Chapter 8

The Staged Model for Open Scientific Data

Vera J. Lipton

This chapter outlines a way forward for open scientific data. Specifically, it evaluates the impact of open data mandates, identifies the problems associated with their implementation, and proposes ways to address them.

The chapter consists of three sections:

1. Before open data mandates

2. The mandates and their impact

3. The staged model for open scientific data

    - Open data and open publications require different approaches

    - One size does not fit all: the concept of research data

    - The need to make choices: the time and resources required

    - Misunderstood incentives: data exclusivity period

    - Proposed scope of the mandate: releasing data along different stages

    - Increased focus on data reusability: more than metadata

    - The need to develop individual and collective incentives

    - Data ownership should be vested in researchers

    - Legal problems with data reuse: text and data mining exemption

Introduction

The previous three chapters have identified the challenges associated with implementing open scientific data in practice at CERN and in the field of clinical trial data. Those chapters also identified emergent best practice in data curation and release. Drawing on the findings of the previous chapters, this chapter evaluates the impact of the open data mandates and proposes a model to address the problems arising in their implementation.

There are three main parts in this chapter. I first outline the ideological and policy setting within which the policies mandating open access to scientific data have emerged. This is followed by an overview of the main features of the mandates and identification of their drawbacks. The final section discusses those shortcomings in more detail and introduces a staged model for open scientific data.

It is argued that the open data mandates have created a momentum for data release globally. At the same time, the mandates alone are insufficient to effectively drive open data into the future because digital curation of research data for public release is both a very recent and a complex function, posing many challenges. The proposed model and its eight recommendations suggest options for dealing with the issues arising in implementation so as to ensure sustainability of open research data into the future.

8.1 Before open data mandates

Open scientific data is largely driven by the emergence of digital science, as outlined in Chapter 2. The transition from modern science[1] to digital science[2] started well before the open access movement. The World Data Center was established in 1955 to archive and distribute data collected during the 1957–1958 International Geophysical Year.[3] As a result, representatives of 13 governments agreed on scientific collaboration enabled by a free sharing of scientific observations and results from Antarctica [451]. In 1966 the Committee on Data in Science and Technology was founded by the International Council for Science to promote cooperation in data management and use [60].

Digital sharing of scientific data builds on these early foundations. It has accelerated in recent years largely due to technological advances in communication technologies and the proliferation of measurement and scientific equipment capable of collecting, processing, and storing vast amounts of data. Such equipment is now more readily available, and the costs associated with automated data harvesting and analysis have dropped significantly. To illustrate this point, I refer back to the Human Genome Project completed in 2003. Decoding the human genome, using the technology available at the time, took 10 years and cost over US$1 billion. Today, complex DNA analyses require only several days at a cost of around US$1000 each.[4]

The year 2003 also loosely marks the emergence of the open access movement, which brought renewed calls for greater availability of scientific data.[5] It was also the year the non-profit Public Library of Science (PLOS) in the United States launched *PLOS Biology* and high-profile journals such as *Nature*, *Science*, and *The Scientist* all published high-profile articles on open access to scientific publications [456–458].

Open scientific data needs to be seen in this historical context. It is not a completely novel concept, and it is not merely an extension of policies mandating open access to publications. Open scientific data is new in that it calls for research data to be freely available for access, reuse, and distribution by anyone—whether as

[1] Thomas Kuhn developed the concept of modern science and elaborated on the concept of scientific revolutions in 1962. Kuhn explains the process of scientific change as the result of various phases of paradigm change. He challenged the Mertonian view of progress in what he called ‘normal science’. He argued for a model in which periods of conceptual continuity in ‘normal science’ were interrupted by periods of ‘revolutionary science’. See Kuhn [100], Chapter 2, Section 2.2.

[2] The term ‘digital science’ is often referred to as ‘open science’ or ‘Science 2.0’. See definitions in Glossary.

[3] Scientists from 67 countries participated in the data collection that year and agreed to share data generated from cosmic ray, climatology, oceanography, earth’s atmosphere, and magnetic research, with a view to make the data available in machine-readable formats. See also Chapter 2, Section 2.2.

[4] Statistics sourced from the International Council for Science [452]. The early economic analysis of the Human Genome Project is included in Chapter 2, Section 2.4.

[5] The calls for enabling open access to research data came from different authoritative sources [45–51, 71, 453–455].

researchers, policymakers, industry partners, or any member of the public. While some scientific articles were previously available for anyone to use freely in digital formats, research data—the 'raw material' necessary to validate the outcomes published in those articles—is only now becoming freely available to the broader public as open data.

Indeed, open scientific data aims to encourage, for the first time in history, the participation in science creation, validation, and dissemination by both scientific and non-scientific actors. The production of scientific knowledge is now more centrally located within social relations—a shift that has been termed as *Mode 2* of knowledge production.[6] This also means that data is viewed in a different way to that found in the previous context of modern science defined by Thomas Kuhn. The key difference is the principle that where data is produced through publicly funded research, then the broader public should have a right to access it. Furthermore, according to the theory of *Mode 2* knowledge production, data is seen as having value through its reuse by a broader range of stakeholders than just the research community that initially collected it.[7]

Open scientific data further highlights the transformative changes in science conduct in the digital era. With increased availability of data in digital formats, computers alone can now validate and generate scientific outcomes—due to advances in artificial intelligence and quantum computing and the development of algorithms capable of solving problems by processing and calculating vast amounts of data. Following on from these developments is the argument that open scientific data challenges established research and science conduct and communication practices, as well as the monopoly of researchers over validating and creating scientific outcomes.

Such profound changes require careful change management and implementation processes. While some researchers welcome these developments and embrace the changes, others are naturally reticent or even sceptical about them. Despite recent progress, the transition to digital science is still in early stages. In some fields of science, especially social sciences, the transition has not even properly started [459]. For these reasons, this book argues the calls for engaging the broader public in science participation may come too early.

The argument draws on the findings of Chapters 5 and 6, which document the experiences with implementation of open data in particle physics and clinical trials. The finding of these chapters is that scientists in both fields are still learning how to implement open scientific data and how to deal with the many challenges associated with the processing, curation, release, and (re)use of open scientific data they produce. Their experiences with open data demonstrate that even a well-established and large data-centric organisation, such as CERN, is still experimenting with the parameters and descriptors that will make its particle physics data available in a form suitable for independent reuse by others.

By contrast, describing, sharing, and reusing clinical trial data in digital formats are a well-established practice in closed scientific circles. However, the free sharing of that data as open data is not developing quickly as a practice, despite the economic and social value the data holdings found to offer society [339, 460, 461]. Instead of looking for ways for facilitating the sharing of data more widely, some members of the research community took the view that disseminating clinical trial data as open data was risky as the data might be used maliciously or to uncover the identity of research subjects [462–464]. Those researchers who were willing to

[6] *Mode 2* is a new paradigm of knowledge production that is characterised as socially distributed, application-oriented, transdisciplinary, and subject to multiple accountabilities ([33], p. 179).

[7] *Ibid*, see also Wessels et al. [89], p. 56, Chapter 2, and Section 2.2.

share data often faced criticism for giving away data that could potentially be used to generate further publications or research revenue for their organisations.

The increased calls for opening up research data come at a time when major governments are decreasing their funding for research[8] and there is an increasing trend in the private sector to draw on public research.[9] Many governments now require publicly funded research organisations to increase the return on the investment in research by generating income through the protection and commercialisation of intellectual property, including though the creation of start-up enterprises [150]. The demand for commercialisation has affected the goals of government research funding. It is causing public sector research agencies to justify the success of research by providing a convincing argument for the future economic value of their science and technology bases [151]. Such agencies are also urged to demonstrate the broader social and environmental benefits of their research.

