IJDC | *Brief Report*

# Uncommon Commons? Creative Commons licencing in Horizon 2020 Data Management Plans

Daniel Spichtinger
Independent researcher / Ludwig Boltzmann Gesellschaft

## Abstract

As policies, good practices and mandates on research data management evolve, more emphasis has been put on the licencing of data, which allows potential reusers to quickly identify what they can do with the data in question. In this paper I analyse a pre-existing collection of 840 Horizon 2020 public data management plans (DMPs) to determine which ones mention Creative Commons (CC) licences and among those that do, which licence types are being used.

I find that 36% of DMPs mention Creative Commons and, among those, a number of different approaches towards licencing exist (overall policy per project, licencing decisions per dataset, licencing decisions per partner, licensing decision per data format, licensing decision per perceived stakeholder interest), often clad in rather vague language with CC licences being "recommended" or "suggested". Some DMPs also "kick the can further down the road" by mentioning that "a" CC licence will be used, but not specifying which one. However, among those DMPs that do mention specific CC licences, a clear favourite emerges: the CC-BY licence, which accounts for half of the total mentions.

The fact that 64% of DMPs did not mention Creative Commons at all is an indication for the need for further training and awareness raising on data management in general and licencing in particular in Horizon Europe. For those DMPs that do mention specific licences, almost 60% would be compliant with Horizon Europe requirements (CC-BY or CC0). However, it should be carefully monitored whether content similar to the 40% that is currently licenced with non- Horizon Europe compliant licences will in the future move to CC-BY or CC0 or whether such content will simply be kept fully closed by projects (by invoking the "as open as possible, as close as necessary" principle), which would be an unintended and potentially damaging consequence of the policy.

*Submitted* date 27 June 2022 ~ *Accepted* date 19 July 2022

Correspondence should be addressed to Daniel Spichtinger, Lange Gasse 26/19 A-1080 Vienna, Austria Author Name, Email: dspichtinger@outlook.com

This article is based on a lightning talk at International Digital Curation Conference IDCC22, online, 13-16 June, 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v17i1.840
DOI: 10.2218/ijdc.v17i1.840

# Introduction

Research data are increasingly conceptualized as inherently valuable products of scientific research, rather than components of the research process that have no value in themselves (Leonelli 2013). Research funders on national and international levels increasingly include requirements for data management, inter alia, the European Union in its Horizon 2020 programme for research and innovation (2014-2020) and its successor Horizon Europe (2021-2027). In both Horizon 2020 (Spichtinger & Siren, 2018) and Horizon Europe, the production of data management plans is a cornerstone of the requirements for research data management and (to a certain extent) data openness.

As policies, good practices and mandates on research data management evolve, more emphasis is, inter alia, put on the licencing of data, which is often considered an important aspect of reusability (the R in FAIR) (Labastida & Margoni, 2020; Vasilevsky et .al 2019). Creative Commons (CC) copyright licenses and related tools allow creators (or licensors) to retain copyright while allowing others to copy, distribute, and make some uses of their work. Creative Commons explains that,

> ‘[e]very Creative Commons license also ensures licensors get the credit for their work they deserve. Every Creative Commons license works around the world and lasts as long as applicable copyright lasts (because they are built on copyright). These common features serve as the baseline, on top of which licensors can choose to grant additional permissions when deciding how they want their work to be used. ‘ (Creative Commons: About the Licenses, 2022)

The licensor can choose between a number of available licences, as depicted in Table 1. Additionally, the CC0 tool allows licensors to waive all rights and place a work in the public domain.

**Table 1.** Creative Commons Licences

| Abbreviation | Title | Description |
|---|---|---|
| CC BY | Attribution | This license lets others distribute, remix, adapt, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials. |
| CC BY-SA | Attribution-ShareAlike | This license lets others remix, adapt, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms. This license is often compared to “copyleft” free and open-source software licenses. All new works based on yours will carry the same license, so any derivatives will also allow commercial use. This is the license used by Wikipedia, and is recommended for materials that would benefit from incorporating content from Wikipedia and similarly licensed projects. |
| CC BY-ND | Attribution-NoDerivs | This license lets others reuse the work for any purpose, including commercially; however, it cannot be shared |

| | | |
|---|---|---|
| | | with others in adapted form, and credit must be provided to you. |
| CC BY-NC | Attribution-NonCommercial | This license lets others remix, adapt, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms. |
| CC BY-NC-SA | Attribution-NonCommercial-ShareAlike | This license lets others remix, adapt, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms. |
| CC BY-NC-ND | Attribution-NonCommercial-NoDerivs | This license is the most restrictive of our six main licenses, only allowing others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially. |

