How to License Research Data

Alex Ball (DCC)

Digital Curation Centre, 2012.
Licensed under Creative Commons Attribution 2.5 Scotland:
http://creativecommons.org/licenses/by/2.5/scotland/
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. 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.

There are many additional reasons why releasing data can be in a researcher’s interests. 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 but also economically and socially.

There may also be circumstances in which researchers find themselves obliged to release data. It is possible that research data held by a UK university may be the subject of a request made under the provisions of the Freedom of Information Act 2000 (FoIA), the Environmental Information Regulations 2004, or their Scottish counterparts and therefore have to be released, even if findings


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


based on the data have not yet been published. This is true even if data were generated and are owned by a third party, such as a university in another country. There are provisions within the FoIA for exempting information that has already been scheduled for release at the time a request is made, implying that it may be possible to avoid the premature release of research data by adopting a policy for releasing it.

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 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.

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. 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. 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 provides a level of consistency among its signatories – 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.

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.

### 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.

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 they did not previously have.

Data automation

It is not only the human reusers of data that need guidance. One of the benefits of working with data is the scope for automation. CrystalEye, 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. The second option is a compromise between the other two, and only works well when data providers use standard licences, and use standard URLs to identify them; methods for doing this are discussed under ‘Mechanisms for licensing data’ below.

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

---

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


2. CrystalEye Website, URL: http://wwmm.ch.cam.ac.uk/crystaleye/.

3. ‘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.’

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. On the other hand, it also maintains some public domain genome sequences as part of the Multinational Brassica Genome Project.

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. In turn, the UKDA makes the data available under one of two bespoke licences: a Special Licence for sensitive data, or an End User Licence for all other data. Similarly, researchers depositing data with the Archaeology Data Service (ADS) are required to sign a deposit licence. Those using data hosted by the ADS do so under both a brief licence and a common access agreement.

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.

**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 they have to work with and aids interoperability. There are circumstances, though, in which it might be worth writing a custom licence: where the data have significant commercial value, 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 texts or templates you could use, or may even offer to write the licence for you. As an example, the Augmented Multi-Party Interaction (AMI) Project at the University of Edinburgh released its AMI Meeting Corpus under bespoke licences written by the university’s Edinburgh Research and Innovation unit.

**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 apart from the Restrictive Licence, each of these licences can be terminated only by expiry of the licensor’s IPR or, for a particular licensee, through breach of terms.

13 Rothamsted Research Website, URL: http://www.rothamsted.ac.uk/.
14 Multinational Brassica Genome Project Website, URL: http://www.brassica.info/.
20 The project uses a dual licensing scheme, with a free, non-commercial licence based on Creative Commons and chargeable commercial licence (see ‘Multiple licensing’ below). AMI Meeting Corpus Website, URL: http://corpus.amiproject.org.
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.21 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.22 While 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, each of which includes the Attribution condition. This allows others to copy, distribute, display, and perform the work as long as the creator is given due credit. The licensor can (and should) specify the way in which credit is given.

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.

Including the Non-Commercial condition in the licence means that the licensee cannot use the work for commercial purposes; otherwise commercial use is permitted.

The Share Alike condition inserts a strong copyleft clause into the licence. This means that all derivative works must be released under the same licence as the original work.23 Licences without this clause permit derivative works to be released under a different licence.

Finally, including the No Derivatives condition means that the licensee is forbidden from altering, transforming or building upon the work; otherwise this is allowed. When the No Derivatives condition is in force, there can be no derivative works to which to apply the Share Alike condition, so the two are mutually exclusive.

The six permutations are therefore

1. Attribution (CC BY);24
2. Attribution Share Alike (CC BY-SA);25
3. Attribution Non Derivatives (CC BY-ND);26
4. Attribution Non-Commercial (CC BY-NC);27
5. Attribution Non-Commercial Share Alike (CC BY-NC-SA);28
6. Attribution Non-Commercial No Derivatives (CC BY-NC-ND).29

As mentioned above, Creative Commons licences are not specifically aimed at data, and their use in this context is not without difficulty. A quite general problem is that the licences are aimed at homogeneous works, and do not cater for the complexities of data: specifically, the distinction between the individual data themselves and the collection/database, and the distinction between using data as part of a new collection/database and using them to generate content (graphs, models, maps, etc.).

Creative Commons

Creative Commons Website, URL: http://creativecommons.org/.

RDF and rights expression languages are discussed under “Mechanisms for licensing data” below.

Creative Commons at a glance

<table>
<thead>
<tr>
<th>Good for</th>
<th>Watch out for</th>
</tr>
</thead>
<tbody>
<tr>
<td>- very simple, factual datasets</td>
<td>- attribution stacking</td>
</tr>
<tr>
<td>- data to be used automatically</td>
<td>- the NC condition: only use with dual licensing</td>
</tr>
<tr>
<td>- the SA condition as it reduces interoperability</td>
<td>- the ND condition as it severely restricts reuse</td>
</tr>
</tbody>
</table>

23 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.

