

Licensing FAIR Data for Reuse postprint

Authors: Labastida, Ignasi, Margoni, Thomas, Labastida, Ignasi

Date: 2022-11-16T00:00:00+00:00

Abstract

The last letter of the FAIR acronym stands for Reusability. Data and metadata should be made available with a clear and accessible usage license. But, what are the choices? How can researchers share data and allow reusability? Are all the licenses available for sharing content suitable for data? Data can be covered by different layers of copyright protection making the relationship between data and copyright particularly complex. Some research data can be considered as a work and therefore covered by full copyright while other data can be in the public domain due to their lack of originality. Moreover, a collection of data can be protected by special rights in Europe to acknowledge the investment in time and money in obtaining, presenting, arranging or verifying the data. The need of using a license when sharing data comes from the fact that, under current copyright laws, when rights exist, the absence of any legal notice must be understood as the default “all rights reserved” regime. Unless an exception applies, the authorisation of right holders is necessary for reuse. Right holders could use any text to state the reusability of data but it is advisable to use some of the existing licenses, and especially the ones that are suitable for data and databases. We hope that with this paper we can bring some clarity in relation to the rights involved when sharing research data

Full Text

Preamble

Addressing FAIR principles

Licensing FAIR Data for Reuse

Ignasi Labastida^{1†} & Thomas Margoni²

¹Learning and Research Resources Centre (CRAI), Universitat de Barcelona, 08007 Barcelona, Spain

²School of Law and CREATE, University of Glasgow, Stair Building 5-9 The Square, Glasgow G12 8QQ, UK

Keywords: Research data; Open content licensing; Copyright; Sui generis database rights; FAIR data; Legal aspects of FAIR (especially for R)

Citation: I. Labastida & T. Margoni. Licensing FAIR data for reuse. *Data Intelligence* 2(2020), 199–207. doi: 10.1162/dint_a_{00042}

Abstract

The last letter of the FAIR acronym stands for Reusability. Data and metadata should be made available with a clear and accessible usage license. But what are the choices? How can researchers share data and allow reusability? Are all the licenses available for sharing content suitable for data? Data can be covered by different layers of copyright protection, making the relationship between data and copyright particularly complex. Some research data can be considered as a work and therefore covered by full copyright, while other data can be in the public domain due to their lack of originality. Moreover, a collection of data can be protected by special rights in Europe to acknowledge the investment in time and money in obtaining, presenting, arranging, or verifying the data. The need for using a license when sharing data comes from the fact that, under current copyright laws, when rights exist, the absence of any legal notice must be understood as the default “all rights reserved” regime. Unless an exception applies, the authorization of right holders is necessary for reuse. Right holders could use any text to state the reusability of data, but it is advisable to use some of the existing licenses, and especially the ones that are suitable for data and databases. We hope that with this paper we can bring some clarity in relation to the rights involved when sharing research data.

† Corresponding author: Ignasi Labastida (E-mail: ilabastida@ub.edu, ORCID: 0000-0001-7030-7030).

1. The Protection of Data, Data Sets, and Databases

European Union (EU) law defines “databases,” but not data sets or, at least for copyright purposes, data. Databases that meet the legal definition can be protected by copyright if they are original. Data sets, if they correspond to the definition of database, are protected by copyright; otherwise, they are not. Data as such are normally excluded from copyright protection [2,3]. It is important to understand that copyright protects original expressions in the “literary and artistic” domain—an expression that has historically included works such as books, musical works, choreographies, cinematographic works, drawings, etc. [4]. Ideas, procedures, methods of operation, or mathematical concepts as such, news of the day, and miscellaneous facts are excluded from copyright protection [4,5,6].

Two main elements are important in this analysis. Copyright only protects original expressions, and these expressions are normally found in the literary and artistic domain. However, the literary and artistic domain should not be

interpreted narrowly, and in recent years new creations have been included, such as computer programs and compilations of data or databases [6]. The latter is particularly important here.

