Chapter 7

Legal Issues Arising in Open Scientific Data

Vera J. Lipton

This chapter aims to answer the question: what are the legal impediments to providing open access to, and the reuse of, research data that is publicly funded?

More specifically, this chapter aims to ascertain the boundaries for the release and reuse of data and/or databases, considering the current and recently proposed legal and policy frameworks and exceptions to copyright infringement.

The chapter consists of the following parts:

1. Copyright in research data
2. Ownership of research data
3. Licencing models for open scientific data
4. Different types of data reuse
5. Privacy and confidentiality issues in research data reuse

Introduction

The preceding chapters examined the many barriers to open data—with the lack of understanding of the concept of data, change of research practices and culture, attendant change management, research data management and funding issues being identified as the most prominent barriers to facilitating open access. There are, however, several legal issues associated with open research data in general and databases in particular. This chapter discusses these issues arising at two critical stages—namely, data release and data reuse. These issues are investigated in two parts.

The first part examines the legal issues arising in data release. The focus is on intellectual property rights, especially copyright in data and databases. There is also the uncertainty around data ownership, which is identified as the root cause of subsequent problems affecting data licencing, the lack of interoperability and clarity around the conditions governing data reuse. The chapter goes on to examine some relevant licencing models.

The second part concentrates on practical matters around data reuse—the need to regard intellectual property rights, where relevant, and the need of governments to facilitate text and data mining. It examines different types of data reuses and whether these can infringe different kinds of rights. Finally, the second part considers the specific issue of the privacy of research subjects and the tensions researchers face between the duty of confidentiality and the requirements to share data.

7.1 Copyright in research data

Data and databases play central roles in facilitating open access to scientific results. Legal protection of them does, therefore, strongly affect how scientists and researchers use data. The question of whether research data falls under intellectual property protection is a complex subject that is dependent on the nature of the data and the conditions under which the data is created, structured, and used. The legal basis for the protection is the existence of international legal frameworks, especially copyright frameworks, which also cover data and collections of data.[1] The international copyright framework is explored in the following sections. This is followed by an analysis of copyright law as it applies to data and data collections in several jurisdictions—Australia, the United States, and the European Union.

7.1.1 The international copyright framework

The scope of copyright protection and associated rights and the extent of the exclusive rights enjoyed by copyright owners are governed by several international treaties. Out of these, the Berne Convention, signed in 1886, is the oldest.[2] The Convention had the objective of providing a solution to the absence of international recognition for the copyright protection regimes of individual countries.

Over time, the Convention has evolved to establish the standards for the minimum level of copyright protection that all parties to it should implement. Those standards have been modified periodically as the notion of property has become more prominent. The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) of 1994 [382] and the WIPO Copyright Treaty (WCT) of 1996 [383] have built on the Berne framework to accommodate advances in technology[3]—including software, databases, and the protection measures that new technologies both enable and require. Consequently, all parties to the Berne Convention—including Australia, the United States, and all member states of the European Union[4]—have used the framework set by the above-mentioned international treaties to develop national copyright law.

The scope of copyright protection in the Berne Convention is defined in Article 2, which includes quite a detailed listing of protected works, including:

> *The expression of 'literary and artistic works' shall include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, such as books, pamphlets and other writings; lectures, addresses, sermons and other*

[1] This distinction between data and collections of data corresponds with the separation in copyright law between ideas that cannot seek protection and the expression of those ideas that can.

[2] See the Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886, last amended in the Paris Act of the Berne Convention on 28 September 1979 (http://www.wipo.int/treaties/en/ip/berne/). In 2017, the Convention had 175 signatory states, according to the World Intellectual Property Organisation (WIPO), the UN agency which administers it.

[3] Other relevant treaties, also administered by WIPO, include the Rome Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organisations (1961), which secures the rights for the performers of artistic and literary works in phonogram and broadcast recordings and the rights of the producers and broadcasters of those recordings (See http://www.wipo.int/treaties/en/ip/rome/), and the WIPO Performances and Phonograms Treaty (1996), http://www.wipo.int/treaties/en/ip/wppt/, which covers, among other matters, sound recordings, broadcasts, and performers' rights. Although these are not the classical form of data, these may represent data in some scientific disciplines.

[4] In the case of the WIPO Copyright Treaty and TRIPs, the EU is a signatory member in its own right.

> *works of the same nature; … photographic works to which are assimilated works expressed by a process analogous to photography; … illustrations, maps, plans …*[5]

Applying this definition to scientific outputs, it follows that scientific publications, regardless of their formats, are subject to copyright protection. However, the situation is not straightforward when it comes to research data that is often just a collection of facts, typically collated using automated or semiautomated instruments or scientific equipment. But, in addition, seemingly uncreative collections of data, such as phone directories, have in recent years sparked litigation and have stimulated policy debates about the extent to which copyright applies (or should apply) to the data.

There are two reasons behind the lack of clarity around the existence of copyright in research data.

The first is that the scope of 'research data' is extremely broad—data can be anything that researchers consider to be the evidence supporting their findings, as discussed in Chapter 4. It can be unstructured data, or it can be a vast dataset, or it can be a figure, a table, or a photograph embedded in these objects. Some of these data elements may be subject to copyright, while others are not.

The second reason is that the application of copyright to data and compilations of data raises many issues. This is largely because the concept of 'data' is a new concept, created in the computer age, while copyright law emerged at the time of printed publications.

At first sight, it may appear that copyright regimes do not apply to data and datasets. Simple facts and ideas do not qualify for copyright protection, whereas the original expression of ideas, classified as 'works', may qualify [384]. Research data in its own right is unlikely to meet the originality standards and, therefore, is unlikely to qualify as a protectable subject matter.

However, copyright can apply to original compilations of data and thus to databases. As discussed in more detail below, courts have confirmed this distinction. Different jurisdictions have assessed the way in which the balance between the 'works' and ideas has been achieved (in the selection and/or arrangement of data) as the test of originality that applies to collections of data, tables, and compilations. The test varies from country to country, as summarised in **Table 5**.[6]

7.1.2 Australia

The position in Australia on the copyright in compilations and databases was settled by the High Court in *IceTV Pty Ltd v Nine Network Australia Pty Ltd* [385] and subsequently by the Full Federal Court in *Telstra Corporation Limited v Phone Directories Company Pty Ltd* [386].

Historically, the common law measure was that originality could be demonstrated by the application of 'skill', 'effort' or 'judgement' (the doctrine of 'sweat of the brow'), as Sackville summarised:

> *The course of authority in the United Kingdom and Australia recognises that originality in a factual compilation may lie in the labour and expense involved in collecting the information recorded in the work, as distinct from the 'creative' exercise of skill or judgement, or the application of intellectual effort [387].*

---

[5] Article 2(1) of the *Paris Act (1971)* of the Berne Convention (http://www.wipo.int/treaties/en/ip/berne/pdf/trtdocs_wo001.pdf).

[6] **Table 5** was prepared by Vera Lipton (the author), and the definitions are based on the references (latest cases) discussed in this section.

Australia	United States	The European Union
Database will be protected by copyright under Australian law if it a) is a literary work b) is expressed in material form c) meets the originality test and d) has a relevant connection with Australia (for example, published in Australia or produced by an Australian resident). Originality involves a triple requirement: 1. the data compilation must not be copied, 2. a human author was involved in reducing or converting the database to a material form, and 3. there be some independent intellectual effort directed to expressing the work in the material form.	A compilation independently created by the author... that possesses at least some minimal degree of creativity (modicum of creativity). Presumably, the vast majority of compilations WILL past this test. The absence of creativity is manifested in an 'entirely typical' or 'garden-variety' end product constructed by processes which correspond to 'an automatic mechanical procedure' or to a so ... 'routine process'.	A broad, dual protection under copyright law and sui generis database rights. Originality defined as the 'author's own intellectual creation'. Originality expressed through the 'choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation.' A small selection of word was found to constitute an intellectual creation. The sui generis database right of protection applies if the creators have invested sufficient time, money, and skill into developing their database. Investment refers to resources used to seek out existing independent materials and collect them in the database.

Table 5.
The criteria determining the existence of copyright in databases.

Earlier Australian cases were considered in *Desktop Marketing Systems Pty Ltd v Telstra Corporation Ltd*.[7] This centred on Telstra's White Pages and Yellow Pages—compilations of names, addresses, and telephone numbers—and the 'headings book', produced by Telstra for use in classifying listings, and whether they constituted original literary works. In its judgement, the court found that this was indeed the case. Specifically, the court found that compilations of facts could qualify as original literary works if skill, judgement, and knowledge were exercised in compiling or arranging the facts or if substantial effort and expense were incurred during that process [388]. Therefore, it was recognised that the originality test was satisfied by this limited form of intellectual input.

