A Digital Curation Centre and JISC Legal 'working level' guide

How to License Research Data

Alex Ball (DCC)

How to License Research Data

Introduction

This guide will help you decide how to apply a licence to your research data, and which licence would be most suitable. It should provide you with an awareness of why licensing data is important, the impact licences have on future research, and the potential pitfalls to avoid. It concentrates on the UK context, though some aspects apply internationally; it does not, however, provide legal advice. The guide should interest both the principal investigators and researchers responsible for the data, and those who provide access to them through a data centre, repository or archive.

Why license research data?

While practice varies from discipline to discipline, there is an increasing trend towards the planned release of research data. The need for data licensing arises directly from such releases, so the first question to ask is why research data should be released at all.

A significant number of research funders now require that data produced in the course of the research they fund should be made available for other researchers to discover, examine and build upon. The rationale given by UK funders is that opening up the data allows for new knowledge to be discovered through comparative studies, data mining and so on; it also allows greater scrutiny of how research conclusions have been reached, potentially driving up research quality.[1] Some journals are taking a similar stance, requiring that authors deposit their supporting data either with the journal itself or with a recognised data repository.[2]

There are many additional reasons why releasing data can be in a researcher's interests.[3,4] The discipline of working up data for eventual release helps in ensuring that a full and clear record is preserved of how the conclusions were reached from the data, protecting the researcher from potential challenges. A culture of openness deters fraud, encourages learning from mistakes as well as from successes, and breaks down barriers to interdisciplinary and 'citizen science' research. The availability of the data, alongside associated tools and protocols, increases the efficiency of research by reducing both data collection costs and the possibility of duplication. It also has the potential to increase the impact of the research, not only academically,[5] but also economically and socially.

Merely releasing data without making clear their terms of use can be somewhat counter-productive, though. The default legal position on how data may be used in any given context is hard to untangle, not least because different jurisdictions apply different standards of creativity, skill, labour and expense when judging whether copyright or similar rights pertain. The situation is

[1] SQW Consulting & LISU. (2008, September). *Open access to research outputs* (§3.10). Swindon: Research Councils UK. retrieved from `http://www.rcuk.ac.uk/RCUK-prod/assets/documents/news/oareport.pdf`.

[2] Examples of journals with such a policy include the *American Economic Review*, the *Journal of Evolutionary Biology*, and *Clinical Infectious Diseases*.

[3] Stodden, V. (2009). Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy*, *13*, 1–25. Retrieved from `http://www.ijclp.net/files/ijclp_web-doc_1-13-2009.pdf`.

[4] *Open to all? Case studies of openness in research*. (2010, September). Research Information Network and National Endowment for Science, Technology and the Arts. Retrieved from `http://www.rin.ac.uk/system/files/attachments/NESTA-RIN_Open_Science_V01_0.pdf`.

[5] Pienta, A. M., Alter, G. C. & Lyle, J. A. (2010, April). *The enduring value of social science research: The use and reuse of primary research data*. Paper from the *Organisation, Economics and Policy of Scientific Research* workshop, Torino, Italy. Retrieved from `http://hdl.handle.net/2027.42/78307`.

complicated by the fact that different aspects of a database – field values (i.e. the data themselves), field names, the structure and data model for the database, data entry interfaces, visualisations and reports derived from the data – may be treated quite differently.[6]

In the US, there is a strong emphasis on creativity, so straightforward tables of, say, sensor data are unlikely to attract copyright. In Australia, creativity is not relevant but originality is. Originality is judged on a range of factors, including skill and labour, but the skill and labour have to relate directly to the work in question: the effort spent compiling a database does not necessarily affect the originality of a report generated from it.[7] Within the EU, the act of compiling a database attracts copyright insofar as the compiler has exercised intellectual judgement in selecting or arranging the data.[8] There is also a separate database right that applies to the contents of a database where a substantial investment was made to obtain, verify or present them. The thrust of the database right is that users may not extract or reuse more than an insubstantial part of the contents without authorisation from the compiler, unless certain exemptions apply. One of the exemptions is for teaching and scientific research, but as the EU Database Directive does not commit Member States to respecting it, it may not apply in all European countries.

Indeed, another potential source of confusion are the variations between jurisdictions in what can be done with copyright material. While the Berne Convention[9] provides a level of consistency among its signatories – which includes most but by no means all countries – there are still variations in the exemptions that each jurisdiction provides, and subtle differences concerning, for example, which acts count as copying, and what constitutes an insubstantial use or extract of a work. The latter is an important point because the exemptions to copyright and database rights permit a dataset to be compiled from insubstantial extracts from a number of other datasets,[10] but the fact of whether the extracts are indeed insubstantial might be contested.

With all these complexities and ambiguities surrounding the rights of database compilers, reusers need clear guidance from compilers on what they are allowed to do with the data.

[6] Data. (2012, June 12). Retrieved from Creative Commons website: http://wiki.creativecommons.org/Data.

[7] *Telstra Corporation Limited v Phone Directories Company Pty Ltd* [2010] FCAFC 149. Retrieved 10 January 2010, from http://www.austlii.edu.au/au/cases/cth/FCAFC/2010/149.html

[8] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. (1996, March 27). *Official Journal of the European Union*, *L077*, 20–28. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML.

[9] Berne Convention for the Protection of Literary and Artistic Works. (1979). Retrieved from World Intellectual Property Organization website: http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html.

[10] Fitzgerald, A. & Pappalardo, K. (2009, November 5). *Creative Commons and data*. Melbourne: Australian National Data Service. Retrieved from http://ands.org.au/guides/cc-and-data.html.

Licensing concepts

The two most effective ways of communicating permissions to potential reusers of data are *licences* and *waivers*. A licence in this context is a legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held. The first thing to note is that only the rights holder (or someone with a right or licence to act on their behalf) can grant a licence; it is therefore imperative that the intellectual property rights (IPR) pertaining to the data are established before any licensing takes place. The second thing to note is that while it is the nature of a licence to expand rather than restrict what a licensee can do, some licences are presented within contracts, and contracts *can* place additional restrictions on the licensee and indeed the licensor.

