CSTP-GSF WORKSHOP: TOWARDS NEW PRINCIPLES FOR ENHANCED ACCESS TO PUBLIC DATA FOR SCIENCE, TECHNOLOGY AND INNOVATION (13 MARCH, 2018)

Draft summary of the workshop proceedings

Paul Uhlir and OECD Secretariat

Introduction

This document aims to capture the issues discussed at the workshop "Towards New Principles for Enhanced Access to Public Data for Science, Technology and Innovation", held at the OECD headquarters in Paris, on 13 March 2018. The workshop gathered ~30 experts, representing governments, the private sector, data repositories, academia, non-governmental organisations, international data networks and libraries (list of speakers in Annex 1). A similar number CSTP and GSF delegates also attended the meeting.

Background

At its 109th session in October 2016, the OECD Committee for Science and Technology Policy (CSTP) discussed and approved a proposed joint development of a possible new overarching recommendation on enhanced access to data, together with the Committee for Digital Economy Policy (CDEP) and the Public Governance Committee (PGC) [COM/DSTI/CDEP/STP/GOV/PGC(2016)1]. CSTP's central instrument in this domain is the Recommendation of the Council concerning Access to Research Data from Public Funding [C(2006)184] (referred to as "the Recommendation" in the text below). A summary of the Recommendation is provided in Table 1.

Table 1: Summary of the OECD 2006 Recommendation Concerning Access to Research Data from Public Funding

A. Openness - access on equal terms for the international research community at the lowest possible cost.
B. Flexibility - adapt to changes in IT, diversity of research systems, legal systems and cultures.
C. Transparency - information on research data to be available in a transparent way, ideally on the internet.
D. Legal conformity - respect legal rights & legitimate interests of all stakeholders in public research.
E. Protection of intellectual property - consider the applicability of copyright or of other IP laws.
F. Formal responsibility - rules and regulations regarding the responsibilities of the various parties.
G. Professionalism - establish and maintain codes of conduct to foster trust.
H. Interoperability - member countries and research institutions should co-operate with international organisations charged with developing new standards, both semantic and technological.
I. Quality - good practices in collection, dissemination and archiving to enable peer review, development of metadata, links to original datasets, data citation practice.

J. Security – protect against loss, destruction, modification and unauthorised access.
K. Efficiency – data management practices to promote cost effectiveness; cost-benefit analysis to be used to decide on maintaining datasets online; documentation of data to avoid duplication; incentives for researchers.
L. Accountability - periodic evaluation by user groups, responsible institutions and funding agencies.
M. Sustainability - administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention.

As described in the Work Plan for the Project on Enhanced Access to Data and Terms of Reference (ToR) [COM/DSTI/CDEP/STP/GOV/PGC(2017)1], the work has the objective to identify gaps in current data governance frameworks, which will enable identifying the common elements that could be further developed as general principles on enhanced access to data, possibly resulting in an OECD umbrella legal instrument adopted by Council to serve as reference for any revision of existing OECD legal instruments or for the development of new ones in other policy areas.

A survey was conducted in mid-2017 to assess the current use of the Recommendation, the results of which are summarised in 'Open Access To Data In Science, Technology And Innovation – Initial Survey Findings' [DSTI/STP(2017)25], discussed by the CSTP at its 111th Meeting in October 2017.
The key issues identified in the survey as requiring policy attention were:

- Data governance for trust - addressing privacy, confidentiality, quality and ethical issues
- Discoverability/findability, machine readability and data standards
- Recognition and reward system for data authors
- Definition of responsibility and ownership
- Business models for open data provision
- Building human capital and institutional capabilities at public agencies, to manage, create, curate and reuse data.

In parallel, the OECD Global Science Forum recently completed two projects to inform policies to promote open data for science. The first of these addresses Business models for sustainable research repositories. The second project focuses on Coordination and support of international research data networks, which are necessary to support a global open science enterprise. These two projects are a follow-up to earlier GSF work on the use of data, including personal data, for social science research . In addition some initial work has been carried out on incentives for sharing data in the specific area of research on dementia [DSTI/STP/MS(2015)16].

The CSTP thus joined forces with the Global Science Forum and organising a workshop back to back with the 112th meeting of the CSTP with the objective to deepen the gap analysis already initiated through the recent survey report [DSTI/STP(2017)25].

The remainder of this report summarizes the presentations and discussions that were organized round six thematic panels of experts. It is structured according to the topics of each panel, beginning with a description of the focus of the panel and followed by the expert comments and discussion highlights.

PANEL 1. DATA GOVERNANCE AND TRUST FOR SCIENCE, TECHNOLOGY AND INNOVATION

Data governance includes the broad range of guidelines, regulations, principles, standards of good practice and processes that determine how data is produced, collected, managed, distributed and re-used.

Sound data governance is needed to ensure trust from both data providers and users and secure accessibility. In addition to provenance and quality, for sensitive data, this requires attention to privacy, confidentiality and ethical issues, including informed consent. Balancing the potential public benefits and risks of sharing data in science, technology and innovation is a critical issue for data governance.

* * *

The context for research data has changed significantly in the past dozen years—both quantitatively and qualitatively - since the OECD Recommendation was first issued. It is still apparent, however, that explicit and formal access arrangements are fundamental conditions to pursue open research data. As open access mandates and rules have drastically increased at the international and national levels in recent years and data management plans have been institutionalized in multiple research organizations, it is relevant to consider outstanding or new policy challenges and opportunities. This could include, for instance, additional policies to motivate appropriate behaviour by researchers and streamlining diverse research funding obligations. Moreover, international cooperation through multilateral organisations, both governmental and non-governmental, and other global policy mechanisms would be desirable in data governance.

In order to secure public trust and accountability, it is worthwhile to monitor the socio-economic impacts of open research data in a more comprehensive way. Over time, such impact assessments could eventually help society evaluate the legitimacy of open data initiatives. Open data policies may in fact create new data management and policy challenges in less economically developed institutions and countries, thus enhancing the need for such socio-economic assessments. The 2006 OECD Recommendation suggested considering a few core aspects for external evaluation, including: overall public investments, management performance of data collection and the extent to which existing data sets are used and reused.