In *IceTV*, the High Court considered copyright in programming guides, the *Weekly Schedules*, produced by the television broadcaster Nine Network Australia. The question of originality was considered in terms of whether taking the time and title data was taking a substantial part of the copyright work.

At first instance, Bennett held that the 'slivers' of information taken were not of a sufficiently substantial quality to be considered a substantial part. Specifically, she held that only the labour and skill involved in putting together the guide (the expression of the information) were relevant, and not the labour and skill involved in the programming decisions (the creation of the information). However, the Full Federal Court took a wider view—it found that data with the time and title was the 'centrepiece' of the guides, and so it concluded that the taking of time and title data amounted to taking a substantial part of the copyright work [385].

In the High Court [389], Gummow, Hayne, and Heydon found that the originality of the weekly programming schedules was in the selection and presentation of the information on times and titles and then packaged with additional program information and program synopses to make up a composite whole. However, the preparatory work involved in producing the time and title information was not relevant to substantiality, and there was left only 'the extremely modest skill and

[7] *Ibid.*

labour'. They also cautioned against reliance on the *Desktop Marketing* emphasis on appropriation of skill and labour, suggesting that the reasoning was out of line with the understanding of copyright law over many years [385].

One year later, in 2010, the Full Federal Court applied these principles in *Telstra Corporation Ltd v Phone Directories Co Pty Ltd* [390, 391]. In this case, Telstra claimed copyright in the content, form, and arrangement for each listing and enhancement in the White Pages and the Yellow Pages; in the overall structure of the listings in both directories; and in the headings, the presentation of the listings under headings, and the cross-referencing in the Yellow Pages. Both the Federal Court at first instance and the Full Federal Court on appeal found that the directories were not original works. A unanimous Full Federal Court affirmed that copyright did not apply to the White Pages and Yellow Pages as compilations because the works lacked 'human authors' who exercised 'independent intellectual effort' to create the form of the directories.[8]

Justice Keane and Justice Perram agreed that it was not necessary to name each author; the only requirement was to demonstrate that authors existed. If individuals had reduced the directories to material form through manual effort or had controlled a computer program in fashioning the form of the work, then the directories would have been original works. On this occasion, however, the task of transforming the information into a form ready for publication was carried out by software alone. Perram held that although humans were ultimately in control of the software, their control was over an automated process, and they did not directly form the material themselves. Therefore, there was no author of the directories and copyright did not exist in them [391].[9]

To summarise, as the consequence of the *IceTV* and *Phone Directories* cases, for a database to be eligible for copyright protection, it must meet the triple requirement that:

1. The data compilation must not be copied.

2. A human author must be involved in reducing or converting the database to a material form.

3. There must be some independent intellectual effort directed to expressing the work in the material form.[10]

Based on these criteria, it appears unlikely that research data created and arranged in databases in Australia would fall under the scope of copyright protection.[11]

Furthermore, copyright owners in Australia also have certain related rights, specifically moral rights—the right of integrity of authorship, the right of attribution of authorship, and the right against false attribution of ownership where copyright exceptions allow certain uses of copyrighted material without the authorisation of rights holders. Australia's copyright system includes an exception for 'fair dealing' for research or study ([392], p. 484). However, since it is unlikely that 'data' and 'databases' produced in Australia are subject to copyright, there is no need to apply the exemption to research data.

7.1.3 The United States

A database is protected by the United States *Copyright Act of 1976* [393] as a compilation, defined as:

---

[8] Fitzgerald and Dwyer [388] at point 13.

[9] per Perram at 101.

[10] See Fitzgerald at point 13.

[11] See Ricketson et al. [384] at point 8.

> *... a work formed by the collection and assembling of pre-existing materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship.*[12]

The concept of originality was further defined by the Supreme Court in *Feist Publications, Inc. v. Rural Telephone Service Company, Inc.* The Supreme Court held that:

> *Although a compilation of facts may possess the requisite originality because the author typically chooses which facts to include, in what order to place them, and how to arrange the data so that readers may use them effectively, copyright protection extends only to those components of the work that are original to the author, not to the facts themselves... As a constitutional matter, copyright protects only those elements of a work that possess more than de minimis quantum of creativity. Rural's white pages, limited to basic subscriber information and arranged alphabetically, fall short of the mark. As a statutory matter, 17 U.S.C. Section 101 does not afford protection from copying to a collection of facts that are selected, coordinated, and arranged in a way that utterly lacks originality [394].*

The Copyright Act is specific in stating that the copyright in a compilation applies only to the compilation itself, and not to the source data ([393], Par. 103 (b)). The decision in *Feist* confirmed that 'raw facts' have no protection under copyright law. Compilations of those facts require the application of a 'modicum' of creativity to be protected by copyright.

The originality requirement does not appear to be particularly stringent:

> *Original requires only that the author make the selection or arrangement independently ... and that it display some minimal level of creativity. Presumably, the vast majority of compilations will pass the test.*[13]

However, the selection in the rural telephone directory did not pass the test, for lack of the 'modicum' of creativity.

The criteria for 'modicum' are established as 'those constituent elements of a work that possess more than a de minimis quantum of creativity'.[14] Even a slight amount of creativity will suffice—'some creative spark, no matter how crude, humble, or obvious it might be'.[15] Furthermore, the modicum of creativity must be 'independently created by the author'.[16] The absence of creativity is manifested in an 'entirely typical' or 'garden-variety' end product constructed by processes which correspond to 'an automatic mechanical procedure'[17] or to a so ... routine process.[18]

Based on this reasoning, copyright law in the United States does not, in theory, appear to prevent the extraction of unprotected data from an otherwise protectable database. However, 'original' compilations of research data are likely to be subject to copyright protection which has repercussions for data licencing and may limit the possibilities for the sharing and reuse of data structured in databases. Only

---

[12] *Ibid*, Par. 101.

[13] Feist at point 25, at 358–359.

[14] *Ibid.*, at 363.

[15] Feist at 345.

[16] *Ibid.*

[17] Feist at 362.

[18] *Ibid.*

copyright holders can licence the data, and, in some cases, there would be multiple owners of copyright in one dataset resulting in copyright co-authorship of the work. That can create problems with data licencing unless all authors agree to the same licence conditions or waive their copyright. Furthermore, the distinction between raw facts (not covered by the protection) and a compilation of raw facts (to which copyright protection extends) is also not clearly delineated, especially in the cases of subsequent copies and derivatives of databases involving the original raw facts.

7.1.4 The European Union

Copyright law in the European Union has developed using the framework established by international treaties, such as the Berne Convention signed by all European Union member states, or by treaties to which the European Union is a signatory member in its own right, such as the WCT and TRIPS. These treaties are implemented through several European Union Directives—namely, the Directive on the legal protection of computer programs (Software Directive) [395], the Directive on rental and lending rights [396], the Directive on satellite broadcasting and cable retransmission [397], the Directive on the term of protection [398], the Directive on the legal protection of databases (Database Directive) [399], the Directive on the harmonisation of copyright and related rights in the information society [400], the Directive on the resale right [401], the Directive on certain permitted uses of orphan works [402], and the recently adopted Directive on collective rights management [403].

The European Union provides the strongest, double layer of protection of databases facilitated by the copyright laws and the Database Directive, which introduced a sui generis database right. As such, databases are, in the first instance, protected by copyright when the selection or the arrangement of the database represents its author's own intellectual creation. This layer of protection covers only the database structure, not its content, as is the position in the United States. The second layer of protection is the sui generis database right, which protects the content of the database—in the cases where there has been a substantial investment in the obtaining, presentation, or verification of the data—from acts of extraction (copying) and reutilisation (redistribution, communication to the public, etc.) of the whole or a substantial part of the contents of the database [404]. If the database meets the requirements for protection under both copyright law and the sui generis database rights, then the two types of protection are cumulative.[19]

With reference to the first layer, the test for originality has been harmonised across the European Union with regard to software[20] and databases[21] and photographic works[22] in the two relevant Directives mentioned previously. The European Court of Justice in *Infopaq International A/S v Danske Dagblades Forening* clarified the requirement of originality as the 'author's own intellectual creation' and established that the originality of a work must be assessed through its 'elements':

> *Regarding the elements of such works covered by the protection, it should be observed that they consist of words which, considered in isolation, are not as such an intellectual creation of the author who employs them. It is only through the choice, sequence*

[19] Article 7(4) Database Directive.

[20] Article 1(3) Directive 2009/24/EC.

[21] Article 3(1) Database Directive.

[22] Article 6, Directive 2006/116/EC of 12 December 2006 on the term of protection of copyright and certain related rights.

