Chapter 7

Promoting data-driven scientific research

This chapter summarises the recent evolution of science – mainly thanks to the advent of data analytics – towards a more open and data-driven enterprise. It examines how new and evolving opportunities for interconnecting and sharing have led to what could be called citizen science. A discussion follows on the various impacts of open access to science, research and innovation on the business and science communities and on citizens. There are examples of organisations involved in open data efforts, and an exploration of the challenges and opportunities presented by data sharing. The focus then shifts to policies and practices in the OECD area and beyond, with the emphasis on infrastructure for data sharing. With unrestricted access to publications and data, firms and individuals may use and reuse scientific outputs to produce new products and services – but do scientists and researchers have the incentives or indeed the skills to perform these tasks?

In the next 50 years, as the technologies of information and knowledge accelerate, the nature of the scientific process will change even more than it has in the last 400 years. (Kevin Kelly, Google TechTalks, 9 May 2006)

Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorisation of reality. (Boyd and Crawford, 2011)

Information and communication technologies (ICTs), new data storage infrastructure and large-scale computing are modifying the way science and research are conducted, disseminated and diffused. (See Chapter 3 on the enabling role of ICTs for data-driven innovation, DDI.) The term often used to describe this transformation of science into a more open and data-driven enterprise is *open science.* All main phases of the data value cycle introduced in Chapter 1 apply and impact this transformation, its scientific processes and outcomes:

- *Data creation and data collection* – ICTs allow the collection of large amounts of real-time data that can serve as the basis for scientific experiments and research, contributing to make science increasingly data-driven. It is now possible to collect, generate, access, use and reuse research and scientific material (articles and data sets but also images or digital lab records) at no or extremely low marginal cost, and speed the transfer of knowledge among researchers and across scientific fields, opening up new ways of collaborating and new research domains. In addition, ICTs are creating new opportunities to organise and publish the content of research projects, scientific publications and large data sets, so as to make it immediately available to other scientists and researchers as well as potential users in the business community and society in general.
- *Data analytics and software* – ICTs are not only modifying the way scientific material is generated, collected and stored, but also helping promote deeper analysis of data through new software and applications that allow a faster and more exhaustive use of data in science and research. New opportunities also arise thanks to text and data mining techniques (see Box 7.1). Some scientific disciplines have historically been at the forefront of data collection and exploitation, by means of complex scientific experiments (as in the case of physics, for example), and have always had a long tradition of data validation and preservation. However, in recent years – and again, thanks to the advent of the ICTs – less historically data-intensive scientific fields (such as the humanities for example) have become increasingly data-driven. Different scientific domains are becoming increasingly interconnected: data generated in one field of research may nowadays be treated with models and techniques traditionally belonging to other fields of research. Typical examples are social science data that today may be treated with algorithms and methodologies traditionally belonging to physics or computer sciences, due to the large scale of the data sets newly generated and collected. Even those disciplines with a long tradition in data collection and lab experiments are moving towards more sophisticated simulations that are validated thanks to the large amount of data digitally available.
- *Scientific output and decision making* – As science becomes increasingly data-driven, its social and economic values are enhanced. Just as business or government data are being used to create new goods and services, scientific data are enhancing the quality and output of scientific research. Open research data allow scientists in one field to exploit data in other fields, and reveal relationships or patterns that were not visible before. It allows them to formulate and test new hypotheses and improve the predictability of scientific models. The value of scientific data is not solely economic; data can help capacity building (education, network formation); public engagement (or public understanding of science and technology); knowledge and material sharing among research institutions; and policy and practice revision.

Box 7.1. **Opportunities and challenges arising from text and data mining (TDM)**

"Text and data mining" (TDM) refers to an ensemble of computer science techniques to analyse and extract knowledge and information from large digital data sets (i.e. big data), by looking for trends and patterns unnoticeable to human eyes. TDM is useful and increasingly used by researchers in all fields, from historians who scan historical documents and archives to medical experts who find common patterns in medical records. TDM is equally well-established in fields such as astronomy and genetics; its methods and techniques are widely used both in the public and the private sector. TDM algorithms investigate large-scale data sets containing not only figures and numbers but also other types of digital records, including text, images and audio files. TDM enables the use of common techniques and makes connections between unconnected fields of research. This represents a huge opportunity for the development of innovation.

The use of TDM has important repercussions in the academic community. With the growing amount of published (and unpublished) academic articles (an estimated 50 million as of 2010), it is becoming impossible for scientists and researchers to manually access, read and analyse publications. TDM provides the potential for accessing, scanning and analysing publications by means of machines.

Research on TDM techniques has also advanced considerably in recent years. The number of academic articles published on TDM since the beginning of the 1990s reveals that the United States has so far produced 46.6% of the publications dealing with TDM, followed by the United Kingdom (11.1%), Chinese Taipei (8.8%), Canada (5.7%) and China (4.6%). Continental Europe lags behind, according to some because of the legal obstacles European researchers face when trying to advance knowledge in this field.

In fact, the growth of open access and open data raises more generally the issue whether researchers using TDM should or should not face restrictions. The uncertainty created by the current legal framework regarding the scope of protection of works and databases is bound to create obstacles to TDM activities. Moreover, transaction costs would rise if researchers had to reconcile the terms and conditions of non-standard or non-interoperable licences.

Within the OECD area, Japan and the United Kingdom have modified the national copyright framework to allow TDM during the last decade. In 2009, Japan introduced an exception to its national copyright law to permit TDM, with the aim of boosting the Japanese digital economy. More recently, the United Kingdom has introduced a specific exception in the Copyright, Designs and Patent Act to allow TDM activities to take place without the rights holder's prior authorisation, under conditions stated in the law. This amendment will enter into force in October 2014. Together with the Japanese provision, the United Kingdom provision is a step towards facilitating scientific research that may in time be followed by other jurisdictions.

Source: European Commission (2014), *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group;* Sergey Filippov (2014), *Mapping Tech and Data Mining in Academic and Research Communities in Europe,* Lisbon Council, 16/2014.

7.1. The evolving scientific enterprise

The evolution of scientific and research processes also has an impact on research addressing global challenges. Global challenges by definition affect many individuals in a large number of countries and cannot typically be addressed or managed by a small team of researchers working in isolation. The data generated at global level that relate to issues of global concern, such as the environment and climate change or the ageing population and health, may be more powerfully exploited if properly interconnected and used by large

networks of scientists and researchers world-wide. This may avoid duplication of efforts and enhance co-ordination in science and research. Recent examples of large-scale research projects with a social challenge focus also include the Human Brain project sponsored by the European Commission[1] and the BRAIN Initiative in the United States.[2] These projects are examples of collective, international, multidisciplinary efforts that combine multiple scientific disciplines from biology, medicine, high-performance computing and robotics, with the aim of advancing understanding of the human brain for the benefit of the society. The potential of open science and open data efforts to advance research to address Alzheimer's disease and dementia has been recently highlighted by the OECD expert consultation "Unlocking Global Collaboration to Accelerate Innovation for Alzheimer's Disease and Dementia" (OECD, 2014; see also Chapter 8 of this volume).

Finally, open science and open data have the potential to strengthen relations between the scientific community and society. The landscape of science communication has changed, and scientists have now a broader range of mechanisms (e.g. through social networks, personal scientific blogs, videos, interviews and discussion forums) to communicate with citizens. These new scientific communication mechanisms can help build public trust in science. In short, the advent of open science has the potential to increase opportunities for the diffusion of research results among the scientific community and to society. In addition, as science is managed and produced in a more open and transparent manner, the basis for "trust" between science and society will be subject to greater scrutiny from citizens. For a similar discussion on the trust enhancing-capacities of open data for government, see Chapter 10.

Data-driven science and research

Data and measurement have always been fundamental to science. The advent of new instruments and methods of data-intensive exploration has prompted some to suggest the arrival of "data-intensive scientific discovery", which builds on the traditional uses of empirical description, theoretical models and simulation of complex phenomena (BIAC, 2011). This could have major implications for how discovery occurs in all scientific fields (Hey and Trefethen, 2003; Jirotka et al., 2006; Anderson, 2004; Bell, Hey and Szalay 2004). Big data science allows the development of scientific experiments as well as computer-based algorithmic simulations, even in those fields that traditionally were less data-intensive than others.

Data analytics tools (such as machine learning and pattern recognition techniques) are increasingly used by scientists to gain knowledge of phenomena and to test or validate models. Big data allow computer-based experiments and simulations even in those fields where traditional lab experiments were impossible or too difficult to organise. Some have even challenged the usefulness of models in an age of massive data sets, arguing that with sufficiently large sets, machines can detect complex patterns and relationships that are invisible to researchers (Anderson, 2008; Bollier, 2010). In addition, big data science and algorithmic-based experiments and research in themselves represent an opportunity for innovation and scientific discovery: fields such as computer or data science are currently exploiting big data as an opportunity to develop new and more efficient algorithms for data analytics, to be used by researchers active in different disciplines and fields (both in the public and private sector).

New instruments such as super colliders and telescopes, but also the Internet as a data collection tool, have been key to new developments in science, as they have changed the scale and granularity of the data being collected. The Digital Sky Survey, for example, which started in 2000, collected more data through its telescope in its first week than had

been amassed in the history of astronomy (*The Economist*, 2010), and the new SKA (square kilometre array) radio telescope could generate up to 1 petabyte of data every 20 seconds (EC, 2010). Furthermore, the increasing power of data analytics has made it possible to extract insights from these very large data sets reasonably quickly. In genetics for instance, DNA gene sequencing machines based on big data analytics can now read about 26 billion characters of the human genetic code in seconds. This goes hand in hand with the considerable fall in the cost of DNA sequencing over the past five years.