In any case, given the growth of the digitally networked data sector, socio-economic impacts of open research data ought to be more broadly assessed, although this is usually expensive and time-consuming. Furthermore, it would be necessary to detect potential risks and uncertainties associated with the digital divide that are exacerbated by data initiatives, whether open or not.

Enhanced access to and use of personal data to progress scientific discovery, technological development and to promote innovation requires trust between all agents (data providers, data holders, researchers, and the general public). 'Trust' in this sense means that all agents have a shared understanding that data will be used for the overall benefit of a population, that any possible harm arising from their use will be outweighed by potential benefits, that *consent for such use has been obtained* and that there is a clear articulation of the purpose of any research using personal data.

In some jurisdictions, privacy is considered a fundamental human right. There are other, perhaps competing, fundamental human rights, that people expect. What is the proper balance between individual rights and other societal objectives?

The issue of consent relates closely to the maintenance of the right to a private life, which requires all agents to respect the privacy of individuals. This is a right recognised in many countries and enshrined in legislation such as the European Union General Data Protection Regulation[1], which states that:

[1] Regulation 2016/679 defines 'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;
http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679

'processing shall be lawful only if and to the extent that …. the data subject has given consent to the processing of his or her personal data for one or more specific purposes'.

However, while 'informed consent' for the further processing of personal data beyond the specific purpose(s) for which the data were originally collected is generally regarded as a *sine qua non*, there are situations where consent is impossible or impractical. This was an important issue considered by a GSF Expert Group[2] convened to make recommendations regarding the ethics of using new forms of data for social and economic research. Among its many recommendations, the Expert Group stressed the need for and outlined the role of independent Ethics Review Bodies (ERBs) in the evaluation of applications for access to publicly funded personal data for research purposes. By establishing and promoting the use of ERBs, trust between all parties with an interest in the research use of personal data would be enhanced, particularly in situations where consent for research use was impractical or impossible. The Expert Group specifically recommended that:

> ***Ethics review bodies*** *should, where consent for research use of personal data is not deemed possible or would impact severely upon potential research findings, evaluate the potential risks and benefits of the proposed research. If the proposed project is deemed ethically and legally justified without obtaining consent, ethics review bodies should ensure that information is made publically [sic] available about the research and the reasons why consent is not deemed practicable and should impose conditions that minimise the risk of disclosure of identities*.

Another important set of actions to enhance trust in research results is to adopt a set of reproducibility enhancement principles. In order to facilitate reproducibility of their research results, researchers need to share the data, software, workflows and details of the computational environment in open repositories. To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code and digital artifacts upon which the results depend. Citation, adequate documentation and open licensing should be standard practice to enable credit for and reuse of shared digital scholarly objects. Scholarly journals should conduct a Reproducibility Check as part of the publication process.

An example of the importance of publicly-funded data was given by Duncan McIntyre, Assistant Secretary, Department of Prime Minister and Cabinet of Australia, regarding the Australian government's efforts to improve Australia's data landscape. They began by releasing the Public Data Policy Statement in 2015, which was a significant step in bringing about long-term transformational change. Since this Statement, Australians have seen a dramatic increase in the number of datasets publicly available. In 2016, the Australian government tasked its Productivity Commission to undertake a 12-month inquiry into Data Availability and Use. The final report contains recommendations to provide citizens with active control over their data and safely enable the sharing and release of all types of data.

Australians can benefit if data are able to be used in new ways — both by using data directly themselves, and from the innovative use of data by businesses, not-for-profits, governments and researchers that ultimately benefits society. To build the required social licence for data use, the Australian government needs to better communicate with the general public to explain what the data agenda means for them and demonstrate the benefits of data openness. They are working towards including all parts of society in this transformation so they feel confident that data are being used to create value for everyone, while managing harm and risk. Australia has done a lot of work on the open data policies, including the open-by-default option of data from public funding.

[2] See http://www.oecd-ilibrary.org/science-and-technology/research-ethics-and-new-forms-of-data-for-social-and-economic-research_5jln7vnpxs32-en

Other policy suggestions offered in this session included the following. Thematic and disciplinary data infrastructures should be provided by governments and maintained by the public sector. Governments ought to create a Chief Data Officer (CDO) position, similar to a Chief Information Officer that many countries have already established, at national level and a network of CDOs at both a national and institutional level could also then be created to promote coordination and cooperation. Categorisation, rather than classification of data is what is needed.

Incentives for structuring, sharing, and opening research databases should be promoted as well. Governments ought to also better communicate the need for data transparency. This means that peer review, unique and persistent identifiers, and other elements of research data should become established practice to consider them as legitimate research production themselves, not secondary ones. This also means that "data literacy" on a disciplinary basis should be promoted. Finally, a focus on the long tail of data, not only on "big data", needs to be pursued.

Some preliminary conclusions:

- Personal data can provide research insights for the biomedical and social sciences. However, their use may pose risks to individuals' **privacy**. Anonymization techniques are not always a guarantee. Text and data mining is not allowed in many countries due to concerns about personal privacy (as well as IPR).
- Evolution of context since 2006 — increased **scope** (including AI) and **scale** of data use (data Moore's law). New or revised rules for data governance thus need to be considered.
- **Trust**: needs to be addressed through rules and regulations (e.g., the General Data Protection Regulation in Europe) and consultations with stakeholders. Research ministries and public research funders need to better communicate to explain what the data agenda means (the Australia example).
- **Reproducibility** of research results: Good practice to make the data that support research results available after the publication of the results to support independent verification and build trust in the process.
- Other approaches for building trust include quality control, certification, greater transparency in data management, use of blockchain, and automation of processes. However, blockchain may prove to be a barrier to the 'right to be forgotten'.
- Consent cannot always be requested (for practical reasons). Strong and properly constituted ethics review boards are needed to be the guardians of good data use and to authorize, where necessary, the use of data without explicit consent.
- Degrees of openness – some data cannot be made public but can be conditionally opened to a specific group of stakeholders. Some countries are introducing open data by default (e.g., France).

PANEL 2. DATA STANDARDS, INTEROPERABILITY AND RE-USE

There is insufficient information on what data are available for and from research, and when data can be found it is not always useable. There is a need for user-friendly and widely accessible catalogues for datasets, services and standards, based on machine readable metadata and common and persistent identification mechanisms. International standards for data documentation have been developed but are not always easy to adopt and thus are variably implemented. At the same time, interoperability and common standards are essential for ongoing efforts to establish open science clouds in Europe, Australia, US (NIH), Africa.