> *and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation.*[23]

In *Football Dataco Ltd et al. v Yahoo! UK Ltd*, the European Court clarified the position with regard to the threshold of originality in databases as follows:

> *... the fact that the setting up of the database required, irrespective of the creation of the data which it contains, significant labour and skill of its author ... cannot as such justify the protection of it by copyright under Directive 96/9, if that labour and that skill do not express any originality in the selection or arrangement of that data [405].*

Furthermore, the Directive does not provide for database right protection to apply to every aggregation of data. For example, databases that arise as a by-product of doing business do not attract database right protection. The sui generis database right of protection applies only if the creators have invested sufficient time, money, and skill into developing their database. The substantial investment must be either in the obtaining, verification, or presentation of the database contents.[24] This requirement was first tested in the *British Horseracing Board Ltd v William Hill Organisation Ltd* [406], in which the European Court of Justice found that 'obtaining' excludes the costs incurred in the creation of new data from being considered relevant to satisfy the requirement of the substantial investment:

> *... the expression investment in ... the obtaining ... of the contents of a database must ... be understood to refer to the resources used to seek out existing independent materials and collect them in the database, and not to the resources used for the creation as such of independent materials. The purpose of the protection by the sui generis right provided for by the directive is to promote the establishment of storage and processing systems for existing information and not the creation of materials capable of being collected subsequently in a database.*[25]

As such, the costs incurred in creating data for a database cannot be considered 'substantial investment'. However, the costs necessary for the verification of the accuracy of the data and for the presentation of such data to third parties do count in the assessment of whether the investment was substantial. This differentiation can also be used to extend the term of the protection granted under the sui generis right. The moment the database is completed or disclosed to the public, this right arises automatically, without any formal requirement. Protection under the database right is limited to 15 years, in theory. However, in practice, it has the potential to be perpetual. If the database is periodically updated, and such updating includes a substantial investment in reconfirming the accuracy of the information contained in it, then the period of protection can be continually renewed.[26] This is because the creator will have a new right to the altered database or its substantial part.

---

[23] *Ibid*, 45.

[24] Article 7, Database Directive.

[25] *Ibid*, 31.

[26] Article 10, Database Directive. Furthermore, Article 24 provides that 'a substantial new investment involving a new term of protection may include a substantial verification of the contents of the database'. See also Davison [407].

From the above, it is apparent that the scope of the sui generis database right goes well beyond the scope of copyright protection. The owner of the protected database has the exclusive right to prevent the extraction and/or reutilisation of the whole or of a substantial part, whether evaluated qualitatively and/or quantitatively, of the contents of that database.[27] Yet enforcing those rights and demonstrating that database rights apply has been a high bar to satisfy before the courts. The above-mentioned *Football Dataco* case to enforce database rights failed. So did the *British Horseracing Board* case.

However, the situation may be changing in the wake of the 2013 decision by the Court of Justice in *Innoweb BV v Wegener ICT Media BV and Wegener Mediaventions BV* [408]. In this case the court held that, in the European Union, the operators of aggregator websites that allow users to search for content on external databases, and provide the same search functionality as those source sites, and then display the found content on the aggregator sites may breach the database rights of the owners of the original content. In other words, in this situation the reutilisation of dataset content offends the protection for sui generis rights of the database creator that is provided under Article 7(1) and (5) [409].

This case is interesting in that it further prevents copying and reusing database content, even though it is questionable whether the original database at issue would have met the substantial investment criteria. The above judgement strengthens the position of database right owners. At the same time, it signals that others, for example, researchers or public libraries, must take care when designing their own search technologies that interrogate the databases created by other parties and then present that information within their own websites.

The broad scope of the sui generis database right and its interpretation by the courts are not a welcome development for open data. In many respects, the sui generis database right in Europe provides database rights holders more protection than the creators of original works can enjoy under copyright law.

Therefore, using somebody else's data produced in the European Union carries an inherent risk of IP infringement, especially as the exceptions to the sui generis database right are extremely limited. The main exception provided by the Database Directive is for the material extracted to illustrate teaching or for scientific research, with due acknowledgement of the source and a limit to the extent that extraction is justified by the non-commercial purpose.[28] Furthermore, there is no right of reutilisation for these purposes—it cannot be redistributed. An additional complication is the uncertainty of its reference to scientific research and whether this signifies 'illustration for scientific research', rather than simply 'scientific research'. Finally, the meaning of 'non-commercial purpose' in a teaching or research environment is also complicated. Finally, this exception is not mandatory, and some European Union countries—including Ireland, France, and Italy—do not have it in national legislation [407].

---

[27] Article 7 offers protection against acts of extraction or reutilisation of the whole or a substantial part of the database, evaluated qualitatively or quantitatively. The same article, in its fifth section, clarifies that the repeated and systematic extraction and/or reutilisation of insubstantial parts of the contents of the database, implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database, shall not be permitted. Extraction of insubstantial parts of the database does not infringe the database right. The sense of this norm is to avoid repeated extraction of insubstantial parts, which leads to the reconstitution of the database as a whole or as a substantial part thereof.

[28] Database Directive, at 26.

Therefore, the data created by European research organisations may need different treatment from data produced in other parts of the world. The strong protection of databases in the European Union appears to be at odds with the commitment to develop a Digital Single Market and data-driven economy—of which open scientific data, particularly via the Open Science Data Cloud, is an important component. Some committees of the European Parliament have called on the European Commission to abolish the Database Directive.[29] The committees have said they believed the Directive was 'an impediment to the development of a European data-driven economy'.[30] The European Commission appears to be aware of the limitations presented by the sui generis database right and has recently reaffirmed its commitment to develop the right environment and conditions for digital networks and services to flourish by providing, among other things, the right regulatory conditions [410]. Over 3 months in 2017, the Commission held public consultations on the application and impact of the Database Directive with the report following in 2018 [411].

As the discussion above shows, all three jurisdictions considered in this study have now adopted a test that requires a level of creativity to determine the existence of copyright in selecting the contents for or arranging a database. With such a test, data produced by researchers, at least in its unstructured or semi-structured form, will most likely fail to qualify for copyright protection. The one difference is in the European Union, where most databases are likely to fall under the provision for sui generis database protection, provided substantial investments in the databases are made. In early stages of the open data process, some international funders have suggested that where research data is protected by copyright law, it is not a proper subject for open access ([412], p. 18). Over time, however, legal mechanisms have evolved that enable the IP issues to be appropriately managed.

So how does the existence of copyright affect open scientific data and how can these issues be managed?

In broad terms, research organisations are familiar with copyright and related rights as they apply to publications, and they are making open research data available on the assumption that copyright also applies to open scientific data. The adopted approach is that the IP issues can be managed through appropriate licencing mechanisms, which would allow research organisations to waive their rights in data and enable others to reuse the content without any restrictions. However, there are issues with this approach.

The first is that the clearance of data rights is far more complex than clearing copyright and related rights in publications. There are two key reasons for this. One is that data owners must be identified in order to waive the rights and, unlike the initial rights in publications, the owners of research data may not be obvious. The second reason is that open data may include embedded objects and composite copyright that may be governed by multiple IP rights and multiple legal regimes. These concerns need to be managed and need to be managed early in the process.

I canvas these matters in the following sections—firstly, discussing data ownership, and then looking evolving licencing mechanisms for open scientific data.

---

[29] See in December 2015, the Committee on Industry, Research and Energy and the Committee on the Internal Market and Consumer Protection have called on the Commission to reconsider the sui generis database right (http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A8-2015-0371&format=XML&language=EN).

[30] *Ibid.*

7.2 Ownership of research data

Anecdotal evidence says that researchers, academics, students, and even academic researchers often believe 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. This view is incorrect. While they are employed to perform research, the data that researchers produce typically belongs to other parties. In most cases of researchers who are employees of a university or a research organisation, the rights to the data they produce is owned by their employers, pursuant to the operation of law[31] or contractual assignment. In sponsored research, the research organisation typically owns the data but leaves the role of data steward to the principal investigator. In industry-funded research, the data typically belongs to the sponsor; however, the right to publish it can also be extended to the investigator. The position with regard to the ownership of research data in the three jurisdictions under investigation is detailed below.

7.2.1 Australia

The ownership of research data in Australia is primarily determined by the organisation where the researchers work. It is currently the policy of the Australian Government to assert its ownership over intellectual property developed with public funding [392, 413]. This extends to apparently copyrighted data. The ownership of intellectual property in publicly funded research organisations is legislated, while most universities in Australia have in place internal procedures and employment contracts with their staff. Such contracts explicitly address the ownership of intellectual property, which also includes data.