'An intellectual property (IP) licence is effectively a promise not to sue for infringement of an intellectual property right (IPR).'

– Irish, V. (2005). *IEE Management of Technology: No. 22. Intellectual property rights for engineers* (2nd ed., p. 173). London: Institution of Electrical Engineers

A waiver, by contrast, is a legal instrument for giving up one's rights to a resource, so that infringement becomes a non-issue. Again, only the entity that holds the rights (or someone with a right or licence to act on their behalf) can waive them. Note that a waiver does not authorise other parties to claim rights – as opposed to freedoms – they did not previously have.

Common terms

Licences typically grant permissions on condition that certain terms are met. While the precise details vary, three conditions commonly found in licences are attribution, copyleft, and non-commerciality.

- An *attribution* requirement means that the licensor must be given due credit for the work when it is distributed, displayed, performed, or used to derive a new work.
- A *copyleft* requirement means that any new works derived from the licensed one must be released under the same license, and only that licence.
- The intent of a *non-commercial* licence is to prevent the licensee from exploiting the work commercially. Such licences are often used as part of a dual-licensing regime (see 'Multiple licensing', below), where the alternative licence allows commercial uses but requires payment to the licensor.

While these all have their uses, they can cause problems in the context of datasets.

Datasets are particularly prone to *attribution stacking*, where a derivative work must acknowledge all contributors to each work from which it is derived, no matter how distantly. If a dataset is at the end of a long chain of derivations, or if large teams of contributors were involved, the list of credits might well be considered too unwieldy.[11] The problem is magnified if different sets of contributors have to be credited in a different way, especially if automated methods are used to assemble the dataset – some of the benefits of automation are lost if attribution conditions have to be inspected manually. Some licenses and licensors tackle this problem by specifying lightweight attribution mechanisms.[12]

The problem with copyleft licences is they prevent the licensed data being combined with data released under a different copyleft licence: the derived dataset would not be able to satisfy both sets of licence terms simultaneously. Some copyleft licences, however, demonstrate a small amount of flexibility in allowing derivative works to be released under a compatible licence, that is, one that applies approximately the same conditions.[13]

Non-commercial licences may have wider implications than intended due to the ambiguity of what constitutes a commercial use.[14] Depending on one's interpretation, it may or may not preclude the data being used in support of works for which an author is given recompense (such as textbooks), and might preclude the data being used in support of works that are sold (such as journal articles) even if the author does not benefit financially.

[11] Protocol for Implementing Open Access Data (§5.3). (2007, December 20). Retrieved from Science Commons website: `http://sciencecommons.org/projects/publishing/open-access-data-protocol/`.

[12] OCLC, for example, builds flexibility into its use of the ODC-By licence by allowing 'in circumstances where providing the full attribution statement . . . is not technically feasible, the use of canonical [dataset] URIs is adequate . . .' alongside examples of acceptable practice (Data licenses and attribution. [n.d.]. Retrieved from OCLC Website: `http://www.oclc.org/data/attribution.en.html`).

[13] For example, the GNU Project maintains a list of licences for code which permit redistribution under the GNU General Public Licence (GPL) and whose terms the GPL can accommodate (Various Licenses and Comments about Them. [2010, August 9]. Retrieved from GNU website: `http://www.gnu.org/licenses/license-list.html`).

Creative Commons maintains lists of licences into which its Share Alike licences may be converted by derived works, but these are currently empty (Compatible Licenses. [n.d.]. Retrieved from Creative Commons Website: `https://creativecommons.org/compatiblelicenses`).

[14] Netpop Research. (2009, September). *Defining 'Noncommercial': A study of how the online population understands 'Noncommercial Use'*. San Francisco, CA: Creative Commons. Retrieved from `http://wiki.creativecommons.org/Defining_Noncommercial`.

Prepared licences

Before considering the licensing options that are available, you should first check whether you are obliged or strongly encouraged to use a certain licence as a condition of funding or deposit, or as a matter of local policy.

Your department or institution may already have a licence prepared for you to apply to your data. Rothamsted Research, a BBSRC Institute, uses several different legacy licences for its own data, each reflecting both a desire to see the data used in current research, and caution against naïve or simplistic interpretation.[15] On the other hand, it also maintains some public domain genome sequences as part of the Multinational *Brassica* Genome Project.[16]

Some data centres have licences that depositors must grant as a condition of deposit. Contributors to the UK Data Archive (UKDA) are required to sign a standard licence agreement that clarifies the respective rights and responsibilities of both parties and permits the UKDA to perform its curatorial functions.[17] In turn, the UKDA makes the data available under various licences depending on the type of data. Open data may use the Open Government Licence, the Creative Commons BY-SA or BY-NC-SA version 4.0 licences, or the World Bank Terms of Use (see 'Standard licences' below). Safeguarded data are made available under one of two bespoke licences: the Special Licence if the data are sensitive, otherwise the End User Licence with or without special (additional) conditions.[18] Similarly, researchers depositing data with the Archaeology Data Service (ADS) are required to sign a deposit licence.[19] Those using data hosted by the ADS do so under both a brief licence and a common access agreement.[20]

Both the UKDA and ADS deposit licences are non-exclusive, which means among other things that granting them does not prevent you hosting a copy of the data yourself and distributing it under a different licence if you wish.

[15] Rothamsted Research Website, URL: http://www.rothamsted.ac.uk/.

[16] Multinational *Brassica* Genome Project Website, URL: http://www.brassica.info/.

[17] Licence Agreement. (2013, December 16). Retrieved from UK Data Service website: http://ukdataservice.ac.uk/media/28102/licenceform.pdf.

[18] Terms and Conditions of Access. (2014, April 9). Retrieved from UK Data Service website: http://www.esds.ac.uk/orderingData/termsandconditions.asp.

[19] ADS deposit licence, URL: http://www.ahds.ac.uk/documents/ahds-archaeology-licence-form.doc.

[20] The Terms of Use and Access to ADS Resources. (n.d.). Retrieved from Archaeology Data Service website: http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess.