* * *

The FAIR[3] principles were recently developed and are now being widely implemented to facilitate open access to data. The FAIR principles have helped communicate and advance the arguments in relation to what is required in order that research data may have the greatest reuse potential. However, there remains considerable effort required particularly in terms of the development of data skills, the roles of data stewards, the expansion of data infrastructure and the use of technical components, particularly identifiers and standards.

A number of reports and studies have attempted to quantify the time spent 'cleaning' data rather than using them. There is a major and fundamental need for greater effort, coordination and investment in the development of standards and vocabularies for data description and interoperability. In order that such investment is not wasted, community processes for the appraisal, adoption and use of standards need to be enhanced: That could include professional codes of conduct about how data are communicated in journals, such as the FAIR principles, to encourage the selection and improvement of standards, and other similar approaches.

Other issues that were discussed in this session included the following:

- One of the major challenges in making data reusable is communicating essential provenance information. There remains a lot we can learn and achieve from the application of the OAIS Reference Model and its principles in the way data is communicated are part of reports on scientific enquiry and scholarly communication writ large.

- Governments should consider creating an accessible national catalogue of data producing organisations (and their data assets) within each country. This would involve liaising with these organisations, connecting them and raising awareness of open research data resources and services that are available. They should also raise awareness of good practices, standards, the FAIR principles and international repositories focused on the long-term preservation of valuable data. This will encourage the organisations to take stock of their research data assets and register them. Governments could also earmark funds for research data work for data owning organisations. Applications could include funds for education and awareness raising for researchers, updating data management practices and infrastructure, cleaning and organising data to ensure that it is of good quality and is discoverable, usable, etc. Funds could also be allocated to studying the economic benefits of these repositories, with periodic evaluations.

- Governments should consider supporting data infrastructure beyond funding short-term projects or initiatives. They could also support the demand for data with funding, and assist with promotion and marketing. Most policies are focused on creating the supply (a dataset) and the assumption is that that the demand will follow. We have however found that this is not the case. Demand takes time to build and needs to be (artificially) created with incentives. This could come, for example, in the form of contests or grants for data use by SMEs.

- The importance of software to achieve the overall goals of Open Science is largely underestimated. It has become clear over the last few years that in order to serve the needs of scientists and allow science reproducibility, it is essential to associate publications, data, and software.

- From the U.S. university perspective, it is beneficial for all parties to have a clear delineation of who is responsible for which items of infrastructure, their development, implementation, and maintenance. It is desirable to have as much specificity from the funding agency as possible. Perhaps these specifics can be

[3] FAIR stands for Findability, Accessibility, Interoperability and Re-use

communicated in the award documents. In summary, policy should call for specifics of infrastructure responsibilities and costs spelled out as much as possible in award documents.

- Also in the U.S., the public access to data policies differ too much from agency to agency. There is too much room for independent, alternative interpretation. This situation creates additional burdens on the researchers and their institutions. The policies of the numerous government funding agencies should be harmonized so that researchers and institutions are dealing with a common set of requirements and guidance. There is also a desire for the required policies, practices and compliance measures to be clearly detailed and communicated to researchers.

- Researchers and their institutions need proper incentives to achieve the aim of making a greater portion of research data publicly accessible. Topics such as academic reward systems and tenure policy, eg with regard to the recognition of data publishing, ought to be addressed.

- Stakeholders in the research enterprise can work together to define impact evaluation methods with a goal of improving the recognition of a researcher who publicly shares data as a scholarly product and aids in measuring its impact. The requirements for data sharing ought to be stated and agreement reached on minimum set of standard practices across universities. These would need to be informed by different disciplines.

- It should be noted that the Research Data Alliance- (RDA) was initiated in 2013 with the goal of building the social and technical infrastructure to enable open sharing of data. The RDA has published recommendations addressing a broad range of issues related to interoperability, data citation, data catalogues, preservation of valuable data, and research data publishing.

- It is important to understand that in some cases, it is difficult for a group of researchers, funded by different agencies that focus on diverse disciplines, to comply with a same standard rule of data usage. There is an evident need for a joint effort across research communities to develop agreement among them regarding the harmonisation of data patterns (e.g., putting a pdf file on the web is sometimes considered as open data).

- It was also stated that more could be done on the demand side. A lot is being considered about the supply of research data, but not the demand. Governments, for example, could use hackathons to foster more data demand. However, the notion of data itself varies a lot across fields. While for archaeology, data could mean digitised findings, bioengineering data is machine readable. Beyond that, the software is also important and should be taken into account when deciding on the quality of data that will be publicly available.

Some preliminary conclusions:

- There is a general lack of information on what databases exist and a lack of adoption of international standards for data documentation and other technical and policy aspects of research data management.
- **Findability**: Data catalogues, or, perhaps more realistically, search engines, should be used to make data more findable. Science clouds are being established with this objective.
- **Accessibility:** As open as possible, as closed as necessary. Not only for academia, needed also for innovative businesses which create value.
- **Interoperability** has 3 aspects: semantic (scientific vocabulary), legal (rights) and technical (machine readability). In particular, interoperability **across disciplines** is a challenge, but it should not be over-specified. Technical issues with regard to interoperability are generally easier to solve than social/cultural issues.

- **Re-use**: Need to focus on machine readable metadata, to make old data understandable in the long term. The Open Archival Information System (OAIS) guidelines are a useful reference.
- **Pace**: Standard setting is a slow iterative process of **reaching agreement** about the content (not as straightforward as TCP/IP for Internet). Private sector may impose their own standards. It is slower than the technology evolution, and therefore certification alone cannot ensure cybersecurity.
- Governments should commit to supporting **data infrastructure in the long term** for sustainability, especially for preserving valuable historical datasets.
- The RDA recommendations address a broad range of issues related to interoperability of data, data citation, data catalogues, various standards and research data publishing. These should be disseminated and adopted more widely.