Many universities have revised their internal IP ownership arrangements after the landmark decision in *University of Western Australia vs Gray* [414]. In this case, the university initiated legal proceedings against an employee to argue that the intellectual property, namely, patents, developed in the course of his employment belonged to the university. Dr. Bruce Gray was appointed as Professor of Surgery at the university in 1985. He carried out research, both before joining the university and after, on the use of microspheres to deliver anticancer agents to the sites of tumours. Dr. Gray filed various patent applications in relation to this work on behalf of a company, Sirtex Medical Ltd., of which he was a director. Subsequently, the company acquired the intellectual property from Dr. Gray. However, the university considered it had some rights to the intellectual property as a consequence of its employment of Dr. Gray to carry out research.

A decision by Justice French was delivered on 17 April 2008. The judgement effectively held that Dr. Gray's employment contract, which included a duty to carry out research, did not include a duty to invent, and accordingly the IP in the inventions Dr. Gray developed was not owned by the university. Justice French also found that the IP regulations of the university, which purported to invest the intellectual property rights of academic staff in the university, were invalid.[32] The university filed an appeal to the Full Bench of the Federal Court, which in its judgement on 3 September 2009 dismissed the appeal and confirmed the earlier decision of Justice French.

---

[31] For example, s 35 (6) Australian Copyright Act (http://www8.austlii.edu.au/cgi-bin/viewdoc/au/legis/cth/consol_act/ca1968133/s35.html).

[32] *Ibid*, FCA 49.

Several issues highlighted in the case can, by extrapolation, also be applied to the ownership of research data. Universities in Australia do not routinely rely on the operation of common law to assert their rights to academic IP. Instead, they make express provision for university ownership, typically by incorporating into academic employment contracts the terms of a university statute or policy to that effect. In the case of *UWA*, French J held that the IP regulations had not been validly passed or incorporated, and therefore the common law applied [415]. Since the decision, universities have amended their policies, and it is therefore unlikely that the common law further applies.

The judgement highlighted the public function of universities. It specifically acknowledged that universities serve the public purpose by offering education, by supporting research facilities, and by awarding degrees. It found, also, that commercial activities performed by universities had not displaced its traditional functions to the extent that it became 'limited to that of engaging academic staff for its own commercial purposes' ([416], p. 184). French further held that academics are to set and pursue research priorities and to publish or share research results. He also said that these freedoms collide fatally with a duty to maintain the secrecy that employer patent ownership inevitably requires. As such, an implied term favouring university ownership would be 'unsupported by a duty of confidence',[33] as in that case it would oddly mean that the academic 'would have been free to destroy the potential patentability of an invention by progressively putting research results into the public domain' ([416], p. 192). Alternatively, this view would be supported by an obligation of confidentiality, which is something so manifestly in opposition with traditional academic freedoms and practices that it cannot be maintained.[34] This judgement explicitly states that the public function of universities comes first and any commercial considerations follow. As such, this position supports the case for open research data.

With regard to the ownership of data produced in publicly funded research organisations, such as the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Section 54 of the *Science and Industry Research Act 1954*[35] provides that 'discoveries, inventions, or improvements' made by CSIRO officers in the course of their 'official duties' are owned by CSIRO, an Australian Government identity. The organisation also takes express assignments of IP in its employment agreements. As a result of the statutory provisions and these assignments, CSIRO controls, under Commonwealth executive approval, all research outputs created in the organisation—whether as data, publications, inventions, or other types of intellectual creations.

However, the CSIRO has not been at the forefront of research data sharing. Some of the data it produces is made publicly available by the organisation on its website, or in other publications, or via researchers (with CSIRO approval). However, only a few data sharing initiatives have emerged from the organisation—with the Atlas of Living, a free, online national biodiversity database being perhaps the best known.[36]

At the same time, the organisation is strongly committed to research commercialisation. In recent years it has had a strict internal policy of confidentiality, and it appears that many researchers fear being criticised for giving away data that could potentially be used to generate revenue for the organisation. The position

---

[33] *Ibid*, 191.

[34] *Ibid*, 192.

[35] This Act established CSIRO and regulates its governance.

[36] https://www.ala.org.au.

for confidentiality was supported by reasoning that the industry funds around 30% of CSIRO research. However, the remaining part is publicly funded, and the Australian Government increasingly expects greater returns from its investments in research.[37]

In March 2017, the Productivity Commission, in its report on an inquiry into data availability, proposed that 'the research community to put its house in order when it comes to data sharing' [392]. It specifically recommended that the data of publicly funded research be available beyond the initial researchers.[38] CSIRO is the largest and the most significant Australian publicly funded research organisation. Its organisational approaches to open research data therefore may need to change as the result of such recent reviews.

Rather than focusing on data ownership, the Productivity Commission preferred to stress the need for greater access. The default position in Australia is that all data created with public money should be publicly accessible within a reasonable time unless there is a compelling reason not to make it available.[39] The Australian Government announced in August 2017 that national, state, and territory governments should provide free and open access arrangements for all publicly funded research within 12 months of publication. This widens the Australian Government policy that presently governs grants from both the Australian Research Council and National Health and Medical Research Council.[40] The Productivity Commission report covers some of this territory, even though the examination is not specific. The report offers innovative approaches to releasing medical data and addresses the issue of the privacy of research subjects, discussed in Section 7 of this chapter.

7.2.2 The United States

The ownership of research data in the United States is typically determined by the employer of the researcher, similar to the position in Australia. As employees, researchers are hired by the university—which, in most cases, retains the rights to the data and other forms of expression. This principle is not open to debate as a legal matter [417]. A natural extension of this principle is that all the data created in the course of employment or with institutional support belongs to the employer. In federally sponsored research governed by the Bayh-Dole Act,[41] the research organisation also owns the data but permits the principal investigator on the grant to control the data [418]. However, the investigator is just a caretaker, not the owner of the collected data. He/she has charge of the collection, recording, storage, retention, and disposal of data.[42] More recently, the wording of research grants and contracts, and of the informed consent forms signed by participants in clinical trials, is also likely to delineate data ownership or disposition. The National Institutes of Health and the National Academies of Science include the requirements for data sharing among the terms and conditions of research grants, as discussed in Chapter 3.

---

[37] I gratefully acknowledge all the generous support, counsel, information, and insights I have received from Mr. Brett Walker, a former CSIRO counsel.

[38] *Ibid.*

[39] *Ibid.*

[40] See Recommendation 16.1, Productivity Commission [392].

[41] The Bayh-Dole Act or Patent and Trademark Law Amendments Act deals with intellectual property arising from federal government-funded research. The Act was adopted in 1980, is codified at 94 Stat. 3015, and in 35 U.S.C. § 200–212, and is implemented by 37 C.F.R. 401.

[42] *Ibid.*

Unlike the established practice in which academic institutions have often waived copyright in the literary and scholarly works of their researchers, universities and research organisations generally do not have an established tradition of abandoning ownership rights to data generated in the course of research by their employees. When faculty members leave an institution, they often negotiate with it to keep their grants and data. In industry-funded research, data typically belong to the sponsor, although in some instances the right to publish the data may be extended to the investigator.[43]

The key focus in the United States has been on consideration of who may access the data developed in the course of scientific research. This was partially driven by the United States patent law, which, until 2014, was based on the 'first to invent' principle. Laboratory notebooks and other evidence developed in the course of academic research were often used as evidence of the inventiveness principle, and academic researchers often appeared as expert witnesses in courts.

The focus on data access is still more dominant than discussion around data ownership. It is generally assumed that research organisations own the data of their researchers. However, the owner of the data does not always have control over it, as is the case in other types of intellectual property to which IP protection can apply. When it comes to research data, other parties may have legal access to it under prescribed conditions and for prescribed purposes. Moreover, data may be taken for public use without the need to seek the consent of the owner—subject to constitutional requirements for due process and fair compensation [419]. Ultimately, in the United States, the question of who owns the data appears to be less of a concern than the matter of the rights and responsibilities of data holders.

Recent years have seen increased calls from patients in the United States to claim rights in data they produce in the course of clinical trials, as previously outlined in Chapter 5. One controversial issue concerning data ownership concerns cell lines and DNA sequences, which can represent 'data' in clinical trials. Controversies have arisen concerning whether research subjects and patients actually own their own tissue or DNA.

Such challenges are not new. A case brought by John Moore against the University of California in the late 1980s raised issues about whether a patient has ownership of his tissue that was used in research to develop a cell line that had commercial interests. In 1976 Moore had gone to the UCLA Medical Center seeking treatment for hairy cell leukaemia. The research performed on cells from his spleen led to the development of a patent 6 years later. Moore sued the University of California Regent as well as the company where his doctor was working, stating that the altered tissue was his own property and that he wanted to recover damages. The claimant also said that he had not been informed about the potential use of his tissue by the researcher. The California Supreme Court held that Moore had a right to sue the doctor for failing to inform Moore of what he intended to do with his cells [420]. However, Moore did not win the right of ownership of his cells nor any entitlement to the data and subsequent financial proceeds that might be generated from the research done using the cells. The court said that if all subjects had the right to their own tissue, it could hinder biomedical research.