International conventions are clear that copyright can protect compilations of data or other material which, by reason of the selection or arrangement of their contents, constitute intellectual creations. But copyright does not extend to the data contained in the database, which may or may not be protected depending on whether they meet the conditions identified above: to be original expressions in their own right.

At this point, it is important to understand that what the law calls data—but does not define—may in fact be quite different from what other disciplines understand by the same term. Databases are defined as collections of independent works, data, or other materials arranged in a systematic or methodical way [1]. This definition means that a protected database can be a systematic or methodical collection of works (e.g., a database of journal articles, movies, or songs), other materials (e.g., sound recordings or broadcasts), and data (which are not defined but certainly include elements such as factual information, measurements, or other non-original information). This situation is fairly consistent at the international level.

It is important to note that in the EU, since the Database Directive of 1996, when data, including non-copyrightable data, are gathered in a non-original database, the maker can claim some rights preventing the extraction and reuse of substantial parts of otherwise unprotected data. This is the so-called *Sui Generis* Database Right (SGDR), not really copyright but similar in some aspects [1].

Accordingly, data understood as factual information—for instance, historical facts or weather measurements—are not protected by copyright or the SGDR as such. When these data are collected and organized in a systematic or methodical way, they will form part of a database. If the selection or arrangement of the data is original in the sense of the author's own intellectual creation, then such selection and arrangement (i.e., the database structure) but not the collected data are protected by copyright [1]. Does this mean that factual data, even when collected following an original selection and arrangement, are not protected by copyright? Yes, unless the single datum itself is protected by copyright because, for copyright purposes, it is not a datum but a work (remember the example of the database of journal articles). Does that mean that there is no form of protection similar to copyright for factual data? No, there is. It is the aforementioned SGDR, which exists when a substantial investment in obtaining, verifying, or presenting the data has taken place. When this happens, the maker of this investment has a right to prevent the extraction (i.e., copy) and reuse of a substantial part of the data—not of the single datum but of a larger amount that will have to be identified on a case-by-case basis.

The Database Directive allows member states to implement some listed limitations to the rights granted to the maker [6], but this has not led to proper

harmonization at the national level [7].

The last aspect to be considered here is that not all types of investment qualify for SGDR. Only investments (time, money, etc.) in the obtaining, verification, or presentation—but not in the creation—of data qualify. However counter-intuitive this may sound, data that are created are not protected by SGDR; only data that are obtained are protected. After all, this is the Database Directive and not the data directive. The initial goal was to incentivize the production of databases, not of data [8]. If there were a right limiting the reuse of data, this would halt, instead of incentivizing, the production of databases.

Even more important, however, is to understand why factual information is excluded from copyright protection. If facts were protected by copyright, it would mean that no one without the authorization of the right holder could reuse that data. It would mean that no one other than the first person or company recording it could use the same measurements of a natural phenomenon such as the temperature of the oceans. No one could reuse factual data such as the performance of the economy or the geospatial coordinates needed to identify a specific point on the Earth. Such a scenario should be seen with suspicion first and foremost by the scientific community as it would undermine scientific freedoms, transparency, and replicability. But it would equally threaten other fundamental rights enshrined in the European Convention on Human Rights and in the EU Charter, such as freedom of information, private property, and freedom to conduct a business [2]. This is why factual information is not protected as such. By affording protection to the obtaining of data, a limited reward for the investment is given to the maker. By excluding protection for the creation of data, the law's goal is to avoid as much as possible so-called "single source databases" due to their anti-competitive and monopolistic nature and to the distortion of scientific freedoms and fundamental rights that they may cause.

Therefore, the legal system has designed a mechanism whereby the basic bricks of knowledge—such as ideas or factual information—are freely available to all in order to learn and advance knowledge in a field. But this is balanced by the possibility to protect some of the results obtained using those ideas and facts: original expressions of unprotected ideas and substantial amounts of obtained data when systematically or methodically collected and arranged.

In the next section, we will discuss the existing options to license data and databases, focusing on reusability. However, there is another aspect of the FAIR principles that must be taken into account before deciding which license to use. This aspect is accessibility. Not all the data can be shared openly, but metadata can be accessible when data must be kept private for security, privacy, or other justified reasons [9]. Metadata are very often a type of factual information and therefore do not qualify as a work covered by copyright. Their compilation, nonetheless, can be protected by existing database rights, as will be seen in one of the cases in Section 3.

