Chapter 4

The Unclear Meaning of Open Scientific Data

Vera J. Lipton

This chapter aims to shed light on the meaning of open scientific data—a term problematic to conceptualise in both policy and practice of open data.

The discussion is structured as follows:

1. What is data?
2. What is scientific data?
3. What falls outside the scope of 'research data'?
4. What is missing in the scope of 'research data'?
5. What makes research data 'open'?
6. The limits of openness

Introduction

The previous chapter found that well-intentioned open data policies do not accommodate the diversity of meanings that can be applied to the term 'data' when used across different scientific disciplines and research projects. Few research funders or publishers define data other than by listing examples of what 'data' might be. The previous chapter also found that such non-exhaustive examples lack the detailed guidance researchers need when depositing data.

This chapter attempts to unpack the notion of 'open scientific data', in all its complexity. It starts by considering the notions of 'data' in the context of scientific enquiry. This is followed by the analysis of the definitions of 'research data' in the open data policies in place. The last part of the analysis centres on the requirements for data 'openness' and 'reuse' and how these terms are evolving as they are adopted by various stakeholders, across different scientific disciplines, and in various contexts. Gaps in the current landscape are identified, along with issues not covered in the definitions and issues falling outside the scope of 'research data' and 'openness'.

Three themes emerge in this chapter—firstly, that the current definitions do not adequately describe open scientific data and the difficulties of conceptualisation, secondly, that the policies create confusion among researchers about the requirements for data deposit and adequate description of the data, and thirdly, that researchers themselves need to be better motivated to take a more active role both in describing the data they produce and in reusing data created by others.

4.1 What is data?

The term 'data' is a plural form of the Latin 'datum'. The term has several meanings in the English language. According to English Oxford Living Dictionaries, 'data' can refer to facts or statistics collected together for reference or analysis, or it can refer to the set of principles accepted as the basis of an argument:

> *... historically and in specialised scientific fields, it is also treated as a plural in English, taking a plural verb, as in the data were collected and classified. In modern non-scientific use, however, it is generally not treated as a plural. Instead, it is treated as a mass noun, similar to a word like information, which takes a singular verb. Sentences such as data was collected over a number of years are now widely accepted in standard English.*[1]

'Data' as a collective noun can refer to a set of known facts or things used as a basis for inference or reckoning.[2]

At the same time, some prominent scholars of digital communications have suggested that data indeed 'are' various 'objects' or 'entities'. Therefore, they continue to treat 'data' as plural.[3]

Another consideration is the very notion of 'scientific data'—which is, as this chapter finds, a term not defined and understood consistently among the key stakeholders. It may therefore be appropriate to approach data as an ever-evolving 'concept' and 'evidence underpinning scientific knowledge' rather than as specific 'objects'. This book adopts the latter approach and, therefore, 'data' is used collectively.

4.2 What is scientific data?

Scholars and researchers tend to interpret 'scientific data' in the context of 'research data' collected in the course of scientific experiments. The terms 'research data' and 'scientific data' are often used interchangeably and irrespectively of the subject collecting the data—whether the subject is a researcher or whether the data collection is semi-automated, such as through online questionnaires, or fully automated, such as data harvested by scientific equipment. 'Research data' may therefore take many forms, come in different formats, and come from various sources.[4] As such, the term 'research data' was meant to be broadly inclusive [230]. Perhaps for this reason, major policies and guidelines define 'research data' by examples.

Some important stakeholders, such as the Research Data Australia Registry developed by the Australian National Data Service, accept data records that their research communities consider to be important, rather than according to an external standard for 'research data' [231]. The reason for this position is simple: there is no established meaning of 'research data'. The analysis below illustrates the diversity of definitions of the term as it has been adopted by key stakeholders.

---

[1] Definition of data in the English Oxford Living Dictionaries [227].

[2] Data defined in [228].