The court reasoned that before a body part is removed, it is the patient who possesses the right to determine the use of that part.[44] However, the court construed that the removal of a body part with informed consent was an 'abandonment' of that part.[45]

---

[43] Fishbein [417] at point 80, 129.

[44] *Ibid*, 500.

[45] *Ibid*, 501.

The judge did not say what rights (if any) others may have in the abandoned body part or whether such 'data' can be used for research purposes and shared subsequently. This issue is of utmost importance and has been brought back to the spotlight in relation to collecting newborn blood samples by some state governments, especially California. While collecting the samples to screen babies for genetic diseases requires the informed consent of the parents,[46] the established practice was to store de-identified samples in a state database and to use them for federal research. In 2017, the Department of Health and Human Services and 15 other federal agencies jointly issued a 'final rule' that:

> *strengthens protections for people who volunteer to participate in research, while ensuring that the oversight system does not add inappropriate administrative burdens [422].*

The effect of this rule is that researchers do not need consent to use de-identified blood spots and, in some cases, can even use identified blood spots without consent.[47] Parents can, however, opt to destroy the blood samples after the newborn test is performed.

7.2.3 European Union

Draft legislation currently being considered by the European Union would specifically regulate ownership in data in general and research data in particular. In the context of the European Commission free flow of data initiative, the agency stated that:

> *... the barriers to the free flow of data are caused by the legal uncertainty surrounding the emerging issues on 'data ownership' or control, (re)usability and access to/transfer of data and liability arising from the use of data [423].*

Data ownership in the European Union was recently considered by a private law firm [424], which found that European Union case law does not explicitly recognise an ownership right in data. However, the European Court of Justice opened the door for a discussion on ownership in intangible assets in its *UsedSoft* judgement issued on 3 July 2012 [425]. In this ruling, the court held that the commercial distribution of software via a download on the Internet involves the transfer of ownership [426]. Specifically, the CJEU held that the copyright holder's exclusive distribution right in a computer program is exhausted upon the first sale of the program, including in a program downloaded over the Internet under a user licence agreement. The court held that such licencing involves the transfer of ownership. Therefore, the owner of copyright in software is unable to prevent a perpetual 'licensee' from reselling the 'used software licences'.

7.3 Licencing models for open scientific data

For data to be released in the public domain as open access, it must meet certain conditions. The Berlin Declaration defines such conditions as:

---

[46] Blood spots are defined as 'human subjects' and require informed consent for federal research. See H. R.1281—Newborn Screening Saves Lives Reauthorization [421].

[47] *Ibid.*

> *The author(s) and right holder(s) of the data grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship (community standards will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use [63].*

For research data to be open, specifically where exclusive ownership rights exist, it needs to be released (published) in the public domain under an open licence. Several licences have evolved over time to meet the specified conditions for 'open scientific data'.

7.3.1 Creative Commons Zero public domain dedication (CC Zero)

Unlike the other six licences developed by the Creative Commons, CC Zero (sometimes presented as CC0) is not a licence but rather a waiver, to the fullest extent permitted by law, of copyright and the full scope of related (or neighbouring) rights. The waiver was developed with an intention to facilitate the sharing of research data. Specifically, the person waiving their rights (the affirmer):

> *... overtly, fully, permanently, irrevocably and unconditionally waives, abandons, and surrenders all of Affirmer's Copyright and Related Rights and associated claims and causes of action, whether now known or unknown (including existing as well as future claims and causes of action), in the Work (i) in all territories worldwide, (ii) for the maximum duration provided by applicable law or treaty (including future time extensions), (iii) in any current or future medium and for any number of copies, and (iv) for any purpose whatsoever, including without limitation commercial, advertising or promotional purposes [427].*

Consequently, the waiver enables users of the data to copy, modify, distribute, and perform the work, even for commercial purposes and to do so without asking permission.

An important point to mention is that, unlike the first three versions of the Creative Commons licences, the waiver covers both copyright and sui generis database rights. Further, CC Zero avoids problems with attribution stacking[48] by removing the legal requirement to give attribution while acknowledging that the scientific community has a well-established culture and norms that encourage the recognition of sources. As such, CC Zero is the recommended tool for releasing research data into the public domain.

7.3.2 Creative Commons 4.0 suite of licences

Creative Commons 4.0 is a suite of standard, globally applicable terms that allow anyone to openly licence all forms of creative works and datasets. These public licences are exceptionally user-friendly and enable copyright owners to licence their works on the Internet and elsewhere. Unless directed otherwise (by research funders, scientific organisations, or other owners of copyright in research data), owners can choose the conditions for the future reuse of the works. These may include Attribution (BY), Non-Commercial Use (NC), No-Derivatives (ND), and

[48] The accumulation of attributions that occurs as each reuse of data incorporates acknowledgements of all prior users.

Share Alike (SA). A tool on the Creative Commons website can generate text, taking into account the conditions selected, by which the copyright owner may grant a worldwide, non-exclusive, perpetual licence to any user to reproduce, display, perform, communicate, and distribute copies of that work.[49] The same licence permits any future reuse of the work according to the stipulated conditions and without the need to contact the copyright owner. The licence applies to all media and formats, whether known now or subsequently devised. All Creative Commons 4.0 licences are irrevocable, meaning that once the licenced work is distributed on the Internet, the author can no longer change the type of licence or withdraw it.

The Creative Commons 4.0 licences are widely used in the context of scholarly publication and the dissemination of research results. The current (fourth) iteration of the licences is recommended, as it also provides for the sui generis rights in the European Union and includes mechanisms to avoid attribution stacking (in the case attribution is selected).[50]

The development of the Creative Commons 4.0 licences has eliminated the need to apply other licences to scientific contents. One such example is the Public Domain Dedication and Licence previously developed by the Open Knowledge Foundation and used in some European Union countries. This strongly resembled the CC Zero waiver; however, it was designed to enable licencing of databases and its contents in the European Union, paying particular attention to the European sui generis database right. With the adoption of the fourth iteration of the Creative Commons licences, many European organisations now recommend solely one suite of licences, namely, the CC Zero waiver and 4.0 licences.

7.3.3 Other licencing issues

Given that copyright is unlikely to apply to data itself, but instead applies to original compilations of the data, it follows that much data is arguably not subject to copyright protection. This is, for example, the current position in Australia with regard to computer-generated data. Another example that would seem to be exempted from copyright protection is unstructured data developed in the course of a research project or harnessed by other means from scientific equipment. The lack of copyright protection in data is the general position with regard to data generated in the United States.

This raises interesting questions about whether any property rights can be claimed in such 'data without author' and what the legal basis for such a claim might be. Arguably, data that is not subject to copyright protection can still constitute 'confidential information' or other forms of 'intellectual property' especially if the data is governed by contractual arrangements with industry or other research collaborations. In these cases, any property rights in such 'data without author' would be most likely determined in the contracts. However, since the data is not subject to copyright protection, and thus does not have an author, issues arise with regard to how to release the data in the public domain. In such cases, when no rights are attached to research data, then there is no ground for licencing the data. Standard copyright licences, such as the Creative Commons licences, are not appropriate.[51]

---

[49] www.creativecommons.org.

[50] Note: Attribution is not a selectable option in CC 4.0.

[51] In the absence of copyright, companies sometimes license data under 'know-how' agreements, and in these cases the ownership of the data is usually vested in the company, or a subcontractor to the company. Adopting this approach may not be appropriate for research data, due to public nature of academic research.

There are two ways institutions have chosen to release research data to which copyright does not apply. Some organisations in the United States release their data in the public domain without a licence. This was the early approach taken by the Harvard-MIT Data Center. Secondly, some organisations in the United States release data under the CC Zero waiver, and this seems to be the recommended practice for sharing research data and databases [180]. Such an approach is preferred because it signals to future data users that the data is without any legal restrictions on reuse.

Creative Commons has also developed the Public Domain Mark (PDM) with a view to enabling marking of materials, including data, which belong to the public domain. Unlike the CC Zero waiver, which can only be used by copyright holders, PDM can be used by anyone.

PDM is not a legal tool in any respect. It was developed with a view to acting as a label, marking material that is free of known copyright restrictions [428]. However, Creative Commons currently does not recommend the PDM for materials for which the copyright status differs from jurisdiction to jurisdiction, even though the tools for marking and tagging such works are currently under development. In the absence of these marking tools, there is a concern that the PDM tool might be used to overwrite the rights of lawful copyright owners. Therefore, using PDM to release research data is not recommended, and the CC Zero waiver has become an established norm around the world.

7.4 Different types of data reuse

The previous section described the challenges associated with data release. I now move to describe the challenges arising in data reuse.