Bespoke licences

Writing a bespoke licence for your data is not a trivial undertaking, and almost certainly unnecessary in the light of the standard licences available (see 'Standard licences' below). Furthermore, using a standard licence helps the users of your data as it reduces the number of licences with which they have to work, and aids interoperability and automation as described above. There are circumstances, though, in which it might be worth writing a custom licence: where the data have significant commercial value,[21] or where you need to clarify your responsibilities and those of reusers in respect of the data.

If you decide to do this, in the first instance you should consult with your organisation's research office, commercialisation services team and/or legal department. At the very least they will be able to advise you on the implications of including particular clauses or using particular wording in the licence; they may have standard

[21] In the UK, examples of public sector data offered commercially under bespoke licences include those from the Ordnance Survey (http://www.ordnancesurvey.co.uk/business-and-government/licensing/licences/) and the Hydrographic Office (http://www.ukho.gov.uk/copyright/).

texts or templates you could use, or may even offer to write the licence for you.

An example of the template approach is the Restrictive Licence (RL)[22] that was developed as part of Queensland's Government Information Licensing Framework (GILF) and later adopted into the Australian Governments Open Access and Licensing Framework (AusGOAL).[23] This licence, intended for government information and data, allows licensors to construct their own custom licence by filling out some simple forms. Left unmodified, the licence does not permit the licensee to do anything beyond what is allowed under copyright law, apart from a few provisions with regard to copying and redistribution. By filling out the licence's schedules, however, one can adjust the copying and distribution permissions, fix the term of the licence, restrict usage geographically, or add specific conditions or permissions. The completed template takes the form of an agreement that both licensor and licensee have to sign, so it cannot be used to give blanket permissions.

An example of fully bespoke licences are the ones used by the Augmented Multi-Party Interaction (AMI) Project at the University of Edinburgh.[24] The project released its AMI Meeting Corpus under two licences written by the Edinburgh Research and Innovation unit. One was a free, non-commercial, copyleft licence,[25] and the other a chargeable commercial licence. This is also an example of a dual licensing arrangement (see 'Multiple licensing' below).

[22] AusGOAL Restrictive Licence template, URL: http://www.ausgoal.gov.au/restrictive-licence-template.

[23] AusGOAL. (2011, May). Australian Governments Open Access and Licensing Framework. (2011, May). Retrieved from Australian National Data Service website: http://www.ands.org.au/guides/ausgoal-awareness.html.

[24] AMI Meeting Corpus Website, URL: http://groups.inf.ed.ac.uk/ami/corpus/.

[25] The AMI Meeting Corpus License is similar but not identical to the Creative Commons BY-NC-SA 2.0 Licence; URL: http://groups.inf.ed.ac.uk/ami/corpus/license.shtml.

'Creative Commons has the option to include commercial uses – we use the Non-Commercial one, though, because some contributors don't want to lose out on what they think more likely revenue (they think companies have money and research groups don't), and because commercial takers can't accept the Share Alike terms.'

– Researcher from the AMI Project, University of Edinburgh

Standard licences

While bespoke licences are useful for catering for very specific circumstances, most research projects would be better served using one of the standard licences. Below is a selection of the standard licences available, along with reasons for and against using each one. Please note that these licences can be terminated only by expiry of the licensor's IPR or, for a particular licensee, through breach of terms.

Creative Commons

Creative Commons is a non-profit corporation set up in 2001 for the purpose of producing simple yet robust licences for creative works.[26] These licences give the creators of such works finer-grained control over how they may be used than simply declaring them public domain or reserving all rights. As well as the legal text, the licences all have quick clear summaries and a canonical URL for use in HTML, RDF and other code. A rights expression language is also provided for use with RDF.[27] While originally aimed at works such as music, images and video, Creative Commons licences have been used widely for most forms of original content, including data.

There are six main Creative Commons licences. While the spirit behind them has remained constant, the wording of their legal deeds has been revised over time, resulting in different *versions*, and adapted to different legal jurisdictions, resulting in different *ports*.

[26] Creative Commons Website, URL: http://creativecommons.org/.

[27] RDF and rights expression languages are discussed under 'Mechanisms for licensing data' below.

Creative Commons at a glance

Good for

- very simple, factual datasets
- data to be used automatically

Watch out for

- versions: use v. 4 or later
- attribution stacking
- the NC condition: only use with dual licensing
- the SA condition as it reduces interoperability
- the ND condition as it severely restricts reuse

Each licence includes the *Attribution* condition. In the version 3 licences and earlier, it is left up to the licensor to specify the way in which credit is given. Recognising the difficulties this may cause in the context of attribution stacking, the version 4 licences can be satisfied by a link to a Web page containing attribution information, though licensors can specify additional, alternative mechanisms.

There are three other conditions that licensors can add, and the various possible combinations produce the six licences. Using just the Attribution condition is known as the CC BY licence.

There is a *Non-Commercial* condition, where commercial is defined as 'primarily intended for or directed toward commercial advantage or monetary compensation'.[28]

The *Share Alike* condition inserts a strong copyleft clause into the licence.[29] The version 1 licences are very strict: derivations may only use the exact same version 1 licence. The version 2 licences onwards, however, allow derivations to use a later version or a different port of the same license. Nevertheless, derivations may not use a Creative Commons licence with a different set of conditions.

Finally, including the *No Derivatives* condition in the version 3 licences and earlier means that the licensee is forbidden from altering, transforming or building upon the work. The version 4 condition is more flexible: it allows these things for private use, but prevents the licensee from sharing the derivations. It and the Share Alike condition are mutually exclusive.

The six permutations are therefore

- Attribution (CC BY);[30]
- Attribution Share Alike (CC BY-SA);[31]
- Attribution No Derivatives (CC BY-ND);[32]
- Attribution Non-Commercial (CC BY-NC);[33]
- Attribution Non-Commercial Share Alike (CC BY-NC-SA);[34]
- Attribution Non-Commercial No Derivatives (CC BY-NC-ND).[35]

The versions of the licences prior to version 4 were not specifically aimed at data, so using them for such presents some problems. The most significant is that they do not explicitly cover *sui generis* database rights such as the one in force in the European Union.[36] This means, for example, that use of substantial portions of a database licensed using the unported terms of version 3 or earlier may constitute a rights infringement in such jurisdictions. The version 4 licences, however, do explicitly include *sui generis* database rights unless the licensor specifically reserves them.

