Licensing Open Data: Resources and Practices

Research into principles and common practices in licensing open government data. Includes description of major licensing options.

Author: Peri Weisberg

Table of Contents

Awesome Resources

- Licensing Open Data: A Practical Guide: Global perspective on what should be licensed and why; strengths and weaknesses of CC, ODB, and GNU licenses, and important considerations for choosing a license.
- Guide to Open Data Licensing from opendefinition.org: Great explanation of rights in data in various jurisdiction; explanation of what makes licenses open. Opendefinition.org also provides a list of conformant licenses.
- Open Data Commons: Comprehensive FAQ about openness and data.
- Creative Commons: Comprehensive and practical FAQ for licensors; plain-language explanations of each license and how to apply it.

Why license?

From the Open Definition's Guide to Open Data Licensing: In most jurisdictions there are intellectual property rights in data that prevent third-parties from using, reusing and redistributing data without explicit permission. Even in places where the existence of rights is uncertain, it is important to apply a license simply for the sake of clarity. Thus, if you are planning to make your data available you should put a license on it — and if you want your data to be open this is even more important.

The interests of open data are best served by standard licenses, rather than bespoke agreements often developed for proprietary services. Use of standard licenses enhances clarity about permissible uses, and promotes interoperability by making it simpler to blend data from different sources.

Intellectual Property in Data

In the U.S., copyright applies to databases if the compilation of data involves some creative expression. Precisely what constitutes creative expression has never been decided. This makes the copyright status of databases somewhat more uncertain than other creative works, which are almost universally protected.

Some works of the federal government are excluded from copyright protection, but state and local data can easily have copyright, provided it meets the originality requirement. Copyright applies whether or not the content creator takes any action.

In the UK, Australia, and other common law countries, the "sweat of the brow" that goes into compiling data merits copyright protection on its own. EU countries also have a sui generis database right, a notion distinct from copyright but very comparable, that applies to databases that do not fall under copyright.

Finally, databases with no copyright protection in the U.S. may qualify for protection in other jurisdictions.

The lesson from all these complicated and ambiguous policies: uncertainty about the existence of copyright protections in data is all the more reason for creators to use an explicit license for any database intended to be open. As a coalition of open data advocates put it, "Data is more valuable when it is clear that there is a green light enabling reuse."

Licensing options

The standard licenses available for protecting data and databases demonstrate a range of openness. Some are conformant to the open definition at opendefinition.org; others are not. Within conformant licenses, there is a particular subset of recommended license. These are recommended by opendefinition.org because they're in wide use and re-usable by any entity.

Spreadsheet

- Open
  - Recommended
    - Creative Commons CCZero (CC0)
    - Creative Commons Attribution 4.0 (CC-BY-4.0)
    - Creative Commons Attribution (CC-BY)
    - Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)
    - Creative Commons Attribution Share-Alike (CC-BY-SA)

        - Open Data Commons Public Domain Dedication and Licence (PDDL)
        - Open Data Commons Open Database License (ODbL)
        - Open Data Commons Attribution License (ODC-BY)
        - Free Art License (FAL)
    - Little used or deprecated
        - GNU Free Documentation License (GNU FDL) - only open if amended slightly
        - MirOS License
        - Talis Community License
        - Against DRM
        - Design Science License
        - EFF Open Audio License
    - Non-reusable
        - UK Open Government Licence 2.0 (OGL-UK-2.0)
        - Open Government Licence – Canada 2.0 (OGL-Canada-2.0)
- Not Open
        - Creative Commons Attribution-NoDerivatives License (CC-BY-ND)
        - Creative Commons Attribution-NonCommercial-NoDerivatives License (CC-BY-NC-ND)
        - Creative Commons Attribution-Noncommercial (CC-BY-NC)
        - Creative Commons Attribution-Noncommercial-ShareAlike (CC-BY-NC-SA)
        - Creative Commons Attribution-Noncommercial-NoDerivatives (CC-BY-NC-ND)
        - Project Gutenberg License

I discuss the licenses from Creative Commons and Open Data Commons below. All of the licenses in these two groups share some basic characteristics:

- international (most current version of CC licenses only)
- non-exclusive
- irrevocable (but can be terminated when the licensee breaches terms)
- non-sublicensable
- disclaims accuracy and completeness, and limits liability

Creative Commons

Creative Commons licenses allow licensors to retain copyright while allowing others to copy, distribute, and make some uses of their work. The licenses are applicable to any type of copyrightable work. They include legal code (example), a summary appropriate for non-lawyers (example), and a machine-readable layer.

