Why data copyright and open licensing matter

Anouk Santos
Informational Resources and Archives Service (UNIRIS)
University of Lausanne
Lausanne, Switzerland
ORCID 0000-0002-1836-0835

A Abstract— *Research data are often not protected by copyright because they lack originality, but are generally made available under a copyright license. This situation is problematic firstly because research data are not always protected, and secondly because the licenses chosen are often restrictive and closed (e.g. those prohibiting modification and/or commercial use); or open, but not very suitable for Open Research Data (e.g. licenses requiring attribution or share alike). For those reasons, and to fully achieve the objectives of the Open Science movement with the least restrictions to impact data reuse, public domain licenses are recommended for sharing research data (e.g. CC0).*

Keywords—copyright, research data, data license, open license, Creative Commons, CC0, public domain.

I. Introduction

I devoted my master's thesis in Information science from the Haute école de gestion of Geneva to the legal framework of research data and data licenses (Santos 2020), mandated by the DLCM project. In this paper, I will present some considerations taken from my thesis. Those are not new findings, but rather reminders of what is needed to understand in terms of copyright and open licensing so as to fulfill the requirements for Open Research Data and, more generally, Open Science.

II. Data copyright

According to the Swiss Federal Act on Copyright and Related Rights (Swiss Confederation 2020), a work must comprise three conditions in order to be protected: to be a creation of the mind, to have an individual character, and to be expressed in one form or another (CCdigitallaw 2020). Having an individual character means that it is impossible for someone performing the same task to create an identical work. Data are not explicitly mentioned in the Swiss copyright law, so their protection must be assessed on a case-by-case basis according to the three aforementioned conditions. The following generic data examples can be mentioned: database, software, data visualizations, metadata or any other data. Thus, to assess if research data are protected by copyright or not is a difficult question to answer.

On the contrary, we know for sure that factual scientific data are not protected. Facts, information, ideas, formulas, algorithms, scientific measurements, etc. are not eligible to copyright protection because they are not considered individual works of authorships. They are discovered and compiled by a researcher's methods and this is something that copyright does not reward (Pantaloni 2017). So, in a dataset, some parts can be protected by copyright while others are not, and sometimes, the entire dataset is not protected. This is something that is important to keep in mind when thinking about data licensing.

III. Open data licensing

Here I will be referring to the Creative Commons (CC) licenses. They are the recommended licenses to use because they are compatible with data, widespread within the scientific field, quite simple to use and well-known. However, CC license are not always fully understood, notably some of their consequences on scientific research. Note that CC licenses are not designed for software or computer code and that there are open source licenses more suitable for this type of data.

One very important definition in order to comprehend open data is *The Open Definition* from The Open Knowledge Foundation (2015) that says: "Open data and content can be freely used, modified, and shared by anyone for any purpose (subject, at most, to requirements that preserve provenance and openness)". In the terms of the CC licenses, to preserve provenance means to give attribution, and to preserve openness is to require that the derivative

work is shared under the same license as the original work (Shared Alike). In fact, according to the open definition, there are only three CC licenses that are truly open and suitable for open data: CC0 (a public domain dedication), CC-BY and CC-BY-SA. All three allow data to be freely used, modify and shared. All the others CC licenses are therefore considered as not open.

CC licenses with the No Derivatives element (ND) are a problem for scientific research because the ND requirement forbids data to be modified, corrected, translated, combined or enriched with any other data, or shared partially (only the full dataset can be shared) (Ball 2014, Kreutzer 2014). Overall, those licenses prevent the creation of derivative works, which is not useful for Open Science because in the research field data generally exist in order to be crossed-referenced with other data, which is impossible to do with such a license. Furthermore, research is very often based on previous scientific works and thus there is a strong need to be able to combine data (Hirschmann 2020).

When a dataset is made available under a CC license with a Noncommercial element (NC) it is not open either. The main problem is to establish what is a noncommercial use because interpretations differ. In fact, a NC license could prevent some common reuse of research data: in a work for which the author receives a financial retribution (for example a published book or the publication of an article in a journal owned by a commercial editor), or also public-private partnerships, which occur regularly. As a result, data must be allowed to be used for any purposes, even commercial ones, to be truly open.

We saw that according to The Open Definition (The Open Knowledge Foundation 2015), CC licenses with Attribution (BY) and Share Alike (SA) are open. But unfortunately, those requirements can be problematic too regarding Open Research Data.

The problem with CC-BY-SA, which is a copyleft license, is the incompatibility with any other copyleft licenses. For example, it is impossible to combine a dataset which is under CC-BY-SA with another one under CC-BY-NC-SA, because both require that the same license is kept afterwards, which is impossible (Ball 2014). For this reason, the SA element affects the interoperability of data and increases the incompatibility of licenses, already initially caused by their proliferation.