Australia is no exception. Many CSIRO researchers work on commercial projects with industry and are under the obligation to maintain confidentiality about the results. Also, all science-intensive research agencies in Australia now have a technology transfer function and try to create revenue from commercialising university intellectual property. However, the vast majority of university research in Australia remains publicly funded, and some 70% of CSIRO research is funded by the government. Thus, there is a strong case for allowing the public to share in the fruits of scientific research by having access to the data these research organisations create.

In recent years, the Federal Court of Australia has upheld the argument that science has a public function. In the *UWA vs Gray* case [414], a dispute over intellectual property rights claimed by a former university employee, Justice French, made specific acknowledgement that the function of universities is to offer education and research facilities and to award degrees and that this amounts to a public function.

Further, he stated that although universities do perform commercial activities, those enterprises had not displaced the public functions of universities in such a way that they became ‘limited to that of engaging academic staff for its own commercial purposes’.[10]

In addition, Justice French held that academic freedoms are incompatible with any duty to maintain confidentiality of the kind required to protect, for commercial purposes, the intellectual property that might result from research activities within a university.[11] In sum, this judgement confirmed the principle that the public

[8] Spending on R&D in government and higher education institutions in OECD countries fell in 2014 for the first time since the data was first collected in 1981. Countries with declining public R&D budgets include Australia, France, Germany, Israel, the Netherlands, Poland, Sweden, the United Kingdom, and the United States. See OECD [146].

In the United States, for the first time in the post-World War II era, the Federal Government no longer funds a majority of the basic research carried out in the country. Data from ongoing surveys by the National Science Foundation show that federal agencies provided only 44% of the US$86 billion spent on basic research in 2015. See Mervis [147].

[9] Chesbrough has shown that technology companies require timely access to knowledge as they increasingly innovate by combining research outputs from external and internal sources and increasingly draw on research from universities and other public research organisations [11, 12].

For example, in the pharmaceutical sector in the United States alone, roughly 75% of the most innovative drugs, the so-called new molecular entities with priority rating, trace their existence to the National Institutes of Health [149].

[10] *Ibid*, FCA 49 at 184.

[11] *Ibid*, at 192.

function of universities is the priority, with commercial considerations subordinate to that.

This position underpins the case for open research data. It is within such an ideological and technological setting that the policies mandating open access to scientific data have emerged.

8.2 The open data mandates

Some of the world's leading research organisations are based in the United States. These were among the earliest institutions anywhere to recognise the potential of open scientific data.

The first policy statement for open access to research data is found in the *Bromley Principles* issued by the US Global Change Research Program in 1991 [169]. Five years later, the *Bermuda Principles*—developed as part of the Human Genome Project—set an international practice for sharing genomic data prior to publication of research findings in scientific journals.[12]

In 2003 open access to scientific data was first codified internationally, in the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* [63].[13] This emerged from a conference hosted by the Max Planck Institute in Munich and represents a landmark statement on open access to scientific contributions[14] including 'original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material'.[15] Research organisations committed to implementing the objectives of open scientific data can sign the declaration, and over 600 have done so already.[16]

Awareness of the need to develop data management infrastructure took a huge step forward in 2010 when the National Science Foundation (NSF) in the United States announced that it would begin requiring data management plans with applications in the grant cycle starting from January 2011.[17] This policy has inspired research funders to introduce similar policies all over the world. The original NSF policy states:

> *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them or their products widely available and usable.*[18]

[12] The Human Genome Project is discussed in Chapter 2, Section 2.5.

[13] The Declaration is analysed in Chapter 3, Section 3.1.

[14] The Berlin Declaration does not use the term 'open research data' but rather refers to 'open knowledge contributions' which represent a broad definition of open research data. See also discussion concerning the definition of research data in Chapter 4.

[15] As of October 2007, there were 240 signatories, in early 2018 over 600 [63].

[16] *Ibid*, [172].

[17] Proposals submitted to NSF on or after 18 January 2011: ... must include a supplementary document of no more than two pages labelled 'Data Management Plan.' This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results [185].

[18] See NSF Award and Administration Guide, Chapter 6—Other Post Award Requirements and Considerations, points 4(b) and (c) [16].

For several years prior to this statement, research funders had required grant recipients to share their data with other investigators. Yet none had policies on how this should be accomplished. The position has changed following the publication of the NSF policy, with many funders now requiring that recipients of grants enable open access to research data and, in many cases, also submit research data management plans at the grant proposal stage. Such policies aim to ensure that data resulting from publicly funded research is retained and can be reused over time—usually for 10 years.

The US government has taken significant steps to enable the dissemination of scientific outcomes arising from public research. In early 2013, the Office of Science and Technology Policy at the White House directed each federal agency with over US$100 million in annual research and development expenditure to develop plans to make 'the results of unclassified research arising from public funding publicly accessible to search, retrieve and analyse and to store such results for long-term preservation'.[19]

The coordinating body for science policy in the United Kingdom, UK Research and Innovation (the successor since April 2018 to Research Councils UK), has had policies on open access since 2005. Its common principles for open data of 2011 [465] take account of the evolving global policy landscape.

The European Commission was among the first of the large funders to test arrangements for encouraging open access to publicly funded research. In 2008, the Commission launched the Open Access Pilot as part of its Seventh Research Framework Programme. That was replaced in 2014, under the Horizon 2020 research and innovation project, with the Open Research Data Pilot for treating the data underlying publications—including curated data and raw data [21]. The Rules of Participation[20] establish the legal basis for open access to research data funded by the European Commission under the Horizon 2020 Work Programme, and the overarching principles are translated into specific requirements in the Model Grant Agreement[21]. The Commission has also developed a user guide that explains the provisions of the Model Grant Agreement to applicants and beneficiaries along with defined exceptions to data sharing.[22]

In addition to the measures taken by the European Commission, individual European countries have taken legislative steps to recognise open access to research outputs. These include Germany,[23] Italy,[24] the Netherlands[25] and Spain.[26]

[19] The White House (2013). *The Results of Federally Funded Scientific Research*. The research results include peer-reviewed publications, publications' metadata, and digitally formatted scientific data. The major shortcoming is that the memo does not mention metadata associated with research data. This omission is unfortunate because, in many cases, scientific data without metadata is unlikely to be reusable.

[20] Article 43.2 of Regulation (EU) No 1290/2013 of the European Parliament and of the Council laying down the rules for participation and dissemination in Horizon 2020, the Framework Programme for Research and Innovation (2014–2020) and repealing Regulation (EC) No 1906/2006.

[21] Multi-beneficiary General Model Grant Agreement, Version 4.1, 26 October 2017 (http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf).

[22] The exceptions include the obligation to protect research results with intellectual property, confidentiality, and security obligations and the need to protect personal data and specific cases in which open access might jeopardise the project. If any of these exceptions is applied, then the data research management plan must state the reasons for not giving or restricting access. (Annotated Model Grant Agreement, Version 1.7, 19 December 2014, 215).

[23] Law October 1, 2013 (BGBl. I S. 3714) amending Article 38 Copyright Act.

[24] Par. 4, Law October 7, 2013, no. 112.

[25] Law June 30, no. 257 amending Article 25fa Copyright Act.

[26] Artículo 37 'Difusión en acceso abierto', Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación.

Elsewhere, significant policy developments are under way in several Latin American countries. The Chinese Academy of Sciences was an early signatory to the Berlin Declaration, and it actively participates in several open data projects.

Australia is hesitant to implement open research data practice, even though the country was one of the first in the world to adopt open access to public sector information. The country's two principal research funders—the National Health and Medical Research Council (NHMRC) and the Australian Research Council (ARC)—mandated open access to peer-reviewed publications in 2012. Starting from 2014, the ARC said that it 'strongly encourages' the depositing of data and any publications arising from a research project in an appropriate subject and/or institutional repository [466]. At the same time, 'research data and metadata' are expressly excluded from the scope of its open access policy.[27] This highlights the need to understand the meaning of 'open research data' within the ARC grants, as pointed out in Chapter 4 and further discussed in Recommendations 1–4 below.