This may not always include collections of data or databases. From Creative Commons: “CC licenses are operative only when applied to material in which a copyright exists, and even then only when a particular use would otherwise not be permitted by copyright… CC licenses do not contractually impose restrictions on uses of a work where there is no underlying copyright. This

feature (and others) distinguish CC licenses from some other open licenses like the Open Data Commons' licenses, which are intended to impose contractual conditions and restrictions on the reuse of databases in jurisdictions where there is no underlying copyright or sui generis database right."

Types of CC licenses

All CC licenses require that the work is attributed to its author. The basic license, CC-BY, has this requirement only. On top of that, licensors can add the following requirements:

- Share-Alike (SA) - Requires that the work or its derivative is shared under the same type of license.
- Non-Commercial (NC) - Does not allow commercial uses of the work
- No Derivatives (ND) - Does not allow adaptations of the work to be shared

The ND and SA requirements cannot be combined, so there are six acceptable combinations.

Creative Commons also offers public domain tools (not technically licenses).

- CC0 allows the holder of copyright or database rights to waive all rights and place the work in the public domain. **Unlike the licenses above, Creative Commons encourages use of this tool for data and databases.**
- The Public Domain Mark identifies a work as free of known copyright restrictions.

Openness

Not all of the Creative Commons licenses are open. When a non-commercial or no-derivatives requirement is used, the license does not meet the standards of opendefinition.org.

Open Data Commons licenses

The Creative Commons licenses were not developed with data and databases in mind. The Open Data Commons licenses were developed in response to this need. Open Data Commons identifies four issues that make creative commons licenses (other than CC0) a poor choice for data:

1. The rights in data(bases) are often significantly different from those in content, both because of the existence of additional IP rights, such as the database right, and because normal copyright applies in a different, and usually 'lesser', fashion to data(bases) as compared to content.
2. In licensing data(bases) one may need to distinguish between the data(base) and its contents. For example, one may have a database consisting of images and wish to (or have to) license the images (the contents) separately from the database itself.
3. The distinction between the data(base) and material (content) generated from it ("produced works") — a distinction which is not relevant when licensing "content". For example, consider using a geospatial database to generate a map (an image). The map is distinct from the database and, as an image, is a classic piece of "content" but is has been generated from that database. This relationship is different from that between the

database and a derivative database (e.g. a database created by adding the locations of post offices to the original database).

4. The relationship and prominence of derivative works. Data(bases) are unlike content (but similar to code) in having a high level of reuse (as opposed to simple use or redistribution). For example, “mash-ups” are all about recombining and reusing data. This fact needs to be borne in mind when designing the licence and especial attention paid to the issue of reuse and derivative data(bases) — for example how must derivative material be made available when applying share-alike provisions.

The Open Data Commons licenses include language to address all of these issues. They also include a clause reserving the right to release the Database under different terms, or to stop distributing or making available the Database.

Types of licenses

- The Public Domain Dedication and License is a tool to puts all material in the public domain.
- The Attribution License (ODC-By) is an open license for data and databases that requires re-users attribute their work to the licensor.
- The Open Database License has the share-alike and attribution requirements of the CC-BY-SA. Like the GPL (or CC Attribution Share-Alike) it requires public reusers of data to share back changes (and attribute).

Openness

Each of these three licenses qualify as open according to opendefinition.org.