[3] For example, Borgman, who consistently uses 'data' to signify plural, has recently defined research data as 'entities used as evidence of phenomena for the purposes of research or scholarship' ([167], p. 29). Borgman cites Rosenberg's historical analysis of the term 'data' who concludes that data remains a rhetorical term, without an essence of its own, neither truth nor reality [229]. However, in the context of scientific communications, and in this book, 'data' is interpreted to be the best possible truth about reality as we know it today.

[4] Borgman [167] at point 3.

Among the earliest and most commonly used definitions is that which appeared in 1999 in the National Academies of Science report:

> *Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors.*[5]

In 2011, the academies clarified that—in addition to all digital representation of literature (whether text, still or moving images, sound, models, games, or simulations)—the term also applies to:

> *... forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines.*[6]

The National Institutes of Health in the United States defines data as:

> *... recorded information, regardless of the form or media on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data [232].*

Such a notion of data spans many fields of science, acknowledging some of the many forms that data can take. In this context, the principles and guidelines developed by the Organisation for Economic Co-operation and Development (OECD) define 'research data' as:

> *... factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated ([68], p. 13).*

The Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, representing three influential research organisations in the United States, defines 'research data' as:

> *... information used in research to generate research conclusions which includes raw data, processed data, published data and archived data and exist in the form of textual, numeric, equation, statistics, images (whether fixed or moving), diagrams or audio recordings.*[7]

From the perspective of researchers, clarity around open scientific data is central to both the conduct of research and preservation of its outputs. As a general guide, the Digital Curation Centre in the United Kingdom recommends that researchers should consider how they will maintain access to any research data that may be necessary for enabling the validation of their research findings [234]. One leading research institution, the University of Glasgow, states:

---

[5] National Research Council (1999), 15.

[6] Uhlir and Cohen (2011) at point 4.

[7] Committee on Ensuring the Utility and Integrity of Research Data in Digital Age ([233], p. 22).

> *... research data should be interpreted as any material (digital or physical) required to underpin research. For different disciplines this may include raw data captured from instruments, derived data, documents, spreadsheets & databases, lab notebooks, visualisations, models, software, images, measurements and numbers [235].*

The Australian Code for the Responsible Conduct of Research provides researchers with the following guidance:

> *... while it may not be practical to keep all the primary material (such as ore, biological material, questionnaires or recordings), durable records derived from them (such as assays, test results, transcripts, and laboratory and field notes) must be retained and accessible. The researcher must decide which data and materials should be retained, although in some cases this is determined by law, funding agency, publisher or by convention in the discipline. The central aim is that sufficient materials and data are retained to justify the outcomes of the research and to defend them if they are challenged. The potential value of the material for further research should also be considered, particularly where the research would be difficult or impossible to repeat [236].*

In line with those guidelines, the Queensland University of Technology defines research data as:

> *... data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media [237].*

In its policy on the management of research data and records, the University of Melbourne identifies 'research data' as:

> *... facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational. Data includes: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples. Provenance information about the data might also be included: the how, when, where it was collected and with what (for example, instrument). The software code used to generate, annotate or analyse the data may also be included.*
>
> *The University of Melbourne makes no functional distinction between physical research products, digital research data and records of research, which can include items such as correspondence, application documents, reports and consent forms [238].*

The Monash University Research Data Policy has a similarly encompassing definition:

> *Research data: the data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed data [239].*

In general, all outputs that are accepted in the scientific community as necessary to validate research findings are included among research outputs. However,

there is no shared understanding of the term 'research output', and stakeholders interpret the term differently. The key point of differentiation appears to be 'interest to the research community'. In broader terms, whatever 'data' is of interest to researchers should be treated as 'research data'.

For example, laboratory notebooks are often considered 'research data', recognising that they are necessary for reproducing research findings, especially in clinical trials. Even so, some funders exclude laboratory notebooks. The OECD stated that access to laboratory notebooks is subject to considerations that differ from those that deal with open data.[8] These include the commercial objectives, as records in laboratory notebooks are often used to establish the novelty principle of an invention, especially in the United States. Another reason why laboratory notebooks are not treated as 'research data' can be format limitation. Research notebooks still come in paper copies rather than in digital formats.