The NHMRC mandate did not extend to open data until early 2018. Australia seemed to be falling behind the rest of the world in terms of open research data, even though the Australian Government was one of the first in the world to develop a national research infrastructure having established the Australian National Data Service as early as in 2008. It was not until 10 years later the NHMRC finally updated its policy, stating that it:

> *… strongly encourages researchers to consider the reuse value of their data and to take reasonable steps to share research data and associated metadata arising from NHMRC supported research [467].*

The introduction of open data mandates by research funders and governments is a welcome development, as Chapter 3 concludes. Research organisations and universities are largely dependent on grant funding. Suddenly, these institutions realised that to enable researchers to successfully compete for grants, they had to provide them with support in the formulation of data management plans. Libraries at many research organisations are now providing these services,[28] and researchers are changing their research data management practices as a result. Within only a few years, the policies introduced by research funders appear to have built a momentum for significant organisational and behavioural changes. Such changes are driving the increased retention and sharing of research data globally.

However, implementation of open data mandates presents many challenges for research organisations, as this book finds. The mandates neither specifically acknowledge nor deal with these challenges. The open data policies are more likely high-level statements of principles and expectations, rather than documents setting out rules and providing detailed instructions to research organisations. These factors make comparative analyses difficult.

To date, there is no agreement on what constitutes 'research data' and, consequently, what is the 'data' that researchers need to release.[29] Only a few of the policies include time limits for data release, and even fewer say what happens if there is no compliance. Very few policies address the funding requirements for

[27] The revised ARC Open Access Policy, version 2017.1, was issued on 30 June 2017 following consultations with the deputy vice chancellors (research) of Australian universities. Other publicly-funded research organisations do not appear to have been consulted.

[28] For example, all large Australian universities provide support to researchers with research data management (RDM). See Chapter 5, Section 5.2 [468–475].

[29] These issues are discussed in Chapters 4, Sections 4.1 and 4.2.

research data and supporting infrastructures, even though some funders include a provision in their grants for data curation for the duration of the relevant research project.[30] However, research data lifecycle generally extends beyond the duration of research projects.[31] Furthermore, the division of responsibilities for data annotation, curation, and preservation is not delineated. Some funders remain silent about the legal and ethical issues arising in research data sharing and reuse. Some appear to hold the perception that appropriate licencing mechanisms can effectively address the issues.[32]

These and other shortcomings and problems with implementation are detailed in Chapters 5–7, which provide a foundation for the development of the staged model for open scientific data that is introduced in the following sections.

8.3 The staged model for open scientific data

8.3.1 Open data and open publications require different approaches

The approach adopted for facilitating open access to scientific data has been strongly influenced by the experiences of research organisation in enabling open access to publications. Chapter 5 argued that research data management cannot be treated simply as a standardised library service for implementing open data mandates in practice. Yet this is exactly the approach taken by universities and many research organisations. While standardised approaches have generally proved to be suitable for developing open access to publications, such approaches are neither suitable nor appropriate for open scientific data. Librarians and research funders, who have played pivotal roles in facilitating open access to scientific publications, tend to apply uniform principles and approaches to open data as well. This creates challenges for researchers, who are required to comply with the open data mandates introduced by research funders but, at this stage, are unable to do so. There are several reasons for the confusion. In particular, there is the need for a more advanced understanding of the different natures of open data and open publications and of the different drivers and processes that have led to both.

Originally, open access was focused nearly exclusively on some 2.5 million articles that appear annually in 25,000 journals around the world, coming from all disciplines [476]. The rationale behind facilitating open access to publications was that, in the digital age, those articles should no longer be accessible only to users at such institutions as could afford the journal subscriptions. Instead, it was argued these articles could be made available to all potential users by depositing them on the web. Institutional repositories were created with open access-compliant software to make the articles interoperable, harvestable, navigable, searchable, and useable as if they were just one global repository—freely open to all.

The message about the feasibility and benefits of open access spreads quickly to academics and researchers, most of whom not only welcomed but gradually also embraced and began to actively promote the concept. Studies have shown that open access publications significantly increase research uptake and impact, as measured by downloads and citations [477–481]. Most publishers endorsed providing immediate open access, and researchers started depositing their articles on the web.

[30] See Chapter 3, especially Section 3.3. For example, the revised Research Councils (UK) Policy includes funding provisions.

[31] See Chapter 5, especially Section 5.1.

[32] See Chapter 3 and Chapter 7, Section 7.5.

However, it soon became apparent that the spontaneous deposit rate was not growing fast enough to make the ever-increasing volume of global annual research output available as open access. Researchers were surveyed, and their responses revealed significant concerns about copyright and about the time and effort that it could take to deposit. The same surveys established that researchers would readily provide open access if their institutions and research funders would mandate it.[33] So the only enablers needed were uniform mandates from research funders and appropriate copyright licencing mechanisms. Once these were introduced, librarians started to implement the new arrangements in collaboration with researchers.

Encouraged by these experiences, the same stakeholders started to call for extending the open access mandate to scientific data. Given the successful implementation of open access to publications, it was thought that mandates from scientific institutions and research funders would be the golden keys to increase the digital sharing of research data.

However, the mandates mushroomed well before any experiences with open data were generated by researchers. Several years down the track, it is becoming obvious that, for the most part, these approaches and assumptions were overly enthusiastic, if not unrealistic—largely because of the different nature of scientific data across different scientific disciplines but also because of the different incentives for collecting and sharing research data. Many of these differences are highlighted below.

For now, I summarise scientific publications and scientific data as two different concepts that require different approaches to their release, management, and curation. In the early stages of the open data debate, these distinctions went unnoticed and only became evident once the open data mandates from research funders became difficult to implement in practice.

8.3.2 One size does not fit all: the concept of research data

Despite the many examples of data provided in the open data policies and the many parameters and conditions that qualify data as 'open', 'findable', and 'intelligible', the term 'research data' (as it is used in practice) conveys different meanings to different people.[34] Research funders, researchers, librarians, and lawyers working in research organisations all approach the term differently. Funders and publishers typically mention research data that underpins publications; researchers talk about files, databases, and spreadsheets they collect and work within the course of research projects; librarians are preoccupied with metadata, data citations, and software; lawyers would like to see 'data' described as facts, raw facts, or compilations of facts in databases.

This can create confusion, as Chapter 4 argues. If researchers are to comply with the policies of funders and publishers, they need to understand what 'data' they need to make available. Similarly, if librarians are to provide effective assistance to researchers with data management, they need to be certain about the research outputs to be considered and how they need to be classified and described.

The nub of the problem with defining 'research data' is that data is a dynamic concept, unlike information.[35] The contents of 'data' vary in the context of its use, as examined in Chapters 4 and 7.[36] What represents 'data' to one researcher may

[33] However, over 90% of the researchers sampled said that if open access was mandated, then they would comply, with over 80% indicating that they would do so willingly [482, 483].

[34] See Chapter 4, Sections 4.2 and 4.3.

[35] See Chapter 2, Section 2.2.

[36] Chapter 4, Section 4.1 and 4.2, and Chapter 7, Section 7.1.

be 'noise'[37] for another researcher working on the same project, as Borgman pointed out [167]. However, the emerging consensus is that the meaning of 'data' needs to be interpreted through the lenses of researchers.[38] Generally, all outputs that are accepted in the scientific community as necessary to validate research findings are included among research data. The terms 'research data' and 'scientific data' are often used interchangeably, irrespectively of the subject collecting the data—whether the subject is a researcher or whether the data collection is semiautomated (such as through online questionnaires) or fully automated (such as data harvested by scientific equipment).