These new developments, affecting all scientific instruments across all scientific fields, indicate the potential for a new era of discovery and raise new issues for science policy. These issues range from the skills that scientists and researchers must master to the need for a framework for data repositories that adheres to international standards for the preservation of data; defines common storage protocols and metadata; protects the integrity of the data; establishes rules for different levels of access; and defines common rules that facilitate the combining of data sets and improve interoperability (OSTP, 2010).

Diversity of scientific data

Scientific research data vary enormously – in type and volume, as well as in use and long-term value (see Chapter 4 for a comprehensive description of the different features of data). Four types of research data in particular are important in research.

Observational data come from telescopes, satellites, sensor networks, surveys and other instruments that record historical information or one-time phenomena (such as astronomical data from the Sloan Digital Sky Survey, SDSS). This category also includes social science research, such as demographic surveys. In many cases these data cannot be replicated and should be retained.

Experimental data may be captured from high-throughput machines (such as accelerators), through clinical trials and biomedical and pharmaceutical testing, or through other controlled experiments. Preservation is particularly important for experimental data where it is not feasible or ethical to replicate data gathering. This includes some data dealing with human subjects and endangered species.

Computational data are generated from large-scale computational simulations. Although such data can be regenerated by rerunning the simulation, there are two reasons why computational data may need to be preserved over the medium term (three or more years). First, the data may be used as the basis for substantive and subsequent analysis, visualisation, or data mining. Second, time on a computer for additional computations may not be available within a sufficiently short delay. This is a common occurrence for very large-scale computations that run on supercomputers shared by the research community, such as those found at US Department of Energy national laboratories and National Science Foundation (NSF) centres.

Reference data sets are highly curated data that are often in high demand by multiple scientific communities. Such data are created for purposes ranging from mapping the human genome and documenting proteins to amassing longitudinal data on economic and social status. The Worldwide Protein Data Bank and Panel Study of Income Dynamics are such reference data sets. With all these data, there is often a need to preserve ancillary materials, such as calibrations of instruments, parameters of experiments, and lab notebooks. While most large research data collections are produced and used by researchers, they are also valuable for public policy. Policy makers' needs for information about climate, seismology, oceanography, clinical trials and social science research surveys, endangered

species, indigenous sites, archaeological sites and sensitive security matters go well beyond the demands of research and become a matter of urgent public priority.

Interaction of data and research materials

Access to research materials and tools are important for advancing research and innovation. As science becomes more data driven, the issue of access to research materials can interact with the issue of open data in complementary or conflicting ways. The first interaction is around the patenting of research materials (e.g. biological materials); insufficient disclosure of difficult (first order results). A second interaction concerns the portfolio patenting strategies of firms or universities. While strategic from a licensing and commercial point of view, such patent portfolios can be structured as de facto propriety databases that could impede open research data efforts in academia, as many researchers rely on (patented) research tools.

Collaborative research platforms and citizen science

As science becomes more data-driven and ICTs offer new possibilities to interconnect and share, international platforms to promote collaborative and networked research have emerged. The goal of these consortia is to promote not only data sharing, but also and more generally information sharing, to facilitate the creation of joint research projects and activities among teams of researchers globally. They are typically either discipline-specific or around broad themes, often related to social challenges such as health or climate change and the environment, and north-south co-operation. In some cases, these collaborative platforms go beyond the involvement of the research community alone and aim to engage business sector actors, the civil society, and individuals more broadly.

Examples of global platforms include the following. The Open Source Drug Discovery (OSDD) is an online platform for drug discovery that brings together scientists from the OECD area and developing countries to develop therapies for diseases endemic in developing countries, such as malaria, tuberculosis and Leshmaniasis.[3] Future Earth is a consortium active in global research on climate change and the environment, managed by the International Council for Science[4] and open to scientists active in all disciplines. DIYbio.org is a platform founded in 2008 with the goal of creating a community of biologists exchanging information and discussing research themes. Other types of collaborative platforms are not restricted to researchers and scientists, but have the goal to reach a broader community in civil society. For example, the BiOS[5] Initiative (Biological Innovation for Open Society) is an effort to promote bio-related innovation (in fields such as agriculture, biotech, water and agronomy) in disadvantaged communities.

Collaborative efforts in science and research reach beyond the research community to increasingly involve citizens and "amateur researchers" at different stages of scientific processes, from the collection of data to the solution of more complex scientific problems. The involvement of non-professional scientific communities in science and research efforts is often referred to as *citizen science*.

The participation of amateurs in scientific processes – the interaction with professional scientists – is not a new concept; it dates from the 18th century and generally related to data collection or observation, in particular in disciplines such as ornithology and astronomy. However, the improvement in communication capabilities, the emergence of mobile devices, the increase in storage capacity for the information collected, the possibility of transmission of not only text but images and sounds, and (certainly) the existence of greater public awareness have all led to the emergence of the citizen science phenomenon.

There are different aspects of citizen engagement. One is related to the degree of public participation, with respect to the kind of role the non-professional is playing. A second aspect is related to the role of citizens in the selection of research streams that will be publicly financed. Finally, citizen science has specific organisational characteristics; these relate to the development of networks of both professional and non-professional personnel through dedicated events, as well as the need for technical support from the scientific to non-scientific communities (Holocher-Ertl and Kieslinger, 2013).

It has been argued that citizen science serves as a means of achieving several different objectives (Riesch, Potter and Davies, 2013); for example, it allows the development of a more democratic environment in science by engaging amateurs as well as professionals in research and scientific efforts. There is a clear willingness on the part of civil society to be directly involved in the scientific process, not only as observers or data collectors, but also as practitioners, planners and evaluators. Society's participation in the process could even lead decision makers to opt for research priorities based on amateur scientist conclusions or to revoke decisions previously taken. Such was the case in the London District of Deptford: the UK Environment Agency revoked a scrapyard's licence after data that citizens collected on noise levels showed that the yard's operation violated noise limits (Gura, 2013). In addition, the involvement of citizens in scientific projects tends to have an educational value, both implicit and explicit. While in the majority of projects the informal learning aspect of adult citizens is addressed, schools are increasingly considered an important target for the introduction and promotion of citizen science. Teachers play a significant role in facilitating the deployment of experiments and transmitting the socio-scientific values of their contributions to the young audience. Citizen involvement in scientific efforts may additionally have positive implications for the development of a scientifically aware culture.

Several activities help promote the engagement of citizens in science. The project "Amateurs as Experts" was a three-year study of volunteer naturalists, biodiversity scientists and policy makers involved in the UK Biodiversity Action Plan process. The project began in October 2002 and lasted three years. The aim was to enrol new actors into the formal UK biodiversity policy process and have them gain experience carrying out social experiments and analysing and assessing their progress, results and problems. Through ethnographic research methods, the study tried to clarify the social and knowledge dynamics while also fostering patterns of interaction between the volunteer naturalists, scientists and policy makers. The project was a cross-disciplinary research study, involving sociologists, anthropologists (Institute for Environment, Philosophy and Public Policy [IEPPP], Lancaster University) and natural scientists (Natural History Museum, London). Its objective was to develop effective biodiversity protection policies, in the United Kingdom and beyond.[6]

Other examples of crowdsourcing for technical skills that can solve scientific problems are, for example, online platforms where solutions to scientific problems are requested from the public. Examples of this kind of website include Kaggle,[7] a web-based platform for predictive modelling and analytics: private companies and research teams publish unsolved problems related to specific data sets, and data scientists from all over the world compete to find the best solutions and highest-performing algorithms. The crowdsourcing approach relies on the fact that there are countless strategies to solve the problem, each with a different computational efficiency. The best strategy wins and receives the prize money advertised by the firm or the research team posting the problem online. It is estimated that Kaggle connects around 200 000 data scientists world-wide.

7.2. Impacts of open access to science, research and innovation

The impact of open science and open data may be measured in multiple ways. On the one hand, greater access to scientific inputs and outputs can improve the effectiveness and productivity of the scientific and research system, by reducing duplication costs in collecting, creating, transferring and reusing data and scientific material; by allowing more research from the same data; by multiplying opportunities for domestic and global participation in the research process; and by ensuring more possibilities for testing and validating scientific results.

On the other hand, increased access to research results (in the form of both publications and data) can not only foster spillovers to scientific systems, but also boost innovation systems more broadly. With unrestricted access to publications and data, firms and individuals may use and reuse scientific outputs to produce new products and services. (See Chapter 4 for a comprehensive discussion on the potential opportunities arising from data reuse.)

Several surveys report access difficulties for academics in the United States and Europe. For instance, according to CED (2012) 15% of US and Canadian scholars from all disciplines reported their level of access not to be satisfactory. Ware and Monkman (2008) found that only 66% of scientists in Europe and the Middle East reported to have good or excellent access (85% in the United States). There were even lower numbers outside those regions. Barriers to access to scientific material for researchers due to the high cost of subscriptions are also reported, by the survey of Rowlands and Nicholas (2005) and Sparks (2005).

Developing countries in particular may benefit from open access to scientific material. Chan, Kirsop and Arunachalam (2005) note figures from a World Health Organisation survey: in countries with an annual GNP per capita of less than USD 1 000, around 56% of medical institutions have no subscriptions to journals; in countries with GNP per capita of between USD 1 000 and USD 3 000, the percentage of medical institutions with no subscriptions was lower, but still as high as 34%. This is why initiatives are in place to improve developing countries' access to scientific material. For example, the Research4Life programme is a public-private partnership among three United Nations agencies, two universities and major commercial publishers that enables eligible libraries and their users to access peer-reviewed international scientific journals, books and databases for free or for a small fee (Royal Society, 2012). Certain access journals have been created in developing countries themselves, such as the *African Journal of Health Sciences*.