At the same time, some funding policies, such as that of the Engineering and Physical Sciences Research Council in the United Kingdom, require that:

> *Publicly-funded research data that is not generated in digital format will be stored in a manner to facilitate it being shared in the event of a valid request for access to the data being received (this expectation could be satisfied by implementing a policy to convert and store such data in digital format in a timely manner) [240].*

Indeed, 'research data' can take any format, even though policies mandating open access to research data primarily focus on research data in a digital, computer-readable format. For example, the Horizon 2020 Open Data Pilot is limited to 'digital research data', defined as:

> *'Digital research data' is information in digital form (in particular facts or numbers), collected to be examined and used as a basis for reasoning, discussion or calculation; this includes statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images [241].*

The digital format has the greatest potential to improve the efficiency of data distribution and its research application, mainly because the cost of transmission through the Internet is negligible. However, open access policies often apply to data that comes in non-digital formats, as documented above, and to data transmitted by means other than the Internet. For example, the OECD Principles could also apply to analogue research data in such instances where the cost of providing access to that data can be held reasonably low.[9]

Another contested area is the division between 'research data' and 'primary materials'. The Australian Code for the Responsible Conduct of Research regards completed questionnaires and recordings as 'primary materials', while transcripts derived from them are 'research data'. Despite this, some researchers have argued that the completed questionnaires and recordings should be treated as research data in terms of the agreed definition [231]. The reasoning used was the questionnaires and recordings qualify as 'factual records, used as primary sources for research'.[10] Consequently, if the research community considers those records as essential for

[8] OECD Principles at Point 10, 14.

[9] OECD Principles at Point 10, 13–14.

[10] *Ibid.*

substantiating research findings, then they also qualify as 'research data' and should be retained for the recommended period.[11]

Data sources also vary widely, as Borgman [167] observed. In the physical and life sciences, researchers gather or produce most data—through observations, experiments, or models. Researchers in the social sciences may gather or produce original data or they may source it from such places such as public records of economic activity. While the concept of 'data' is least well-developed in the humanities, the growth in digital research is leading to the more common usage of the term. Typically, humanities data is taken from cultural records—archives, documents, and artefacts.[12]

4.3 What falls outside the scope of 'research data'?

Some policies define 'research data' by limiting the entities that cannot be treated as research data. For example, in the United States, the Office of Science and Technology Policy states:

> *... [data] does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens [242].*

Similarly, the OECD Principles explicitly define what falls outside the scope of research data:

> *... laboratory notebooks, preliminary analyses and drafts of scientific papers, plans for future research, peer review or personal communications with colleagues, or physical objects (e.g. laboratory samples, strains of bacteria, test animals such as mice.*[13]

However, some researchers might argue that laboratory notebooks or preliminary drafts fall under the scope of data because of the importance for their research. The above definitions prove that the notion of 'research data' depends on the context. Those definitions help to explain why 'research data' often depends on interpretation. As Borgman put it, one researcher's signal—or data—may be someone else's noise [243].

4.4 What is missing in the scope of 'research data'?

Only one of the above definitions, namely, the definition of research data developed by the University of Melbourne, explicitly includes metadata as a component of research data. Metadata is structured information associated with an object for purposes of discovery, description, use, management, and preservation.[14] In short, metadata is data about data—information about information. In the context of scientific data, metadata is even more important because it provides the context needed to make sense of what would otherwise be a collection of

---

[11] *Ibid.*

[12] Borgman [167] at point 3.

[13] OECD Principles at point 10, 14.

[14] University of Melbourne at point 16.

numbers. Without metadata, any data is unlikely to be reusable and almost certainly would not allow for the research to be reproducible. For this reason, all stakeholders involved in open research data need to explicitly acknowledge that 'research data' includes metadata.