'Research data' may therefore take many forms, come in different formats, and arrive from various sources. In the physical and life sciences, researchers typically generate data from their own experiments or observations. In the social sciences, data can either be generated by the researchers themselves or sourced from elsewhere, such as from statistics collected by government departments. The notion of 'data' is least well-established in the humanities, although the rapid development of digital research in those disciplines has seen the use of the term become more common. In the humanities, the source of data is generally cultural records—archives, published materials, or artefacts [167, 484]. This variety of research practices across different disciplines results in a variety of practices for the collection and preparation of open access. The research community has yet to come to a uniform understanding of these matters [485].

Another facet of 'research data' is the sharing of it at various stages of granularity and processing levels. These range from top-level data underpinning scientific publications to various working versions incorporating different levels of analysis, cleaning, reorganising, and processing; to raw data collected in field research or harvested by scientific equipment.[39]

The open data mandates fail to acknowledge this fact, which is unfortunate, because agreement on the stages at which data needs to be shared across scientific disciplines would instantly assist researchers to make the data management task easier. In general terms, the lower the level of granularity of the data shared, the greater the possibilities for research reproducibility and data reuse. But this is conditional—the data must be supported by rich metadata and detailed description of the assumptions made by the original data collectors along the different levels of their research and data analysis and with the statistical methods used to analyse and aggregate the data and the methods used to clean the data and reduce 'noise'.[40]

Finally, there is consideration of the varying level of control of research data. Scientific organisations around the world implement numerous approaches and models of research data with varying levels of access control. At one end of the spectrum is the sharing of research data by anyone and with everyone. On the other end is a complete ban on data sharing gathered as part of certain research projects or across entire disciplines or institutions. Even though it is now generally accepted that sharing of publications is desirable and should be encouraged and pursued to

[37] Data noise is additional meaningless information included in data, for example, duplicate or incomplete entries. 'Noise' also includes any data that cannot be understood and interpreted correctly by machines, such as unstructured text.

[38] For example, the Australian National Data Service accepts records of data that are considered to be important to the Australian research community [231].

[39] For example, the sharing of clinical trial data can happen at the stage of the raw data collected in case report forms during trials, to the coded data stored in computerised databases, to the summary data made available through journals and registries. See also Chapter 6, Section 6.3.

[40] See definition of 'noise' at point 52 above.

the maximum extent possible, such an agreement is yet to emerge on the scope for the open sharing of research data.

Recommendation 1

The open data policies must incorporate the various facets of scientific data—that is, data which is heterogeneous, is complex, and differs across various scientific disciplines, various levels of granularity, and various levels of processing and control.

Research funders, publishers, and learned societies should, in close collaborations with researchers, facilitate the discussion to clarify the notion of data, its stages of processing, and the requirements for data sharing at each of these stages.

8.3.3 The need to make choices: the time and resources

Unlike academic publications, in which the objective is to publish as open access as many peer-reviewed outputs as possible, simply publishing more open data is unlikely to yield the same benefits. Choices need to be made about what data to keep and to preserve into the future and why. There are several reasons for this.

Firstly, preserving and curating all data collected in scientific experiments are not possible at this stage of technological development and at recoverable cost. This is because the burden of preparing and maintaining usable open access data repositories requires far more effort and resources than preparing publications for digital release. In this context, research data needs to be treated as an independent 'product', rather than part of research. What is more, the development of infrastructures is required to make the data discoverable, retrievable, interpretable, and usable.

The additional time and effort required from researchers cannot be overestimated. This is an important point of difference between open publications and open data. Publications are generally readily available in digital formats, and releasing them in electronic formats does not require any additional effort from researchers.

Preparing data for digital release is far more labour-intensive, especially in organisations implementing controlled access to data. Data curation requires detailed description of the datasets and the methods used to process it. The stages of receiving and processing applications for data release, then developing agreements on its use and related contracts, then producing and transferring data, and finally responding to any subsequent requests for clarification involve a diverse range of people throughout the data sharing organisation [366].

Research organisations typically have limited resources to handle these requests, which can result in clashes with other demands on staff time, such as research tasks. In the absence of support from research funders to prepare the datasets, some research bodies may request that applicants pay the cost of the staff time required to fulfil requests for sharing data or that they cover it from their own research budgets.

The sharing of scientific publications is generally straightforward and uniform across the world. Publications may exist in many copies and in many collections but need to be catalogued only once. Libraries are well-experienced in doing this and share such digital services across institutions. University libraries make agreements about what publications each will collect, promoting the concentration of resources and providing access to community members ([167], p. 75). While the same can be done with data collections, the experiences with data use and reuse are only just starting to emerge. Research data is more analogous to archival materials—each set is unique and requires its own metadata and provenance records.[41] Data is only meaningful and reusable if supported by properly recorded metadata and other data descriptors.

[41] *Ibid*, 307.

More work is required to describe unique items or to merge them into common structures. And even more work is required to keep the data collections up to date. While the effort associated with curation of publications is generally complete at the stage of release, data also requires post-release curation and tracking of issues such as software versions and other data processing systems.

Another reason why choices need to be made about what data to keep and curate as open data is the limitation on available computing power, data storage facilities, and other resources. Even CERN, an organisation at the forefront in the development of quantum computing in the world, has to make many hard choices about what 'data' to preserve into the future. The CERN processes and data decision points are detailed in Chapter 5.

Clearly, the resources required to curate and preserve open scientific data are immense and go well beyond the resources required to develop digital data repositories in the same manner as providing open access to publications. The open data mandates fail to recognise the resource implications, especially the efforts required from researchers to prepare data for release and the time required for any subsequent consultations with other researchers wishing to reuse the data. These efforts need to be recognised and rewarded.

Recommendation 2

Research funders and policymakers should allocate funding for the documentation, curation and preservation of research data that requires additional effort and time from researchers.

Choices need to be made about what data to preserve and why. Researchers are best positioned to make such choices provided data sharing is properly resourced.

8.3.4 Misunderstood incentives: data exclusivity period

A striking difference between open publications and open data is that increased impact is not the primary incentive for publishing research data. While it is true that the release of research data can increase use of the resulting scientific publications,[42] the purpose of the data itself is a prerequisite for conducting the research and writing the publication. Stevan Harnard summarised these differences well as early as in 2010 when he said:

> *Scientists and researchers are not data gatherers, they are analysers and interpreters of the data. They do gather and generate data and often at the cost of much time and effort. But researchers do so in order to be able to able to exploit and mine the data they have gathered or generated. What they publish in articles are the results of these analyses, that is why they are researchers and it is on that result that their careers and rewards depend [489].*

While researchers are generally keen to make their refereed articles available for open access immediately after publication, researchers are generally reticent to share their data freely immediately after gathering the data or immediately upon publication of the first data analyses. The reasons for this are many and are not well understood. Those known include the incentives for career progression and the prospects for scrutinising and, in some cases, even maliciously challenging research findings published in articles.

However, the most important reason is the significant time and effort required to process and describe the data. For these reasons, it has become obvious that

[42] A number of studies across several scientific fields have shown increased impact of publications supported by data [115, 486–488].

researchers in most disciplines insist on exclusive exploitation rights over their data, even if the data collection process is publicly funded. The period of exclusive exploitation required to produce the necessary publications varies by research disciplines and even by research projects.

At this time, most research organisations remain silent about the length of the exclusivity period required for their data. The lack of discussion on this issue is a significant impediment to open scientific data and needs careful consideration and negotiation between researchers, research funders and research organisations.

Setting unrealistic deadlines may achieve no more than setting no deadlines. This is especially the case as researchers themselves often do not have rights to the data they collect—a situation that is different from the legal rights they have, at least initially, in publications. The length of the period needs to be agreed at the beginning of research project. One way of achieving agreement would be to negotiate the length of exclusivity at the stage of preparing data management plans.

The stakeholders in the process also need to recognise that circumstances can arise for which the immediate release of research data is required in the public interest—such as to assist in dealing with a public health emergency or a national security interest.

