Author: Arnulf Christl

2012-01-23

Originally published in OSGeo FOSS4G Proceedings 2009

# Getting a Grip on Open Data

Abstract: This article introduces the core concepts of Open Data and highlights the different approaches required by specific communities. It advocates stronger licensing by arguing that this will result in better metadata and will ultimately generate more value.

## The Definition of Open Data

The Open Knowledge Foundation [1] defines Open Data as:

> "A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike." Source: [2]

Academics, politicians and the wider Internet community increasingly embrace Open Data. In academia, open source (software) has always been a natural prerequisite for any scientific work. Markus Neteler summarizes this in the statement:

> "In our view, the explicit use of Free and Open Source Software (FOSS) with availability of the code is essential for completely open science." Source: [3]

The same applies to data; if it is not openly available the validity of a proof cannot be confirmed or rejected. Unfortunately, even in the scientific domain, a tendency to close down access to data and use closed source algorithms to protect the intellectual property of researchers often prevails. This habit may be explained – but not excused – by the current scarcity of funding.

More recently governments have also started to publish more and more data under an Open Data regime.

The wider internet community ranges from those who champion proprietary software and data to groups like OpenStreetMap who rely on crowd sourcing and open availability of all their data.

This article aims to explain several aspects of how Open Data is defined by governments and open communities, what role licenses play, and why metadata is crucial for professional use of Open Data.

## The Open Definition

The Open Knowledge Foundation maintains the Open Definition [4] which states that:

> "A work is open if its manner of distribution satisfies the following conditions:"

1. Access
2. Redistribution
3. Reuse
4. Absence of Technological Restriction
5. Attribution
6. Integrity
7. No Discrimination Against Persons or Groups
8. No Discrimination Against Fields of Endeavor
9. Distribution of License
10. License Must Not Be Specific to a Package
11. License Must Not Restrict the Distribution of Other Works

Other bodies like the Sunlight Foundation [5] provide similar definitions in their "Ten Principles for

Opening up Government Information" [6].  The focus of these definitions is on the openness aspect of data. In contrast some businesses need to restrict access in order to preserve the value of their data by making the access to it exclusive.

## *Some Background Philosophy*

Open Data appears to contradict the perceived common mechanisms for generating value. Instead of reducing the availability of the product or service on the market and thereby raising its value, Open Data does the opposite. But Open Data is not a material 'good'; it is a virtual asset. As such it has quite a few similarities to Open Source software which has been shown to generate a lot of value **because** it is open. The secret to the success of 'open' digital assets is that they increase their value as they get shared.

The term "sharing" is not ideal as applied to virtual or digital goods because we normally apply it in a material context [7]. A cake is 'shared' by cutting it into 12 slices and eating it. It may have been shared between 12 people but it is now gone. Each slice may have some value and be bought but there is a finite amount of cake which must be consumed to be useful.

Data, on the other hand, can be multiplied by sharing. In fact it may even be enhanced, the quality may improve and new data may be derived from combining it with other data.

In short it can be said that data has little intrinsic value when it is not used (or shared). However, in use it will improve, multiply and generate new value.

## *Public Open Data*

Government open data sites are appearing everywhere. We believe this is a good thing. But the way in which it is done may give rise to much uncertainty over licensing and copyright issues and incompatibility between different Open Data licenses can render interoperating impossible. Technical interoperability can be achieved by adhering to standards such as OGC [8] and INSPIRE[9].  Legal interoperability is challenging in cases where government Open Data definition is less restrictive than Open Database License (ODbL) used, for example, by OpenStreetMap.  This means that the two sets of data cannot be readily combined to create a derivative and value added work.

A very thoughtful and comprehensive overview of government open data issues has been published by Socrata in the Open Data Field Guide [10].

## *Community Driven Open Data*

OpenStreetMap was originally described as a "crowd sourced". Later the term Volunteered Geographic Information (VGI ) was coined.  Both terms miss some 'community' aspects of such projects: all contributors earn the irrevocable right to reuse their collaborative work.

OpenStreetMap is a free, editable map of the world and is sometimes compared to Wikipedia which may be the closest comparison for free and openly available information. But OpenStreetMap offers more; it gives users the freedom to adapt the geographic data in many different ways to create new data and applications.

There are five core principles underlying the concept of community driven open data [11]:

- Spatial data is collected and maintained by a community;
- There is a clear and accepted license for reuse;
- Any individual or organization can use, modify and redistribute the data;
- Derived products (for example maps) can be copyrighted.
- The data stays open and publicly available.

OpenStreetMap, for example, was originally made available under the CreativeCommons Share Alike license designed for works of art and not well suited for data. In 2012 it resolved to relicense all data under the new ODbL and use the CreativeCommons license for the resulting map images.