PANEL 3. DEFINITION OF RESPONSIBILITY AND OWNERSHIP

Issues of copyright and intellectual property over data play a big role in open data access, especially with regard to text and data mining (TDM) in the context of scientific research and innovation. Issues of ownership can complicate data sharing and re-use even amongst different public-sector actors. Cooperation with the private sector is an additional challenge to be addressed, with sensitive issues of data ownership for data created through public private partnerships or data from public research being offered on a private platform. In this respect it is worth noting the specific case of the EU, which has created an exclusive "sui generis" right for database producers to protect the investment of time, money and effort, irrespective of whether a database is innovative or not.

* * *

The legal and policy context for research data, especially from public funding, has changed substantially since 2006. Even where governments and funders are providing funding for data to be free they may not require managing the data properly to ensure that the data are kept legally open for access and re-use by anybody.

Copyright and related intellectual property (IP) laws have become stronger, broader, and longer in the last few decades, while the speed of change in research databases and other information products and technologies has greatly increased. Scientific databases, software, and other types of information advance rapidly and many become obsolete in a few years, but statutory copyright protects them for a minimum of the life of the author, plus 50 years, and the 1996 EU directive on the legal protection of databases is potentially not limited in time.

The rise of digital networks has put database and other information product producers and users in privity, which some legal observers have called "the power of the two-party deal".[4] This has resulted in the broad use of licenses and contracts to modulate copyright law, either substantially strengthening it for commercial products (e.g., through End User Licensing Agreements – EULAs) or weakening it through common-use licenses and waivers (e.g., Creative Commons, http://creativecommons.org). Private-law mechanisms, such as licenses and agreements, therefore, have largely superseded public-law instruments, such as copyright and other IP statutes, for disseminating databases and other information products, whether in the private or the public sectors.

[4] J.H. Reichman and Jonathan H. Franklin. Privately Legislated Intellectual Property Rights: Reconciling Freedom of Contract with Public Good Uses of Information. University of Pennsylvania Law Review 147:875 (1998). Available at: http://scholarship.law.upenn.edu/penn_law_review/vol147/iss4/2/.

There has been a concomitant rise in data-intensive research, which has been called "the fourth paradigm" of science (theory, observation, experiment, and now data science).[5] Many reports, including by the OECD,[6] have demonstrated both theoretically and empirically that there is significant value and benefits to providing research data from public funding as openly and freely as possible, with the least amount of restrictions on reuse, subject only to well-justified, legitimate restrictions. Thus, as proposed by the EC, research data from public funding should be "as open as possible and as closed as necessary."[7]

If data are publicly funded, there are several principles or characteristics that are helpful in guiding whether the data should be open or not. One is economic. Data and other forms of information are a "quasi-public good". Unlike a private good, a pure public good cannot be exhausted by use. In fact, it gains in value the more that it is used. It is also not excludable, in the sense that people cannot be prevented from using it. Data from publicly-funded research cannot be exhausted but can be excluded, although it is inefficient to do so.

A second characteristic is legal, which is the concept of public domain. Many data are not copyrightable or protected by statute. Publicly-funded data can be placed openly in the public domain for anyone to re-use, either through a statutory exemption to copyright (as in the U.S.) or by a waiver of rights (such as CC0), as noted above.

A third characteristic that is important to public research data is a political or policy principle. It is essential to protect and serve the public interest whenever public expenditures are used to develop public-sector products and services.

Proprietary data ownership can be one of the main impediments to realizing the potential of big data to gain important insights into human health and behaviour, and many important social problems. Even where governments and funders are promoting research data to be free for use, they need to ensure the rights that are applied will keep the data open for use by anybody.

At the same time commercial sector owners of data, the GAFAs of the world, justifiably point to the fact that owning the "digital crumbs" data is a crucial asset of their companies, enabling them to provide "for free" the services they provide. We may need new data access/ownership policies, analogous to intellectual property solutions, whereby certain data can remain proprietary for a specified period of time and then has to be made available for research proposes.

Research sector initiatives, such as the development of Artificial Intelligence, integrated global change research, or exploring complex social and environmental sustainability goals, will raise more questions for data management and analysis. Any kind of principles or guidelines that are developed now need to be flexible enough to adapt to changing contexts in the future.

[5] Tony Hey, Kristin Tolle, and Stewart Tansley. The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research (2009). Available at: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[6] OECD. Making open science a reality. OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris (2015). Available at: http://dx.doi.org/10.1787/5jrs2f963zs1-en.
OECD. Business models for sustainable research data repositories. OECD Science, Technology and Industry Policy Papers, No 47, OECD Publishing, Paris (2017). Available at: http://dx.doi.org/10.1787/302b12bb-en.

[7] European Commission

The issue of Text Data Mining (TDM) in publishing and public-private partnerships (PPPs) in research are not problematic if there is the right legal context, but there was a general view that IP laws need to be reviewed and changed, if necessary, to enable that.

The other major issue in this area is the roles and responsibilities of the different players in the research enterprise. The threshold question is to determine who is responsible in first place. Good Research Data Management (RDM) principles are needed and researchers should be supported by communities, institutions and funders to set up coherent data management plans. RDM is gaining rapidly in importance and could have a more prominent part in any OECD principles and guidelines. Good RDM practices would resolve many problems because researchers would need to think these matters through at the beginning of any research project. [Other Panels also discussed the roles and responsibilities for publicly-funded research data, especially Panels 5 and 6.]

Co-operation with the private sector is an additional challenge to be addressed, with sensitive issues of data ownership for data created through public-private partnerships (PPPs), e.g. data from public research offered on private platforms. There is a hidden but increasing risk of 'privatisation' of data arising from publicly-funded research, especially in government, since data previously openly accessible by the public may be contracted to private-sector data collectors. There might also be an opportunity to consider combining public research data with those of the private sector. After an embargo period you can open data of private sector, based on the 2006 OECD recommendation.