21 Creative Commons Website, URL: http://creativecommons.org/.

22 RDF and rights expression languages are discussed under ‘Mechanisms for licensing data’ below.

24 CC BY, URL: http://creativecommons.org/licenses/by/3.0.

25 CC BY-SA, URL: http://creativecommons.org/licenses/by-sa/3.0.

26 CC BY-ND, URL: http://creativecommons.org/licenses/by-nd/3.0.

27 CC BY-NC, URL: http://creativecommons.org/licenses/by-nc/3.0.

28 CC BY-NC-SA, URL: http://creativecommons.org/licenses/by-nc-sa/3.0.

29 CC BY-NC-ND, URL: http://creativecommons.org/licenses/by-nc-nd/3.0.
The Attribution condition should not be problematic if the data are to be combined with data from only a small number of other sets. At the other extreme, it should not be a problem for a dataset constructed from insubstantial extracts from a large number of other datasets, due to copyright/database right exemptions, though having to judge whether a use is substantial – and hence whether an exemption applies – will likely be offputting to reusers. Between these two extremes, compiling a dataset from many others is likely to be unfeasible due to the administrative burden of crediting each individual contributor to the superset in the manner of their choosing. This problem is sometimes known as ‘attribution stacking’.

Similarly, as the Share Alike condition requires the licensee to release any derived dataset under the same licence (and only that licence), it prevents 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. Note that this is true even within Creative Commons: a derived dataset cannot contain both CC BY-SA-licensed data and CC BY-NC-SA-licensed data. Having said that, some copyleft licences 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.

The No Derivatives condition requires that the licensed data is used ‘as-is’, though precisely what this means in practice is a matter of some debate. It would likely restrict the use of the data to such cases as checking that subsets of data within the licensed set derive from each other as claimed. Most substantive types of data reuse would be forbidden, however.

The Non-Commercial condition does not cause any problems as regards combining data, but may have wider implications than intended due to the ambiguity of what constitutes a commercial use. The advice from Creative Commons is that commercial means ‘primarily for monetary compensation or financial gain’. 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 may even preclude the data being used in support of works that are sold (such as journal articles) even if the author does not benefit financially. The Non-Commercial condition is often used as part of a dual-licensing regime (see ‘Multiple licensing’, below).

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 was set up in 2007 to develop a successor to the Talis Community Licence (TCL). 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

---


32 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.](http://www.gnu.org/licenses/license-list.html)).

33 ‘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.’

---

34 Open Data Commons Website, URL: [http://opendatacommons.org/](http://opendatacommons.org/).

35 TCL, URL: [http://tdnarchive.capita-libraries.co.uk/tcl](http://tdnarchive.capita-libraries.co.uk/tcl).
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.\(^{36}\) If content is generated from the data, that
content should include or accompany a notice explaining that
the database was used in its creation. 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.\(^{37}\) 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 permits compatible licences to be used instead, though,
as an aid to integrability. 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. The Open Data Commons Database Contents Licence
(ODC-DbCL) may be used in conjunction with the ODbL to waive
copyright for the contents of the database.\(^{38}\)

Being written in database terms, these licences are better
suited to research data than the CC equivalents; the ODC-ODbL
copyleft condition is also more flexible than CC’s Share Alike. Note,
however, that these licences are specific about how attribution
should be accomplished, whereas the CC licences leave it to the
licensor to specify an appropriate mechanism.

Open Government Licence

The Open Government Licence (OGL) was released as part
of the UK Government Licensing Framework in September
2010.\(^{39}\) It is intended for UK public sector 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 inclusion of UK
legislation among its terms.

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. The licence contains
some additional conditions, however:

- derivative works must not be represented as official, or as
  endorsed by the licensor;
- the resource must not be used in a way that misrepresents
  the licensor or the resource;
- the resource must not be used to mislead others; and

---

\(^{36}\) ODC-By, URL: http://opendatacommons.org/licenses/by/.

\(^{37}\) ODC-ODbL, URL: http://opendatacommons.org/licenses/odbl/.

\(^{38}\) ODC-DbCL, URL: http://opendatacommons.org/licenses/dbcl/.

\(^{39}\) Open Government Licence for public sector information, URL:
• use of the resource must not breach the Data Protection Act 1998 or the Privacy and Electronic Communications (EC Directive) Regulations 2003.

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, e.g. ‘Contains public sector information licensed under the Open Government Licence v1.0.’

A non-commercial variant was introduced in July 2011. 40

Restrictive Licence

The Australian Governments Open Access and Licensing Framework (AusGOAL) was launched in 2011. It encourages suppliers of publicly funded information to use one of seven different licences (an eighth is recommended for software). The first six are the Australian Creative Commons licences, while the seventh – the Restrictive Licence (RL) – is in fact a template for constructing a bespoke licensing agreement. 41

Unlike the other standard licences discussed here, the RL is an agreement that both licensor and licensee have to sign. As the name implies, it grants very few permissions by default. Indeed, the standard text 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. The standard text is, however, made adaptable through a set of schedules that allows the licensor to fix the term of the licence, charge a fee (or not), restrict usage geographically (or not), and specify different copying/distribution permissions for confidential/personal data and other data. Notably, the licence does not directly forbid modifications to the data, though some of the options with regard to copying would indirectly forbid them. Attribution is required by the licence inasmuch as it does not attempt to alter the effect of the licensor’s moral rights under Australian law. As well as the options provided, the schedules also provide scope for the licensor to apply additional usage conditions or grant additional permissions.