In addition to core research data and metadata, the final component required for reproducibility by research funders is the code or algorithms used to undertake the analyses [244]. In a substantial number of cases, the interpretation and analysis of data are dependent on the availability of software. In some cases, sharing of software code may not be permitted, as the code may be a commercial application or it can be protected by intellectual property such as patents. Yet in all instances where the code *can* be shared, it *should* be shared, and research funders need to include statements in their policies to that effect. The code and the relevant algorithms need to become integral to the term 'research data'.

Furthermore, most definitions of 'research data' do not address the degree of processing of the data that is shared. Research data can refer to 'top-level data' typically underpinning scientific publications or to 'working versions' incorporating different types of analyses, cleaning, and processing steps or to 'raw data' collected in research or harvested by scientific equipment. The various levels of data processing and control are discussed in the context of research data management, in Chapters 5 and 6.

Researchers like to organise their data in a 'dataset', but this is another term subject to dispute. A dataset might consist of a large or small spreadsheet, a text file, a set of files, or all of these. At present, researchers can share open data anything they like, and there are no criteria for assessing the quality of the data being shared other than by checking the parameters for data identification and discoverability. The potential of the data to be reused thus cannot be easily established. Some of the data released under the open access mandate may not be reusable.

4.5 What makes research data 'open'?

A major goal of open scientific data is to increase the sharing of the data by making it available to anyone who seeks it, regardless of location or affiliation, with minimal barriers for access and reuse.

In general, 'data sharing' implies the release of data in a form that others can use [245]. Data sharing thus encompasses many means of releasing data, open data being one such means. Other forms of release include the private exchange between researchers, publication of datasets on websites, the deposit in archives, repositories, domain-specific collections, library collections, or data papers and, finally, attachment as supplementary material in journal articles [246]. The data shared by any of these means would meet the criteria of openness provided that the data:

1. Is freely available on the Internet

2. Permits any user to download, copy, analyse, reprocess, pass to software, or use for any other purposes

3. Is without financial, legal, or technical barriers other than those inseparable from gaining access to the Internet itself [247]

Examples of open data include repositories and archives (including Zenodo, GenBank, Figshare, Dryad), data networks (such as Global Biodiversity Information Facility), virtual observatories (such as Digital Earth), domain repositories

(such as PubMed Central), and institutional repositories. The list of major data network is provided at Appendix B.

The Open Knowledge Foundation has developed an extensive definition of 'open works', which also applies to 'open data'. For a work to be open, it must have an open licence, be accessible at a fair reproduction cost, or be freely available on the Internet along with the necessary information on compliance with the work's licence [248].

The aim of these licences is to allow free reuse and redistribution of all, or parts of, the work. The licence must also allow for derivatives of the work to be made and to be subsequently distributed or compiled with any other works. Also, the licence must allow the use, redistribution, modification, and compilation for any purpose. The rights attached to the work must apply to anyone who receives it redistributed, without the need to agree to any additional terms. There may be some clauses that ask for attribution for those who produced the work. There is often a share-alike clause that requires copies or derivatives of a licenced work to remain under a licence that is the same as, or similar to, the original. In general terms, this approach requires a licence to avoid discrimination against any person or group and must ensure that the works are free, so that there are no royalty charges or fee arrangements of any sort.[15]

Nevertheless, levels and standards for openness vary among different repositories and datasets.

For example, some open data repositories permit contributors to maintain their copyright and control over deposited data, which poses challenges to reuse. Furthermore, over half of seemingly 'open' datasets do not include any express licence,[16] which also limits the potential for data reuse. In some instances, data is open but cannot be reused without proprietary software—which, again, limits the potential for reuse and, in that circumstance, the dataset may fall under the protection of copyright law.

Conversely, data generated by open-source software may not be available for reuse in its modified form, if this involves a 'modicum' of creativity and, as discussed in Chapter 7, thus becomes a form of intellectual property. Openness may be tied to funding streams and business models (e.g. charging for value-added data services), as the OECD recently noted ([249], p. 27).