Source: Creative Commons, 2022a

A number of organisations, such as the Digital Curation Centre (Ball, 2014), OpenAIRE (OpenAIRE, 2019) and others (e.g., universities) have in recent years developed specific guidance on research data licencing for researchers and research data managers and similar professionals. In Horizon 2020, licencing of research data was mentioned in several guidance documents, most notably the Annotated Model Grant Agreement (AGA) which contains the recommendation to provide an "appropriate" Creative Commons licence for data, with CC-BY or CC0 being mentioned specifically (European Commission, 2019a). Similarly, the Guidelines to the Rules on Open Access to Scientific Publication an Open Access to Research Data in Horizon 2020 outline the requirements to first deposit and then in a second step "take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data" (European Commission, 2017) under the overall principle of "as open as possible, as closed as necessary". In this context, the guidance states: "one straightforward and effective way of doing this is to attach Creative Commons Licences (CC BY or CC0) to the data deposited" (European Commission, 2017).

In the new Horizon Europe programme this recommendation has been upgraded into a legal requirement of the Model Grant Agreement (that is the legally binding document that all EU funded projects must sign). Article 17 of the Grant Agreement states:

> The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:
> …
> as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary'… (European Commission, 2022)

The same article also states that metadata of deposited data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded). The addition of these requirements is a major step forward for licensing (and thus, ultimately reusability) and also in line with the general approach of the European Commission which adopted Creative Commons licences for their own documents in 2019 (European Commission, 2019b).

But what is happening on the ground? How many Horizon 2020 projects have used Creative Commons licences for their data, which ones and in which context? Those are the questions I investigate in this paper. From this data grounding I also draw conclusions as regards implications for CC licencing in the new Horizon Europe programme.

# Methodology

This research was originally conducted as an assignment submitted for the Creative Commons Certificate, an in-depth course about CC licenses, open practices, and the ethos of the Commons (Creative Commons, 2022b). For the final assignment of this course I looked into Creative Commons and copyright in DMPs . In doing so, this project builds on previous research, in which a vetted collection of data management plans was established as part of the DMP Use Case project which was undertaken for the University of Vienna on behalf of the EU funded OpenAIRE Advance project. During the project a curated collection of 840 DMPs from the EU funding programme Horizon 2020 was provided in a publicly accessible repository and analysed (Spichtinger 2021).

Taking this pre-existing dataset, I identified which of these 840 DMPs mention Creative Commons and among those, which specific CC licences are being used by Horizon 2020 projects. To do so, an automated search for the terms “creative commons”, “CC-BY”, “CC-BY-SA”, “CC-BY-NC”, “CC-BY-NC-ND”, “CC-BY-NC-SA”, “CC-0” was conducted amongst the 840 DMPs in the collection. This was initially conducted with the seek fast software tool (full licence), which allows users to quickly and easily search text in files, in this case pdf. However, when double checking it was found that this software was very restrictive with regard to finding different variants of spelling such as whether a hyphen was present or not (e.g., CC-BY or CC BY). The search was therefore repeated with the windows file manager search function, which was able to identify more spelling variants and was therefore able to cover the content of the DMPs more accurately for the purpose of this research. The results were then manually double checked to ascertain the context in which the search terms were used. Additional manual double checking was also carried out to also catch further variants (e.g., when an abbreviation such as “SA” was spelled out as “share alike”). The outcome of the research was added to an excel file that lists the 840 publicly accessible DMPs. Further worksheets were then added to this excel file for each of the licence types that were queried (Spichtinger 2022).

# Findings

## General Findings

From the 840 DMPs, 35.9% (n=302), contain some reference to Creative Commons. When manually looking at these references it became apparent that projects used a wide variety of approaches. While some DMPs define a policy for the project as a whole, other DMPs leave dataset licencing up to the individual project partners, since they are the owners of the results. In some cases, partners were given a choice between different CC licences, and, in one case, a CC licence was only applied to jointly owned results.