This study has identified three such issues: firstly, the inability of data users to perform automated analysis and mining of digital data; secondly, ensuring the ethical use of data and limiting the risks of inaccurate interpretation; and lastly, ensuring the privacy and confidentiality of research subjects involved in clinical trials. These challenges are examined below.

7.4.1 Text and data mining

Legal uncertainty remains with regard to certain data uses and reuses in the digital environment. Typically, linking and mining of data and text are necessary to extract value and insights from datasets or other forms of data. However, such uses may constitute copyright infringement.

This uncertainty stems from several factors. Firstly, databases and some forms of data may be protected by copyright, as discussed above. Secondly, such data may be available in the public domain, but is not open, meaning that a prospective user can access the data but may not be able to reuse it or is unaware of the terms under which the data may be reused. Thirdly, new types of data uses, such as linking and mining, may cover several data sources and span several jurisdictions. Making temporary copies of the data is usually necessary to perform large-scale data analyses. Yet the act of copying is not clearly covered in the scope of exceptions and limitations to copyright infringement. Moreover, the scope of these exceptions varies from jurisdiction to jurisdiction, and this may hinder data interoperability and reusability.

Text and data mining generally involves automatically collecting information and extracting data and insights from digital data by means of software. Citing various legal and literature sources, the European Parliament has defined the process of text and data mining in these terms:

TDM works by:

1. *Identifying input materials to be analysed, such as works, or data individually collected or organised in a pre-existing database*

2. *Copying substantial quantities of materials which encompasses*:

    a. *Pre-processing materials by turning them into a machine-readable format compatible with the technology to be deployed for the TDM so that structured data can be extracted*

    b. *Possibly, but not necessarily, uploading the pre-processed materials on a platform, depending on the TDM technique to be deployed*

3. *Extracting the data*

4. *Recombining it to identify patterns into the final output ([429], p. 5)*

The nub of the problem with text and data mining is the requirement to create a temporary copy of the data. While data itself is not protected by copyright and/or the sui generis right, a database might be, especially if substantial parts of the original database are extracted for purposes other than research or learning.

Publishers have typically taken a sceptical approach to allowing text mining, even for research purposes, and instead have promoted obtaining a licence on a case-by-case basis. This is time-consuming and involves high transaction costs. Some academic journal publishers, such as Elsevier and Oxford University Press, allow text and data mining for non-commercial use [430]. This permission overrides the need to seek permission from the publishers to reuse the content. However, permissions that specifically address data mining are uncommon at this time.

There are two principal ways to ensure that text and data mining does not infringe copyright law. The first is the fair use doctrine enshrined in the United States copyright law; the second is the system of exceptions and limitations embedded in Australian and European Union law. The United States system is considered more favourable to text and data mining due to its inherent flexibility. Many scholars and policymakers have argued that Europe lags behind the United States in unlocking the value of data because of its inflexible copyright laws.[52]

7.4.2 The fair use system in the United States

The fair use doctrine is stipulated in Paragraph 107 of the United States Copyright Act [393]. The application of the doctrine requires consideration of several factors to determine whether a certain use of copyrighted works indeed constitutes 'fair use'. These include factors such as:

> *... the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes; the nature of the copyrighted work; the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and the effect of the use upon the potential market for or value of the copyrighted work [431].*

[52] See, for example, the discussion of the text and data mining exemption in the European Parliament here.

The factors are weighed as a whole, and so the claimant need not win on every factor for a court to rule in favour of fair use.

More recently, the use of text and data mining was considered in the cases involving the Google Books Library Project, especially in the *Authors Guild, Inc. et al. v. HathiTrust* [432]. In this matter, Google had created digital copies of books held in university libraries and then provided digital copies to HathiTrust Inc., which developed a searchable database for use by researchers and scholars. The search results included 'snippets' of text. Judge Chin held that the digitisation of books by Google was 'highly transformative' as it adds value, serves several important educational purposes, and may enhance the sale of books to the benefit of copyright owners. In this reasoning, the judge explicitly referred to 'text and data mining' as a new area and method of research.[53] A similar judgement by the district court for the Southern District of New York explicitly referenced the benefit of Google Books to TDM, noting that it 'transformed the book text into data for the purpose of substantive research, including data mining and text mining in new areas' [433]. While the consideration of fair use varies from case to case, the previous judgements indicate that text and data mining is likely to be considered 'fair use', especially if undertaken in the course of research.

7.4.3 Australia

As it stands, Australian copyright law does not currently allow text and data mining of large datasets. Australia does not have a text and data mining exemption but has, on several occasions,[54] considered introducing a fair use system similar to that of the United States in place of the current 'fair dealing' system. However, the response of the Australian Government to these reviews is lacking. The current 'fair dealing' system allows certain limited exceptions for use of copyrighted works for criticism and review, research and study, reporting the news, use in judicial proceedings, and parody and satire.

The 2014 review by the Australian Law Reform Commission specifically considered the effects of text and data mining in the context of copyright law. In that review, it was concluded that where the text or data mining involves the copying, digitisation, or reformatting of copyright material without permission of the copyright owners, it may give rise to copyright infringement [434], especially if the whole dataset needs to be copied and converted into a suitable format (such as XML format). In such cases, the copying would exceed a 'reasonable portion' of the work and so fall under the scope of infringement. The inquiry also said that it 'seemed unlikely' that text and data mining might fall under the temporary reproduction of work exception. The recommendation was to introduce the 'fair use' system based on the United States system. However, this recommendation has not been adopted by the government, which—in the words of the then Attorney General—'was still to be persuaded that the adoption of fair use was the best direction for Australia law'.[55] This position has not changed under the current Australian Government, and, as a result, copyright law poses challenges to data reuse. The recommendation of this study in this regard is provided in Chapter 8.

---

[53] *Ibid.*

[54] In between 1998 and 2018, eight reviews considered introducing fair use, and six reviews explicitly recommended it.

[55] Senator the Hon. George Brandis QC, Attorney General and Minister for the Arts (2014). Address at the opening of the *Australian Digital Alliance fair use for the future. A practical look at copyright reform forum.* Canberra, 14 February.

7.4.4 The European Union

The threshold for copyright protection in data, even raw data, is relatively low in the European Union. In *Infopaq*,[56] the Court of Justice held that even a short sequence of 11 words may be subject to copyright if it reflects a sufficient level of creative choices leading to an 'own intellectual creation'.[57] Multiple extractions of text from the same source, such as the systematic mining of a blog, further increase the risk of infringement. It seems clear that, as a general principle, relatively small takings of data can raise copyright issues.

However, the European Union is currently considering broad and ambitious reform to the European Copyright Directive that was adopted in 2001 and is now considered outdated. The current package of proposals[58] has been developed over several years and includes a new copyright exception for text and data mining. Such an exception is necessary to ensure harmonisation of laws across the European Union. Some member states have, however, recently proceeded to introduce national text and data mining exemptions.

The first country to do so was the United Kingdom, following the recommendations of the Hargreaves Review ([435], p. 47) and so adopting a text and data mining exemption on 19 May 2014.[59] The exception only applies to non-commercial research. According to the amended legislation ([436], Par. 29A):

> *... the making of a copy of a work by a person who has lawful access to that work does not infringe copyright if it is made so that that person can carry out a computational analysis of anything included in that work for non-commercial research purposes.*[60]

France, Estonia and Germany have also introduced text and data mining exceptions. The French exemption is extremely narrow and covers only reproduction from 'lawful sources' made available with the consent of the rights holders, as well as the storage and communication of files created in the course of performing text and data mining activities.[61] The scope of the exemption adopted in Estonia is similar to the United Kingdom's law and is limited to text and data mining performed by any person but only for non-commercial purposes.[62] Germany is the latest European country to introduce the text and data mining exception, in March 2018. It covers the act of reproduction necessary to undertake text and data mining for non-commercial purposes.[63]

The package proposed for the entire European Union is currently being considered by the European Parliament and includes various changes to the scope of the proposed exception. Of particular interest is the enabling of researchers and

---

[56] CJEU, 16 July 2009, case C-5/08, *Infopaq*.

[57] *Ibid*, 48.

[58] See Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016PC0593&from=EN).

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

[60] *Ibid*, Par. 29.1.

[61] See European Parliament at point 2, 17.

[62] *Ibid*, 18.

[63] *Ibid.*

businesses to harness the benefits of data mining. A specific case was put forward for including start-ups in the scope of the exemption. It was argued that the exception would allow start-ups to increase the European Union competitiveness and knowledge leadership in the field of big data analytics, as desired by the Commission [437]. Another reason put forward for including non-commercial use in the exemption was reasoning that nearly all research today includes multi-parties—public, private, and not-for-profit, among others. As such, limiting the scope of the exemption to non-commercial research may not cover any data uses by parties other than researchers working in publicly funded research organisations.[64]

7.5 Privacy of research subjects

Protecting data, including research data, is an increasingly important topic for research and regulatory agencies, especially those involved in clinical trials. People participating in clinical trials have a right to expect that their personal data and the information shared with their doctors will remain confidential. Health services depend on trust, and trust depends on confidentiality [438].