AusGOAL is the result of expanding Queensland’s Government Information Licensing Framework (GILF) to the whole of Australia; GILF itself had been developed over a period of about five years by QUT and the Queensland Government, with support from


Restrictive Licence at a glance

Good for
• Australian public sector databases and datasets
• confidential or sensitive data
• valuable information

Watch out for
• attribution stacking
• default restriction to non-commercial uses
• options to restrict copying and redistribution
• ties to the Australian legal context

the Cooperative Research Centre for Spatial Information amongst others. The Australian National Data Service is a strategic partner of AusGOAL, and provides support for implementing it in the research context.42

Design Science Licence

The Design Science Licence (DSL) was written by Michael Stutz between 1999 and 2001.43 It is focused on content with a source/rendering separation (e.g. software, \LaTeX documents) although it indicates how it might be used with images and audio files.

It is comparable with the CC BY-SA licence, requiring that copyright and licensing information is included with all redistributed copies, and that derived works indicate authorship of the parts of a new work that derive from the original. The copyleft condition makes no compatibility concessions, but does not extend to entire compilations of which the work forms a part. In addition, derived works should be given a new title to prevent confusion with the original. The DSL is also an open source licence, requiring that anyone receiving the rendered/compiled/compressed version must also be given access to the source data.

While the source/rendering distinction can be understood in the data context to refer to source data and visualisations such as graphs or maps, the licence does not take special notice of the distinction between a database and the data it contains. It should also be noted that anyone using the data to, say, produce a graph would also take on a responsibility to redistribute the entire set of licensed data.

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.

Creative Commons Zero (CC0) is the Creative Commons tool for dedicating works to the public domain.44 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.

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.45 The motivation for the tool is to allow public domain works to be more easily discovered and recognised as such,46 but it should not be used for waiving rights. CC0 and CC PDM together replace the Creative Commons Public Domain Dedication and Certification (CC PDDC) tool, which is now deprecated and should no longer be used.47

The Open Data Commons Public Domain Dedication and Licence (PDDL) accomplishes much the same thing in much the


43 DSL, URL: http://www.gnu.org/licenses/dsl.html.

44 CC0, URL: http://creativecommons.org/publicdomain/zero/1.0/.

45 CC Public Domain Mark, URL: http://creativecommons.org/publicdomain/mark/1.0/.


same way as CC0, but worded specifically in database terms.48

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.49 These express the same ideals as the corresponding licence, but in the form of a code of etiquette rather than a legal obligation.50

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, 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.51 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.52

**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 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.53 In terms of the discussion of Creative Commons licences above, 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.54 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

48 PDDL, URL: http://opendatacommons.org/licenses/pddl/.

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

50 Creative Commons is also working on a set of community norms that could be associated with public domain works. (Public domain guidelines. [2010, October 8]. Retrieved 18 November 2010, from Creative Commons: http://wiki.creativecommons.org/Public_Domain_Guidelines)


example, core Mozilla project source code can be licensed under the Mozilla Public Licence (MPL), the GNU General Public Licence (GPL) or the GNU Lesser General Public Licence (LGPL).\textsuperscript{55}

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\textsuperscript{56, 57} 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.

**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 they all 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/\textit{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 machine-readable equivalent of this would be a Resource Description Framework (RDF) statement such as the one in Figure 2.\textsuperscript{58}

The rights statement should be displayed prominently, so that any user of the data (whether human or an automated tool) will realise that they are licensed or public domain. If possible


you should include the rights statement within each data file; the following table 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 the one 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** Add the attributes `version="XHTML+RDFa 1.0"` and `xmlns:d="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"`.

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 table.

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

**METSRights** 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.

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

---

**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 8 November 2010, from http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf).


60 METS Website, URL: http://www.loc.gov/standards/mets/.

61 METSRights schema, URL: http://www.loc.gov/standards/rights/METSRights.xsd.

62 MODS Website, URL: http://www.loc.gov/standards/mods/.
Add the (plain text) human readable rights statement to `<Collection> > <DefaultAccess> > <AccessConditions>`.

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>` elements 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.

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).

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.

Another option that is not explored here is using a Rights Expression Language such as MPEG-21 REL,

63 DDI Website, URL: http://www.ddialliance.org/
64 XFDU Website, URL: http://sindbad.gsfc.nasa.gov/xfdualliance.org/
66 IMS Content Packaging Website, URL: http://www.imsglobal.org/content-packaging/
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.

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:

- 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.

---

68 ODRL Initiative Website, URL: http://www.odrl.net/.

69 METSRights schema, URL: http://www.loc.gov/standards/rights/METSRights.xsd.


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


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


Follow the DCC on Twitter: @digitalcuration, #ukdcc

Revised: 20 June 2012