A Brief Introduction to

Open Data, Open Source Software and Collective Intelligence for Environmental Data Creators and Users

T. (Tom) Hengl, I. (Ichsani) Wheeler and R. A. (Bob) MacMillan

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Rationale

"Wikipedia is something special. It is like a library or a public park. It is like a temple for the mind, a place we can all go to think and learn."

Jimmy Wales, Wikipedia Founder

"Human beings have an innate inner drive to be autonomous, self-determined, and connected to one another. And when that drive is liberated, people achieve more and live richer lives."

Daniel H. Pink

Using the term "Open data" has become quite fashionable, but using it without clear specifications is misleading i.e. it can be considered just an empty catch phrase. Probably even worse is using the term "Open Science" — should not ALL science be open? Are we reinventing something that should be obvious from the start? Here are some common misconceptions about Open Data and Open Source software that we commonly encounter:

- *Open Data is any data that is available publicly via internet, for example Google maps.*
  NO: Google neither allows full access to the data (it only permits access to data visualizations e.g. PNG files) nor does Google make data it's available via any of the Open Data licenses. In fact, Google Maps terms of use are quite restrictive[1]. Open data is restricted to data registered under one of the commonly accepted licenses e.g. listed under https://opendefinition.org/licenses/

- *There is no need to define license terms for materials that are available publicly on web because I am a government employee.*
  NO: Not specifying a license for materials you publish is NOT good practice and does not make your materials more Open. Any material you put online and which is intended for longer-term use should have at least: (a) a license, (b) terms of use, and (c) contact person.

- *Once you decide to distribute Open data and make Open Source software it means that you are not interested in obtaining any commercial benefit from your work.*
  NO: In the next 2–5 years, FOSS will likely become the dominant platform for Big Data analytics, cloud computing and server OS. Even dominantly proprietary software development companies such as Microsoft are now investing aggressively in FOSS. Android

[1] https://maps.google.com/help/terms_maps.html

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

is currently the biggest Open Source software project in the business world but there are many new examples.

- *Non-Commercial licenses e.g. CC-NC help promoting Open Data.*
  NO: NC licenses are not Open Data licenses at all. In fact, an NC license indicates that the data producers likely intend to restrict uses in order to commercialize the data.

- *Businesses are not interested in Open Data because they are not able to make a profit from it.*
  NO: Currently many businesses have managed to greatly increase their profits by releasing large amounts of their data and software as Open Data / Open Source (e.g. Android).

With these common misconceptions in mind, this guide tries to clarify some key aspects of Open Data, Open Source Software and Crowdsourcing using examples of projects and businesses. It aims to help you understand and appreciate the complexity of Open Data, Open Source software and Open Access publications. It was specifically written for producers and users of environmental data, however, the guide will likely be useful to any producers and users of data .

This guide aims to help you:

1. Understand why releasing your data as Open Data is important and what the benefits of Open Data are,
2. Understand why releasing your code as Open Source software is important, and how this can be considered a positive business model for your company or organization,
3. Select the most optimal data licensing system and data production workflows for your organization, and
4. Realize the benefits of contributing to "Collective Intelligence" (and benefiting from the same).

The fundamental principles of Open Data and Open Source Software discussed in this guide are based on **the Free Culture Movement**[2], **the Open Source Software initiative**[3], **Berlin Declaration on Open Access to Knowledge**[4], **Global Open Data Initiative**[5], and USA's Freedom of Information Act i.e. **U.S. Federal Open Data Policy**[6]. For more info about these initiatives, please refer to the "Further reading" section at the end of this document. The document is divided into three main sections. The first section discusses Open Data and issues connected specifically with Open environmental data. The second section discusses Open Source Software, both from perspective of data producers and users, mostly in the domain of environmental science. The final section is dedicated to Crowdsourcing and how is it connected to Open Data and Open Source Software projects.