All versions of the licences treat datasets and databases as a whole: they do not treat the individual data themselves differently from the collection/database. This might be considered an advantage in terms of simplicity, but means they cannot be used without difficulty in certain complex cases such as collections of variously copyrighted works.

Similarly, the licences do not distinguish using data as part of a new collection/database from using them to generate content

[28] Frequently Asked Questions (section entitled 'Does my use violate the NonCommercial clause of the licenses?'). (2014, June 24). Retrieved from Creative Commons wiki: `http://wiki.creativecommons.org/Frequently_Asked_Questions`.

[29] The strength of a copyleft clause refers to the range of derivations to which it applies, with weaker clauses applying to a narrower range. For example, giving a software library a weak copyleft licence means that all future versions/modifications of that library inherit the licence, but software that merely depends on that library does not.

[30] CC BY, URL: `http://creativecommons.org/licenses/by/4.0`.

[31] CC BY-SA, URL: `http://creativecommons.org/licenses/by-sa/4.0`.

[32] CC BY-ND, URL: `http://creativecommons.org/licenses/by-nd/4.0`.

[33] CC BY-NC, URL: `http://creativecommons.org/licenses/by-nc/4.0`.

[34] CC BY-NC-SA, URL: `http://creativecommons.org/licenses/by-nc-sa/4.0`.

[35] CC BY-NC-ND, URL: `http://creativecommons.org/licenses/by-nc-nd/4.0`.

[36] More precisely, the ports of the version 3 licences to European jurisdictions fully waive the *sui generis* database right, while all other ports and the unported versions fully reserve it.

(graphs, models, maps, etc.). This means the Share Alike and No Derivatives conditions might have further reaching consequences than intended. Indeed, the No Derivatives condition would likely disallow most substantive types of reuse, leaving only such cases as checking that data within the set derive from each other as claimed. It should therefore be avoided.

In addition to the six main licences, Creative Commons provides tools for entering works into the public domain, or certifying works as already being in the public domain (see 'Public domain', below).

Open Data Commons

The Open Data Commons Project[37] was set up in 2007 to develop a successor to the Talis Community Licence (TCL).[38] The first licence to be produced was a public domain dedication for databases. The project transferred to the Open Knowledge Foundation in 2009 and has produced two further licences having some of the character of the Creative Commons licences, but designed specifically for databases. All three follow the Creative Commons model of providing a clear summary and canonical URL alongside the full legal text.

The Open Data Commons Attribution Licence (ODC-By) allows licensees to copy, distribute and use the database, to produce works from it and to modify, transform and build upon it for any purpose.[39] If content is generated from the data, that content should include or accompany a notice explaining that the database was used in its creation.[40] If the database is used substantially to create a new database or collection of databases, the licence URL or text and copyright/database right notices must be distributed with the new database or collection.

The Open Data Commons Open Database Licence (ODC-ODbL) is the same as ODC-By but for a couple of additional conditions.[41] It adds a copyleft condition that applies to new databases derived from the database (but not collections of databases or non-database content produced directly from it); this condition would be satisfied by future versions of the same licence or a compatible one as judged by the licensor. The other condition is that technological restrictions such as Digital Rights Management (DRM) mechanisms can only be applied to the database or a new database derived from it if an alternative copy without the restrictions is made equally available.

Being written in database terms, these licences are suited to a wider range of research data than the Creative Commons equivalents. The ODC-ODbL copyleft condition is also slightly more flexible than Creative Commons' Share Alike, though the ODC attribution requirement is slightly less flexible.

Open/Non-Commercial Government Licence

The Open Government Licence (OGL) was released as part of the UK Government Licensing Framework in September 2010; version 2 was released in June 2013.[42] It is intended for UK public sector

Example
In 2010, OpenStreetMap changed its licence from CC BY-SA 2.0 to ODC-ODbL 1.0 because ODbL
- handled database rights;
- enforced copyleft for derived data but not derived maps;
- allowed the project to speak for all contributors.

[37] Open Data Commons Website, URL: http://opendatacommons.org/.

[38] TCL, URL: http://tinyurl.com/p3ag72b.

ODC-By at a glance
Good for
- most databases and datasets
- data to be used automatically
- data to be used for generating non-data products

Watch out for
- attribution stacking

[39] ODC-By, URL: http://opendatacommons.org/licenses/by/.

[40] Example notice: 'Contains information from ⟨*database*⟩ which is made available under the ODC Attribution License.'

ODC-ODbL at a glance
Good for
- most databases and datasets
- data to be used automatically
- data to be used for generating non-data products

Watch out for
- attribution stacking
- the copyleft condition as it reduces interoperability
- the DRM clause as it may put off some reusers

[41] ODC-ODbL, URL: http://opendatacommons.org/licenses/odbl/.

[42] Open Government Licence for public sector information, URL: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/.

A machine-readable version of the Open Government Licence is available at http://reference.data.gov.uk/id/open-government-licence.

and government resources, particularly datasets, source code and collected or original information; that it cannot be used by licensors outside the UK is not directly stated, but is implied by the wording of its exemptions.

The terms of the licence are similar to CC BY in that attribution is required, derivative works and commercial uses are explicitly allowed, and there is no copyleft condition. Version 1 of the licence contained some additional conditions; most of them have been removed from version 2, except that derivative works must not be represented as having official status.

There are also categories of information for which the licence explicitly does not permit use:

- personal information;
- unpublished information, other than that disclosed under information access legislation (FoIA, etc.);
- public sector logos, armorial bearings, etc. other than as an integral part of a document or dataset;
- military insignia;
- identity documents;
- information subject to patents, trademarks, design rights, third party copyright (unless authorised), etc.