Clearly, there is a discrepancy between 'ideal openness', espoused in policies and in an array of criteria for making data 'open', and 'actual openness', which may only be 'semiopen' or otherwise flawed and may not allow for unfettered data reuse. Or the reuse can still be possible but with questionable legality under copyright law. In some instances, 'open data' is even interpreted as controlled access or restricted access to full datasets.[17]

The OECD specified 13 conditions for open data, yet in any particular situation, only a few are likely to be satisfied.[18]

The *Science as an Open Enterprise* report defines 'intelligent openness' as:

a. **Accessible:** Data must be located in such a manner that it can readily be found. This has implications both for the custodianship of data and the processes by which access is granted to data and information.

---

[15] *Ibid.*

[16] Initial results from the Global Open Data Index 2016/17 show roughly that only 38% of the eligible datasets were openly licenced. Available at: https://index.okfn.org/ [Accessed: 10 June 2018].

[17] *Ibid.*

[18] OECD Principles [51] at point 10. See also Borgman [167] at point 3.

b. **Intelligible:** Data must provide an account of the results of scientific work that is intelligible to those wishing to understand or scrutinise it. Data communication must therefore be differentiated for different audiences. What is intelligible to a specialist in one field may not be intelligible to one in another field. Effective communication to the wider public is more difficult, necessitating a deeper understanding of what the audience needs in order to understand the data and dialogue about priorities for such communication.

c. **Assessable:** Recipients need to be able to make some judgement or assessment of what is communicated. They will, for example, need to judge the nature of the claims that are made. Are the claims speculations or evidence-based? They should be able to judge the competence and reliability of those making the claims. Assessability also includes the disclosure of attendant factors that might influence trust in the research.

d. **Usable:** Data should be able to be reused, often for different purposes. The usability of data will also depend on the suitability of background material and metadata for those who wish to use the data. They should, at a minimum, be reusable by other scientists ([250], pp. 14–15).

This articulation of the parameters of openness became seminal and was adopted by key stakeholders including the European Commission, which incorporated it and slightly expanded on it the *2014 Guidelines on Data Management in Horizon 2020.* The definition also highlighted the fundamental problem in understanding the differences between data 'use' and 'reuse'.

Pasquetto et al. have clarified the difference.

In the first instance, data is collected by a researcher or a team of researchers, and the first (data) 'use' is by that individual or research team. If the data originator (s) use(s) the same dataset for any later purpose, relating to the original project or not, that too would count as a 'use'. If the data is shared, including as open data, that would be considered a 'reuse'. In other words, 'reuse' implies a subsequent use of the data by someone other than the originator(s).[19]

In practice, it can be difficult to monitor data reuse, mainly because researchers rarely cite the repository [251]. At the same time, data originators themselves inconsistently cite data they deposit for reuse. Encouraging consistent data citation practices might increase dissemination,[20] yet the factors that motivate researchers to reuse data deposited by others are not well understood.

With advances in technology providing better instrumentation and techniques for gathering data, the quantity of data available for reuse is increasing. At the same time, the reuse and sharing of data are becoming prominent in disciplines where these practices were once uncommon [252]. Data reuse is common in geospatial sciences, astronomy, clinical research, social media, and genomic research, among other areas.

Many obstacles to data reuse remain. They arise largely due the fact that 'releasing data and making it usable are quite different matters'.[21] Successful data reuse necessitates detailed documentation of the original data collection and processing steps in the language and in the context that would enable interpretation of the data by any subsequent user. Such data is often referred to as metadata.

---

[19] Pasquetto et al. [245] at point 31, 3.

[20] *Ibid.*

[21] Borgman [167], p. 40.

Yet the requirements for data appear to be far broader than is currently captured by the meaning of metadata. For example, detailed description of the unique methods by which data was collected, processed, cleaned, analysed, grouped, and interpreted in statistically correct ways may all be necessary to enable reuse. Information about the software used, including the software version, may also be required, especially in cases where reproducibility is the desired objective. The software used by the data originator may not be available freely or may require upgrading, which can decrease the possibilities for data reuse.