Considerations for open data

The report “Licensing Open Data: A Practical Guide” outlines additional considerations that organizations should consider. The following is lifted directly from that document:

- **Interoperability.** Compatibility issues can arise between CC licences. For example, whilst the CC BY SA licence is more open that the CC BY ND, it is less interoperable. Similarly, CC BY NC SA and CC BY SA licensed data can only be blended with themselves (and not each other) or with CC BY licensed data (or equivalent licensed data) or with data released under CC0.
- **Attribution Stacking.** This occurs when data licensed under of the CC licences is blended with similarly licensed data leading to the build up and impracticalities of required attribution information whenever the data is used or reused.
- **Irrevocability.** At a strategic level, committing to irrevocable terms raises issues of broader access and commercial goals for organisations. The use of [irrevocable] licences should be a policy decision and should form part of the overall strategic direction relating to rights management, use and exploitation where the full implications can be examined and understood.
- **Third party rights.** Standard licenses are usually not suitable where third party rights issues are present and require additional clearances. Most do not guarantee to provide

any information about what content does contain third party materials, nor any indemnities for the user in the case that they do – leaving the licensee taking all the risk.

In general, licensors should be aware that an attribution requirement can potentially lead to attribution stacking, and a share-alike requirement can potentially compromise interoperability.

Licensing elsewhere

In practice

DataSF's current catalog specified licensing in the "about" section of each dataset.

Creative Commons 1.0 Universal	348
Creative Commons Attribution 3.0 Unported	2
Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported	6
Creative Commons Attribution-Share Alike 3.0n Unported	83
Open Database License	1
Public Domain	145
None	296

I visited 16 other city portals to research their approaches to licensing. I visited between three and eight datasets per portal to get a sense of their approach. In almost all cases when a city or portal had a written policy regarding licensing, the practice did not match that policy precisely.

Lexington and Boston seem to have applied the most effort to clarifying licensing.

- In Lexington's portal, all datasets carry Open Data Commons Public Domain Dedication and Licence (PDDL), and badge is prominently displayed next to each.
- Part of Boston's terms of use involve the application of a bespoke license agreement that covers all of their data(the license terms are relatively flexible - the only clauses that I could identify as unique are those that restrict users from re-assigning or sublicensing, and place other restrictions on sharing data with unlicensed users. The city may also change the license with notice). However, it's not clear from their individual datasets that this license applies. Some datasets say nothing about licensing, and some use CC0.

Seattle, Salt Lake City, and Oakland are like San Francisco. Instead of a blanket license, their datasets are labeled with one of a handful of terms - usually "Public Domain," "CC0" or nothing at all.

Finally, eleven portals have no indication at all about licensing, either in their terms of use or in regards to a particular dataset.

- NYC

- Chicago
- Palo Alto
- Philadelphia
- Asheville
- Louisville
- Sacramento
- DC
- Austin
- Miami
- Houston

In policy

Sunlight Foundation collects data from 38 legislative or administrative policies; twenty have some provision referring to licensing or terms of use.

Of those with language regarding licenses, San Francisco's allows for the most licensing options. It calls for "reasonable, user-friendly" requirements. It is joined in doing so by Nashville, TN.

The most common policy approach is to mandate that data be "made open" or "made available without any registration or license requirement or restriction on use." Twelve policies use this language, which does not prescribe a specific licensing scheme but limits permissible licensing options to an open license, or no license at all. Six of these specify that an allowable license requirement would be an attribution or share-alike requirement.

Four policies specifically mandate an open license. Tulsa, OK, and San Mateo County, SA, simply direct the city to explore options for open licensing.

Finally, just one - Jackson, Michigan's - mandates data be entirely license-free via a CC0 waiver.

Principles

Many of the shared principles of open data licensing have been expressed above. Key shared principles include the need for a clear indicator of allowable uses of data, and the value of using standard licenses when possible.

Some but not all advocates go so far as to propose higher standards of openness than that put forth by opendefinition.org. A statement from a coalition of advocates including Govtrack, OKFN, Sunlight Foundation, EFF, and more asserts that "When [copyright] applies to public information, it should be waived so that the public can use it without restriction."

The authors of Licensing Open Data: A Practical Guide encourage organizations to consider the potential risks of the standard open licenses and to view openness as continuum rather than an absolute standard.

The same authors also point out potential for the attribution and share-alike requirements that are common to open license to hinder usability. Share-alike requirements can make it difficult to make derivative works using multiple datasets, which may compromise some of San Francisco's open data goals.