At the same time, sharing patient information for research purposes is an important prerequisite for advancing public science and the well-being of all citizens. Therefore, the practice of research requires a careful balancing of the respective interests in both data protection and data sharing. For these reasons, stakeholders who advocate the sharing of scientific data refer to it as 'responsible sharing'.[65] In this context, the tasks of maintaining confidentiality and safeguarding the privacy of research subjects are viewed as the requirements of research conduct, rather than barriers to data sharing. This is an important distinction and one that implies that sharing clinical trial data without compromising the privacy or confidentiality of research subjects is not only desirable but is also possible and can be achieved through transparent and open data sharing practices championed by institutions such as the EMA.

7.5.1 The sources of confidentiality

Researchers and research investigators have the primary responsibility for maintaining confidentiality and safeguarding the privacy of people participating in their research.[66] They are also responsible for collecting informed consent and informing participants about data use and how confidentiality will be maintained. Obligations of confidence stem from diverse sources of law and have been extended to various areas—including privacy, confidentiality, trade secrets, data protection, labour law, and professional and research ethics, among others. This section considers the key effects of these laws on the release of open data and the latest practice guiding the responsible sharing of clinical trial data.

In Australia, the obligations of confidentiality generally arise under the common law system, as:

- Implied by operation of our common law through the equitable doctrine of confidence

---

[64] *Ibid.*

[65] Institutes of Health (2017).

[66] Universities in particular require ethics committee approval for undertaking research involving human beings or human activity.

- Expressed through a contractual obligation
- Imposed through operation of legislation (e.g. disclosure of sensitive information)

The first doctrine is commonly implied through a relationship between the party disclosing information and the person to whom it is disclosed—for example, through a doctor-patient relationship or employer-employee relationship. In this regard, the employee has a duty of fidelity to the employer, who can prevent disclosure of information acquired in the course of employment ([439], pp. 860, 867).

Secondly, an obligation of confidentiality may arise from various contracts that govern the disclosure of confidential information—such as trade secrets, confidential agreements, or non-disclosure agreements. In the public research setting, such arrangements are typical in contracts with industry research sponsors who explicitly or implicitly require that confidentiality. Release of research data that contains confidential information is effectively prohibited unless the industry partner provides explicit permission.

Lastly, the obligation to maintain confidentiality can stem from statutes such as the *Privacy Act 1888* (Australia), the General Data Protection Regulation (European Union),[67] and the *Health Insurance Portability and Accountability Act of 1996* (United States) [440] or from various professional code of conduct principles enshrined in legislation. Under such legislation, and due to the recent changes to the global regulatory framework for the sharing of clinical trial data introduced by the European Medicines Agency (EMA) in 2014 [441], the practice of clinical data sharing has transformed quite dramatically in recent years. Legislative and policy changes require drug regulatory agencies to redact the records and/or data they share to de-identify personal details and to remove commercial confidential information. Given the global reach of research based in, or funded, by the European Union, the developments occurring on the continent are likely to influence the global practice of sharing clinical trial data as open data, with efforts mounted to drive the adoption of the interoperability of standards.

7.5.2 Latest approaches to the protection of privacy and sharing sensitive data for research

There have been significant recent developments in the European Union data protection law that will have an impact on research data sharing. The General Data Protection Regulation (GDPR) took effect on 25 May 2018.[68] For the first time, the Regulation is directly enforceable across the European Union and replaces transposition of the Directive at the national level, as was the case with the previous Directive.[69] However, the European Union member states are permitted minor differences in interpretation, with the European Court of Justice as the ultimate arbiter.

The principal tenets of the Regulation with regard to the processing and sharing of sensitive data in scientific research are as follows:

---

[67] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) [42, 43].

[68] See point 4 above.

[69] Directive 95/46/EC.

- A risk-based and context-specific approach to data processing, aimed at ensuring that appropriate data protection measures are employed in data processing.[70]

- A highly decentralised approach to data handling and processing, vesting responsibilities for data processing in data controllers[71] and providing for decentralised accountability.[72] Data controllers need to adopt a proactive approach to data protection and are responsible for the assessment, implementation, and verification of the measures to ensure compliance with the Directive.

- The Directive specifically enables the processing of sensitive data for scientific research in the 'public interest',[73] requiring organisational and technical measures such as 'pseudonymisation'[74] and the designation of a data protection officer in the cases of large-scale and systematic processing of sensitive data.[75]

- Maintaining the broad notion of informed consent required to process data for future uses, which may not have been known at the time of obtaining informed consent.

The term 'scientific research' is not defined in the Regulation, yet a recent report of the GDPR Working Group[76] clarified that it means 'a research project set up in accordance with relevant sector-related methodological and ethical standards'.[77] Moreover, the processing of personal data for scientific research purposes 'should be interpreted in a broad manner'.[78] Recital 33 states:

> *It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.*[79]

The Working Group further clarified that scientific research projects can only include personal data on the basis of consent if they have a well-described

---

[70] Enshrined in Article 25, in the 'data protection by design and default principle'.

[71] Articles 5(2) and 24. Controllers are defined as the persons, companies, associations, or other entities that are in control of personal-data processing.

[72] Article 40.

[73] Article 9.2.j and 9.2.g.

[74] Article 4(5) defines pseudonymisation as 'the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'.

[75] Articles 37 and 39.

[76] The Working Party on the protection of individuals with regard to the processing of personal data. Guidelines on Consent under Regulation 2016/679 adopted on 28 November 2017.

[77] *Ibid*, 27.

[78] Directive at point 4, Recital 159.

[79] *Ibid*, Recital 33.

purpose[80] and if processing of the data is compatible with the initial purposes for which personal data was originally collected.[81] If purposes are unclear at the start of a scientific research program, controllers will have difficulty pursuing the program in compliance with the Directive, which has introduced criteria for compatibility assessment. These aim to determine, on a case-by-case basis, whether further processing of personal data would meet the requirement of compatibility.

The Working Group also mentioned that transparency is an additional safeguard when the circumstances of the research do not allow for specific consent. A lack of purpose specification may be offset by controllers providing regular information on the development of the purpose as the research project progresses so that, over time, the consent will be as specific as possible. In that context, the data subject should have at least a basic understanding of the state of play, allowing that person to assess whether or not to use, for example, the right to withdraw consent pursuant to Article 7(3) of the Directive.

The processing of sensitive research data should be subject to appropriate safeguards for the rights and freedoms of the data subject, and so the Directive mentions techniques such as data minimisation, anonymisation, and data security.[82] Anonymisation is the preferred solution, provided that the purpose of the research can be achieved without the processing of personal data.

Similar decentralised approaches to data de-identification are currently being pursued in the United States. Policy 45 CFR part 46, known as the 'Common Rule' [442], requires de-identification of data prior to release for further research.

The HIPAA Privacy Rule[83] defines the direct personal identifiers (see **Table 6**)[84] and outlines two approaches commonly applied—firstly, expert determination, and secondly, safe harbour.

The first approach requires a statistical expert to apply statistical methods to render data not individually identifiable. This method often results in excessive information loss that can wipe out the analytical utility of the dataset [443].

The safe harbour approach is consistent with the de-identification approach pursued in Europe and requires masking of both direct and indirect identifiers. This process can be automated to a large degree.

7.5.3 Open sharing of sensitive commercial documents

An important aspect of the de-identification process is not only to safeguard the privacy of the research subject but also to enable publishing of the de-identified results online so as to enable the transparency of pharmaceutical research, particularly for regulatory approvals. Championed by the EMA, this approach to open access—in addition to safeguarding the privacy of research subjects—requires the redaction of confidential commercial information.

The requirement by the EMA for the public release of clinical summary reports submitted to it for gaining marketing authorisation or additional market exclusivity

---

[80] Report of the Working Party at 30.

[81] Directive at point 4, Article 6.4.

[82] The processing of personal data for scientific purposes should also comply with other relevant legislations such as on clinical trials (see Recital 156 of the Directive at Point 4).

[83] Arising from the *Health Insurance Portability and Accountability Act of 1996* to provide data privacy and security for medical information (https://www.hhs.gov/hipaa/for-professionals/privacy/index.html).

[84] Source: *Health Insurance Portability and Accountability Act of 1996* (https://www.hhs.gov/hipaa/for-professionals/privacy/index.html).