Sharing data has always been considered a crucial activity for scientific research and widely accepted by the scientific community (Fienberg, Martin and Straf, 1985). There is some evidence that, as with open access to scientific publications, sharing data can increase the citation rate of scientific papers (Piwowar, Day and Fridsma, 2007; Piwowar and Vision, 2012) and it fosters good scientific behaviour (Mooney, 2011). Sharing data allows the use and reuse of data from other researchers and individuals (Groves, 2009), would also protect against faulty behaviours and fraud in science and research, and may help improve data collection and management. For all these reasons, data-sharing practices are often regarded positively by the research community (Cragin et al., 2010).

Impacts on the scientific community

Data sharing allows not only verification of scientific results but also the re-analysis of data for different purposes from the ones originally conceived. This enhances the utilisation of data, promotes the competition of ideas and research (Gardner et al., 2003),

and fosters collaboration (Brase et al., 2009; Piwowar and Chapman, 2008; Murray et al., 2009). Data sharing also reduces duplication of effort from different researchers attempting to collect the same data sets (Kowalczyk and Shankar, 2010).

Lakhani et al. (2007) found that disclosing information relating to scientific problems to a large group of outside solvers is an effective means of solving those problems. In addition, disclosure of problem information facilitates problem solving at the boundary or outside the discloser's fields of expertise, thanks to the transfer of knowledge from one field to another. Williams (2010) researched material relating to efforts to decode the human genome. She found that articles dealing with openly available genome sequences led to 30% more articles than those focused on sequences protected by IPRs. The advantage in publications and commercialisations generated by open sequences was notable. Hardisty and Haaga (2008) conducted research on medical articles and found that open access articles were read twice as often by mental health practitioners. In addition, a practitioner's reading of the open access article was associated with their recommending a more cutting-edge treatment. As for scientific publications, data sharing is especially important for researchers in developing countries with fewer means to undertake expensive and time-consuming data collection efforts (Arzberger et al., 2004).

Impacts on the business community and on citizens

Scientists and academics are not the only groups that can potentially benefit from more open data. The demand from the business sector and individual citizens to access research results in the forms of data is significant. For example, the usage data from PubMedCentral show that 25% of the daily unique users are from universities, 17% are from companies, 40% are individual citizens and the rest are government or other categories (UNESCO, 2012). A recent study on R&D-intensive SMEs in Denmark (Houghton, Swan and Brown 2011) found that 48% of those SMEs consider research outcomes very important for their business activities, and more than two-thirds reported difficulties in accessing research material. Ware (2009) conducted a survey on UK small and medium-sized enterprises and found evidence that the equivalent of 10% to 20% of articles were not easily accessible to his survey respondents. Finally, it has been argued that making research data publicly available may promote the public understanding of science, evidence-based practices, and citizen science initiatives (Kowalczyk and Shankar, 2010).

Quantifying impact

Several studies have attempted to estimate the impact of greater access to data on the economy. A recent analysis of UK organisations (Royal Society, 2012; CEBR, 2012) estimated that data was worth approximately GBP 25 billion to UK private and public sector organisations in 2011. The estimates are the cumulative results of GBP 17.4 billion GDP gained in business efficiency, GBP 2.8 billion derived by business innovation, and GBP 4.8 billion derived by business creation.

In the United States, data released by the National Weather Service are estimated to contribute to the development of the private sector meteorology market in an amount corresponding to approximately USD 1.5 billion (Spiegler, 2007). In 2008, the NASA Landsat satellite imagery of the Earth's surface environment became freely available on the Internet. The usage of this database increased from 19 000 scenes per year (when scenes were sold for USD 600 each) to 2.1 million scenes per year. Leading Silicon Valley companies such as Google (in particular Google Earth) use these images, and the open release is estimated to have generated direct benefits of more than USD 100 million per year

to the US economy. According to a recent estimate of the US Open Data initiative (data.gov), open data has the potential to generate more than USD 3 trillion per year in additional value in sectors such as finance, consumer products, health, energy and education.[8]

The European Commission's open data initiatives are expected to generate a yearly income of EUR 140 billion (EC, 2012). In addition, the OECD (2015a) estimated that the public sector information (PSI) market for the OECD area could be around USD 500 billion plus an additional USD 200 billion if barriers to use were removed, skills enhanced, and data infrastructure improved.

The calculation of estimates for the economic value specifically of research data and the related contribution to economic development is more problematic. Available estimates include that of Houghton and Sheehan (2009), who analyse the effects of increasing accessibility to public sector research outputs in Australia, and estimate that increased accessibility generates a return of AUD 9 billion over 20 years. Houghton et al. (2010) estimate that the open access archiving mandate for US Federal Research Agencies over a transitional period of 30 years may be worth around USD 1.6 billion, and up to USD 1.75 billion if no embargo period is in place. Around USD 1 billion would benefit the US economy directly, and the remaining amount would translate into economic spillovers to other countries. These figures would be significantly higher than the estimated cost of implementing open access archiving. JISC (2014) conducted a study on the economic impact of three UK data centres (the Economic and Social Data Service, the Archaeology Data Centre and the British Atmospheric Data Centre) and estimated that each of them could bring a twofold to tenfold return on investment over 30 years.

Open data and the involvement of supranational entities

International organisations play a critical role in promoting co-ordination, at international level as well as in the adoption of standards and norms related to the interoperability of data-sharing platforms. They are also involved in the promotion of an open data culture among scientists and researchers. Examples of international organisations playing these roles are listed in Box 7.2. In addition, international governmental organisations (IGOs) such as the OECD, UNESCO, the EU and the World Bank have been active in recent years in promoting open science efforts of their respective member and, in some cases, non-member countries. The OECD has been active in developing guidelines and principles on open science-related themes, including access to public sector information (see Box 7.3, and Annex of Chapter 10 for more information) and in research data (see Box 7.4) or. At European level, the European Union has adopted and promoted open data efforts in the most recent Framework Programme for Research and Innovation, Horizon 2020 (Box 7.5).

Box 7.2. **Examples of organisations involved in open data efforts**

The *International Council for Science* (ICSU, at www.icsu.org), is a non-governmental organisation gathering members of national scientific bodies and international scientific unions worldwide, representing 140 countries. ICSU was founded in 1931 to promote international scientific activity. The organisation's current mission is to promote international science co-operation for the benefit of society. ICSU identifies and address major issues of importance to science and society; facilitates interaction among scientists across all disciplines and countries; provides independent advice to stimulate dialogue among the scientific community and governments, civil society and the private sector. The ICSU 2012-17 strategic plan has identified the following priorities: i) international research collaboration; ii) science for policy (making); and iii) the universality of science. ICSU has recently published its statement on Open Access Principles.[1]

Box 7.2. **Examples of organisations involved in open data efforts (cont.)**

The *Committee on Data for Science and Technology* (CODATA, at www.codata.org), is an interdisciplinary scientific committee of ICSU. CODATA works to improve the quality, reliability, management and accessibility of science and technology data. It also promotes awareness and cross-border co-operation of scientists. The committee was established in 1966 by ICSU to promote globally the compilation, evaluation and dissemination of reliable numerical scientific data. Legally independent from ICSU, CODATA has 23 members across different continents. Country membership often takes place through national research councils. CODATA activities include both technical discussion on standards and interoperability and policy-level discussion on data issues. The committee works on different aspects of data, from research data to social science data, and public sector information (PSI) including government data. CODATA is concerned with all types of data resulting from experimental measurements, observations and calculations in every field of science and technology, including the physical sciences, biology, geology, astronomy, engineering, environmental science and ecology. Special emphasis is placed on data management problems common to different disciplines, and to data used outside the field in which they were generated.

The *Research Data Alliance* (RDA, at rd-alliance.org) has the goal of promoting data sharing to accelerate data discovery, use, reuse, standards and harmonisation. RDA is organised into working groups and interest groups around different themes; the groups comprise experts from different countries and belonging to different communities (academia, the business sector, governmental agencies). RDA was created in 2013 by a core group of organisations: the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. Individuals may also apply for membership; today, RDA counts around 1 600 members from more than 70 countries.

The *EMBL-EBI* (www.ebi.ac.uk) – The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL), a non-profit organisation and basic research institute funded by 20 member states in Europe, Israel, and one associate member (Australia). EBI is a major European laboratory for the life sciences; it provides freely available data from life science experiments in the field of molecular biology. EBI maintains the world's most comprehensive collection of freely available up-to-date molecular databases. Its services allow scientists to share data, perform complex queries, and analyse the results. Database users can generally work locally by downloading EBI data and software, and can access different resources through EBI web services. EBI serves millions of researchers world-wide who are active in multiple fields of the life sciences, from clinical biology to agri-food research. EBI also offers training programmes to maximise the benefits of data available in the life sciences to researchers in academia and the business sector. Some 20% of EBI users are engaged in industrial R&D, and EBI has developed an Industry Programme to collaborate specifically with firms active in bio-informatics. EBI addresses the specific needs of industry in other ways: from public-private partnerships to develop better and safer medicines for patients (the Innovative Medicines Initiative), to the provision of data infrastructure and services to SMEs, enabling bio-informatics spin-offs from EMBL and facilitating key pre-competitive research projects with industrial partners. EBI is located on the Wellcome Trust Genome Campus in the United Kingdom.

The *European Organization for Nuclear Research* (CERN, at web.cern.ch) is an international research laboratory containing the world's largest and most complex scientific instruments to study fundamental particles. CERN was founded in 1954 and is located on the Franco-Swiss border. It was one of Europe's first inter-country joint ventures, and it has now 21 member states. CERN actively supports open access efforts. In particular, since 1 January 2014, CERN hosts the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3). Supported by partners in 24 countries, SCOAP3 works in collaboration of over one thousand libraries, library consortia and research organisations to make available free of charge scientific articles in the field of high-energy physics. The consortium benefits from the support of funding agencies and has been established in co-operation with the publishing industry. As a result its efforts, articles are open access, the copyright stays with the author(s), and licence agreements allow text and data mining.