The attribution condition is couched in flexible terms so as to mitigate the problem of attribution stacking. In cases of data being drawn together from many different datasets, a simple generic statement will satisfy the licence terms.[43] Furthermore, if a derived dataset is released under CC BY version 4 or ODC-By, users complying with that licence's attribution requirement automatically satisfy those of the OGL.

A non-commercial variant was introduced in July 2011,[44] where commercial uses are understood to be 'primarily intended for or directed toward commercial advantage or private monetary compensation'. The current version retains some of the additional conditions from OGL version 1 not present in version 2:

- the resource must not be used to mislead others; and
- use of the resource must not breach the Data Protection Act 1998 or the Privacy and Electronic Communications (EC Directive) Regulations 2003.

Notably, while the licence as a whole is not copyleft, the non-commercial aspect of it is. In other words, it requires that any derivations are released under a non-commercial licence.

Public domain

The most permissive way of releasing data is under a dedication to the public domain. This is where all copyright interests and database rights are waived, allowing the data to be used as freely as possible. Dedicating a work to the public domain is not as simple as it sounds, which is why Creative Commons and Open Data Commons have produced special tools for the purpose.

OGL at a glance

Good for

- UK public sector databases and datasets
- data to be used automatically

Watch out for

- attribution stacking if used with differently licensed data
- categories of data that cannot be licensed in this way
- ties to the UK legal context

[43] 'Contains public sector information licensed under the Open Government Licence v2.0.'

NCGL at a glance

Good for

- commercially valuable UK public sector databases and datasets
- data to be used automatically

Watch out for

- attribution stacking if used with differently licensed data
- restrictions on uses: only use with dual licensing
- categories of data that cannot be licensed in this way
- ties to the UK legal context

[44] Non-Commercial Government Licence for public sector information, URL: `http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/non-commercial-government-licence.htm`.

A machine-readable version of the Non-Commercial Government Licence is available at `http://reference.data.gov.uk/id/non-commercial-government-licence`.

Public domain at a glance

Good for

- most databases and datasets
- data to be used by anyone or any tool
- data to be used for any purpose

Watch out for

- lack of control over how database is reused
- lack of protection against unfair competition

Creative Commons Zero (CC0) is for dedicating works to the public domain.[45] It works on two levels: as a waiver of a person's rights to the work, and in case that is not effective, as an irrevocable, royalty-free and unconditional licence for anyone to use the work for any purpose. The rights waived include database rights, so CC0 is suitable for use with data.[46]

There is also the Creative Commons Public Domain Mark (CC PDM), a tool that anyone can use to assert that a work is already in the public domain.[47] The motivation for the tool is to allow public domain works to be more easily discovered and recognised as such,[48] but it should not be used for waiving rights.

The Open Data Commons Public Domain Dedication and Licence (PDDL) accomplishes much the same thing in much the same way as CC0, but is worded specifically in database terms.[49] (It should not be confused with the deprecated Creative Commons Public Domain Dedication and Certification [CC PDDC] tool.) The PDDL explicitly provides for a set of community norms to be associated with a database, such as the Open Data Commons Attribution-Sharealike Community Norms.[50] These express the same ideals as the corresponding licence, but in the form of a code of etiquette rather than a legal obligation.There is also the Open Data Commons Database Contents Licence (ODC-DbCL), which waives copyright for the contents of the database without affecting the copyright or database right of the database itself.[51]

Given that dedicating data to the public domain involves permanently relinquishing so many rights and protections, including protection against unfair competition, it is perhaps an unattractive option for data whose creators have yet to fully exploit them, either academically or commercially. Nevertheless, it does resolve many of the ambiguities surrounding data use and reuse – to which parts of a database copyright applies, the extent to which database rights apply, what constitutes fair or insubstantial use, what constitutes commercial use – and greatly simplifies integration with other data.

While community norms documents have no legal force, unlike copyright and licences, they can still be effective if the target community shares the values reflected and incorporates the norms into its governance mechanisms. The paradigmatic example is the prohibition of plagiarism, which as a community norm has arguably a greater moral force than copyright law.[52] In the data context, Polar Science is a field in which community norms are being used to ensure both high quality contributions and respectful reuse of data without resorting to legal measures.[53]

Multiple licensing

In cases where none of the above licences are entirely satisfactory, it may be possible to use a multiple licensing approach. This would allow recipients of the data to choose from a specified set the licence under which they use the data.

Multiple licensing is usually used in the open source software world to achieve one of two aims. The first is to control, rather than freely permit or forbid outright, use of the software in

[45] CC0, URL: http://creativecommons.org/publicdomain/zero/1.0/.

[46] Peters, D. (2009, March 11). Expanding the public domain: Part zero. Retrieved from http://creativecommons.org/weblog/entry/13304.

[47] CC Public Domain Mark, URL: http://creativecommons.org/publicdomain/mark/1.0/.

[48] Peters, D. (2010, October 11). Creative Commons launches Public Domain Mark: Europeana and Cultural Heritage Institutions lead early adoption. Retrieved from http://creativecommons.org/press-releases/entry/23755.

[49] PDDL, URL: http://opendatacommons.org/licenses/pddl/.

[50] ODC Attribution-Sharealike Community Norms, URL: http://opendatacommons.org/norms/odc-by-sa/.

[51] ODC-DbCL, URL: http://opendatacommons.org/licenses/dbcl/.

[52] Murray, L. J. (2008). Plagiarism and copyright infringement: The costs of confusion. In C. Eisner & M. Vicinus (Eds.), *Originality, imitation and plagiarism: Teaching writing in the digital age* (pp. 173–181). Ann Arbor, MI: University of Michigan Press.

[53] Appropriate Behavior when Contributing and Using PIC Data. (n.d.). Establishing the framework for the long-term stewardship of polar data and information. (n.d.). Retrieved from Polar Information Commons website: http://web.archive.org/web/20140720090800/http://www.polarcommons.org/ethics-and-norms-of-data-sharing.php.

commercial or proprietary applications, thereby providing a means of generating income from the open source code. The second is to resolve the compatibility problems that exist between copyleft licences.[54] In the language of the Creative Commons licences, it allows owners of source code to address the issues associated with the Non-Commercial and Share-Alike clauses, respectively.