2. Suitable Options for Licensing Data and Database Rights

When looking for existing contractual solutions for sharing rights, one can think of the most well-known set of licenses for open content: The Creative Commons (CC) licenses [10]. Initially, these licenses were not drafted to specifically cover data, which normally are not protected by copyright. This is why, for example, earlier versions of the licenses did not cover the SGDR. The current version 4.0 includes the SGDR in its scope (which means that it will follow the license conditions, e.g., BY, SA, etc.). Creative Commons has also drafted a specific legal tool called CC0 aimed to enable scientists, educators, artists, and other creators and owners of copyright or database-protected content to waive those interests, including SGDR.

Besides the tools provided by CC, there are other legal instruments created especially for data, like the ones developed by the Open Data Commons, or by some national governments.

2.1 The Options Provided by Creative Commons

When Creative Commons launched the first suite of licenses almost two decades ago, the licenses were drafted with the US Copyright Act as the reference model, and national versions of the licenses were developed to better address local legal issues. Quickly, with rising international interest, it was seen that the licenses would need to change in order to cover different features of other copyright frameworks in a more systematic way. Among other issues, Creative Commons had to face the inclusion of the SGDR into the scope of the license. This topic was not approached uniformly in the porting process by all EU-affiliated institutions at the time. Some of them did not mention it, others simply waived the SGDR, and others included it in the license grant.

In 2013, when Creative Commons launched the last version of its license suite, it was decided to end the porting of the licenses. Instead, the decision was to have a single legal text that could fit all legal specificities and could be suitable for any jurisdiction. Those updated legal texts could be translated into any language, but there was no need to port them.

The current legal texts of the Creative Commons licenses include the SGDR in the scope of the license, meaning that the SGDR is treated exactly as any other licensed right. This means that the conditions of the license (e.g., Attribution-BY, Share Alike-SA, Non-Derivative Works-ND, Non-Commercial-NC) will also apply to the SGDR. Therefore, when the maker of an SGDR-protected database uses a CC license, they grant to the public the right to extract and reuse the whole or a substantial part of its contents. Those rights can be restricted to non-commercial uses if the database maker chooses a Non-Commercial license such as CC BY-NC, CC BY-NC-SA, or CC BY-NC-ND. Moreover, the creation of a new database using the whole or a substantial part of the contents of the licensed database can be restricted by using a Non-Derivative license such as CC

BY-ND or CC BY-NC-ND, or it can be conditioned by a copyleft practice by means of CC BY-SA or CC BY-NC-SA that carry the Share Alike clause. This latter condition requires that any derivative database must be licensed under the same license or a compatible one. A derivative database must be understood as a new database that includes the whole or a substantial part of the contents of the original licensed database.

CC0 is a legal tool that was introduced by Creative Commons in 2008 as a demand from researchers to have a license to waive any database right and especially the SGDR [11]. Its creation followed the Protocol for Implementing Open Access Data launched by Science Commons at the end of 2007. Science Commons was a Creative Commons project created in 2005 and led by John Wilbanks to develop strategies and tools for opening research.

Sometimes CC0 has been viewed with some concerns among European scholars because it is understood only as a waiver of all copyright and related rights. This full waiving of rights is not easy to fit into many continental copyright laws. However, CC0 is more than a waiver; it has three different layers of action. First, the right holder waives any copyright and related rights that can be waived in accordance with applicable law. Secondly, if there are rights that the right holder cannot waive under applicable law, they are licensed in a way that mirrors as closely as possible the legal effect of a waiver. And finally, if there are any rights that the right holders cannot waive or license, they affirm that they will not exercise them and will not assert any claim with respect to the use of the work, once again within the limits of applicable law. Therefore, in the case of moral rights, in countries where they do not exist (mainly the US, with the exception of the categories covered by the Visual Arts Rights Act [12]), CC0 operates smoothly. In countries where moral rights exist but where they can be waived or not asserted, they are waived if asserted (e.g., the UK). In countries where they cannot be waived, they will remain in full effect in accordance with applicable law (think of France, Spain, or Italy where moral rights cannot be waived).