Recommendation 3
Research funders and publishers need to seek consensus among research stakeholders that:

a. Researchers who generate original data will have the right of exclusive first use for a reasonable period.

b. The length of the period of exclusive use will vary by research discipline and even by research project and should be determined at the outset of each project in consultation between researchers and research funders.

c. The length of the agreed period of exclusive use should not exceed the maximum limits defined in the commonly agreed community norms and protocols for each scientific discipline.

d. Exceptions to this period of the exclusive use of data will apply in circumstances that are of urgent public interest—for example, in the case of a public health emergency.

8.3.5 Scope of the mandate: releasing open data along different stages

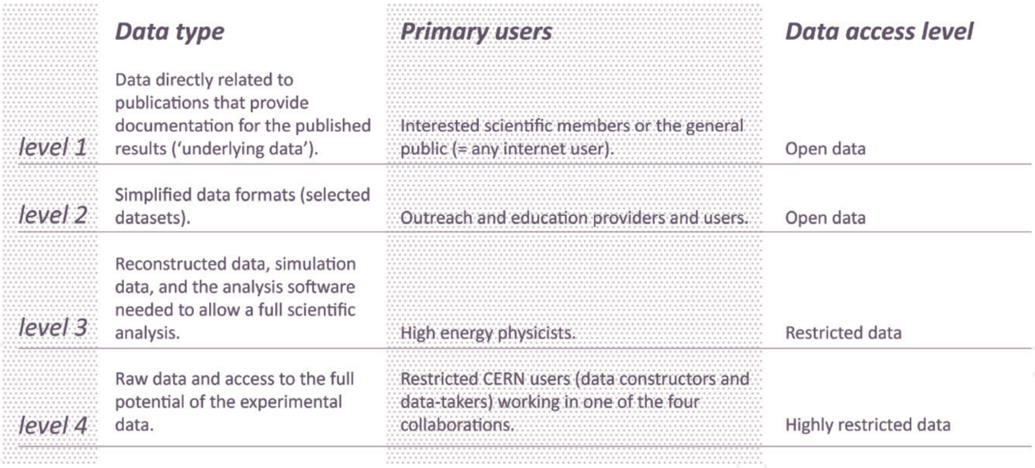
The key impediments to the practice of open data are the lack of recognition for the various types of research data and the lack of recognition that research data can be shared at various stages of processing and granularity. These issues were outlined in the previous sections and are discussed in detail in Chapters 5 and 6.[43] In this section, I introduce a staged approach for enabling open access to data that addresses the gaps—a modified version of the approach to research data as it has evolved at CERN. This approach can also be adopted to open research data in other organisations.

CERN has classified its data along four different levels of processing, which are summarised in **Table 7**.[44]

Level 1 data, the data underpinning scientific publications, is available simultaneously with publications and is mandatory.

[43] See Chapter 5, Section 5.3.2.

[44] This table is based on the four Open Access Policies in place across CERN [287, 293, 304, 310].



	Data type	Primary users	Data access level
level 1	Data directly related to publications that provide documentation for the published results ('underlying data').	Interested scientific members or the general public (= any internet user).	Open data
level 2	Simplified data formats (selected datasets).	Outreach and education providers and users.	Open data
level 3	Reconstructed data, simulation data, and the analysis software needed to allow a full scientific analysis.	High energy physicists.	Restricted data
level 4	Raw data and access to the full potential of the experimental data.	Restricted CERN users (data constructors and data-takers) working in one of the four collaborations.	Highly restricted data

Table 7.
Data processing levels at CERN.

Level 2 data consists of carefully selected and highly pre-processed datasets, such as those where students can search for the Higgs Boson. These datasets are released sporadically, mostly for educational purposes. CERN found its outreach education programs utilising *Level 2* data were highly successful and popular among high school students in many countries. This engagement has helped to develop data literacy and to promote awareness of particle physics among students.

Level 3 data is data ready for scientific analyses and processing and requires expert use. The data is 'reconstructed'—the level of processing that would roughly correspond to data cleaning and removing 'noise' in datasets in other research organisations.

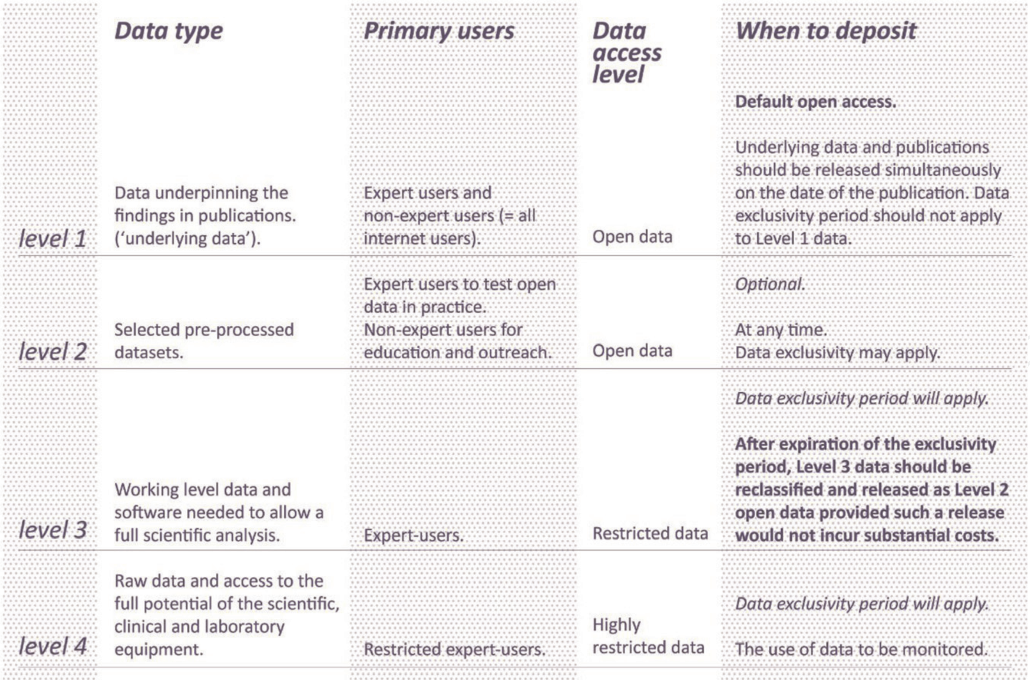
Level 4 data is called experimental data in the field of particle physics. It is the data collected from the Large Hadron Collider with minimal processing steps. This data is highly restricted and requires enormous computing power and resources for processing and descaling. CERN is, however, open to the possibility of sharing selected experimental datasets with expert users.

The data classification at CERN highlights another difference between open data and open publications. Access to open publications is generally available to anyone, whether as a member of the general public or of a scientific audience. Any person of reasonable intelligence can read the publication and is able to interpret and to assimilate the knowledge included in the publication to a certain degree.

This is not the case with open data in general and open scientific data in particular. A person of reasonable intelligence is unlikely to be able to interpret and to adequately utilise lower-level scientific data, even if the data is properly described and supported by relevant software. Freely accessible research data across all scientific disciplines may not be of widespread interest to the general public.

On occasion, good reasons may exist for restricting access to scientific data, especially raw data, to those scientists capable of using it in line with precisely defined research methods and established principles for research ethics. At the same time, the arguments presented by researchers against data sharing need careful examination before accepting any exceptions for not sharing data.

The key issue to keep in mind is that both open access publications and open access data collections gain in value as they grow ([167], p. 67). Therefore, many of the benefits of large open data collections will also only be discovered as the collections grow. This presents opportunities for broadening open access to lower-level data. However, at this point, neither the experimental data is described to the level of detail that would enable independent reuse, nor are the non-expert users able to process the data outside CERN.