1. More information available at: www.icsu.org/general-assembly/news/ICSU%20Report%20on%20Open%20Access.pdf.

Box 7.3. **The OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information**

This OECD recommendation was developed by the OECD Committee for Information, Computer and Communication Policy and by its Working Party on the Information Economy. The recommendation was adopted by the OECD Council in 2008. The recommendation refers to the following items:

- *openness* – maximise the availability of public sector information for use and reuse, by taking into account limitations related to privacy and security, and in accordance to copyright
- *access and transparent conditions for reuse* – promote use and reuse by removing unnecessary restrictions. Improve access over the Internet and in electronic form
- *asset lists* – strengthen awareness of the kind of public sector information that is available for reuse
- *quality* – ensure the use of methodologies to enhance the quality and reliability of public sector information
- *integrity* – maximise the integrity and availability of information
- *new technologies and long-term preservation* – improve the usage of interoperable systems, and develop the necessary skills for preservation and access
- *copyright* – intellectual copyright should be respected
- *pricing* – price public sector information transparently when it is not provided for free
- *competition* – ensure that pricing strategies take into account considerations of unfair competition when both public and business users provide value added services
- *redress mechanisms* – provide appropriate transparent complaint and appeals processes
- *public-private partnerships* – promote public-private partnerships where appropriate and feasible
- *international access and use* – seek consistency in access regimes to facilitate cross-border use, improve interoperability, and support international co-operation and co-ordination
- *best practices* – share best practices and exchange information on implementation, the education of users, cost and pricing models, copyright handling, monitoring performance and compliance, and the wider impacts – on innovation, entrepreneurship and economic growth, as well as socially.

Source: Adapted from OECD, 2008.

Box 7.4. **The OECD principles and guidelines for access to research data from public funding**

In 2004, ministers of science and technology of OECD countries met in Paris and discussed the need for international guidelines on access to research data. At that meeting a *Declaration on Access to Research Data from Public Funding* was adopted by OECD countries. Following that meeting, the OECD's Committee for Scientific and Technological Policy (CSTP) launched a project to develop a set of principles and guidelines. Those that resulted from this project were approved by CSTP in October 2006 and then endorsed by the OECD Council. The principles can be summarised as follows:

- *Openness* – Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.
- *Flexibility* – Flexibility requires taking into account the rapid and often unpredictable changes in ICTs, the characteristics of different research fields, and the diversity of research systems, legal frameworks and cultures among member countries.
- *Transparency* – Information on research data and data-producing organisations, documentation on the data, and conditions attached to the use of data should be internationally available in a transparent way, ideally through the Internet.
- *Legal conformity* – Data access arrangements should respect the legal rights and legitimate interests of all stakeholders in the public enterprise. Restriction to access may be for reasons of national security; privacy and confidentiality; trade secrets and intellectual property rights; protection of rare, threatened or endangered species; or legal processes.
- *Protection of intellectual property* – Data access arrangements should consider the applicability of copyright and other intellectual property laws that may be relevant to publicly funded research databases (as in the case of public-private partnerships).
- *Formal responsibility* – Access arrangements should promote the development of rules and regulations dealing with the responsibilities of the various parties involved; should be developed in consultation with representatives of all parties affected; and should be responsive to factors such as the characteristics of the data and their potential value for research purpose. Data management plans and long-term sustainability should also be considered.
- *Professionalism* – Institutional arrangements for the management of research data should be based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.
- *Interoperability* – Access arrangements should consider the relevant international data documentation standards.
- *Quality* – The value and utility of data depend to a large extent on the quality of the data. Particular attention should be paid to ensuring compliance with explicit quality standards.
- *Security* – Attention should be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of research data.
- *Efficiency* – One of the central goals of promoting data access and sharing is to improve the efficiency of publicly funded scientific research so as to avoid expensive and unnecessary duplication of effort. This also involves cost and benefit analysis to define data retention protocols, the engagement of data management specialist organisations, and the development of new reward structures for researchers and database producers.
- *Accountability* – The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and research funding agencies.
- *Sustainability* – Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure.

Source: OECD, 2007.

Box 7.5. **Open Data under Horizon 2020**

The new EU framework programme for research and innovation 2014-20 includes a pilot project on open research data. Researchers involved in projects participating in the pilot will be asked to make publicly available the data constituting the basis of the project research results; these can then be used by other researchers and projects, innovative industries and citizens. The researchers will also be asked to develop data management plans. Over 2014-15 the Open Research Data Pilot will receive around EUR 3 billion. The pilot project targets all key thematic areas of Horizon 2020 (future and emerging technologies, research infrastructure, leadership in enabling and industrial technologies, societal challenges, and science with and for society).

Researchers participating in the pilot have the possibility of opting out of it to protect intellectual property or personal data; for security concerns; or if the main objective of their research can be compromised by making data openly accessible. The pilot has the goal of providing a better understanding of what supporting infrastructure is needed, as well as the role of limiting factors such as security, privacy and data protection that could induce researchers to opt out. The Pilot also aims to contribute to a better understanding of the best mechanisms to define the right incentives for researchers to curate and share the research data they produce. The pilot will be closely monitored during the implementation phase of Horizon 2020.

Source: http://europa.eu/rapid/press-release_IP-13-1257_en.htm.

Data-sharing challenges and opportunities

An essential element for the usefulness of data-sharing efforts is the quality of the publicly released data. In many scientific communities there is as yet no standard data quality assessment protocol, as there is for scientific publications (Brase et al., 2009). As highlighted for example in Royal Society, 2012 and Dallmeier-Tiessen et al., 2012, data have little value if they do not meet minimum quality criteria. Data quality here implies being not only accessible (for example available on the Internet), but also intelligible, assessable, trustworthy and, of course, reusable. In this respect the development of detailed data-sharing information and metadata is essential for the further use of the same data by multiple teams of researchers. Additional challenges relate to privacy, or the definition of the ownership of the data itself. The OECD Global Science Forum has recently identified nine challenges related to data sharing (see Box 7.6).

As in the case of access to publications, data collection, curation and sharing vary by scientific discipline. Some fields have been traditionally more data-intensive than others, especially those making use of large-scale experiments managed by teams of hundreds of researchers (for instance the case with data generated through particle accelerators at CERN), or making use of machine-collected or -generated data. Researchers belonging to other scientific disciplines, notably in the social sciences and humanities, traditionally collect and built their own data sets, in some cases manually or by developing surveys and questionnaires. That makes this kind of the data set more tied to the individual researcher, and therefore less easily ready to be shared without proper curation, cleaning and metadata compilation (see Box 7.6, Challenge 4).

As highlighted by the OECD Global Science Forum (Box 7.6, Challenges 7 and 9), scientists and researchers do not have necessarily the incentives or the skills to perform these tasks, since proper curation and dissemination of data sets are costly and time-consuming, and can be even considered another type of scientific output (Uhlir, 2012). In addition, scientists and researchers traditionally compete to be the first to publish scientific results, and may not see the benefits of disclosing information on the data they want to use to produce as yet unpublished research outcomes.

Box 7.6. **Data sharing: Nine challenges identified by the OECD Global Science Forum**

The OECD Global Science Forum has recently identified a number of challenges related to data-driven and evidence-based research.

Challenge 1 – Massive amounts of digital data are being generated at unprecedented scale, partly thanks to the advent of ICTs. The reliability, statistical validity and "generalisability" of new forms of data are not yet fully understood.

Challenge 2 – While administrative, survey and census data are widely collected by national statistical agencies and government departments, micro-data records are much less available.

Challenge 3 – New forms of personal data, such as social networking data, are increasingly created and collected. The use of those data may generate risks to individuals' privacy.

Challenge 4 – Barriers to legal, cultural, language and proprietary rights of access hinder cross-national collaboration and international data exploitation, especially in the social sciences.

Challenge 5 – Global research agendas require increasingly interdisciplinary and international co-ordination.

Challenge 6 – Collaboration and experience sharing across countries in the development of comparable data resources is necessary to fully exploit the potential of data sets.

Challenge 7 – Researchers often lack the resources or the skills to make sure that the data they use, gather and produce are available for reuse.

Challenge 8 – National investments in skills and infrastructure related to data creation and curation are essential to avoid the risk of data loss or degradation.

Challenge 9 – Researchers need to have the right set of incentives to ensure effective data sharing.

Source: Adapted from OECD 2013.

A possible solution to the above-mentioned disincentives is data citation: researchers wishing to be acknowledged for their work could release data sets through mechanisms similar to the one already in place for citations of academic articles (Mooney and Newton, 2012; CODATA-ICSTI, 2013). Data citation is not, however, necessarily a standardised or widely accepted concept in the academic community. Some scientists see data citation as limiting citation to scientific articles. Funding agencies in some cases question the idea of recognising individuals as data authors, and traditional bibliometrics indicators are not yet taking into account non-article citations (Costas et al., 2013). There are in addition technical barriers restricting the development of data citation and related metrics. These include incompatibility – in machines and software, data file structures, data storage and management (Groves, 2010). Some organisations, such as DataCite (www.datacite.org; see Box 7.7), have been active in promoting conditions that will enable data citations, such as unique data object identifiers for data sets.

Box 7.7. **Organisations promoting data citations**

DataCite (www.datacite.org) is an international non-profit organisation established in London, United Kingdom since 2009. DataCite has the aims of promoting access to research data through the Internet; supporting open data archiving; and enabling verification of scientific results and the reuse of data for further studies. To facilitate data release, DataCite helps researchers with the unique identification and attribution of data sets for citation purposes, and supports journal publishers in establishing linkages between published articles and data sets. In addition, the organisation supports data centres by providing identifiers for data sets and defining workflows and standards for data publication.