[2] https://en.wikipedia.org/wiki/Free_culture_movement
[3] https://en.wikipedia.org/wiki/Open_Source_Initiative
[4] http://openaccess.mpg.de/Berlin-Declaration
[5] http://globalopendatainitiative.org/
[6] https://opengovdata.io

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Disclaimer

NOTICE, The authors of this document are not lawyers and this document does not purport to offer legal advice. If you have any legal concerns about open activities or open products that you wish to contribute, you should seek legal advice appropriate to your particular situation and country.

What the authors do offer, in this document, is a summary of their personal experiences working with open data and open source software and contributing to the open source scientific community. Our experiences and observations may help you to find some answers to questions you have about Open Data and Open Software. Hopefully, our experiences, as related here, will also encourage you to decide to become a practitioner of open science and to contribute to open source initiatives.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Acknowledgements

The authors are grateful to several former ISRIC colleagues, especially J. (Jorge) Mendes de Jesus for comments and suggestions on the first draft of the text, and to David G. Rossiter for ongoing interesting discussions and comments.

Every effort has been made to trace copyright holders of materials used to produce this document. Should we, despite all our efforts, have overlooked contributors please contact the authors at Envirometrix and we shall correct this unintentional omission without any delay and will acknowledge any overlooked contributions and contributors in future updates.

OpenGeoHub is a not-for-profit research foundation with headquarters in Wageningen, the Netherlands (Stichting OpenGeoHub, KvK 71844570). The main goal of the OpenGeoHub is to promote publishing and sharing of Open Geographical and Geoscientific Data and using and developing of Open Source Software.

Cite this document as:

Hengl T, Wheeler I, MacMillan RA. (2018) A brief introduction to Open Data, Open Source Software and Collective Intelligence for environmental data creators and users. PeerJ Preprints 6:e27127v2 https://doi.org/10.7287/peerj.preprints.27127v2

In Wageningen, October 19th 2018

Basic definitions

Intellectual property: Intellectual property (IP) is a term referring to creations of the intellect for which a monopoly is assigned to designated owners by law e.g., copyrights, patents, industrial design rights and similar. Categories of Intellectual property related to environmental mapping activities include e.g.:

- primary data (and metadata): e.g., environmental point observations and environmental samples (databases / tables with observed values of environmental variables), polygon and raster maps,
- derivative data: e.g., compilations and harmonizations of environmental data, translations of environmental data in different systems and languages,
- publications and reports: books, posters, manuals and reports,
- cartographic (maps) and similar visualizations

Derivative work: A derivative work is a work based on or derived from, one or more preexisting works. Derivative work can be, for example[7] adaptation, translation, compilation or similar. A compilation of environmental point observations combined from various source environmental point databases (*the building blocks*) can be considered a derivative work. An **adaptation** is a work based on one or more preexisting works. What constitutes an adaptation depends on applicable law, however translating a work from one language to another or creating a film version of a novel are generally considered adaptations[8]. The copyright in a derivative work covers only the additions, changes, or other new material appearing for the first time in the work.

Facts: Plain tabular scientific data without any packaging / mapping / analysis are facts. Facts are not copyrightable[9] and are not protected by the copyright laws (regardless of what the vendor may claim). There is no permission required to generate new IPs by using factual data. However, some facts might be business secrets, some might contain private information (e.g. heavy metal concentrations in a farm).

Free cultural works: Free cultural works are works or expressions which can be freely studied, applied, copied and/or modified, by anyone, for any purpose. Any Open Data, Open Source Software

[7] http://copyright.gov/circs/circ14.pdf
[8] http://copyright.gov/
[9] http://www.pddoc.com/copyright/not_copyrightable.htm

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

or publications published under Open licenses such as GPL, Apache, CC0, CC-BY and CC-BY-SA or alike can be considered Free cultural works.