	Data type	Primary users	Data access level	When to deposit
level 1	Data underpinning the findings in publications. ('underlying data').	Expert users and non-expert users (= all internet users).	Open data	**Default open access.** Underlying data and publications should be released simultaneously on the date of the publication. Data exclusivity period should not apply to Level 1 data.
level 2	Selected pre-processed datasets.	Expert users to test open data in practice. Non-expert users for education and outreach.	Open data	*Optional.* At any time. Data exclusivity may apply.
level 3	Working level data and software needed to allow a full scientific analysis.	Expert-users.	Restricted data	*Data exclusivity period will apply.* **After expiration of the exclusivity period, Level 3 data should be reclassified and released as Level 2 open data provided such a release would not incur substantial costs.**
level 4	Raw data and access to the full potential of the scientific, clinical and laboratory equipment.	Restricted expert-users.	Highly restricted data	*Data exclusivity period will apply.* The use of data to be monitored.

Table 8.
Staged model for facilitating open access to research data.

With this in mind, it is important for research funders across the different scientific disciplines to ascertain the levels at which scientific data is generally collected and processed across each scientific discipline. The funders should then set the boundaries for the levels at which the data holds the highest potential to be reused by other researchers (expert users) and by other interested users (non-expert users). The staged model summarised in **Table 8** can serve as a guideline for such deliberations.

The model puts a renewed emphasis on mandatory sharing of 'underlying data' that should be released concurrently with publications.

There should be no delays in releasing the *Level 1* data. A period of data exclusivity would not apply, because *Level 1* data represents highly selected and highly processed subsets of the lower-level research data. *Level 1* data is directly related to the results published. Once the findings are in the public domain, the reasoning that data underpinning those results can have a commercial value may not be plausible, as recently tested in cases to which the European Medicines Agency was a party.[45] Therefore, *Level 1* data should be released on the date of publication in all instances.

Level 2 data would be optional and would allow researchers as well as non-expert users to experiment with research data, enabling them to explore ways for reusing data produced by others and for embedding the data in their own research practice (see also Recommendation 6).

Level 3 data would be shared among expert users during a data exclusivity period, a situation which is not too dissimilar from the current practice among expert users in clinical trials and in particle physics experiments. Under this scenario, expert users would be authorised to access and to freely utilise the data and support tools directly in institutional repositories, or the data would be shared under data use agreements.

[45] See Chapter 7, Section 7.5.3.

However, after expiry of the exclusivity period, *Level 3* data would be published as open data and reclassified as *Level 2* data. Research funders along with librarians working in research organisations should be responsible for monitoring the expiry of the exclusivity period and release the data as open data when appropriate, provided there would be no substantial additional costs.

The need for sharing *Level 3* data after the expiry of the exclusivity period is especially relevant to those scientific disciplines where data infrastructures are well-developed and where open data is already embedded in research practice—such as in geospatial and earth sciences, materials sciences, biomedical research, computational engineering, and digital humanities.

Level 4 data can be governed by the same access mechanisms as *Level 3* data. However, the data would not be reclassified or released as open data after the expiry of the exclusivity period unless there would be a compelling business case for curating and preserving the data. This is because the curation and preservation of *Level 4* is costly and extremely labour-intensive.[46]

Recommendation 4

1. The open data mandates should:

    a. Put a renewed emphasis on mandatory sharing and unlimited use of the data underpinning the results published in scientific publications ('underlying data').

    b. Simultaneously develop transparent norms and protocols that would govern the levels of processing, dissemination, and reuse of 'working to raw level data' (*Level 3* and *Level 4* data) in each scientific discipline.

2. Researchers and learned societies should play a key role in coordinating the development of the open data norms and levels of data access for both scientific and non-scientific users in each discipline.

3. Open sharing of 'working level data' (*Level 3* data) should be the default practice in those scientific fields in which data infrastructures are well-developed and where open data is already embedded in research practice—such as in clinical and biomedical research, geospatial and earth sciences, materials sciences, computational engineering, and digital humanities.

4. The data exclusivity period would apply to releasing all but 'underlying data' (*Level 1* data).

8.3.6 Increased focus on data reusability

Chapter 2 found that the theories advocating open data release—namely, the theories of knowledge-based society[47] and science production in the digital era[48]—fail to recognise that data reuse is necessary for the envisaged benefits of open data to accrue. These theories of knowledge production and dissemination envisage that mere data release will bring out the desired economic and social benefits of open science.

[46] See Chapter 5, Section 5.2.4.

[47] According to Castelfranchi, a knowledge society generates, shares, and makes available to all members of the society knowledge that may be used to improve the human condition [32, 89, 95, 490].

[48] See Gibbons [33] at point 9.

The staged model proposed in this chapter rebuts this argument, positing that simply providing *access* to data in the public domain is useless to society unless that data is *reused*. In fact, facilitating open access to data is a potential burden to society if substantial costs in curating data are required and the data is not subsequently reused or produces other benefits. The crucial importance of data reuse in realising the benefits of open data does not appear to figure in the understanding of open data by research funders, even though reusability of open data is one of the conditions typically placed on open data.

Reusability can be achieved by providing rich metadata with attendant software and algorithms. However, there is little understanding of what makes metadata rich and how exactly metadata facilitates reusability. Experiences at CERN and with clinical trials both confirm that there is far more to metadata than computer-automated reports and that substantial human inputs are required to describe the data and all the steps taken to process and analyse it.

Based on the CERN experience, the notion of metadata needs to be expanded to include detailed documentation of all assumptions underpinning the data-gathering process, the cleaning and processing of the data, and the statistical and mathematical methods used to analyse the data—including all the decisions made along the different stages. Only researchers who collect and process the original data are capable of furnishing such descriptions. What is more, these steps need to be recorded at the time of data collection and analysis and, as such, need to be embedded in the research workflow.

Open data in large research organisations cannot be treated just as a 'product' resulting from research. Open data is an essential part of that research. It took CERN several years to define and fine-tune the parameters that make its particle physics data reusable. In particular, there was the need for data format and software version control.

The library team at CERN conducted several pilot studies and collected information about how researchers record their research workflows [301]. This was followed by an extensive consultation process and testing that eventually resulted in the new library service, which captures each data processing step and the resulting digital objects [302]. To facilitate future reuse of multiple research objects, researchers at CERN need to plan data preservation from an early stage of their experiments. For this reason, the decisions about recording 'metadata' in research organisations should also be made early in the research process.

Another area not yet explored by research funders that requires further attention is the nature of the factors that would motivate researchers to reuse the open data produced by others.

There appears be to an assumption, among both researcher funders and scientists, that once data is released, it will be reused by interested parties, as happens with open publications. While a correlation exists between the increased citations of publications supported by research data,[49] the incentives for data reuse are not well understood.

In some cases, researchers may opt to combine data from different sources, but some may prefer to collect their own data even if data produced by others is readily available as open data. This is because embedding open data in research practice is not yet common and requires new approaches and new reward mechanisms, as canvassed in the following section.

[49] *Ibid*, p. 61.

Recommendation 5
To ensure the maximum value from open data:

a. The potential for reusability should be the top criterion for evaluating any deposit of open data and when making decisions about investing resources in further curation or preservation.

b. Metadata and/or other detailed annotation and description of open data should form a mandatory part of every research data file submitted to repositories.

c. Software (code) and algorithms used to process the data should also be properly documented and shared wherever this is feasible.

8.3.7 The need to develop individual and collective incentives

The future success of open data practice lies primarily in the development of incentives that would motivate researchers both to release their own data and to reuse data produced by others. While many new metrics are currently under consideration—for example, altmetrics discussed in Chapter 6—all the new metrics are based on the measurement of 'data impact'. There are, however, several problems with this approach. The first is that researchers are rewarded for their 'publication impact', not 'data impact'.

The second problem is that 'data impact' does not lead to career progression. It follows that increased impact is not the key incentive for publishing research data (see discussion on Recommendation 3) and, therefore, data citations are unlikely to sufficiently motivate researchers to curate and release open data. So how could we better motivate researchers to put substantial time and effort into curating data?