Some preliminary conclusions:

- Ownership and the division of responsibility for different data management functions must be addressed, particularly where institutions have created services or resources. Lack of clarity can complicate data sharing for text and data mining, in particular. It is important to identify the incentives of the owners of repositories, be they public or private.
- Legislation and other rules for managing research data should be **harmonised** as much as possible. Data custodians often operate under varying legal frameworks that govern the collection and use of research data. Research funding agencies also have differing policies (e.g., in the U.S.). Coordinated top-down and bottom-up approaches are needed. One also needs to account for a lot of data that is not copyrightable or licensable.
- **Exceptions** from copyright law are granted by some countries, but generally only for non-commercial purposes. This does not fully consider value creation goals.
- Private-law mechanisms, such as **licenses** and **agreements**, have superseded public-law instruments, such as copyright and other IP statutes, for disseminating databases and other information products, whether in the private or the public sectors.
- Good Research Data Management (RDM) principles are needed and researchers should be supported by communities, institutions and funders to set up coherent data management plans in order to deal with these different challenges.
- Text and data mining in PPPs is not problematic if there is the right legal context. There is a view that Europe over regulates research data through its intellectual property laws – in particular through database rights and copyright. This is potentially holding back innovation in Europe.
- Licensing information should be included with the data, so that machines can automatically detect the authorised use.
- If publicly funded, research data have public good (economic), public domain (legal) and public interest (political or policy) characteristics.

PANEL 4. RECOGNITION AND REWARD SYSTEMS FOR DATA PROVIDERS AND STEWARDS

Data sharing requires cultural change among researchers in many fields of science. Perceived barriers and risks of providing open access to data need to be counterbalanced by appropriate acknowledgement and reward systems. Researchers have incentives to publish scientific results, preferably positive ones. Incentives to publish data are less developed, and usually seen as a constraint imposed by funding agencies (threat of discontinuing funding) and/or publishers (data statements required). Data citation has not been widely implemented, and the prerequisites for it - standard formats, citation metrics - are not being broadly adopted. Open Science needs to be embedded in evaluation systems to ensure that researchers who provide high quality research data are rewarded.

* * *

Governments tend to focus on the FAIR principles and FAIR metrics, but these are not sufficient because the sustainability dimension is not addressed. For instance, data stewards in research data repositories and active data curation communities support the stewardship of research data and play a critical role to enable Open Science. However, such communities and activities are not always well organised or sustainable. They should be valued, better recognised and supported by governments and this can be achieved by endorsing and supporting efforts leading to increased trustworthiness of research data repositories, such as through certification. Certification of organisational entities (e.g., data repositories) can bring better recognition for data stewards (by publishers and funders) and additional reward for researchers publishing data into a certified trustworthy repository (in the same way as publishing in reputable journals). For a discussion of "trust" in research data, see the Panel 1 discussion, above.

Governments can also encourage science publishers, research funders, research communities and data stewards to align their requirements and expectations regarding data (and research outputs) availability and management. There is currently a proliferation of sometimes incompatible funder and journal mandates that could create obstacles to securing funding and publishing data. This also affects promoting and implementing reward mechanisms such as data citation because of the lack of common approaches.

To accelerate research progress, making data open as a default rule should be applied to research outcomes (papers, reports), content (data, materials, protocols), and process (registrations, pre-analysis plans). For example, although oriented toward scholarly journals, the Transparency and Openness Promotion (TOP) Guidelines provide an actionable framework for policy implementation by all stakeholders in research (publishers, institutions, funders, government): http://cos.io/top/ Policymakers could consider implementing their specific expectations from TOP as policies, or require that the agencies/communities that they support/govern use the TOP framework to define locally appropriate policies.

Research funding agencies could reward open research practice when deciding on assigning funding including it in the evaluation criteria for proposed research. This could include taking into account pre-prints, software products, data, protocols, or other research outputs and related impacts, such as data citations and re-use. A "data paper" should be seen as a legitimate scientific publication. Open science can only reach its full potential when major incentive and reward changes are realised and professional Data Stewards are trained and supported.

Open research practices should be assessed and rewarded in the evaluation of performance and career development, and throughout the researcher's career. The same criteria should also be used when hiring new researchers.

It is important to reward and publicly recognise champions of open research in various research communities and at different levels to encourage multiplying good practice and avoiding bad practice; i.e. champions who stimulate changing scholarly communication behaviour amongst faculty. Researchers should be helped to open their data on request. There are many cases where the researcher does not have enough motivation or time for data management because of the short duration of funding and the pressure to publish research results in articles, but not appropriately curate the data.

A detailed example of recognition and reward policies from a research funder for data work comes from the U.S. National Science Foundation (NSF), which has had an incremental strategy for public access this decade. The NSF public access activity began in 2011 with a requirement that every new grant proposal contain a data management plan (DMP). These DMPs, which undergo rigorous evaluation during the merit review process, provide the research community with the opportunity review and comment on a researcher's plan for sharing data. An individual's 2-page biosketch, required with every grant proposal, underwent change in 2013. The category for the top five *publications* became the top five *products*. "It's just a recognition of a broadening of what could be put into the *biographical sketch*," said *NSF* senior policy specialist Beth Strausser in a 2013 interview. "*Products* of research are not just publications." In 2016, NSF added a requirement to the proposal section on prior NSF funded research, which includes discussing evidence of research products and their availability, including, but not limited to data. Finally, NSF funds the ongoing operation of community repositories, and through division-level Data Management Plan guidance, encourages researchers to use the repositories. These policy actions bring visibility to data products emerging from NSF funded research and form a cornerstone of incentive structures. It is also evident that trying to reproduce the work of researchers is difficult and research will rely increasingly on machines. Hence, it is important for the data to be easily tracked and analyses reproduced.

The complexity and interplay in the US between copyright, licensing, and other intellectual property laws for research data can create confusion for NSF-funded investigator when deciding what to do with the data produced under a research grant. NSF policy is a balanced stance between retention of legal rights and data sharing. It is NSF policy to "normally allow grantees (institutions) to retain principal legal rights to intellectual property developed under NSF grants." But, as the NSF policy goes on, "such incentives do not […] reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers". The investigator's decision is additionally influenced by publishers and data repositories on matters of licensing and copyright.

Finally, there is considerable discussion now in the open science community and among federal agencies in the U.S. over uniform (globally unique) identifiers for investigators, awards, and institutions. NSF is moving forward on one piece of this: allowing an investigator to optionally enter their unique ORCID ID. The adoption and use of globally unique persistent IDs (PIDs) for data is a necessary pre-condition for achieving even modest gains in realizing the potential of data discovery and use for scientific data.

