenced when the data are reused.

58. 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.

59. 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[15] 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.

60. 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 work flow.[16]

61. 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.[17] 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.[18]

---

[14] Examples include: Nature Scientific Data, CODATA Data Science Journal, Wiley - Geoscience Data Journal.

[15] http://www.esrc.ac.uk/research/our-research/secondary-data-analysis-initiative/

[16] 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]

[17] for example, in the UK see http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

[18] for example EPSRC dipstick testing - https://www.jisc.ac.uk/guides/meeting-the-requirements-of-the-EPSRC-research-data-policy

Appendix to follow

Working Group members

Dominique Babini (ISSC & Argentina)

Geoffrey Boulton (ICSU & UK, chair)

Simon Hodson (CODATA)

Jianhui LI (IAP & China)

Tshilidzi Marwala (TWAS & South Africa)

Maria Musoke (IAP & Uganda)

Paul Uhlir (ICSU & USA)

Sally Wyatt (ISSC & The Netherlands)