A better incentive might be to acknowledge the original data creators as 'co-authors' of any publications arising from the reuse of their original data. Such an acknowledgement would have an immediate impact on researchers' career progression and would also stimulate collaborations among researchers, especially as early experiences with open data suggest that their benefits can be maximised in consultation with the original data creators.

However, the above recommendation highlights another problem that the current research performance metrics are biased in favour of individual performance, encouraging researchers to compete rather than to collaborate with each other. This approach is not appropriate to promote collaborations in the digital era that often require input from researchers across several disciplines and across different organisations. The research performance metrics need adjustment to promote and reward collective efforts.

The approach championed by CERN can serve as inspiration for other organisations. Large research teams always publish collectively—it is not unusual for a research publication to list over 3000 authors. The key to managing performance in such teams is the control over who is entitled to be considered a member and, consequently, to be included as an author in the publication. CERN has developed detailed guidelines for the approval process, and these incentives are definitely working. The spirit of collaboration is present in all communications with CERN.

An extension of this approach could be joint PhDs, a concept already allowed by some higher educational institutions but still uncommon in the scientific community.

Recommendation 6

a. Ensure that acknowledgement of the original creators of open data as 'co-authors' is included in any publications arising from reuse of the data.

b. Develop other performance metrics that will encourage researchers to curate and release research data and metrics that encourage the reuse of data developed by others.

c. Design such metrics so as to promote the formation of collaborations and collegial working relationships among researchers.

8.3.8 Uncertainty surrounding data ownership and confidentiality

An issue that arises from facilitating open access to, and the reuse of, publicly funded research data has highlighted the need to determine the legal ownership of data and to provide clarification on who should have the right to restrain unauthorised disclosure of confidential information, as Chapter 7 concludes.

This need arises because of two reasons.

Firstly, the various types of research data can be protected by copyrights and only data owners can licence the data under open licences. The uncertainty about data ownership has been identified as the root cause of subsequent problems affecting data licencing, the lack of interoperability, and the lack of clarity around the conditions governing data reuse.

Secondly, most researchers employed in research organisations have a duty of fidelity to their employer that prevents them from disclosing information acquired in the course of their employment ([491], 13.2 and 13.7). In Australia, this duty offers extensive protection for the employer and can include research data, especially in those research organisations that engage in collaborations with industry.

In such cases, the duty of confidentiality may also arise under a contract signed between the organisation and the industry partner where there usually is a term to prevent unauthorised disclosure of information.[50] Under these arrangements, the decision to release research data may be vested in a 'data steward'—the researcher or data manager with the responsibility to assess whether such release would constitute an authorised disclosure of confidential information—rather than be a decision for the owner of research data.

The effect of these provisions on researchers is that they often do not know who can clear the data for release or they are simply afraid to share research data, even in those cases where the data is not subject to any confidentiality provisions. A recent authoritative survey of researchers identified intellectual property and confidentiality as the top reasons for not sharing data.[51] Researchers are indeed afraid to share data when they are unsure whether it is appropriate for them to do so.

With regard to ownership of research data, there are two key legal regimes governing its ownership. The first is the copyright regime under which, as a general rule, the owner of the copyrighted work is the person who creates it by translating

[50] See Monotti [492] and *Ormonoid Roofing and Asphalts Ltd v Bitumenoids Ltd* [1930] NSWStRp 88; (1930) 31 SR (NSW) 347.

[51] In 2014 the publisher Wiley conducted an extensive survey of researcher attitudes to data sharing. The company contacted 90,000 researchers across many research organisations and received 2250 responses. Of those, 42% stated that they are hesitant to share their data because of intellectual property or confidentiality issues. See further discussion in Chapter 6, Section 6.3.6.

the idea into a fixed, tangible expression.[52] The second regime involves various contractual arrangements that may transfer or assign ownership of research data. The most common contractual arrangements guiding the ownership of research data are employment agreements and research funding agreements.

As employees of a university or a research organisation, researchers in most cases assign the rights to the data they produce (in the course of their employment) to their employers. In sponsored research, the research organisation typically retains ownership of the data but grants the role of data steward to the principal investigator.

In industry-funded research, the data typically belongs to the sponsor, although the right to publish the data can also be extended to the investigator. Where publicly funded research data is created under research collaboration between researchers working in different organisations, data ownership becomes even more unclear. Collaboration may involve a number of organisations, external researchers, funding bodies, government agencies, and commercial entities. The data ownership policies of the collaborating parties might be different or even conflicting.

The situation is also complicated because many researchers assume (often wrongly) that they own the data they collect in the course of their research. This position stems from their understanding that data and databases can be subject to copyright and, therefore, researchers are the legitimate owners because they have ‘created’ it—similar to the position with academic publications. However, only students and external visiting researchers typically own copyright that they create in the course of research or studies [492]. Likewise, researchers who create copyright outside their employment own it. But if the research is performed in the course of employment and the research organisation contributes resources, then the resulting data is likely to be owned by the organisation.[53]

Regardless of the legal position on data ownership, all researchers seem to maintain a sense of ownership over the data they produce. The role of researchers is also crucial in managing and documenting research data along the various stages of its processing and curation. Given these additional responsibilities placed on researchers for documenting and curating research data, vesting ownership of the data in researchers (or even better research teams) would be a logical step. The right of ownership would enable them to exercise greater autonomy over that data.

However, the prevalent view is that research data should belong to organisations, not individuals or research teams, since only organisations can be responsible data custodians and guarantors of data security and preservation. This notion of ownership is at odds with the open data mandates that place the responsibility for data deposit with researchers. Since researchers are not the legitimate owners of research data, they may be unable to fulfil this requirement and share the data under a licence, especially if the data was created in a joint project. In such a case, data release may be dependent on the consent of all co-owners.[54]

Another relevant point is that much scientific data is computer-generated and therefore it is unlikely to be subject to copyright protection and so should be placed in the public domain. Accordingly, researchers and research organisations need to become aware of the fact that determining copyright ownership may be irrelevant

[52] In copyright legislation this general rule is usually qualified by a specific rule that gives the employer copyright in certain circumstances and the crown under the crown copyright provisions.

[53] *Ibid.*

[54] For example, in Australia a co-owner of copyright is unable to exploit (copy or reproduce), grant an exclusive licence, or assign the copyright work without the consent of the other co-owner.

to computer-generated data. Furthermore, research organisations need to ensure that the data release in the public domain actually occurs.

To sum up, there is a need to delineate the notion of data ownership and confidentiality and to clearly define the attendant responsibilities for data management and sharing. It is not desirable for both research funders and organisations to be silent on these issues. While this book does not offer a recommendation in this regard, it highlights the importance of addressing uncertainties surrounding confidentiality and ownership of research data. Ultimately, data generated using public funds should be public property, and everyone has a responsibility to ensure that maximum value is derived from it. Data ownership needs to be managed so as to balance the interests of all—scientists, research funders, research organisations, and society as a whole.

Recommendation 7

Policymakers should Commission further research into data ownership and confidentiality with a view to achieving greater sharing of research data as open data.

Large research funders such as the European Commission and the National Institutes of Health are best positioned to provide direction for the research.

8.3.9 Introducing text and data mining exemption into copyright law

With the increasing availability of research data in the public domain, various types of reuse of that data will inevitably come to the forefront of the open data debate. Text and data mining,[55] often referred to as data analysis, is necessary to extract value and insights from large datasets. Such processes typically involve accessing the materials, extracting and copying the data, and then recombining it to identify patterns [494]. In Australia, subsequent to the definition of 'originality' established by the courts in the proceedings described in Chapter 7, such extraction of data and facts from protected work should not be subject to copyright protection.[56] However, since data mining typically requires the making of a (temporary) copy of the data, it is likely that this act would classify as copyright infringement.