Some preliminary conclusions:

- **Cultural change** is a long process (researchers tend to be possessive about data), with many legitimate and some spurious reasons for withholding data. Perceived barriers and risks of open access to data need to be mitigated by appropriate acknowledgement and reward systems.
- **Incentives** exist to publish research results, but the incentives to release or publish data are much less developed.
- **Data citation** practices and protocols are now being developed but are not yet widely implemented, and the supporting elements for it are still mostly missing: no standard formats, no citation metrics.

- The panel suggests making it mandatory to **reward** research data work and to include such activities in **evaluation** and **recruitment** criteria.
- Incentive mechanisms for recognition and reward of data work include:
  - empowering researchers to assert control over the data produced while promoting data sharing,
  - requiring the use of Persistent Identifiers (PIDs) for supporting data citation,
  - recognising data management skills and data products at a par with publications
  - encouraging data publishing in journals,
  - enabling the organisation and sharing of data by financially supporting infrastructure and services
  - tracking data use metrics.

PANEL 5. BUSINESS MODELS FOR OPEN DATA PROVISION

Costs of provision of open data are often borne by the providing institution, but benefits accrue to stakeholders around the world. Business models for research data repositories are restrained by mandates and incentives. The OECD Global Science Forum has recently published a report[8] on this subject and the field is changing rapidly with new private sector actors competing with and/or complementing public repositories.

* * *

In the 2006 Recommendation there are general observations concerning long-term sustainability and cost effectiveness. Twelve years later we are now much better informed about the challenges and possible approaches of these issues. The 2017 OECD report on Business Models for Sustainable Research Data Repositories provides us with the latest insights in these areas. Based on the outcomes of that work additional policy principles for sustainability and cost effectiveness include alignment of regulation, mandates and incentives and recognising the importance of economies of scale.

A number of issues were raised in this topical area.

Policy recommendations should be agnostic with regard to business models – the aim is to support the goal of open data provision.

Governments, funders and societal organisations dealing with data-intensive approaches should consistently invest in the required infrastructure for open science. This is a prerequisite for implementation of the FAIR data and services discussed earlier in this report. Long-term funding for data infrastructure - technical and personnel - is an imperative.

Governments should make it clear to the stakeholders in the research enterprise that 'open research data' does not include scientific publications (unless explicitly published under an unrestricted open access license). However, scholarly publishers also have a role to play in open data provision, especially related to the data supporting the research results that they publish.

[8] OECD (2017), "Business models for sustainable research data repositories", *OECD Science, Technology and Industry Policy Papers*, No. 47, OECD Publishing, Paris. http://dx.doi.org/10.1787/302b12bb-en

'Open data by default' needs to balance the costs of data provision between data providers (who are most of the time also the reusers) and other reusers of the data. Governments should mandate open research data by default.

A large proportion of the value rests with the reuse of data. Cost-sharing may be applicable in some circumstances, such as when data have no or limited reuse potential or there is a public-private partnership (PPP) for an activity. For example, data markets already exist for some PPPs, including those for research purposes.

Most repositories have a long-term vision, but the services they provide are based on project funding while long-term funding is frequently an issue.. Repositories play a role in assuring quality-control of the data they store. A mixture of project and organisational core funding is required in order to balance incentives and sustainability.

Open data is a powerful lever to attract new stakeholders in the innovation ecosystem. Some private companies are opening their data to get returns on investment (e.g., for recruiting talent, for setting up new innovative partnerships, for improving their image, and the like). Interesting public-Private Partnerships can be struck around this concept. For instance, in medical research one may want to combine data about people's medical history, genomics, food intake and mobility. Here, medical and genomic data (in the non-US context) would come from the public sector, while mobility and food data would come from private sector. Public-private partnerships built around such themes, should be encouraged to support infrastructure development and the creation of value-added services built upon open data.

Some preliminary conclusions include:

- **Costs** of provision of open data are **borne by the providers**, but the **benefits accrue mostly to users**. In addition, public financing occurs mostly through **short-term** project funding, while responsibilities for data preservation are **long term**.
- Increasingly we will not be able to store data at the pace we produce it, particularly in some data-intensive fields.
- Business models for cost recovery are in many cases not well developed, especially since the expectation is often that **open access = free access**. Access does not always have to be free, but in many cases should be free **at point of use.**
- More data has led to more repositories that demand to be supported. Different business models have been used and these should be evaluated for cost-effectiveness.
- Centralisation allows taking advantage of **economies of scale**, but networked repositories offer better **ownership** and flexibility.
- **Cost/benefit analysis** is crucial, since some data may be too costly to curate and provide on open data platforms. However, there is no standardised method of calculating the cost-benefit of the curation of a dataset
- The automation of certain data curation functions needs to be evaluated and implemented where appropriate.

PANEL 6. BUILDING HUMAN CAPITAL AND INSTITUTIONAL CAPABILITIES

Researchers often lack data management skills (although funding agencies can make this a requirement for recurrent funding). Users (who may be from different sectors of academia or the private sector) do not

always have appropriate skills for correct interpretation and analysis. Technical staff in data repositories require training in data curation and stewardship. Specific curricula including statistical skills, computer science and information science are needed. Many countries report limitations of current curricula in addressing these skills needs. Professional staff dealing with data can be further broken down into various categories, including data engineers, data analysts and data stewards that require distinct skill sets.

* * *

Skills and competences will continue to play a significant role as the use of data to conduct and underpin science grows. The demands will expand as interdisciplinary collaborations, innovative sub-divisions of traditional domains, and new demands on policy makers and strategists all conspire to necessitate new approaches to skills and competences. Data science, in the widest sense of the term, requires technical skills – both IT, and mathematical and statistical - and increasingly competences in unstructured data too, as well as communication and leadership skills. Some universities (Belgium was provided as an example) have made it mandatory for research students to possess some advanced data management skills.

Our ability to harness the value of data is contingent on the development of a data literate workforce – a workforce that will encompass a range of competencies, such as skills in utilising data as an information and economic asset, skills in cross-disciplinary data integration, and skills in innovating with data. Policy recommendations need to differentiate between changes to digital education and training for researchers, and for data and eResearch professionals. With regard to the former, data literacy and competency should be a key part of any high-quality research training system. On the latter point, professional data and eResearch services need to become integral parts of the research system, with appropriate status, career paths, credit and recognition for those who provide these services.