This points to the more general fact that in many cases there was not just one CC licence but rather a number of CC licences that were used. The rationale for which licence was applied to different data varied among projects; in some project DMPs, more restrictive licenses were used for data which was deemed commercially valuable. In other cases, commercially sensitive data was not opened at all – in line with the EC’s approach of “as open as possible, as closed as necessary" but it was clarified that CC licences were only applied to those datasets that are being made public.

Others do not explicitly state this but use vague phrases such as CC licences being used "where appropriate", "where possible" or, in one case, "unless this hampers the business model of our partners".[1] Ambiguous phrasing is often used in the DMP rather than specifying the exact CC licence(e.g., by stating that "a" CC licence, but not which specifically, will be used). This could also reflect the fact that some of the DMPs that were submitted were from projects that were just beginning. DMPs sometimes state that CC licences will be used for "most" of the data without making clear which data this applies to. In some cases, licences were determined based on the type of output (sometimes even the data format) or the presumed target groups for different datasets.

In many cases, the DMPs did not prescribe a licence but rather "recommended" the use of a CC licence by the project partners. Several DMPs also indicated that the issue of licencing was still under discussion and that the project had not arrived at a final decision.

While this paper has a focus on datasets, several DMPs also – or even exclusively – mention Creative Commons licences for scientific articles. A number of DMPs simply quote the Commission guidance without indicating which approach the project has chosen.

Some DMPs stated that Share Alike (SA) licences would not be used, since this would hamper the further choices of future data reusers. Other DMPs contain statements to the effect that those (CC) licences which allow the broadest possible reuse will be used. Several projects also made the conscious choice not to use CC licences for software, which is in line with the guidance from Creative Commons.

In some cases, the DMP did not contain a generic policy on licencing but did list the CC licences for specific datasets produced by the project. In other cases, the project did not use CC licencing for its own results but rather reused content from public sources that already had a CC licence and mention Creative Commons in this context (i.e., as a justification that they have the right to reuse external data).  In some projects very similar wording was used in their DMPs which suggests the use of a template.

The issues described above need to be borne in mind for the further analysis of the specific licences used. For this analysis, the following decisions were taken:

- If a project applies more than one licence, each licence type will be counted (but only once)
- If a project only reiterates the EC requirements without indicating which licence(s) it has chosen, this will not be counted
- If a project only refers to a CC licence for publications and not for data, this will not be counted
- If a project used vague wording such as "as far as possible" "for most data" etc. this will be counted
- If a project does not have a generic licencing policy but the DMP includes concrete datasets which are licenced, these licences will be counted
- If a project indicates that its policy has not been finalised but has recommendations on which licences to use, these will be counted
- If a project mentions CC licences for public data that it reuses but does not indicate a CC licence policy for its own data, this will not be counted

[1] For details concerning the examples quoted see the underlying dataset (Spichtinger 2022).

- A generic mentioning of using the "CC licence family" or similar is not sufficient to be counted for each specific CC licence; rather the specific CC licence has to be at least mentioned as being under consideration
- For the sake of simplicity, the version number of the CC licence(s) has not been included in the findings.

## Specific licences used

Bearing in mind the factors and caveats described above, the following distribution of specific CC licences emerges from the analysis of a subset of 250 DMPs:

- CC-BY: 125
- CC-BY-SA: 38
- CC-BY-NC: 21
- CC-BY-ND: 7
- CC-BY-NC-ND: 10
- CC-BY-NC-SA: 25
- CC-0: 24

The following pie chart show visualises that CC-BY licences constitutes by far the largest number of specific CC licences, followed by CC-BA-SA, CC-BY-NC-SA, CC0, CC-BY-NC, CC-BY-NC-ND and CC-BY-ND, the least popular licence. The total mention of specific licences amounts to 250.