Some countries, such as the United States, consider an activity such as making a copy as falling under the scope of the 'fair use' doctrine[57] of copyrighted works. Meanwhile, the United Kingdom has recently introduced a text and data mining exemption that covers such data uses but only for non-commercial research.[58]

The scope of the exemption in the United Kingdom is quite narrow, and it has the effect of hindering the realisation of the full value of open research data. A similar exemption is currently under consideration in the European Parliament, and the scope of the proposed exemption is broader than that in the United Kingdom—if

[55] The Australian Law Reform Commission defines data mining as 'automated analytical techniques that work by copying existing electronic information, for instance articles in scientific journals and other works, and analysing the data they contain for patterns, trends and other useful information' [493].

[56] Such uses would be classified as non-expressive use. The key principle here is that copyright law protects the expression of ideas and information and not the information or data itself.

[57] Par. 107 of the US Copyright Act 17 USC. The fair use requires a consideration whether the use of a work adds value to the original, for example, if used as raw material, transformed in the creation of new information, new aesthetics, new insights, and understandings.

[58] See Regulation 3 of the Copyright and Rights in Performances (Research, Education, Libraries, and Archives) Regulations 2014, No. 1372, adding Article 29A to the UK Copyright, Designs and Patents Act 1988. The Regulations came into force on 1 June 2014.

adopted, it would allow any Internet user to perform text and data mining for any purpose, whether commercial or non-commercial [429]. The proposed exemption cannot be overridden by contract, and some scholars have suggested that this principle should be extended to technology protection measures [495].

France, Germany, and Estonia have recently also introduced similar text and data mining exemptions, albeit more limited in their scope.

In Australia, text and data mining is not covered by the existing exemptions and could be considered copyright infringement if a substantial part of the text/data is reproduced. Limited text mining may be covered by the fair dealing exception if conducted for the purposes of research or study. However, the copying of an entire dataset would exceed a 'reasonable portion'[59] of the work and constitute infringement.

Australia currently does not have a text and data mining exemption but has, on several occasions, considered introducing a fair use system similar to that of the United States in place of the current fair dealing system. Despite that interest, action on the proposals is lagging, and, consequently, Australian research organisations seem disadvantaged. In the 2013 enquiry conducted by the Australian Law Reform Commission, the CSIRO argued that:

> *... if laws in Australia are more restrictive than elsewhere, the increased cost of research would make Australia a less attractive research destination [497].*

Furthermore, the CSIRO was of the view that

> *... the commercial/non-commercial distinction is not useful, since such a limitation would seem to mean that 'commercial research' must duplicate effort and would be at odds with a goal of making information (as opposed to illegal copies of journal articles, for example) efficiently available to researchers.*[60]

In line with this reasoning, it is proposed that—in the absence of fair use—a text and data mining exemption should be introduced into the *Copyright Act 1968* [496] (Cth).

Recommendation 8

Introduce the text and data mining exemption into copyright law—to enable data users to access, extract, combine, and mine data and datasets that currently are governed by various licence, contractual, copyright, technological protection, and legal regimes.

The exemption should eliminate legal uncertainty regarding the various data reuses associated with text and data mining. Such data reuses should be allowed to take place without the right holder's prior authorisation under conditions to be specified in the law.

8.4 Conclusion

In this chapter, I have argued that facilitating open access to research data requires vastly different approaches from those for enabling open publications. This is because research data are heterogenous, complex and differ across various scientific disciplines, various levels of granularity and various levels of processing and control.

[59] *Copyright Act 1968* [496] (Cth) s 40(5), setting out what is a 'reasonable portion' with respect to different works.

[60] *Ibid*, 11.69.

Open data mandates as they stand today fail to acknowledge that diversity and the fact that research data can be shared as open data at any point. The staged model proposed in this chapter calls for discussion across scientific disciplines to define the content of the data they hold and the stages of its processing. In the case of CERN and clinical trial data, the stages of data processing and sharing are well defined, and it is hoped that the proposed model can stimulate discussion about the levels of data processing in other research disciplines.

Rigorous data management practices and input from researchers are required to prepare the data for reuse for unknown audiences and for unknown purposes. However, these requirements should not be excuses for not sharing data. The proposed model calls for default open access to data that underpins results published in scientific publications (Level 1 data). Such data should be deposited in online repositories concurrently with publications, and research funders should take measures to ensure that their open data mandates include specific provisions to that effect. In cases of clinical data, the mandates that specifically required data archival in repositories along with a data accessibility statement included in the manuscript achieved the highest deposit rates.

The proposed model recognises the value open data can deliver if it is used for education and outreach purposes, as demonstrated with Level 2 data, especially the data showcasing the Higgs boson recently discovered at CERN. The related open dataset has reached thousands of high school and university students, and it has been used as a case study to promote data literacy and the development of computing skills among budding scientists. Such uses also help other research organisations in particle physics to replicate the experiments conducted at CERN and learn from them.

The proposed model encourages organisations to showcase their own research and to encourage the general public and expert users to reuse open data in innovative ways. Those experiments are necessary to promote the use of open data, embed it in research practice, and discover new reuses of open data in collaborative spaces.

The proposed model also recognises that not all research data can be of interest to the general public and that there are certain risks associated with sharing of some types of data—especially risks of breaching the privacy of patients involved in clinical trials and the risks of data misuse and misinterpretation of the original research. The staged model also recognises that lower-level data (Level 3 and 4 data) may not be shared immediately after the publication of research results and that such data may only be competently reused by expert users. For these reasons, the proposed model calls for clarification of the data levels that should be made available as open data to these two types of users—expert and non-expert users.

At the same time, the model makes the case for greater transparency in enabling access to low-level research data to experts, immediately after the expiry of a data exclusivity period. The length of that period would vary among scientific disciplines and even research projects and needs to be negotiated between researchers, their organisations, and research funders. Generally, it should not exceed the maximum limits defined in commonly agreed community norms and protocols.

If implemented, the proposed model would instantly improve access to high-level research data as open data, thus enabling any Internet user—whether researcher or non-researcher—to access and reuse the data for any purpose. By clearly defining the required competencies, skills, and attributes necessary to effectively reuse research data, the model would also lead to more transparent and improved data sharing among experts.

Specifically, it is anticipated that the proposed model would lead to a more nuanced discussion about the conditions and parameters that would qualify experts to promptly access low-level research data without restrictions, such is the case with

Level 3 LHC data that CERN makes available to physicists around the globe. In the field of clinical trials, a promising development on the same level of data access (Level 3 to Level 4) would be to enable the sourcing of background data for clinical trials directly from patients' electronic health records, smartphones, health insurance data, and other government databases.

The future of open data is in its use. The only way to make open data successful is to reuse it and prove that it can deliver the envisaged benefits. For this potential to be realised, open scientific data must be embedded in research practice and reused by other researchers or non-researchers. The proposed model posits that the potential for reusability should be the key criterion for evaluating any deposit of open data and when making decisions about investing resources in further data curation and preservation.

The future of open scientific data therefore lies in the hands of researchers. Only they can prove the value of the data by its reuse. Research funders and research organisations need to encourage them by developing appropriate incentives for forming collaborations and then sharing and reusing research data developed by others.

Finally, the law should not stand in the way of scientific progress, and it should not pose challenges in data release and reuse. Every dataset needs to have a clear owner so that the data can be properly licenced and be capable of reuse by others without any restrictions. Researchers should not be afraid to share, mine, and analyse research data in their quest to unearth new scientific knowledge. It is important for policymakers to ensure that data can be reused freely.

The proposed text and data mining exemption holds a great potential to enable Australian research organisations, businesses, and the broader public to reap the benefits of open scientific data.

Author details

Vera J. Lipton
Zvi Meitar Institute for Legal Implications of Emerging Technologies,
Harry Radzyner Law School, IDC Herzliya, Israel

*Address all correspondence to: vera.lipton@bigpond.com