2.2 Other Legal Tools

In 2007, Jordan Hatcher and Charlotte Waelde created the project Open Data Commons to provide legal solutions to the issue of data licensing. Later, the project was transferred to the Open Knowledge Foundation. Open Data Commons launched the first-ever open data license: the Public Domain Dedication and License (PDDL) in 2008, and afterward they provided two other legal tools: the Attribution License (ODC-By) and the Open Database License (ODC-ODbL).

The PDDL is a tool aimed at placing databases in the public domain by waiving all rights. Like CC0, it has several layers of action: a dedication to the public domain, a waiver of any copyright rights over the work, a license of any non-waivable right, and an assertion of not claiming any remaining rights for the

use of the work to the extent allowed by applicable law. ODC-By and ODC-ODbL are licenses only designed for databases, meaning that only copyright in the database (the structure of the database) and the SGDR are licensed, whereas the content of the database (in the example of a database of journals, the copyright in the single journal articles) is not. A specific second license for the database content should be applied if needed.

By means of ODC-By or ODC-ODbL, the rights holder of a database grants the extraction and reuse of the whole or a substantial part of the contents of the licensed database, along with the authorization of the creation of derivative and collective databases and the exploitation of the original licensed database. Those rights are granted subject to the proper attribution of the rights holder and the original source of the licensed database and, in the case of an ODC-ODbL licensed database, the use of the same license or a compatible license in any derivative database created. This latter condition is equivalent to the Share Alike requirement in Creative Commons licenses, and they are both an application of the copyleft practice.

There are also some legal instruments created by national governments under their open data projects aimed at sharing governmental data without any restriction. Among them, we can mention the French Open License [13] and the British Open Government Licence [14]. Both licenses grant licensees all rights to exploit works subject to an acknowledgment of the authorship of the content licensed. They both mention compatibility with licenses such as CC BY and ODC-By. It should be noted that these licenses are intended for use only by the public bodies that have developed them, so whereas researchers should feel absolutely free to reuse content under, and in compliance with, those licenses, they should not choose those licenses to license their own contributions.

3. Some Cases of Licensing Data Rights

In 2011, Europeana, the European digital platform for cultural heritage, decided to adopt a new data exchange agreement requiring its data providers to release all their metadata under CC0. Metadata sent to Europeana was seen as not copyrightable individually, but the complete set of metadata could fall into the scope of the SGDR. In order to avoid managing all the required permissions to extract the metadata for its inclusion in the portal or in any project, Europeana adopted the CC0 solution. This approach has been used by other cultural and academic institutions like libraries. National libraries have shared catalogue records, bibliographic files, or even metadata from repositories under CC0. Europeana has employed an interesting approach to request attribution: they ask you kindly—or better, as an accepted community norm in the scientific field. There is no legal requirement to do that, just an ethical or community one [15].

OpenStreetMap (OSM) is a collaborative project to create a free editable map of the world that started as a consequence of the restrictions on using some of

the public maps available. All the OSM data is licensed under the ODbL since 2012 when it changed from its initial CC BY-SA license. The change was made because at the time CC licenses did not include SGDR (but the current version 4.0 does). The use of this license allows broad reuse of the data, and there are many adopters of OSM data, with Apple being probably the most well-known among them.

In 2011, an English chemist working in Australia, Matt Todd, launched a collaborative research project to share new discoveries on fighting malaria. The project received the name of Open Source Malaria and provided a platform to share research data. Data can be reused by anyone under the terms of the CC BY license and, in some cases, under CC0. In 2018, Matt Todd launched a similar project now focused on curing eumycetoma: Open Source Mycetoma.