ORCID (www.orcid.org) is a non-profit, community-driven organisation whose aims are to create and maintain a registry of unique researcher identifiers, and to link research activities and outputs on the basis of these identifiers. Research identifiers can be assigned not only to scientific articles but also to other forms of research output, including equipment, experiments, patents and data sets. Apart from the registry, ORCID provides application programme interfaces that support system-to-system communication and authentication. ORCID codes are available through open source licence.

Sources: DataCite website, at www.datacite.org; Orcid website, at www.orcid.org.

According to the EU FP7 research project Opportunities for Data Exchange, or ODE (Kotarski et al., 2012), there are certain unique features of data citation, owing to the particular properties of data sets. For instance, data sets may be of very different sizes and it is not always clear to what specific elements inside the data sets scholars are referring to – or, in the case of updates to the data sets, which version to cite. According to the ODE project, some of the good practices/challenges related to data citations are as follows:

- citation of the data set with identifier should be listed in a work's reference/bibliography, to enable tracking of citation metrics
- publishers need to provide authors and referees with guidance on data citation
- there is no clear agreement on the persistence or longevity requirements for data sets to be considered citable or cited
- there is lack of clarity and agreement on what authorship of data set means
- researchers need to promote awareness in their communities of the benefit of data citation and follow agreed data citation guidelines.

Other possible vehicles for publishing data sets are data journals, i.e. collections of scientific articles that specialise in publishing data papers. Data papers are articles with the primary purpose of describing data sets, rather than reporting on scientific investigation and analysis. Data papers contain facts about and descriptions of data. These papers aim to be a citable source of information on data that brings credit to the scholars who produced and described the database, disclosed detailed information on a data set, and brought the existence of the data to the attention of the scientific community (Chavan and Penev, 2011). Data journals may target broader scientific areas as well as specific domains, such as earth system science or geoscience.

7.3. Policies and practices: OECD countries and beyond

OECD and non-member countries are increasingly developing frameworks, guidelines and initiatives to encourage greater openness in science, both at national and supranational level (see Boxes 8.3 and 8.4 for recent OECD principles and guidelines related to access to research data or public sector information). Most of the respondent countries to the *OECD Science, Technology and Industry Outlook 2014* policy questionnaire highlighted recent changes in their policy framework for open science (OECD, 2014b). However, most open science efforts are broadly targeted, with the focus on developing the infrastructure necessary to promote sharing of research content and the collaboration platform of scientists and academics, or on requiring that the results of publicly funded research (mostly in the form of papers or publications) be made available on line, free of charge to the reader.

In other cases, open research data initiatives are nested into broader national open government or PSI strategies; they mainly have to do with releasing large sets of publicly collected data (such as weather or GIS data), which may or may not include research data. Some OECD countries and non-members, however, have adopted specific initiatives to promote open research data. Finally, a number of countries are modifying national copyright frameworks to promote greater use and reuse of data and research materials in science and research. This section will briefly review examples of infrastructure development for data sharing and open government initiatives, and then will cover more in detail those initiatives in OECD or non-member countries more specifically targeting open research data, the development of data analysis skills, and the involvement of business sector actors.

Developing the infrastructure for data sharing

Several countries are developing the infrastructure necessary to collect, store and disseminate research results (both articles and data). Examples of these initiatives include the creation of online repositories, databases, archives and digital libraries, as well as platforms containing information on R&D projects and researchers' CVs.

Argentina developed the SICyTAR (Sistema de Información de Ciencia y Tecnología Argentino) database containing information on the CVs, publications and affiliations of researchers. Colombia, Estonia, Greece, Mexico, South Africa, Korea and Poland have created national networks of repositories and digital libraries. Finland has launched an infrastructure roadmap to promote open science. People's Republic of China (hereafter 'China') has developed online platforms for data and publication archiving. Australia has been developing the eResearch infrastructure to help research organisations address the issues of data storing, accessing, analysing, modelling, sharing and manipulating. Germany also supports the development of open access libraries and archives, and has launched many initiatives on open access infrastructure for research outputs. In New Zealand, the Kiwi Research Information Service is the most comprehensive selection of publicly available research papers and related resources; materials include papers, conference materials and PhD theses, but also data sets. In France, many e-libraries and infrastructures for sharing research outputs have been developed, such as the National Hyper Articles Online Platform (HAL). In the United Kingdom, the E-infrastructure Leadership Council (ELC) advises the government on e-infrastructure aspects such as networks, data stores, computers, software and skills. It also advises BIS (business, innovation and skills) ministers on its antecedent implementation and

development. It works in partnership with stakeholders across the academic community, industry, government and society.

The European Commission has also been active in promoting the development of EU and member country repositories and platforms. In Poland, the new research data centre OCEAN has been announced, and will be operational in 2015 with the aim of providing the e-infrastructure for storage of open research data, as well as facilities and the expertise to undertake big data analytics. The target population of this centre is the entire Polish research community.

Including research data in broader open government/PSI agendas

Open government initiatives are high on the agenda of many OECD countries. The release of data collected by public administration may in some cases include research data or data on R&D activities. In addition, data provided through open government initiatives is contributing to open science, by giving to researchers and academics the opportunity to use, analyse and reuse those data, and potentially to advance in scientific discoveries and innovations.

The open government initiative in Canada is committed to making government data available and to promoting open data as a vehicle to further enhance the commercialisation of public research. In Sweden, the government supports the public release of data sets. In France, ETALAB co-ordinates the French open government data efforts, but it does not focus specifically on research data. In Norway, data.norge.no is the national portal for the storage and dissemination of publicly available data sets. A special Norwegian licence for public data was also developed in 2011 to facilitate reuse of data.norge data.

In the United Kingdom the Open Data Institute (ODI) has been created with the aim of promoting an open data culture to create economic, environment and social value. ODI helps to unlock data supply, generate demand and create and disseminate knowledge. ODI supports the UK Open Data agenda, and works in close co-operation with research centres and the business sector to advise how best to use and manage available data. Other open data initiatives in the United Kingdom focus on the personal data of consumers (for more informed choices), health and social welfare data, transport data, and education data. The US Open Government Initiative aims to improve access to federal data, including research data, to the public, through searchable, machine readable formats. Data are accessible through the Data.gov portal that also provides descriptions (metadata) on how to access and use the data sets. In Poland, access to public sector information is regulated by the Act of 2001. A new 2014 regulation defines the information assets as being places in dedicated repositories and the frequency of updates.

Targeting of research data

Some OECD and non-member countries have been recently implemented initiatives specifically to promote research data-sharing practices. The European Commission has in addition launched an open research data pilot project, under the new framework programme 2014-20, Horizon 2020 (see Box 7.5).

In Chile, since 2010, the National Commission of Technological Research (CONIYT) has been working on three main initiatives to generate access to research data from public funding: i) the development of a programme to manage research data and scientific information produced with public funding; ii) an analysis of the state of the art of

accessing and managing research data and scientific information; iii) the design of an institutional policy of management of research data and scientific information.

In China, the first pilot project of the Scientific Data Sharing programme was launched in 2001. At first restricted to meteorological data, the programme is now offering data-sharing opportunities in 24 scientific sectors, including the environment, agriculture, population and health. Major Chinese scientific institutions are users of the programme, which has promoted a data-sharing culture in Chinese research institutions.

The Finnish Research Data Initiative (preparatory phase 2009-11, project phase 2011-14) is part of the broader Finnish Open Science Roadmap, launched in 2014. The initiative was established to develop policy guidelines, improve ICT interoperability, establish collaboration platforms and provide service related to research data storage, long-term preservation, metadata, etc. The initiative was launched following completion of a survey of the status of research data management, which highlighted the need for action in these areas. The initiative identified a list of principles and guidelines related to research data, including:

- The principle of openness governing not only research data but also research methods such as computer models.
- Openness adhering to ethical principles and legal frameworks.
- Creative common licences are recommended to allow machine readability.
- Long-term preservation of data and infrastructure.
- Research data need to be documented and described.
- The need for each organisation to develop a data policy and related guidelines.
- Making references to data and methods developed by others is encouraged.
- Professional skills related to data need to be developed.
- Text and data mining should be allowed.
- Interoperability of international standards will be followed when possible.
- Researchers will be requested to consider the ownership of the data at the early stages of their research projects.
- Data management plans will be requested in every research plan.
- Copyright legislation will be amended to allow text and data mining for research purposes.

The United Kingdom is promoting the open release of data emerging from publicly funded research, as stated in the 2011 Research Strategy for Growth. Research Council guidelines on open access encourage the disclosure of research data and information on how to access it; however, data are not mandated to be open. A Data Capability Strategy has also been developed. The strategy focuses on three overarching aspects: i) human capital, a skilled workforce and data-confident citizens; ii) the tools and infrastructure necessary to store and analyse data; iii) data themselves as an enabler of research, innovation and economic development. In addition, a network of Administrative Data Research Centres (ADRN) has been created to promote research from interlinked administrative data. In 2012 the government agreed to introduce an exemption to the Freedom of Information Act to prevent the premature disclosure of research data and consequently having non-peer-reviewed data or results interpreted incorrectly by

journalists and other communities. In 2012, the government also established the Research Sector Transparency Board, which advises the government on how to increase access to research data. In addition, the UK copyright law framework has been recently amended (in 2014) to introduce targeted copyright exceptions that allow text and data mining for non-commercial research without specific authorisation.