Policy change is needed to both respond to and facilitate the development of systemic approaches to digital skills and training, to deliver a flexible, agile, and professional workforce with the capacity and skills needed to maximise the massive potential gains of open data policies in the research sector and industries of the future.

Several issues were identified in this context.

The main elements in the science data ecosystem have been present for some time: provenance, sharing, metadata, regulations and standards, repositories and institutional responsibilities. However, the key missing elements in this narrative are the creation and capture of scientific and monetary value from the data. This will be particularly relevant in the context of partnerships and exchanges between public and private funding streams.

Enabling discoverability requires data-sets to be published with unique and citable permanent digital identifiers. Researchers should then cite their own as well as other sources of data in publications. A reproducibility check for journals should also become a standard procedure; they should run the code with the data to check if the results are actually capable of being reproduced.

As we move further towards a data-driven ecosystem for scientific research, where data, metadata and associated algorithms and other software applications, as well as the infrastructure (storage, processing, networking and cloud services), all become critical components, so the risk landscape becomes correspondingly complex. Another level of complexity is added by the interdependency of the organisations conducting research. Essentially, science becomes a distributed supply chain as researchers outsource services and utilities. How and where is the risk managed and mitigated?

Data engineers (“tool innovators”), data scientists (mainstream researchers primarily using data for their work) and data stewards (or “research data administrators”) are different fields of expertise, but they will

need to cooperate. The main challenge for a data steward is "translating" between the science and technology. Different roles are also emerging around the skillsets needed to provide services over the data, to maximise data usage and analysis, including research software engineers (see, e.g., https://software.ac.uk/blog/2018-04-05-research-software-engineers-and-data-scientists-more-common).

Data skills are not just for data stewards – training needs to be self-evident in any research curriculum — but stewards face different challenges depending on the scientific discipline, though common principles obviously must be applied. What are the most efficient methods for educating young research professionals? Is it a module? A training course? Lifelong learning?

A detailed example was provided concerning the situation in South Africa. Three main points were raised.

1. Researchers require training in and incentivisation to undertake a basic level of content curation and data stewardship.

This issue pertains to capacity building on the part of individual researchers as well as the provision of infrastructure and platforms that can facilitate this practice. The establishment of institutionally-based secure data archives, virtual research environments and collaborative repositories not only serve as tools for better data management; their use also prompts researchers to consider foundational curatorial issues in the course of the research process. In instances where under-resourced institutions are not able to afford this infrastructure, consortium and regional initiatives are crucial to capacitate marginalised academics and institutions. An example of such an initiative is the Data Intensive Research Initiative of South Africa (DIRISA) pilot which is rolling out a nationally-subsidised network of Figshare instances at South African universities to address data management and the fulfilment of data-sharing mandates.

The South African case has made it evident that mandates around data sharing are necessary but not sufficient to stimulate better data management and sharing. Researchers need to be capacitated and incentivised to incorporate a more professional approach to data stewardship, an area of work which is very new to certain disciplines. Embedding a curatorial mindset in the academic process will go a long way towards stimulating activity in this regard, and in making the process less taxing for researchers.

2. Intermediaries are critical in the open data sharing process.

The importance of intermediaries in the open data ecosystem has been widely acknowledged by advocates and practitioners. These intermediaries are typically professionals trained with data stewardship and publication expertise who (a) alleviate the load of data curation and publication tasks faced by researchers; or (b) provide value-add services (such as analytics and visualisations) that enrich the third-party user experience.

A study on open data in the governance of South African higher education (Van Schalkwyk, Willmers & Czerniewicz, 2014) found that open data intermediaries increase the accessibility and utility of data; provide both supply-side as well as demand-side value; can assume the role of a 'keystone species' in a data ecosystem; and have the potential to democratise the impacts and use of open data in that they play an important role in curtailing the 'de-ameliorating' effects of data-driven disciplinary surveillance.

In many developing country contexts there is an imperative to provide capacity building initiatives that address foundational research skills and introduce academics to the principles and processes of open data sharing. Many institutions face severe challenges around skills and infrastructure deficits as well as language barriers in accessing the support required to develop new scholarly communication skills. In these contexts, intermediaries such as data curators have a crucial role to play in undertaking the data preparation and de-identification work required in order to publish data. They also have a crucial role to

play in working with researchers to build the capacity and confidence of researchers in the data-sharing process, building trust in the process and liberating them from misconceptions around publication scooping and exploitative, unethical third-party data use.

The recognition and support of data-publication intermediaries is made more acute in South Africa and other developing countries where libraries are under enormous pressure to transform into entities that can provide research-oriented support services in addition to their traditional focus on undergraduate resource provision. Many libraries in African universities face budget and skills deficits as they grapple with the challenge of addressing bourgeoning student populations with increasing demands for e-learning and remote access. This makes it challenging for African librarians to play the role of open access and open data service providers, as is the case in many developed countries.

3. A lack of cohesion in policy frameworks hinders open sharing by researchers.

The ability of researchers to share outputs arising from their work is dictated by institutional intellectual property (IP) policies, which are in turn largely influenced by national copyright acts. In the African context, many universities have nascent policy environments, meaning that they may not have an IP policy or it is out of date and inadequate to cover the intricacies of online content sharing, particularly as relates to open data transfer and publication. This situation makes for confusion on the part of academics in terms of what their actual rights are in terms of data sharing … or, in some cases, may lead to flagrant disregard for policies and mandates.

A review of South African universities' IP policies revealed that even though all 25 South African universities are public state-funded institutions, they each have their own IP policies, which provide different prescriptions regarding copyright ownership. The survey was done with a focus on the provisions of sharing teaching and learning materials, but it is reasonable to infer that similar discrepancies will occur in relation to data ownership and sharing. The University of Cape Town IP Policy, for instance, promotes open content sharing and the use of Creative Commons licensing, ceding copyright of research data to academics; while the Stellenbosch University Policy on the Commercial Exploitation of Intellectual Property is strongly focused on commercialisation and states that the university owns copyright over all outputs produced by academics in the course of their work, including raw data created during research.