In the first case, a typical scenario would be for the owners of the source code to release it under an open source licence with a strong copyleft clause, such as the GNU General Public Licence (GPL). At the same time, they offer the source code under an alternative licence without the copyleft clause, and charge a fee for the use of this less-demanding licence.[55] This dual licensing regime gives developers the choice of using the code for free in free, open source software, or paying a fee to use the code in closed source, possibly commercial software.

In the second case, the owners of the source code allow developers to use it under one of several open source licences, broadening the range of code with which it can be combined. For example, the source code of the SeaMonkey Internet application suite is triple-licensed under the Mozilla Public Licence (MPL), the GNU General Public Licence (GPL) and the GNU Lesser General Public Licence (LGPL).[56]

While multiple licensing can be a useful strategy, there are some issues that need to be borne in mind. The option to multiply license a dataset is certainly available to you if you hold all the rights that pertain to the dataset: that is, you hold rights over the dataset, and any aspect of the data for which you do not hold rights is public domain or exempt from copyright/database right restrictions. If this is not the case then what you can do is, of course, determined by the terms of the licensed data that contributes to your dataset:

- If the licence applies a copyleft condition to derived works/databases, you must respect that and license the derived dataset in the same way.
- If the licence applies a non-commercial condition to uses of the licensed data, then you should not charge others for any of the licences under which you release your derived dataset, though this does not prevent you using multiple licensing as a compatibility strategy.

In any event, whenever licensing a dataset containing data licensed to you, you should be careful not to claim rights you do not hold.

Multiple licensing works both ways, of course. If the ability to license your derived dataset as you please is important to you, you may be able to negotiate a special licence or contractual arrangement with the other rights holders that allows you to do this, in which case the rights holders are setting up a multiple licensing regime of their own. Another, more extreme, possibility is to negotiate a rights assignment.[57,58]

By way of illustration, a dual licensing model working within these constraints is shown in Figure 1. This model was devised with software development in mind, though it could be applied to situations where a data resource is expanded by many contributors over time.

[54] Blanco, E. (2012, September 9). Dual-licensing as a business model. Retrieved from OSS Watch website: http://oss-watch.ac.uk/resources/duallicence2.

[55] Välimäki, M. (2003). Dual licensing in open source software industry. *Systemes d'Information et Management*, *8*(1), 63–75. Retrieved from http://ssrn.com/abstract=1261644

[56] SeaMonkey Legal Resources. (2012, May 7). Retrieved from SeaMonkey Project website: http://www.seamonkey-project.org/legal/.

[57] Meeker, H. (2005, April 6). Dual-licensing open source business models. Retrieved from http://linux.sys-con.com/node/49061/print.

[58] When a Company Asks For Your Copyright. (2010, October 3). Retrieved from GNU Project website: http://www.gnu.org/philosophy/assigning-copyright.html.

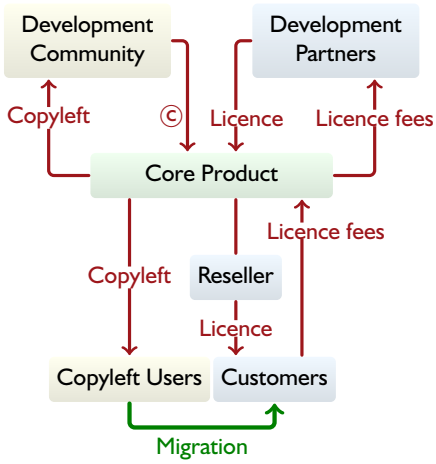


Figure 1: Licence streams of a core product in a simplified dual licensing model (adapted from Välimäki, 2003).[55]

Mechanisms for licensing data

Once you have decided on a suitable licence, all that remains is to attach that licence to the data. There are a few different ways of doing this, but mostly they involve a *statement* that the data is released under a particular licence or public domain dedication, and a mechanism for retrieving the *full text* of the licence itself. As an example, the suggested text for attaching the Open Data Commons PDDL to a database is as follows.

> [This database is/These data are/⟨*name of dataset*⟩ is] made available under the Public Domain Dedication and License v1.0 whose full text can be found at: `http://opendatacommons.org/licenses/pddl/1.0/`

The rights statement should be displayed prominently, so that any user of the data will realise that they are licensed or public domain. It is important to note, though, that the first inspection of the data might be done by an automated tool rather than a human. CrystalEye,[59] for example, is a database of crystal structures compiled by automatically parsing journal articles and other data sources. The problem for such efforts comes when the tool has to review the IPR status of a data source, examine any available licence terms, and decide whether to accept them. There are three possible ways to overcome this difficulty:

1. a human could review each data source before letting the tool use it;
2. a human could decide in advance under which licences the tool would be allowed to use data, and the data provider could label the data source in such a way that a tool could tell under what licence it is released;
3. tool authors and data providers could agree a common vocabulary for describing the capabilities of tools, and data providers could associate with the data a machine-readable list of operations that are, or are not, permitted.

The first of these is not scalable. The third requires extensive co-ordination and places limits on the capabilities an automated tool can have, but once set up requires very little human intervention. On a technical level it can be achieved through use of a Rights Expression Language such as MPEG-21 REL,[60] Open Digital Rights Language,[61] or METSRights.[62] Permissions and restrictions written in such a language represent an arrangement in their own right: strictly speaking they can only be used as an alternative to, or replacement for, an actual licence, not as a machine-actionable 'explanation' of one. The exception to this is the Creative Commons Rights Expression Language, which delegates the precise definition of its terms to the respective full legal codes of the Creative Commons licences.[63,64]

The second option is a compromise between the other two; it only works well when data providers use standard licences identified by standard URLs. For example, the machine-readable equivalent of the ODC PDDL statement above would be a Resource Description Framework (RDF) triple such as that shown in Figure 2.[65]

[59] CrystalEye Website, URL: `http://wwmm.ch.cam.ac.uk/crystaleye/`.