In the United States, since the COMPETES Reauthorization Act of 2010, the US Office of Science and Technology Policy (OSTP) co-ordinates with US federal agencies to develop policies to promote public access to the results of federally funded research, including digital data. In 2011, two working groups on access to digital data resulting from government-funded research and access to scholarly publications were created. As a result of the efforts of these working groups, in February 2013 the OSTP issued a memorandum to federal government science agencies requesting the development of plans for increasing public access to the results of federally funded research, in particular to scientific publications and digital data. The OSTP memorandum establishes the following principles for agency policies:

- maximise free of charge general public access to digitally formatted scientific data collected with the support of federal funds, while protecting confidentiality and personal privacy, recognising intellectual property rights and preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden.
- make sure that researchers develop data management plans, including long-term preservation, or explain why long-term preservation and access are not possible
- allow the inclusion of appropriate costs for data management and access in proposals for federal funding of scientific research
- ensure evaluation of the data management plans submitted
- promote the storage of data in publicly accessible database repositories
- promote co-operation with the private sector, including by means of public-private partnerships to improve data access and compatibility
- develop mechanisms for data attribution and identification, to acknowledge researchers' open data efforts
- Support training, education, and workforce development related to scientific data management, analysis, storage and preservation.

In addition, individual American research institutions have developed open data policies. The US National Institutes of Health (NIH) developed a data-sharing policy as early as 2003 to encourage NIH-funded researchers to share scientific data sets. The policy requires of applicants requesting USD 500 000 or more of funding to include a data-sharing plan in the grant application procedure, or to justify why data sharing is not possible. Data-sharing plans should include a description of whether and how data will be made available, including how to account for protection of privacy, confidentiality, security and intellectual property rights; a description of the data to be shared; the timeline of sharing; data formats; procedures related to data-sharing agreements; and limitations on the use of data. The policy requires that data be shared no later than the acceptance for publication of the main findings from the final data set. Other NIH data-sharing policies are specifically developed to target different types of scientific data, collected during different projects researching different aspects of medicine, health and biological research.[9]

Investing in skills

In order to promote a transition towards data-driven research and innovation, a number of countries are investing in the skills necessary for data analytics. Data science training can be promoted at the postgraduate level, in order to provide PhD students or researchers with the skills necessary to make more extensive use of data analytics, or to promote the development of data management plans. Such training can also be promoted at the undergraduate level with the creation of new or adapted university curricula focused on science, technology, engineering and mathematics (STEM) skills, with the aim of delivering data scientist or data engineer degrees (see Chapter 6 on skills implications of DDI).

The European Commission, for example, estimates that Europe alone will face a shortage of up to 900 000 ICT professionals by 2020, due to a severe skill mismatch. To overcome this problem, in March 2013 the European Commission launched the Grand Coalition for Digital Jobs, a multi-stakeholder partnership to promote collaboration among business, education providers and public and private actors, to attract the young to ICT curricula and retrain unemployed people.

Countries severely hit by the financial crisis, such as Portugal, see the re-skilling or training of personnel in data-intensive areas – such as big data, data management and business analytics – as an opportunity to reduce the high domestic unemployment rates. Other countries are promoting the development of skills or ad hoc training in the broader context of open data and PSI agendas. The Italian Agency for Digital Italy has developed national guidelines for the exploitation of open data sets, bringing awareness of open access to data to academics, students, researchers and citizens. In Finland, in 2013, a working group established by the Finnish Ministry of Education and Culture produced a Data Management Guide that covers various aspects of that responsibility. The guide contains a checklist to assist researchers in data management planning, and provides information on available related services.

Other countries are developing specific training centres or higher education programmes. Poland has established a research data centre to provide extensive training curricula for big and open data management, and analysis to support interdisciplinary research in different scientific areas. In the United Kingdom, many initiatives are devoted to providing skills training in numerical data subjects, to teach students and academics how to use the big data sets emerging from open data efforts. The UK Engineering and Physical Sciences Research Council has announced a number of new Centres for Doctoral Training focusing on big data to be developed in several universities, including the University of Nottingham, the University of Edinburgh and the University of Oxford. In addition, the UK Economic and Social Research Council, the Higher Education Funding Council for England, and the Nuffield Foundation are supporting delivery of data science undergraduate programmes in 15 universities across the country, to promote quantitative social science training.

In the United States, the 2013 Office of Science and Technology Policy memorandum (see previous section) encourages the support of federal science agencies for training, education and workforce development related to scientific data management, analysis, storage, preservation, and stewardship. In addition, individual institutes have developed ad hoc skills policies. For example, the National Institutes of Health focuses on skills development through the initiative Big Data to Knowledge. The goal is to develop teams of researchers skilled in the science of big data and to increase the level of competencies in data usage and analysis across the biomedical research workforce. Governments also

focus on helping the ICT sector to better understand labour market shortages. Canada's Sectoral Initiatives Program, for example, includes training programmes to address industry skill mismatches by better aligning the skills of ICT and data specialists with the needs of employers (see Chapter 6).

The role of the business sector

Business sector actors are involved in data-intensive scientific activities, in several ways. Business organisations can be the actors providing the infrastructure to store, maintain and curate large-scale data sets and related services, such as the provision of processed data and of specific database extractions that may be relevant for the research community. With respect to open access to scientific publications, private scientific publishers have traditionally been the ones offering the services of article peer review (to guarantee quality) and digital and paper publication. Given the increasing data intensity of science, the range of services offered by business organisations to the research community are likely to diversify and expand. Some new services have already emerged, like those based on software and applications to catalogue and organise scientific articles and libraries or the data sets containing bibliometric records (see Box 7.8).

Box 7.8. **Start-ups for open data: The case of Figshare**

Figshare is an online digital repository of research data that includes figures, images and videos. Figshare was launched in 2011 by a PhD student in London. Figshare users can make their research outputs available to other researchers or users in a "citable, sharable and discoverable manner". This means that they can easily share data, search for data sets, and get credit for the data sets they upload on the website (through data set citations). Figshare allows users to upload any file format.

Figshare has recently established partnerships with other open science business actors, such as the open access publishing company PLOS ONE, the Nature Publishing Group, Taylor and Francis and F1000. These partnerships allow authors to directly upload data sets linked to papers online and ORCID (see Box 7.2), a service promoting data set citations. In addition, Figshare tracks the number of downloads of research materials, and it is often used as a source of alternative metrics. All files uploaded on Figshare are released under a creative commons licence.

Figshare stores more than 1.5 million files. From its original location in the United Kingdom it has today expanded to the United States and Romania. Users can sign in and upload or download content on Figshare for free. The company does however charge for premium services (such as larger private online storage space or private collaborative spaces) to individual researchers, and for services offered to publishers. In addition, it recently launched Figshare for Institutions, a service that is explicitly designed for research institutions around the globe.

Source: The Figshare website, at www.figshare.com.

The business community is also involved in open science and open data, as private firms can be the beneficiaries of open access publications and data that they use to develop new products and services and promote innovation more generally. In addition to entirely privately funded initiatives, joint public-private initiatives have begun to emerge for the delivery of services related to open science and open data. In other cases, the business community may hold and collect culturally and scientifically valuable material. There may be a

public interest in curating and preserving that material, and public-private partnerships can be developed to support preservation. This is the case, for instance, with the US National Film Preservation Board and the National Recording Preservation Board, which each year select items (often belonging to private organisations) to be added to their national registries (Blue Ribbon Task Force, 2010). Other examples of public-private partnerships involved in open science have recently been developed in Finland (Box 7.9).

Box 7.9. **Public-private partnerships for open science: The Finnish SHOK and DIGILE**

The Strategic Centres for Science, Technology and Innovation (SHOK in Finnish), established in Finland, are new public-private partnerships aiming to speed up innovation processes through renewal of clusters and the development of radical innovations. SHOK centres develop and apply new methods for co-operation, co-creation and interaction.

One such strategic centre is directly involved in open science: DIGILE, whose mission is to create digital business ecosystems that enable new global growth business for DIGILE's owners and partners. DIGILE aims not only to bring together R&D communities, but also to make sure that the results of scientific processes are understood, applied and adopted by companies. There are over 30 partners, including companies, research institutes and universities. DIGILE's strategy for 2015 focuses on data sharing, management and reusing as well as innovative data-intensive business models and services.

Source: DIGILE's website at www.digile.fi, Finland country note in OECD (2015b), and Myllymäki, P. (2013), "Data to Intelligence (D2I) Research Programme on Intelligent Data-Driven Services" presentation on www.digile.fi.

In addition to public science and innovation actors, private non-profit organisations and foundations may play a significant role in developing, raising awareness of and encouraging an open science culture. They may not only fund open access research and introduce requirements in grant agreements, but also develop and facilitate the creation of networks of stakeholders worldwide.

Open research data and IP protection

The expansion of open access policies to publicly funded research data raises a number of legal and policy issues that are often distinct from those concerning the publication of scientific articles and monographs. Since open access to research data – unlike publications – is a relatively new policy objective, less attention has been paid to the specific features of research data.[10]

Internationally, the protection afforded to databases (as collections of data or other elements) is established – or confirmed – by both Art. 10(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) and in the almost identical Art. 5 of the WIPO Copyright Treaty (WCT). According to the former:

> *Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creation shall be protected as such... (Art. 10(2) of the TRIPS Agreements)*

Databases are a particular subject matter that is protected by copyright under certain circumstances, but that in some areas – namely within the European Union, Japan and South Korea – is also protected by a so-called sui generis database right (SGDR). This additional layer of protection is found in some countries and is afforded to databases regardless of the intellectual creation (i.e. "selection or arrangement") that may or may not be present. What is protected instead is the investment in making the database, i.e. in the obtaining, verification or presentation of the data. This type of right is typical, for example, of the EU Database Directive. It should be borne in mind that while the original protection afforded to databases focuses on the arrangement or selection without extending to the content of the database, the SGDR offers protection against the copy of substantial parts of the database – that is to say it extends, at least to some extent, to the data themselves.

With respect to research data in particular, the complexity of their rights status in Europe and other jurisdictions arguably has the potential to adversely affect the reuse opportunities of the collections of scientific data, given the difficulty – both for research institutions making the database available and for prospective reusers – in determining each time whether a certain database is covered by a sui generis right and in which measure reutilisation and extraction can take place freely. It is uncertain whether the use of compilations or databases for purposes of research and private study in general, and text and data mining in particular, is covered by any relevant exception on copyright or the database right. The use of Creative Commons licences 4.0 may alleviate the uncertainty, by clearly stating what can and cannot be done with the licensed material.