## *Licensing*

Open Data does not mean "unlicenced" or "un-copyrighted". On the contrary, Open Data without the need for attribution loses a lot of its value because it lacks the link to its source. This may render copies of the data worthless for professional or scientific use. Even if data is licensed correctly, including reference to proper attribution, it may have experienced changes which – if not explicitly declared – can become a significant problem.

## The UK Open Government License

The UK Open Government License [12] was created and is maintained by The National Archives (TNA) [13]. It states that you are free to copy, publish, distribute and transmit the Information; adapt the Information; and exploit the Information commercially by combining it with other Information, or by including it in your own product or application. However you must abide by conditions: acknowledge the source of the Information and where possible, provide a link to this license; don't imply any official status or endorsement; don't mislead or misrepresent the Information; and don't breach Data Protection or Privacy laws.

The problem with this UK OGL is its failure to ensure that changes to the data have to be made visible. It is therefore difficult to assess whether any information derived from the source is accurate and based on the original data.

## The Open Database License (ODbL)

The Open Knowledge Foundation [14] has taken many of these aspects into account and created the ODbL [15]. This offers the same opportunities but also specifies that a derivative 'produced work' must be accompanied by information that shows (or enables the user to find) all of the changes that have been made to the original database.

This license is currently the best fit for Open Data and also complies with European Database law which is complicated in having to reconcile the different concepts of copyright, in common law jurisdictions, and author's rights in civil law jurisdictions.

One important aspect clarifying this passage and making it fit for reuse in a commercial context is covered in 4.6:

> *4.6 Access to Derivative Databases. If You Publicly Use a Derivative Database or a Produced Work from a Derivative Database, You must also offer to recipients of the Derivative Database or Produced Work a copy in a machine readable form of:*
>
> > *a. The entire Derivative Database; or*
> >
> > *b. A file containing all of the alterations made to the Database or the method of making the alterations to the Database (such as an algorithm), including any additional Contents, that make up all the differences between the Database and the Derivative Database.*

In the opinion of the author this paragraph is 'best practice' because it allows for the reconstruction of the whole process chain that resulted in the data at hand. It also ensures that OpenStreetMap data, for example, can be combined with private data for internal use and still provide full data protection and privacy.

## Case Study: SplashMaps

SplashMaps, the start-up maps-on-fabric business, has found that the Open Data licenses overcame significant barriers on the road to successful innovation. The Open Data from Ordnance Survey and other public authorities gave the company the confidence to develop their own products. David Overton, managing director said, "We are able to develop and test our products directly on the market without fear of contravening licenses. This means that we are enabled to pursue the propositions *we* choose, rather than those that are "allowed" under a typically restrictive license." It also means that there is less erosion of margins by royalty payments, so the company can plough back more of its proceeds into further developments.

SplashMaps have successfully combined OpenStreetMap data with complementary Ordnance Survey Open Data. This fills certain 'gaps' and enables tailoring of content to specific users' needs, while also sharing edits on the map with the rest of the community. The maps encourage feedback on the data and could even grow the community contributing to the OSM.

## Summary

Open Data is here to stay. Even in times of austerity with continuing privatization and persistent opposition to "opening up", Open Data has proven to be a valuable asset for many special interest communities. The broad application of 'openness' to data is still in its early days and there are still many relevant issues to be fully addressed. However it has already been conclusively demonstrated that strong licencing and copyright are not a contradiction to openness but actually support important aspects like metadata, and improve the quality, longevity and reliability of the original open data.

This article is licensed following the Copystraight [16] paradigm.

## References

[1] http://opendefinition.org
[2] http://okfn.org/opendata/
[3] Rocchini, D., Neteler, M. (2012): Let the four freedoms paradigm apply to ecology. Trends in Ecology and Evolution, , Vol 27(6), pp 310-311, doi:10.1016/j.tree.2012.03.009
[4] http://opendefinition.org/okd/
[5] http://sunlightfoundation.com/
[6] http://sunlightfoundation.com/policy/documents/ten-open-data-principles/
[7] http://arnulf.us/Share
[8] http://www.opengeospatial.org/standards
[9] http://inspire.jrc.ec.europa.eu/
[10] http://www.socrata.com/open-data-field-guide/
[11] http://metaspatial.net/conferences/gwf2013_opendata.html#/4/4
[12] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/1/open-government-licence.htm
[13] http://www.nationalarchives.gov.uk/
[14] http://okfn.org/
[15] http://opendatacommons.org/licenses/odbl/summary/
[16] http://arnulf.us/Copystraight