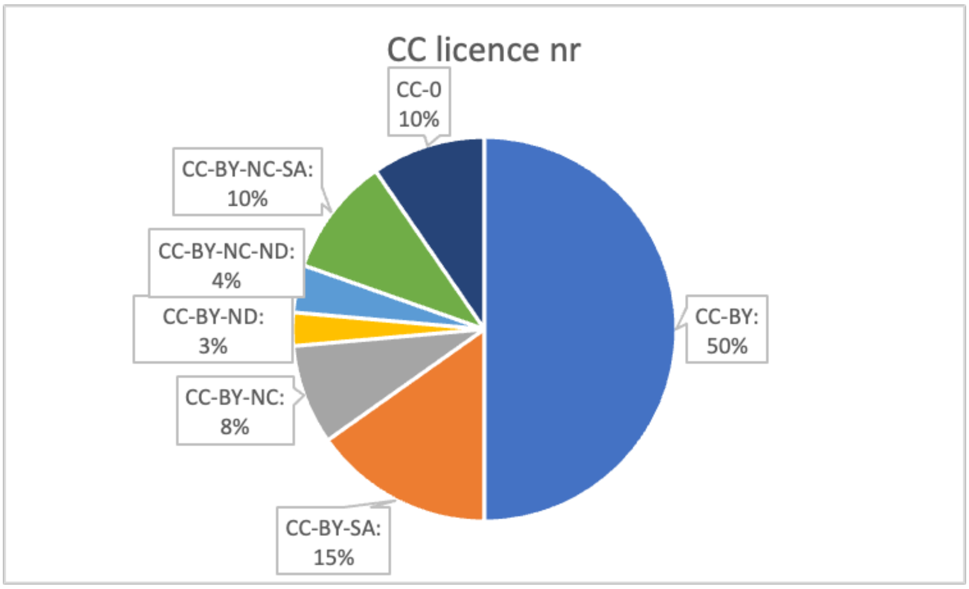

**Figure 1.** CC licence types mentioned in DMP sample (N=250)

# Discussion and Conclusions

On the one hand, the project findings show a certain amount of what in German one could call "Wildwuchs", to be roughly translated as "uncontrolled development or rank growth" (often used with plants). With this I mean the large and varying number of licensing policies or lack of them among the projects: only 36% of DMPs mention Creative Commons at all and those who do provide a number of different approaches (overall policy per project, licencing decisions per dataset, licencing decisions per partner etc), sometimes rather vague in the form of recommendations or suggestions.

On the other hand, among those DMPs that do mention specific CC licences, a clear favourite emerges: the CC-BY licence, which accounts for half of the total mentions of a specific licence. Given the currently available data we can only speculate[2] why this is the case. Many projects do state that they want to enable the widest possible reuse and in many cases a CC licence is only applied to data which the project intends to make public anyway. The fact that CC-BY-ND was the least popular licence may reflect the fact that, for scientific work, some sort of derivation is often required,(although the number of licences that do not allow derivations raises to 17, if we count not only ND but also NC-ND). Moreover, a number of licences included restrictions for commercial reuse – counting them all together (that is NC + NC-ND + NC-SA) we get 56 licences that bar commercial reuse. Finally, it should also be noted that the popularity of CC-0 has been underwhelming, in particular given that CC-BY and CC-0 were specifically named in the EC guidance. Furthermore, a number of CC-0 licences applied only to metadata.

## Implications for Horizon Europe

What does this mean for Horizon Europe? While it goes without saying that we cannot retroactively apply the Horizon Europe requirements to Horizon 2020, the data from Horizon 2020 can provide us with a grounding in what to watch out for in Horizon Europe, which I summarise as conclusions below:

- Conclusion 1: the fact that only 36% of DMPs mention Creative Commons means that a lot of projects were either not familiar with them or did not consider them relevant). This indicates a need for further training and awareness raising which is in line with previous findings which, inter alia, highlight the need for more support for data management (e.g., through a dedicated one stop shop on Horizon Data Management, similar to the already existing IP Helpdesk (Spichtinger 2021)). The Horizon dissemination booster could potentially provide a best practice example in this context.