Similarly, the potential for reuse is decreased unless data is documented in the course of the original research project by those with the expertise of data collection and analysis to describe it [167]. Often, however, researchers are preoccupied with writing publications. They are not rewarded for documenting data. Therefore, questions of responsibilities for data documentation and curation lie at the core of our ability to reuse data.[22]

Many stakeholders in research and academia are exploring the options for overcoming the challenges for data reuse, especially those challenges that can be solved with technology.

In 2014, the deliberations of a workshop in Leiden on fair and safe data stewardship and sharing saw the emergence of the notion that by defining and reaching general agreement on certain principles and practices, then all interested parties would find it easier to access and reuse the data that contemporary science generates [253].

From that meeting came a draft set of principles—that all research objects should be findable, accessible, interoperable, and reusable (FAIR) by both people and machines. Subsequently elaborated, these are the FAIR Guiding Principles, summarised in **Table 1** [31].

The FAIR principles apply to data repositories and incorporate the total 'research object'—code, data, and tools for interpretation [254]. They are the most advanced technical standards for open scientific data to date. In the context of this study, the ideal 'open scientific data' is in repositories or archives that apply the FAIR standards, recognising that some data already in the public domain does not meet the standards. Yet every lesson learnt from imperfect open data brings us one step closer to making open scientific data a reality.

4.6 The limits of openness

While open scientific data is desirable and should be pursued to the maximum extent possible, there are some restrictions. The European Commission Horizon 2020 Model Grant Agreements [255] comprehensively state the legal limitations on the 'openness' of data. The European Union (EU) is a significant funder, distributing over €7 billion for research annually. However, this funder limits, quite substantially, the possibilities for sharing the research data resulting from its projects:

> **Article 27—Protection of results—Visibility of EU Funding**
> **27.1 Obligation to protect the results**
>
> Each beneficiary must examine the possibility of protecting its results and must adequately protect them[23]—for an appropriate period and with appropriate territorial coverage—if:

[22] *Ibid.*

[23] Protection may be sought through patent, trademark, industrial design, trade secret or confidentiality.

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Table 1.
The FAIR principles for open scientific data.

a. the results can reasonably be expected to be commercially or industrially exploited and

b. protecting them is possible, reasonable and justified (given the circumstances).

When deciding on protection, the beneficiary must consider its own legitimate interests and the legitimate interests (especially commercial) of the other beneficiaries.[24]

Article 36—Confidentiality
36.1 General obligation to maintain confidentiality

During implementation of the action and for 4 years after the period set out in Article 3, the parties must keep confidential any data, documents or other material (in any form) that is identified as confidential at the time it is disclosed ('confidential information').

If a beneficiary requests, the [Commission][Agency] may agree to keep such information confidential for an additional period beyond the initial 4 years.

[24] Ref. [241] at point 51.

If information has been identified as confidential only orally, it will be considered to be confidential only if this is confirmed in writing within 15 days of the oral disclosure.

Unless otherwise agreed between the parties, they may use confidential information only to implement the Agreement.

The beneficiaries may disclose confidential information to their personnel or third parties involved in the action only if they:

a. need to know to implement the agreement and

b. are bound by an obligation of confidentiality.

This does not change the security obligations in Article 37, which still apply.[25]

There are, however, some exceptions to the obligation of confidentiality. The conditions set out in Article 4 of the Rules for Participation Regulation [193] require the Commission to make available information on the results to other European Union institutions, bodies, offices, or agencies and to Member States or associated countries.

Since the Commission is also committed to developing an Open Science Cloud to support open science and innovation [144], perhaps the results might be available to selected European Union users as open data via this means.