Another legal issue that comes into play in the context of open data, but that is less sensitive in the case of open access to scientific publications, is privacy and personal data protection. Data gathered in the course of research often contain personal data (e.g. medical records), and so opening such data has to respect the rights of data subjects (Lane et al., 2014). This does not mean that the data cannot be opened, but it does calls for implementing protective procedures (see Chapter 5). One of them is anonymisation, which may lead to the inapplicability of the whole personal data protection regime. In some cases however, research results depend on personal data and personal characteristics, and a complete anonymisation is therefore not always possible. Additionally, not all anonymisation techniques are effective (Narayanan and Shmatikov 2008).

Hence, the open science movement faces more challenges with regard to data than to scientific publications. While sui generis rights issues are to a large extent addressed in model licences (Creative Commons, Open Database Licenses), there are still no standardised procedures to follow with regard to privacy. Access and reuse of PSI also remain evolving subjects.

7.4. Key findings and policy conclusions

The scientific enterprise is evolving and becoming more data-intensive. Recent advances in technology are radically changing the way in which data are collected, stored and used. Data are collected and generated more quickly and in larger volumes than ever before. While not entirely novel – some scientific fields have always relied on large data sets – the increase in existing data as a source of inference for scientific evidence is spreading and deepening across scientific fields, including the social sciences. Another characteristic is the role of machines and algorithms to make sense of the large amounts

of data. Consequently, science requires continued investment in infrastructure, both hard and soft (e.g. data science skills).

However, several barriers to data sharing still remain. Although there is a clear potential to improve science and innovation systems, barriers still remain with respect to data-sharing efforts. Some barriers are of a technical nature, such as issues related to storage, the technical infrastructure to allow data sharing, interoperability and standards. Other types of barriers are related to the lack of an open data culture or the disincentives that researchers and scientists face with respect to the disclosure of data sets, especially relative to research at the pre-publication stage. This raises the question of the "optimal" level of openness to boost research and innovation without discouraging data collection from individual researchers.

Open science and open data are hot topics in many OECD countries and beyond. Many OECD countries and non-members have recently adopted and developed initiatives to promote greater openness and sharing of publicly funded research outputs, in the form of both articles and data. As in the case of policies to promote access to publications, the policy measures to promote open data may be developed and adopted by diverse sets of actors at both national and sub-national levels, as well as at the institutional level (universities, public research institutes). Policy measures may include efforts and initiatives such as mandatory rules, incentive mechanisms or enablers:

- *Mandatory rules* are often implemented in the form of requirements in research grant agreements, or in some cases defined in national strategies or institutional policy frameworks.
- *Incentive mechanisms* may take the form of financial incentives to cover the release of data sets. They may also be in the form of proper acknowledgment of open data efforts of researchers and academics, for instance in the form of data set citations or career advancement mechanisms partly based on metrics that take into account open science or data-sharing initiatives.
- *Enablers* are for example the infrastructure developed to share data; initiatives undertaken to develop an open science and open data culture; amendments to the legal framework to make it increasingly open science-friendly; or development of the skills necessary for researchers to share and reuse the research outputs produced by others.

(Cross-)subsidising the production of scientific knowledge requires picking winners (users or applications) by assessing (social) demand for a public good such as scientific knowledge based on the (social) value it creates. Governments can support the production of public goods i) by directly producing these goods, or ii) by supporting private firms' production of public and social goods through (e.g.) research grants, procurement programmes, contracted research and tax incentives. All these strategies raise a number of issues, including but not limited to difficulties in picking winners and losers, and the fact that resources are limited. Open access regimes can be a more efficient and politically attractive "indirect intervention" to support the production of public and social goods (see Chapter 4). As Frischmann (2012) highlights, "commons management is not a direct subsidy to … users who produce public or social goods, but it effectively creates cross-subsidies and eliminates the need to rely on either the market or the government to 'pick winners' – that is, to prioritise or rank … users worthy of access and support".

There is *tension between open research data and IPRs,* and a balance must be struck between efforts to promote open data in science and efforts to promote commercialisation

of public research, especially in the case of public-private partnerships involving companies. The tension can however be lessened by policies that clarify IP ownership and promote non-exclusive licensing possibilities, as well as by greater IP awareness among researchers, including of copyright and databases.

There is an *increasing need to develop skills* related to data curation, cleaning and preservation, both in the research community and in society in general. A number of countries have begun to address the shortage of data management skills by requiring researchers to develop data management plans in grant agreements or by developing training programmes or new academic curricula. There is a general need to understand the demand for those skills and the type of skills currently lacking in the research community and beyond to fully reap the benefits of data sharing.

Coherent guidelines are needed to promote better access to data across the economy. Many of the barriers to data access in science are common to those encountered in other domains, including open government data (see Chapter 10). Existing frameworks that promote better access to data, some of which are sector specific, may need to be reviewed and eventually consolidated to foster coherence among public policies related to data access, linkage and reuse. This would also include the OECD Council Recommendations promoting better access to data, including in particular the OECD (2008) *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* of 30 April 2008, and the OECD (2006) *Recommendation of the Council concerning Access to Research Data from Public Funding* of 14 December 2006, both of which are currently under review.

Notes

1 More information is available at: www.humanbrainproject.eu.

2 More information is available at: www.nih.gov/science/brain/.

3 More information is available at: www.osdd.net.

4 More information is available at: www.icsu.org/future-earth.

5 More information is available at: www.bios.net.

6 More information is available at: www.lancaster.ac.uk/fss/projects/ieppp/amateurs/.

7 Kaggle has partnered with large organisations including NASA and Deloitte, see www.kaggle.com.

8 An ongoing study of the US GovLab Academy at New York University – an online community that uses technology and innovation to solve public domain problems – is attempting to understand how US companies use open government data, through the Open Data 500 project. The project is analysing US-based companies (including international companies with a major presence in the United States) using open government data, a critical resource for their business. Most of the companies in the study belong to the technology, financial and business/legal services industries. According to the study, the most widely used data originates from the Department of Commerce, followed by the Department of Health and Human Services and the Securities and Exchange Commission.

9 See www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html; http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm; and http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm.

10 This section is derived from a background paper prepared for the OECD by Lucie Guibault and Thomas Margoni, "Legal aspects of open science and open data", 2014.

References

Anderson, C. (2008), "The end of theory: The data deluge makes the scientific method obsolete", *Wired*, 23 June, www.wired.com/science/discoveries/magazine/16-07/pb_theory/, accessed 15 April 2015.

Anderson, W.L. (2004), "Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data", *Data Science Journal*, 30 December, pp. 191-202, www.jstage.jst.go.jp/article/dsj/3/0/191/_pdf.

Arzberger, P. et al. (2004), "Promoting access to public research data for scientific, economic and social development", *Science*, 3 November, pp. 1777-78.

Bell, G., T. Hey and A. Szalay (2009), "Beyond the data deluge", *Science*, 6 March, pp. 1297-98, www.cloudinnovation.com.au/Bell_Hey_Szalay_Science_March_2009.pdf, accessed 15 April 2015.

BIAC (2011), "BIAC thought starter: A strategic vision for OECD work on science, technology and industry", Business and Industry Advisory Committee to the OECD, 12 October, http://biac.org/wp-content/uploads/2014/05/02-FINAL-11-10-_THOUGHT_-STARTER-_-VISION_-PAPER.pdf, accessed 6 May 2015.

Blue Ribbon Task Force (2010), "Sustainable economics for a digital planet: Ensuring long-term access to digital information", Blue Ribbon Task Force on Sustainable Digital Preservation and Access, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf, accessed 15 April 2015.

Bollier, D. (2010), *The Promise and Peril of Big Data*, Aspen Institute, Washington, DC.

Brase, J. et al. (2009), "Approach for a joint global registration agency for research data", *Information Services and Use*, Vol. 29, No. 1, pp. 13-27.

CEBR (2012), "Data equity: Unlocking the value of big data", Centre for Economics and Business Research Ltd, London, www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf, accessed 15 April 2015.

Chan, L., B. Kirsop and S. Arunachalam (2005), "Open access archiving: The fast track to building research capacity in developing countries", *SciDevNet*, November.

Chavan V., Penev L., (2011), "The data paper: a mechanism to incentivize data publishing in biodiversity science", *BMC Bioinformatics*, **12**(Suppl 15):S2, doi:10.1186/1471-2105-12-S15-S2.

CODATA-ICSTI (Committee on Data for Science and Technology – International Council for Scientific and Technical Information) (2013), "Out of cite, out of mind: The current state of practice, policy and technology for the citation of data", *Data Science Journal*, 13 September, www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article.

Committee for Economic Development (CED), (2012), *The Future of Taxpayer-Funded Research: Who Will Control Access to the Results?*, Washington D. C.

Costas, R. et al. (2013), "The value of research data: Metrics for datasets from a cultural and technical point of view", Knowledge Exchange report, www.knowledge-exchange.info/datametrics, accessed 15 April 2015.

Cragin, M.H. et al. (2010), "Data sharing, small science and institutional repositories", *Philosophical Transactions of the Royal Society A*, Vol. 368, No. 1926, pp. 4023-38.

Dallmeier-Tiessen, S. et al. (2012), "Compilation of results on drivers and barriers and new opportunities". Retrieved from www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf , accessed 15 April 2015.

EC (2010), "Riding the Wave: How Europe can gain from the rising tide of scientific data", Final report by the High-level Expert Group on Scientific Data, European Union, October, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf, accessed 15 April 2015.

EC (2012), "Public sector information – Raw data for new services and products", http://europa-eu-audience.typepad.com/en/2010/08/ec-public-sector-information-raw-data-for-new-services-and-products.html, accessed 15 April 2015.