As for attribution, there are two main problems that raise voices against it for research data. The first one is known as “attribution stacking”: when reusing or combining a lot of datasets that have a lot of authors, you must cite each of them correctly. It can be time consuming and difficult to achieve as the datasets are reused (Ball 2014). The second one is more an ethical consideration: can researchers articulate community norms, here peer citation, as a legal form? The answer is no: attribution cannot be legally binding by a license if the data are not protected by copyright… and as a result not eligible to licensing (Ball 2014). Thus, for instance, with factual scientific data that are not protected, or works that are into the public domain, someone doing that would be overriding his rights to that content.

Consequently, we are left with CC0, a public domain dedication, as the best choice. Here are some of the reasons why:

- CC0 solves the problem of licenses’ incompatibility: placing data into the public domain means that anyone can reuse them for any purpose. It avoids creating data silos that are incompatible with each other (Lämmerhirt 2017).
- CC0 achieves legal interoperability as it is an answer to the ambiguity of data copyright: there is no need to know which data are protected or not because all of them are placed into the public domain (Fortney 2016). CC0 allows legal interoperability by waiving patrimonial and moral rights of the data that are protected (to the extent allowed by law).
- There is a certain logic to put publicly funded data into the public domain. It is also coherent with the general sharing and reuse ethics which prevail normally within the scientific community (Murray-Rust et al. 2010).
- Open Science is easier to achieve with the least restrictions to impact data reuse (Labastida, Margoni, 2020).

IV. CONCLUSION

In conclusion, if non-open restrictions are put in place in licenses for research data, they must be carefully considered and justified because of their consequences. Such licenses can be very concrete barriers to the reuse of the data, and more globally the sharing of scientific research. Therefore, in my opinion, there is a need to raise awareness about data copyright and advocate for open licenses, and preferably for the public domain, in order to ensure that the principles of the Open Science movement are preserved. For the researcher sharing data openly also has its advantages: if data are reused, the data creator will be cited for his work, thus the visibility and discoverability of his

research will increase. He will potentially create opportunities for new research collaborations and demonstrate his integrity and the robustness of his work (Leeming 2017).

REFERENCE

Ball, A. (2014). *How to License Research Data.* Digital Curation Centre. https://www.dcc.ac.uk/resources/how-guides/license-research-data.

CCDigitalLaw. (2020, April 1st). *2.1 Protected work.* CCdigitallaw. https://ccdigitallaw.ch/index.php/english/copyright/2-what-work/21-urheberrechtlich-geschuetztes-werk.

Fortney, K. (2016, September 15). CC BY and data: Not always a good fit. *Office of Scholarly Communication of the University of California.* https://osc.universityofcalifornia.edu/2016/09/cc-by-and-data-not-always-a-good-fit/

Hirschmann, B. (2020, January 20). *Creative Commons Licenses*. Manual Research Collection (ETHZ Library). https://documentation.library.ethz.ch/display/RC/Creative+Commons+Licenses.

Kreutzer, T. (2014). *Open Content: A Practical Guide to Using Creative Commons Licences.* German Comm. for UNESCO. https://meta.wikimedia.org/wiki/File:Open_Content_-_A_Practical_Guide_to_Using_Creative_Commons_Licences.pdf.

Labastida, I. & Margoni, T. (2020). Licensing FAIR Data for Reuse. *Data Intelligence*, 2(1-2), 199-207. https://doi.org/10.1162/dint_a_00042

Lämmerhirt, D. (2017, December). Avoiding data use silos: How governments can simplify the open licensing landscape. *research.okfn.org.* https://research.okfn.org/avoiding-data-use-silos/

Leeming, J. (2017, June). Ask not what you can do for open data; ask what open data can do for you. *Nature jobs*. http://blogs.nature.com/naturejobs/2017/06/19/ask-not-what-you-can-do-for-open-data-ask-what-open-data-can-do-for-you/

Murray-Rust, P., Neylon, C., Pollock, R. & Wilbanks, J. (2010, February 19). *Panton Principles: Principles for Open Data in Science.* Panton Principles. https://pantonprinciples.org/

Open Knowledge Foundation. (2015). *Open Definition 2.1*. http://opendefinition.org/od/2.1/en/.

Pantaloni, N. (2017, Decmber 12). Copyright and data curation. *Indiana University Libraries Blogs.* https://blogs.libraries.indiana.edu/scholcomm/2017/12/12/copyright-and-data-curation/.

Santos, A. (2020, August 14). *Données de la recherche : cadre juridique et licences.* Geneva: Haute école de gestion. Master's thesis. https://doi.org/10.5281/zenodo.3967402.

Swiss Confederation. (2020, April 1st). *Federal Act on Copyright and Related Rights (Copyright Act, CopA) of 9 October 1992*. https://www.admin.ch/opc/en/classified-compilation/19920251/index.html