1.	Name
2.	Geographic subdivisions smaller than a state. The initial three digits of a ZIP code can be retained if certain criteria are met.
3.	With the exception of year, all elements of dates directly related to an individual (such as birth date, admission date, discharge date, date of death).
4.	Telephone numbers
5.	Fax numbers
6.	Email addresses
7.	Social security numbers
8.	Medical record numbers
9.	Health plan beneficiary numbers
10.	Account numbers
11.	Certificate/licence numbers
12.	Vehicle identifiers and serial numbers, including license plate numbers
13.	Device identifiers and serial numbers
14.	Web Universal Resource Locators (URLs)
15.	Internet Protocol (IP) addresses
16.	Biometric identifiers, including finger and voice prints
17.	Full-face photographs and any comparable images
18.	Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of HIPAA Safe Harbor section; and
19.	The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Table 6.
Ensuring privacy—HIPAA 18 direct identifiers.

has met the resistance of pharmaceutical companies. A number of them have objected to the disclosure of the documents and initiated legal proceedings against the EMA. In February 2018, the General Court delivered judgements in the cases brought by Pari Pharma [444], PTC Therapeutics International [445], and MSD Animal Health Innovation [446]. The court dismissed all three cases as it considered that the companies had failed to provide any concrete evidence of how the disclosure of the contested documents would undermine their commercial interests.

These cases tested, for the first time, the application of the EMA policy on access to documents [346] in the context of the European Union Transparency Regulation [447]. That policy enabled the release of documents that the companies considered were submitted on a confidential basis, and these cases were the first to challenge the legality of the transparency of the EMA approach. Specifically, the EMA submitted that the balance between the commercial interests of the companies and the interests of the general public and public health should lead to disclosure as a default position, except in the cases where the company would clearly demonstrate that such disclosure would undermine its commercial interest.

To implement the policy, the EMA had developed a robust document redaction process and consulted the companies whose documents it sought to release. However, the EMA resisted the claim that the entire documents should be protected from disclosure. The arguments put forward by the EMA included that some of the contents were available in the public domain. The companies counterargued that

their compilation of public and non-public data might enable competitors to gain a market advantage.

The Pari Pharma case was the first considered, and the resulting judgement framed the results in the other two. Specifically, the court dismissed the claim that the published documents were presumed confidential. It said that the documents could be subject to a presumption of confidentiality if there existed ongoing judicial or administrative proceedings, but in this case there were none. With regard to the substance of commercially sensitive information, the court said that these could include 'considerations relating to an inventive strategy'[85] or a 'new scientific conclusions'.[86] However, Pari Pharma failed to make the case that any individual pieces of information included in the report should be protected from disclosure.

In particular, the court held that Pari Pharma failed to 'describe in specific terms the professional and commercial importance of the information'[87] along with 'the utility of that information for other undertakings which are liable to examine and use it subsequently'[88] and that the company had failed 'to show specifically and actually how, once the documents have been disclosed, competitors would be able to enter the market'.[89]

Pari Pharma then tried to argue that there was no overriding public interest in disclosure as it was already served in another report. But the court said that, having concluded that the contested information was not commercially confidential, the EMA did not need to determine whether there was or was not an overriding public interest in disclosure. So the claims failed on all accounts.

In the meantime, the EMA continues to disclose reports submitted as part of the regulatory process. In light of this practice, companies continue to argue for maximum redaction and have refined their approach to submitting evidence presented to the EMA. However, last year the EMA rejected 76% of the requests by pharmaceutical companies to redact what they claimed was confidential information [448] and published over 1.3 million pages in 2017 alone.

7.5.4 Approaches to data sharing, managing privacy, and confidentiality in Australia

The current Australian Government has taken an active role in developing an integrated data system across the economy and has, at the time of finalisation of this book, introduced a roadmap towards a new data regulatory mechanism with a view to improving Australia's ability to capture the social and economic benefits from the existing data [449].

The proposed mechanisms aim to improve access to and derive value from public data by introducing a new data regulatory mechanism. The key elements of the proposed framework relevant to the sharing of research data include:

- Taking a risk-based approach to releasing available publicly funded datasets.
- Streamlining and standardising data sharing arrangements.

---

[85] Pari Pharma at 25, at 78 and 79.

[86] *Ibid*, 77.

[87] *Ibid*, 108.

[88] *Ibid.*

[89] *Ibid*, 118.

- Accredited Data Authorities will engage with data custodians and users on matters relating to data availability and use. The authorities will make decisions on the data to be shared openly and that which requires restricted sharing. The authorities would also certify 'trusted users'.

- Data sharing agreements between data custodians, Accredited Data Authorities, and data users will be a key part of the governance framework.

- Development of National Interest Datasets across and between sectors, including public, private, not-for-profit, and academia.

- Introducing a *Data Sharing and Release Act*, which will set clear rules and expectations for data sharing and release, including making clear when data can be shared and embedding strong safeguards for sensitive data and effective risk management practices.[90]

While the objectives of the Australian Government are laudable, there are, however, significant problems with the proposed approach of 'balancing data sharing with secrecy' and adopting centralised and 'standardised' approaches to data sharing. A particular issue explored in detail in this book is that standardised approaches to data sharing have not been effective drivers of increased data availability and reuse. Similarly, developing closed and rigid communities of 'trusted users' is unlikely to achieve the desired spillover of data and knowledge to enable harnessing of the economic benefits of data. The proposed approach fails to recognise that privacy and security concerns only apply to highly sensitive datasets, which represent only a small subset of national datasets. Most data can be shared freely without any restrictions. However, as the proposal stands now, it appears that the Australian Government has adopted screening approaches across the whole board.

One of the defining features of our time is that the Internet and communication technologies have led to the reconfiguration of power structures and have promoted the rise of distributed social and research networks. In this environment, balancing data sharing with secrecy cannot be a zero-sum game. Any attempts to centrally regulate and restrict data release and use will be met with resistance from Australian citizens and researchers who currently control data and use it on a daily basis to extend the boundaries of science. These are important considerations for the Australian Government to incorporate into the proposed governance structures.

Conclusion

This chapter outlined the legal issues arising in stages of data release and data reuse, focusing on the recent developments and legislative proposals aimed at enabling researchers to share and reuse data while also respecting the emergent rules for responsible data sharing.

The examination found that copyright law poses serious challenges to data release and reuse in all three jurisdictions under examination—the United States, Australia and the European Union. The problems arise due to uncertainty surrounding the scope of copyright protection as it applies to the various forms of data, especially databases. The situation is even more complicated in the European Union which provides a double layer of sui generis and copyright protection. Therefore,

[90] *Ibid.*

using the data created by European research organisations carries an inherent risk of IP infringement. Another source of legal uncertainty is the ownership of data and the inability of users to identify data owners, which poses challenges to data licencing and subsequent reuse due to lack of clarity around the conditions governing data reuse.

Various mechanisms have emerged to deal with the challenges. A particular focus has been placed on enabling greater access to data produced by publicly funded research organisations. The question of data ownership appears to be less of a concern to researchers than the matter of the rights and responsibilities of data holders.

This is particularly the case with clinical trials, which collect vast amounts of data from patients and other research subjects. The sharing of the data requires informed consent, and recent years have seen patients demanding a greater say over the use of the data generated in clinical trials. The prevalent view in all jurisdictions is that privacy rights need to be balanced with the benefits accrued from public research and that in the cases where patient consent for future data reuse cannot be foreseen, the data may be used for research purposes in the public interest. This is the position taken by the General Data Protection Regulation.

The European Medicines Authority has championed a novel approach to publicly releasing data after redacting confidential information, and recent judgements have affirmed such sharing of clinical trial data and summary reports in the public interest.

The centralised data-screening approach proposed by the Australian Government seems to go in the opposite direction, despite the fact that [450] was largely modelled around the European approaches to data protection valid at the time. Centralised approaches to data sharing and vetting of prospective data users will be costly and are unlikely to bring about the desired benefits of increased data availability and reuse. An approach with restricted data sharing, too many review boards, too many arguments to be made for gaining access to data, and too many conditions placed on data reuse cannot lead to increased innovation and data uptake.

In this study, it has been shown that decentralised governance mechanisms have been central to the rise and uptake of open data and its reuse by stakeholders. For example, this has been the prevalent approach shaping European science policy, especially biomedicine and medical research, which have advanced as a result of the concerted efforts of heterogeneous stakeholders directly involved in the research conduct. Experiences with open data from CERN and from the EMA confirm that the benefits of open data can be best harnessed by allowing research and regulatory agencies themselves to set the rules for data sharing.

Furthermore, the European data system has primarily relied on trust among stakeholders and on soft-rule instruments, such as codes of professional conduct and research ethics, rather than on more rigid forms of legislative interventions. These three key elements—decentralisation, trust in data holders, and reliance on soft instruments—have been integrated into the new General Data Protection Regulation, which is arguably the most stringent piece of privacy legislation in the world. And yet, the approach adopted in Europe to data sharing is highly decentralised and open.

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

IntechOpen