[60] ISO/IEC 21000-5:2004. *Information technology – Multimedia framework (MPEG-21) – Part 5: Rights Expression Language*. International Organization for Standardization.

[61] ODRL Community Group, URL: `http://www.w3.org/community/odrl/`.

[62] METSRights schema, URL: `http://www.loc.gov/standards/rights/METSRights.xsd`.

[63] Abelson, H., Adida, B., Linksvayer, M. & Yergler, N. (2008, March 3). *ccREL: The Creative Commons Rights Expression Language*. Version 1.0. Creative Commons. Retrieved from `http://wiki.creativecommons.org/images/d/d6/Ccrel-1.0.pdf`.

[64] CC REL by Example. (n.d.). Retrieved from Creative Commons website: `http://labs.creativecommons.org/2011/ccrel-guide/`.

[65] Manola, F. & Miller, E. (Eds.). (2004, February 10). *RDF primer*. W3C Recommendation. W3C. retrieved from `http://www.w3.org/TR/rdf-primer/`.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="" xmlns:dc="http://purl.org/dc/terms/">
    <dc:license rdf:resource="http://opendatacommons.org/licenses/pddl/1.0/"/>
  </rdf:Description>
</rdf:RDF>
```

Figure 2: A rights statement encoded in RDF/XML. Note that the `rdf:about` attribute should identify the data to which the statement applies. In the context of an XMP packet, this attribute is left blank to identify the resource in which the packet is embedded (*Extensible Metadata Platform (XMP) specification, part 1: Data model, serialization, and core properties*. San Jose, CA: Adobe Systems. Retrieved from `http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf`).

Again, this should be made available somewhere the tool would look when downloading the data, such as within a dataset catalogue record or landing page. If possible you should also include the rights statement within each data file – the following list indicates how this may be done for some common data formats:

XML Find a point in the document at which arbitrary XML can be embedded and insert an RDF/XML block similar to that shown in Figure 2.

MS Excel Add the human-readable statement to the Comments document property.

MS Access Add the human-readable statement to the Comments database property.

XHTML[66] Add the attributes `version="XHTML+RDFa 1.0"` and `xmlns:dc="http://purl.org/dc/terms/"` to the root `<html>` element. Add the human-readable statement somewhere in the document, marking up the link to the full licence text as an `<a>` element with the attribute `rel="dc:license"`.

[66] Adida, B. & Birbeck, M. (Eds.). (2008, October 14). *RDFa primer: Bridging the human and data Webs*. W3C Working Group Note. W3C. retrieved from `http://www.w3.org/TR/xhtml-rdfa-primer/`

Failing that, you should incorporate the rights statement when packaging data; indeed, it is good practice to do this anyway. The following table shows where the statement should be added for some common packaging standards. In most cases, the insertion points specified permit arbitrary XML to be included; the simplest option is therefore to insert an RDF/XML statement like that in Figure 2 within the specified element, though in future it may be possible to include an XHTML/RDFa fragment instead, along the lines of the XHTML method given in the above list.

METS[67] In the manifest file, add the rights statement (or a link to it) to the `<rightsMD>` element in the Administrative Metadata section.

[67] METS Website, `URL: http://www.loc.gov/standards/mets/`.

METS+METSRights[68] Within the `<rightsMD>` element in the Administrative Metadata section of the manifest file, add the hierarchy `<mdWrap>` › `<xmlData>`. Within that, add a `<mr:RightsDeclarationMD>` element with its `RIGHTSCATEGORY` attribute set correctly. Within that, add a `<mr:RightsDeclaration>` element containing the (plain text) human-readable rights statement; you should also add a `<mr:RightsHolder>` element.

[68] METSRights schema, `URL: http://www.loc.gov/standards/rights/METSRights.xsd`.

METS+MODS[69] In the manifest file, add the rights statement (or a link to it) to the `<mods:accessCondition>` element in the Descriptive Metadata section.

DDI[70] Add the (plain text) human readable rights statement to `<Collection>` › `<DefaultAccess>` › `<AccessConditions>`.

XFDU[71] In the Metadata section of the manifest file, add a `<metadataObject>` element with attributes `category="PDI"`, `classification="OTHER"` and `otherClass="ACCESS RIGHTS"`. Within that, add a `<metadataWrap>` element with attribute `textInfo="license"` or `textInfo="Public Domain declaration"`. Within that, add the rights statement within an `<xmlData>` element. To link to the rights statement instead, use the `<dataObjectPointer>` element (if it is in the XFDU Package Interchange File) or the `<metadataReference>` element (if elsewhere) instead of the `<metadataWrap>` element.

MPEG-21[72] In the DIDL file, within the `<Item>` element containing the data, add a `<Description>` element, and within that, a `<Statement>` element with the attribute `mimeType="text/xml"`. Within that, add an `<r:license>` element with the attribute `xmlns:r="urn:mpeg:mpeg21:2003:01-REL-R-NS"`. Within that, add an `<r:otherInfo>` element and to that add the rights statement (or a link to it).

IMS CP[73] In the manifest file, add the rights statement to the `<metadata>` element directly within the `<resource>` element containing the data.

If the data are to be packaged informally (in a ZIP or TAR file, or an ordinary directory, for example) the rights statement should be included in an obvious introductory document, such as a `readme.txt` file, at the top level of the directory structure.

In addition to these methods, it is also a good idea to ensure the rights statement is clearly displayed on pages from which the data may be downloaded. You might consider introducing a click-through notice, so that whenever someone requests the data, they are asked to assent to the licence terms before the transfer will proceed, but bear in mind this interferes with the ability of automated tools to access the data.

The example rights statements shown above both use URLs to specify the full legal text of the licence, but there is a question as to whether they should use the canonical URL for the licence, or point to a file within the package that contains the full text. The latter option is legally more robust, but canonical URLs have the advantage of being easier for automated tools to recognise. If you do include a copy of the licence with your data, it is customary to include it in a file named 'license' at the top level of the directory structure.

[69] MODS Website, URL: `http://www.loc.gov/standards/mods/`.