This uneven policy context is not only confusing for academics to navigate, but introduces highly challenging legal constraints for open data sharing. Grant agreements do increasingly provide exceptions and caveats to restrictive IP policies, but these agreements are often not adequately scrutinised by researchers and the lack of cohesion between institutional policies and the dictates of funding entities serves to amplify the distrust of open data practice.

National and regional initiatives to assess and revise institutional IP policies so that they are conducive to open data sharing (or establish these policies in cases where they do not exist), would be extremely valuable in terms of promoting open data practice and ensuring a clear, cohesive approach to the legal and ethical aspects of the process – the uncertainty of which often inhibits researchers' practice in this regard.

Some preliminary conclusions:

- **Researchers** often lack data management planning and curation skills (but funding agencies can make this a requirement for recurrent funding).
- **Users** do not always have appropriate skills for correct interpretation and analysis of the data that they access, or the data are insufficiently organized, documented and curated.

- **Technical staff** in data repositories need training on data standards, and good policy and practice. Specific curricula including statistical skills, computer science and information science are needed, either at the tertiary education level or in subsequent training programs. Many countries, however, report limitations of current curricula in addressing those skills needs.
- A lack of skills breeds lack of **trust** – in particular, academic and private-sector stakeholders should be encouraged to trust the system through incentives (see also Panel 1).
- More higher education curricula and young and mid-career (lifelong) training in data management are needed to maximize the value of preserved research data. Data skills are not just for data stewards – training needs to be self-evident in any curriculum. Such courses should be officially recognized.
- Some examples and models exist for research data education and training curricula, but these need to be better evaluated and implemented more broadly.
- Data engineers, data scientists and data stewards are different fields of expertise, but they need to cooperate. The main challenge for a data steward is "translating" between the science and technology.
- Research managers and administrators need to care about data scientists' career paths. In some universities and research domains data scientists are integrated within research teams with related career prospects and rewards but this is frequently not the case.
- To maximise the benefits that can be gained from open data, it is also necessary to understand the need for development of skills in developing sustainable software for analysis of the data, and to consider this in tandem with discussion on how to develop a data literate workforce.

Annex 1

Chair	**Dominique Guellec**, *Head of Science and Technology Policy Division, OECD*
Introductory remarks	**Andrew Wyckoff**, *Director of Science, Technology and Innovation Directorate (STI), OECD*
Access to public data for science, technology and innovation - Summary of learnings from OECD project	**Alan Paic**, *Senior Policy Analyst - Science and Technology Policy Division, OECD*
Panel 1. Data governance and trust for science, technology and innovation *Moderator:* **Duncan McIntyre**, *Assistant Secretary, Department of Prime Minister and Cabinet of Australia, Australian Government*	**Alan Paic**, *Senior Policy Analyst - Science and Technology Policy Division, OECD* **Marin Dacos**, *Open Science Counsellor, French Ministry of Education, Research and Innovation* **Andreas Ebert**, *CTO for Western Europe, Microsoft* **Eunjung Shin**, *Associate Research Fellow, Korea Science and Technology Policy Institute (STEPI)* **Peter Elias**, *Professor, Institute for Employment Research, University of Warwick*
Panel 2. Data standards, interoperability and re-use *Moderator:* **Ross Wilkinson**, *Australian National Data Service & Research Data Alliance*	**Simon Hodson**, *Executive Director, CODATA* **Tyler Walters**, *Dean, University Libraries and Professor, Virginia Tech* **Thordis Sveinsdottir**, *Senior Research Analyst, Trilateral Research* **Jean-Francois Abramatic**, *Consultant, INRIA and IBM (formerly)* **Stavroula Kampouridou**, *Fintech Senior Advisor to the Governor, Bank of Greece*
Panel 3. Definition of responsibility and ownership *Moderator:* **Ingrid Dillo**, *Deputy Director, Data Archiving and Networked Services, RDA and Leiden University*	**Stan Matwin**, *Director, Institute for Big Data Analytics, Dalhousie University* **Paul Uhlir**, *Consultant - Ex Officio Member, Board on Research Data and Information, National Academy of Sciences* **Ben White**, *Head of Intellectual Property, British Library* **Marie Timmermann**, *EU Legislation & Regulatory Affairs, Science Europe*
Panel 4. Recognition and reward systems for data providers and stewards *Moderator:* **Simon Hodson**, *Executive Director, CODATA*	**Vanessa Proudman**, *Director, SPARC Europe* **Brian Nosek**, *Professor of Psychology, School of Medicine, University of Virginia* **Mustapha Mokrane**, *Executive Director, World Data System* **Samuel Goeta**, *Co-founder, Datactivi.st* **Beth A. Plale**, *Science Advisor for Public Access, Office of Advanced Cyberinfrastructure, National Science Foundation*

Panel 5. Business models for open data provision *Moderator:* Carthage Smith, *Senior Policy Analyst - Lead co-ordinator - Global Science Forum, OECD*	**Barend Mons**, *Professor of BioSemantics, Leiden University* **Jean-Marc Lazard**, *Chief Executive Officer, Opendatasoft* **Ingrid Dillo**, *Deputy Director, Data Archiving and Networked Services, RDA and Leiden University* **Duncan Campbell**, *Director, Global Sales Partnerships, John Wiley & Sons*
Panel 6. Building human capital and institutional capabilities *Moderator:* Bart Dumolyn, *Assistant to the Director - Department of Economy, Science and Innovation, Flemish Government*	**Kazuhiro Hayashi**, *Senior Research Fellow - National Institute of Science and Technology Policy, NISTEP Japan* **Michelle Barker**, *Deputy Director, Research Software Infrastructure, Nectar/University of Melbourne* **Michelle Willmers**, *Curation and Dissemination Manager, University of Capetown* **Pedro Fernandes**, *Training Coordinator, Instituto Gulbenkian (Elixir project)* **Victoria Stodden**, *Associate Professor, School of Information Sciences, University of Illinois* **Steve Brewer**, *Research Staff - Electronics and Computer Science, University of Southampton*
Concluding remarks	**Christian Reimsbach-Kounatze**, *Internet Economist and Policy Analyst, Digital Economy Policy, OECD* **Edwin Lau**, *Head of Reform of the Public Sector Division, OECD* **Dominique Guellec**, *Head of Science and Technology Policy Division, OECD*