Fienberg, S.E., M.E. Martin and M.L. Straf (1985), *Sharing Research Data*, National Academies Press, Washington, DC.

Filippov, S. (2014), "Mapping tech and data mining in academic and research communities in Europe", *Lisbon Council Newsletter*, Issue 16/2014.

Frischmann, B.M. (2012), *Infrastructure: The Social Value of Shared Resources*, Oxford University Press.

Gardner, D. et al. (2003), "Towards effective and rewarding data sharing", *Neuroinformatics*, Vol. 1, No. 3, pp. 289-95.

Groves, T. (2010), "The wider concept of data sharing: View from the BMJ", *Biostatistics*, Vol. 11, No. 3, pp. 391-92.

Groves, T. (2009), "Managing research data for future use", *The BMJ*, 338, b358

Gura, T. (2013), "Citizen science: Amateur experts", *Nature*, No. 496, pp. 259-261, naturejobs.com, http://dx.doi.org/10.1038/nj7444-259a.

Hardisty, D.J. and D.A.F. Haaga, (2008), "Diffusion of treatment research: Does open access matter?", *Journal of Clinical Psychology,* Vol. 67, No. 7.

Hey, T. and A. Trefethen (2003), "The data deluge: An e-science perspective", in F. Berman, G.C. Fox and A.J.G. Hey (eds.), *Grid Computing: Making the Global Infrastructure a Reality,* John Wiley & Sons, Ltd., Chichester, England, pp. 809-24, http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf.

Holocher-Ertl, T., B. Kieslinger (2013), "Towards a better society of empowered citizens and enhanced research", Green Paper on Citizen Science, Citizen Science for Europe, Socientize, Citizen Science Projects.

Houghton, J. and P. Sheehan (2009), "Estimating the potential impacts of open access to research findings", *Economic Analysis and Policy*, Vol. 29, No. 1, pp. 127-42.

Houghton, J., A. Swan and S. Brown (2011), "Access to research and technical information in Denmark", Technical Report, School of Electronics and Computer Science, University of Southampton.

Houghton, J. et al. (2010), "Economic and Social Return on Investment in Open Archiving: Publicly Funded Research Outputs", report to SPARC (Scholarly Publishing and Academic Resources Coalition), Center for Strategic Economic Studies, Victoria University.

JISC (2014), "The value and impact of data sharing and curation: A synthesis of three recent studies of UK research data centers", JISC (UK Joint Information Systems Committee), http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf, accessed 15 April 2015.

Jirotka, M. et al. (2006), "Special issue: Collaboration in e-research", *Computer Supported Cooperative Work (CSCW)*, Vol. 15, No. 4, pp. 251-55, http://dx.doi.org/10.1007/s10606-006-9028-x.

Kotarski, R. et al. (2012), "Report on best practices for citability of data and on evolving roles in scholarly communication", *Opportunities for Data Exchange*, www.stm-assoc.org/2012_07_10_STM_Research_Data_Group_Data_Citation_and_Evolving_Roles_ODE_Report.pdf, accessed 15 April 2015.

Kowalczyk, S. and K. Shankar (2010), "Data sharing in the sciences", *Annual Review of Information Science and Technology*, Vol. 45, pp. 247-94.

Lakhani, K.R. et al. (2007), "The value of openness in scientific problem solving", www.hbs.edu/faculty/Publication%20Files/07-050.pdf, accessed 15 April 2015.

Lane, J. et al. (eds.) (2014), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press.

Mooney, H. (2011), "Citing data sources in the social sciences: Do authors do it?", *Learned Publishing*, Vol. 24, No. 2, pp. 99-108.

Mooney, H. and M.P. Newton (2012), "The anatomy of data citation: Discovery, reuse and credit", *Journal of Librarianship and Scholarly Communication*, Vol. 1, No. 1, eP1035.

Murray, F.P. et al. (2009), "Of mice and academics: Examining the effect of openness on innovation", National Bureau of Economic Research, www.nber.org/papers/w14819 (accessed 15 April 2015).

Myllymäki, P. (2013), "Data to Intelligence (D2I) Research Programme on Intelligent Data-Driven Services", presentation, www.digile.fi.

Narayanan, A. and V. Shmatikov (2008), "Robust de-anonymization of large sparse datasets", SP 08 Proceedings of the 2008 IEEE Symposium on Security and Privacy, IEEE Computer Society, Washington, DC.

OECD (2015a), "Assessing government initiatives on public sector information: A review of the OECD Council Recommendation", OECD, forthcoming.

OECD (2015b), *Making Open Science a Reality*, OECD, forthcoming.

OECD (2014a), "Unleashing the power of big data for Alzheimer's disease and dementia research", Main points from the OECD Expert Consultation on "Unlocking Global Collaboration to Accelerate Innovation for Alzheimer's Disease and Dementia", DSTI/ICCP/IE(2013)13/FINAL, OECD Publishing, Paris.

OECD (2014b), *OECD Science, Technology and Industry Outlook 2014,* OECD Publishing, Paris.

OECD (2013), "New data for understanding the human condition: International perspectives", OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences, OECD Publishing, Paris.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, C(2008)36, 30 April 2008, OECD, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris.

OECD (2006), Recommendation of the Council concerning Access to Research Data from Public Funding, 14 December 2006, [C(2006)184], OECD Publishing, Paris.

Piwowar, H.A. and W.W. Chapman (2008), "Identifying data sharing in biomedical literature", AMIA Annual Symposium Proceeding Archive, pp. 596-600.

Piwowar, H.A. and T.J. Vision (2013), "Data reuse and the open data citation advantage", PeerJ 1:e175, https://dx.doi.org/10.7717/peerj.175.

Piwowar, H., R.S. Day and D.B. Fridsma (2007), "Sharing detailed research data is associated with increased citation rate", *PLOS ONE*, Vol. 2, No. 3.

Riesch, H., C. Potter and L. Davies (2013), "Combining citizen science and public engagement: The Open AirLaboratories Programme", *Journal of Science Communication,* Vol. 12, No. 03.

Royal Society (2012), "Science as an open enterprise", Royal Society Science Policy Centre Report, February.

Rowlands, I. and D. Nicholas (2005), *Institutional Repositories for the Research Sector: Feasibility Study*, National Library of New Zealand, Wellington.

Sparks, S. (2005), *JISC Disciplinary Differences Report*, Rightscom, London.

Spiegler, D.B. (2007), "The Private Sector in Meteorology – An Update", http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-88-8-1272.

The Economist (2010), "Data, data everywhere", 27 February.

Uhlir, P.F. (2012), "For Attribution – Developing data attribution and citation practices and standards: Summary of an international workshop", National Academies Press, Washington, DC.

UNESCO, (2012), *Policy Guidelines for the development and promotion of Open Access*, UNESCO Publishing, Paris.

Ware, M. (2009), "Access by UK small and medium-sized enterprises to professional and academic literature", Publishing Research Consortium, April.

Ware, M. and M. Monkman (2008), "Peer review in the scholarly journals", Publishing Research Consortium, January.

Williams, H. (2010), "Intellectual Property rights and innovation: Evidence from the human genome", National Bureau of Economic Research, NBER Working Paper 16123.

Further reading

Bjork, B.-C. et al. (2013), "Anatomy of green open access", *Journal of the American Society for Information Science and Technology*, www.openaccesspublishing.org/apc8/Personal%20VersionGreenOa.pdf, accessed 15 April 2015.

Coe, D.T. and E. Helpman (1995), "International R&D Spillovers", *European Economic Review*, Vol. 39, No. 5, pp. 859-87.

Coe, D.T., E. Helpman and A.W. Hoffmaister (2009), "International R&D Spillovers and Institutions", *European Economic Review*, Vol. 53, No. 7, pp. 723-41.

Dallmeier-Tiessen, S. et al. (2011), "What scientists think about open access publishing: Highlights from the SOAP (Study of Open Access Publishing) project survey", Cornell University Library, http://arxiv.org/abs/1101.5260 (accessed 15 April 2015).

EC (2014), *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group,* European Commission, http://lalist.inist.fr/?p=12043 (accessed 15 April 2015).

Hall, B.H., J. Mairesse and P. Mohnen (2009), "Measuring the Returns to R&D", National Bureau of Economic Research, NBER Working Paper 15622.

Ipsos MORI (2011), "Public Attitudes to Science 2011", report to the Department for Business, Innovation and Skills, May.

Jaffe, A.B. (1989), "Real effects of academic research", *American Economic Review*, Vol. 79, No. 5, pp. 957-70.

Khan, M. and K.B. Luintel (2006), "Sources of knowledge and productivity: How robust is the relationship?", OECD *Science, Technology and Industry Working Papers,* 2006/6, OECD Publishing, Paris.

Luintel, K.B. and M. Khan (2011), "Basic, applied and experimental knowledge and productivity: Further evidence", *Economics Letters*, Vol. 111, No. 1, pp. 71-74.

Parcher, J. (2012), "Benefits of open availability of Landsat data", www.oosa.unvienna.org/pdf/pres/stsc2012/2012ind-05E.pdf, accessed 15 April 2015.

van der Meulen, B. and A. Rip (2000), "Evaluation of societal quality of public sector research in the Netherlands", *Research Evaluations*, Vol. 8, No. 1, pp. 11-35.

Verspagen, B. (2004), "Innovation and economic growth", in J. Fagerberg, D.C. Mowery and R.R. Nelson (eds.), *Handbook of Innovation*, Oxford University Press, Oxford.

From:

Data-Driven Innovation

Big Data for Growth and Well-Being

Access the complete publication at:
https://doi.org/10.1787/9789264229358-en

Please cite this chapter as:

OECD (2015), "Promoting data-driven scientific research", in *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264229358-11-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.