Furthermore, the obligations for confidentiality do not apply if:

a. the disclosing party agrees to release the other party;

b. the information was already known by the recipient or is given to him without obligation of confidentiality by a third party that was not bound by any obligation of confidentiality;

c. the recipient proves that the information was developed without the use of confidential information;

d. the information becomes generally and publicly available, without breaching any confidentiality obligation; or

e. the disclosure of the information is required by EU or national law.[26]

Restrictions on open sharing of data proposed by the European Commission are for the protection of security in relation to disclosure and subcontracting.

Article 37—Security related obligations

The beneficiaries [of grants] must comply with the 'security recommendation (s)' set out [by the Commission].

For security recommendations restricting disclosure or dissemination, the beneficiaries must—before disclosure or dissemination to a third party (including linked third parties, such as affiliated entities)—inform the coordinator, which must request written approval from the [Commission][Agency].

---

[25] *Ibid*, 264.

[26] Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 of December 2013, 81.

Finally, personal data cannot be shared, and this imposes a limit on openness. The restrictions for the processing of personal data as set out in Article 39 of the model grant agreement are canvassed in more detail in Chapters 6 and 7.

The key restrictions to data sharing cited above—the obligation to commercialise and protect the research results with intellectual property and the obligation to maintain confidentiality, the security obligations, and the handling of personal data —are significant impediments to open scientific data. The scope of these restrictions is not yet clearly defined and requires further conceptualisation.

Conclusion

This chapter showed that the terms 'data', 'research data' and 'open data' may hold different meanings for different stakeholders and across different research disciplines, different levels of processing, different data repositories and even in the eyes of individual researchers working on the same project. The only agreement emerging on these definitions is that no single definition will suffice.

Alternatively, no single definition is necessary because the meaning of 'open scientific data' depends both on the context for the use of that data and on the subject using it.

Key stakeholders—research funders, researchers, librarians, and lawyers—may approach the term differently. Funders typically mention 'research data' that underpins research outcomes; researchers talk about databases and spreadsheets, while librarians tend to be preoccupied with metadata and citations. This creates confusion. If researchers are to comply with the policies of funders and publishers, they need to understand what 'data' they should make available. Similarly, if librarians are to provide effective data management services, they need to be certain about what 'data' should be considered and what would make the data 'findable', 'accessible', 'interoperable' and 'reusable' for others.

The diversity of the definitions of 'data' makes any attempts to specify the meaning of 'open scientific data' extremely difficult. Yet this effort is necessary to identify how to best document and curate scientific data to facilitate reuse. The key argument emerging in this chapter is that that even though defining open scientific data is a challenging task, more research effort and resources should be dedicated to this area. Only an improved understanding of the parameters that can make data findable, useful, and reusable can assist in realising the benefits of open scientific data. This chapter finds that the recent FAIR standards are a very helpful contribution to the conceptualisation debate.

Another key point highlighted in this chapter is the necessity of data 'reuse' to realise the benefits of open research. For data to be reusable, it needs to be meticulously documented. In this sense, data documentation is a broader concept than metadata and requires a detailed description of the unique methods by which data was collected, processed, cleaned, analysed, grouped, and interpreted in statistically correct ways. Researchers and data scientists need to tackle this challenge in their research practice, as they are developing improved ways to describe data and are becoming more skilled in reusing data created by others.

The initial focus of the open scientific data movement was on ensuring the release of 'data' into the public domain. Now it is necessary to provide further guidance to research organisations with regard to possible methods of reuse of open research data.

At present, there appears to be a high degree of discrepancy between 'ideal openness' (as espoused in policies and in an array of criteria for making data open) and 'actual openness'. The data available in the public domain may only be

'semiopen' or flawed in some respects and may not allow for unfettered data reuse. Or the reuse in practical terms can still be possible, but the legality of such reuse may be questionable, as further discussed in Chapter 7. Only the practice of open data can help narrow the distance between espoused openness and the way open data is practiced at present. The early experiences with the implementation of open data at CERN and in clinical trials data are discussed in the following two chapters.

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