[70] DDI Website, URL: `http://www.ddialliance.org/`.

[71] XFDU Website, URL: `http://sindbad.gsfc.nasa.gov/xfdu/`.

[72] Bekaert, J., Hochstenbach, P. & Van de Sompel, H. (2003, November). Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, *9*(11). `doi:10.1045/november2003-bekaert`

[73] IMS Content Packaging Website, URL: `http://www.imsglobal.org/content/packaging/`.

Where a signed licensing agreement is used instead of an open-ended licence, it is less critical for data and data packages to be marked up with licensing information as the licensee's data management regime should enforce compliance with the agreement.

Licensing related information

If released data are to be as useful as possible, they need to be supported by additional information. A comprehensive set of such information might include[74]

- details of how the data have been encoded (database structures, file formats);
- a list of software known to work with the data and their supporting information;
- indications of how the data relate to other data assets;
- administrative information (identifiers, checksums);
- explanations of what the data represent (e.g. for sensor data, what the sensor was measuring and in what units);
- the processing history of the data (how they were generated and subsequently transformed, when and by whom);
- a narrative describing the context (why the data were generated/collected, what methodology was used and why).

The last three types of information are particularly important for users as they interpret the data, and determine whether and how they can be integrated with other data.

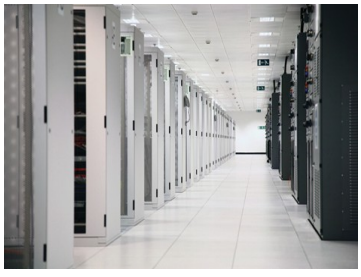
If any of this information exists in the form of further datasets, it should be released under the same licence or dedication as the main data, unless there is a compelling reason to do otherwise. This helps both parties to avoid confusion, and reduces the likelihood of data becoming separated from the supporting data on which they rely.

For information in the form of documents, it is not so critical to apply a licence, as there are long-established community norms for citing, quoting from and paraphrasing earlier written works. Having said that, applying a licence may (depending on the one you choose) provide users of the data with more flexibility with regards redistributing your documentation with their derivative datasets, or quoting substantial portions of your documentation within their own. If you do license your documentation, choose a licence that reflects how you want it to be used. As this may be quite different to your intentions for the data, you need not use the same licence for both.

[74] Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)*. Magenta Book. Also published as ISO 14721:2012. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf.

'We believe the concept of open data . . . goes beyond making data freely accessible. Data should also be free to distribute, copy, re-format, and integrate into new research, without legal impediments. . . . Therefore, to eliminate potential legal impediments to integration and reuse of data, specifically, and to help enable long-term interoperability of data we believe an appropriate licence or waiver specific to data should be applied, and made explicit by the authors and publishers.'

– BioMed Central's Position Statement on Open Data (Draft). (2010, September 2). Retrieved from BioMed Central Blog website: http://blogs.biomedcentral.com/bmcblog/files/2010/09/opendatastatementdraft.pdf



Acknowledgements

Thank you to Margaret Henty (ANDS), Jason Miles-Campbell (JISC Legal), and Angus Whyte and Lorna Brown (DCC) for helpful comments.

Further information

Three other DCC guides, each by Mags McGinley, cover this topic:

Awareness Level: *Legal Watch: Creative Commons licensing*

Awareness Level: *Legal Watch: IPR in databases*

Awareness Level: *Legal Watch: Science Commons*

Barlas, C. (2006, July). *Digital Rights Expression Languages (DRELs)*. London: JISC. Retrieved from http://www.webarchive.org.uk/wayback/archive/20130607115257/http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0603.aspx

Guibault, L. & Wiebe, A. (Eds.). (2013). *Safe to be open: Study on the protection of research data and recommendations for access and usage*. Universitätsverlag Göttingen. Retrieved from http://webdoc.sub.gwdg.de/univerlag/2013/legalstudy.pdf

Harris, L. E. (2009). *Licensing digital content: A practical guide for librarians* (2nd ed.). Chicago, IL: American Library Association.

Jasserand, C. (2011). Creative Commons licences and design: Are the two compatible? *JIPITEC*, *2*, 131–142. Retrieved from http://nbn-resolving.de/urn:nbn:de:0009-29-30856

Korn, N. & Oppenheim, C. (2011, June). *Licensing open data: A practical guide*. London: HEFCE and JISC. Retrieved from http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

Korn, N., Oppenheim, C. & Duncan, C. (2007, May). *IPR and licensing issues in derived data*. London: JISC. Retrieved from http://www.jisc.ac.uk/media/documents/projects/iprinderiveddatareport.pdf

Korn, N., Oppenheim, C. & Picciotto, S. (2007, May). *Other types of IPR and their impact on JISC projects*. London: JISC. Retrieved from http://www.jisc.ac.uk/media/documents/projects/othertypesofip.pdf

Murray-Rust, P., Neylon, C., Pollock, R. & Wilbanks, J. (2010, February 19). Panton principles for open data in science. Retrieved from http://pantonprinciples.org/

Data Re-Use and Licensing Frameworks. (n.d.). Retrieved from Australian National Data Service website: http://www.ands.org.au/publishing/licensing.html

WIPO Lex. (n.d.). Retrieved from World Intellectual Property Organization website: http://www.wipo.int/wipolex/en/. (Database of national intellectual property laws and treaties.)

Creative Commons licences. (2009, March). London: JISC. Retrieved from http://www.jisc.ac.uk/publications/briefingpapers/2009/bpcreativecommons.aspx

Diagnostic Tools. (2010). Retrieved from Open Educational Resources Intellectual Property Rights Support Project website: http://www.web2rights.com/OERIPRSupport/diagnostics.html

Starter Pack. (2010). Retrieved from Open Educational Resources Intellectual Property Rights Support Project website: http://www.web2rights.com/OERIPRSupport/starter.html

Pollock, R., Gray, J. et al. (n.d.). Guide to open data licensing. Retrieved from Open Definition website: http://opendefinition.org/guide/data/

Please cite as: Ball, A. (2014). 'How to License Research Data'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Follow the DCC on Twitter: @digitalcuration, #ukdcc

Revised: 17 July 2014