- Conclusion 2: as Horizon Europe mandates CC-BY or CC-0, over half of the 250 H2020 DMPs that mentioned Creative Commons (59,6%) would be compliant with the requirement. However, we must bear in mind that most DMPs do not mention CC licences at all (see conclusion 1 above). Furthermore,

- Conclusion 3: Within Horizon 2020 projects, a number of different CC licences and approaches were used (e.g., on a per dataset, per stakeholder or per partner basis). By contrast, in Horizon Europe, the Grant Agreement only allows a choice between CC-BY and CC-0. The new Horizon Europe mandate is therefore useful in that it (if properly implemented) does away with the variety of different, often vague, policies and non-binding recommendations that we find in Horizon 2020 DMPs, which I referred to

[2] This is to a certain extent speculation since the project resources did not allow for follow up interviews to clarify the underlying rationale for choosing a specific licence. This potentially provides room for further research.

earlier as "Wildwuchs". However, the new mandate begs questions about what will happen to the content that can't be shared more openly as it has been licenced more restrictively in Horizon 2020 DMPs (e.g., CC-BY-SA, CC-BY-NC CC-BY-ND, CC-BY-NC-ND and CC-BY-NC-SA).

Following up on conclusion 3 two potential outcomes seem to exist:

1. Such content will, in the future, be made available in a more open manner, through the use of the prescribed licences CC-BY or CC-0 (intended consequence, positive)

   or

2. Such content will, in the future, be completely closed off, with projects citing the "as open as possible, as closed as necessary" principle and preferring to keep such content closed (unintended consequence, negative)

Given that Horizon Europe only started in 2021 there is currently not enough available data to say whether outcome 1 or 2 is more likely. However, given that outcome 2 is at least theoretically possible, one wonders whether it would not have been preferrable to provide projects with a larger choice of permissible Creative Commons licences in Horizon Europe. The issue should be closely monitored as part of Horizon Europe data and intelligence gathering from the side of the European Commission.

Finally, further work on the dataset used in this project and/or other Horizon 2020 DMPs could focus on differences between thematic areas. For instance, in some domains the use of CC licences may be more common than in others however, this aspect was beyond the scope of the current project.

## Acknowledgements

The participation fee for the Creative Commons Certificate as well as some of the working time involved in the course was covered by the Ludwig Boltzmann Gesellschaft.

## References

Ball, A. (2014). 'How to License Research Data'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Retrieved from https://dcc.ac.uk/guidance/how-guides/license-research-data

Creative Commons: About The Licenses. (2022) Retrieved from https://creativecommons.org/licenses/?lang=en

European Commission (2017) H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2. p.10. Retrieved from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

European Commission (2019a) H2020 Programme: AGA – Annotated Model Grant Agreement. Version 5.2. p253. Retrieved from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

European Commission (2019b) Commission Decision adopting Creative Commons as an open licence under the European Commission's reuse policy. Retrieved from https://ec.europa.eu/transparency/documents-register/detail?ref=C(2019)1655&lang=en

European Commission (2022) Horizon Europe (HORIZON) Euratom Research and Training Programme (EURATOM): General Model Grant Agreement. EIC Accelerator Contract. Version 1.0. p.109 Retrieved from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/agr-contr/general-mga_horizon-euratom_en.pdf

Labastida, I. & Margoni, T. (2020) Licensing FAIR data for reuse. Data Intelligence 2, 199–207. doi: https://doi.org/10.1162/dint_a_00042

Leonelli, S (2013) Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology. Bull Sci Technol Soc; 33(1–2): 6–1 Retrieved from https://journals.sagepub.com/doi/10.1177/0270467613496768

OpenAIRE (2019) Research data how to license. Retrieved from https://www.openaire.eu/research-data-how-to-license/

Spichtinger, D & Siren, J (2018) The Development of Research Data Management Policies in Horizon 2020. In: Kruse, F., Thestrup, J.B., (eds.) Research Data Management - A European Perspective pp. 11–23.: De Gruyter SAUR, Berlin/Boston Retrieved from https://zenodo.org/record/1188886#.YPbWxcRxdFo

Spichtinger, D. (2021) Data Management Plans in Horizon 2020: what beneficiaries think and what we can learn from their experience [version 2; peer review: 2 approved, 1 approved with reservations]. Open Res Europe 2022, 1:42. Retrieved from https://doi.org/10.12688/openreseurope.13342.2

Spichtinger, D. (2022) Uncommon Commons? Creative Commons licencing in Horizon 2020 Data Management Plans [Data set]. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6685131

Vasilevsky, N. & Carbon, S., Champieux, R., McMurry, J., Winfree, L. Wyatt, L.& Haendel. M (2019). Evaluating scientific data licensing with the (Re)usable Data Project. In PLOS ONE (Vol. 14, Number 3, p. e0213090). Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3497130