4. Conclusion

Researchers have a broad set of options when sharing data. However, it should be clarified that in many instances there will be no copyright or related rights on data. In these cases, a CC0 or a PDDL, eventually with a request to give credit, could be the best option between unrestricted access and the promotion of a fair community practice that acknowledges the provenance of data because it is ethically—not legally—required. If particular needs are present, CC BY 4.0 or ODC-By and ODC-ODbL will also work well, but researchers opting for these solutions should carefully assess why they do that. In the generality of cases, Open Science is easier to achieve if fewer restrictions are present for the reuse of data. This does not mean that you should not ask for attribution for your data. It means carefully weighing the pros and cons of requiring attribution. This will allow you to make the best choice in most cases. Finally, clauses such as non-commercial or non-derivatives should be avoided as they are not Open Access compliant and severely restrict the reuse of knowledge.

These may be difficult choices for researchers, and a number of resources have been made available for guidance and support [16].

Author Contributions: Both I.L. (ilabastida@ub.edu) and T.M. (Thomas.Margoni@glasgow.ac.uk) contributed significantly and equally to this article.

Acknowledgements: Thomas Margoni co-coordinates the legal task force of OpenAIRE Advance, a project funded under the H2020 programme of the European Commission, project n°: 777541. We would like to thank the support of the project.

References

- [1] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [Database Directive]. Available at: <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML>.

- [2] B. Hugenholtz. Against “data property” . In: H. Ulrich, P. Drahos & G. Ghidini (eds.) *Kritika, Essays on Intellectual Property*, Vol. 3. Cheltenham, UK: Edward Elgar Publishing, 2018, pp. 48-71.
- [3] M. W. Carroll. Sharing research data and intellectual property law: A primer. *PLoS Biology* 13(8) (2015), e1002235. doi:10.1371/journal.pbio.1002235.
- [4] World Intellectual Property Organization. Berne convention for the protection of literary and artistic works (as amended on September 28, 1979). Available at: <https://wipolex.wipo.int/en/text/283698>.
- [5] World Intellectual Property Organization. Marrakesh treaty to facilitate access to published works for persons who are blind, visually impaired or otherwise print disabled (adopted on June 27, 2013). Available at: <https://wipolex.wipo.int/en/treaties/textdetails/13169>.
- [6] World Intellectual Property Organization. WIPO Copyright Treaty (adopted in Geneva on December 20, 1996). Available at: <https://wipolex.wipo.int/en/treaties/textdetails/12740>.
- [7] Communia Association. Policy paper in reaction to the public consultation on the Database Directive. August, 2017. Available at: <https://www.communia-association.org/policy-papers/policy-paper-12-reaction-public-consultation-database-directive/>.
- [8] The British Horseracing Board Ltd and Others v William Hill Organization Ltd, Case C-203/02 of 9 November 2004, ECLI:EU:C:2004:695. Available at: <http://curia.europa.eu/juris/liste.jsf?num=C-203/02>.
- [9] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The “A” of FAIR –As open as possible, as closed as necessary. *Data Intelligence* 2(2020), 47–55. doi: 10.1162/dint_a_{00027}.
- [10] T. Margoni, D. Peters. Creative Commons Licenses: Empowering Open Access. *Editorial Office News* 9(2) (2016). Available at: https://cdn.ymaws.com/www.ismte.org/resource/resmgr/EON_{Shared}{Articles}/Creative{Commons}_{L
- [11] J. Wilbanks. Public domain, copyright licenses and the freedom to integrate science. *Journal of Science Communication* 7(2) (2008). doi: 10.22323/2.07020304.
- [12] 17 U.S. Code § 106A. Rights of certain authors to attribution and integrity. Available at: <https://www.law.cornell.edu/uscode/text/17/106A>.
- [13] Etalab. Open License/Licence Ouverte 2.0 (2017) Available at: <https://www.etalab.gouv.fr/wp-content/uploads/2018/11/open-licence.pdf>.
- [14] United Kingdom Government. Open government licence for public sector information, version 3.0 (2014). Available at <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>.
- [15] Europeana. Usage guidelines for metadata. Available at: <https://www.europeana.eu/portal/rights/metadata>
- [16] OpenAIRE. Guides for researchers. Available at: <https://www.openaire.eu/guides>.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv –Machine translation. Verify with original.