Open data: Open data and content can be freely used, modified, and shared by anyone for any purpose. An open work must satisfy the following requirements in its distribution:

- **Open license** i.e. must allow free use of the licensed work; it must allow redistribution of the licensed work, including sale, whether on its own or as part of a collection made from works from different sources etc.,
- **Open Access** i.e. the work must be available as a whole and at no more than a reasonable one-time reproduction cost, preferably downloadable via the Internet without charge, and
- **Open Data Format**[10] i.e. the work must be provided in a convenient and modifiable form such that there are no unnecessary technological obstacles to the performance of the licensed rights.

Spatial predictions: Spatial predictions are newly derived data from point data and covariates (rasters) and can be considered results of meta-analysis of point and raster data. Spatial predictions are new IP for two reasons: (1) because model fitting and generation of new predictions is a creative process and different groups working with the same data could produce very different maps, (2) spatial predictions are often neither replacements for, nor simple translations of, point data. In the majority of cases, it is impossible for the end user to back-derive values of the original source point data or use spatial predictions at the locations of the original point data as replacements for these. Only if the spatial predictions were made with methods that permit the original point to be recovered from the prediction (e.g., Thiessen or Voronoi polygons) can spatial predictions be used to re-generate point values. Spatial predictions can therefore be compared to digitizing geomorphological boundaries in a (3D) landscape using Google Earth and visual (expert-based and hence creative) photo interpretation. Although Google Earth data is clearly copyrighted[11] and Google's Geoguidelines[12] also might indicate that literally digitizing content (without creative interpretation) is against the terms of use:

> *"I'd like to trace a map using your imagery. Can I?*
>
> *You may not use Google Maps or Google Earth as the basis for tracing your own maps or other geographic content."*

[10] http://opendefinition.org/od/
[11] http://www.google.com/intl/en/help/legalnotices_maps.html
[12] https://www.google.com/permissions/geoguidelines.html

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

digitizing polygons and lines as an expert using Google Earth and then distributing these under an independent license is not an issue; Jochen Albrecht[13]: *"the (geomorphological) boundaries that you are digitizing are not part of the source material but your (creative) interpretation. You may, however, not copy; that is, you may not re-assemble the Google Earth tiles."* There is still a wide grey area where it is not entirely clear whether, or how, data be used or what the exact producers' and users' rights are.

Crowdsourcing[14] is the process of obtaining content (data), services, funds and/or ideas by soliciting contributions from a large group of people, especially from an online community. Crowdsourcing often involves soliciting rather tedious *"microtasks"* that can be performed in parallel by large, paid crowds (e.g., Amazon Mechanical Turk[15]), unconsciously by unpaid crowds (e.g. Facebook), and/or by volunteers (to develop common goods e.g. Wikipedia). Crowdsourcing is also closely related to activities such as **crowdfunding**, and **collaborative editing**. If members of the *"crowd"* are financially rewarded for their *"microtasks"*, then this is also referred to as the **cloud labor**. Crowdsourcing for collecting geographical data (on-field) can also be referred to as volunteered geographic information.

Citizen science: Is a special type of crowdsourcing and collective knowledge building where primarily non-professionals (non specialists) are involved in data collection, analysis and interpretation. Citizen science can generate accurate data, but it requires that interfaces for data collection both educate and guide users during data collection to avoid any systematic bias or misuse.

Web 3.0: Web 3.0 is a web-technology that allows users to read, write and execute operations, hence it includes editing tools, web services and various API's in addition to web content. Web 3.0 system is usually a requirement for any crowdsourcing system leading to collective knowledge.

Collective intelligence: Can be defined as information mixed with different points of view producing synergies and developing complementary aspects (even among geographically distant people). Collective intelligence is the intelligence of the crowd that can be used for problem solving and joint creation etc. It can be considered the ultimate goal of any crowdsourcing activity aiming at community building and/or knowledge building.

[13] https://stat.ethz.ch/pipermail/r-sig-geo/2014-November/022075.html
[14] Term coined by Jeff Howe (https://en.wikipedia.org/wiki/Crowdsourcing#cite_note-archive.wired.com-18)
[15] https://www.mturk.com/mturk/welcome

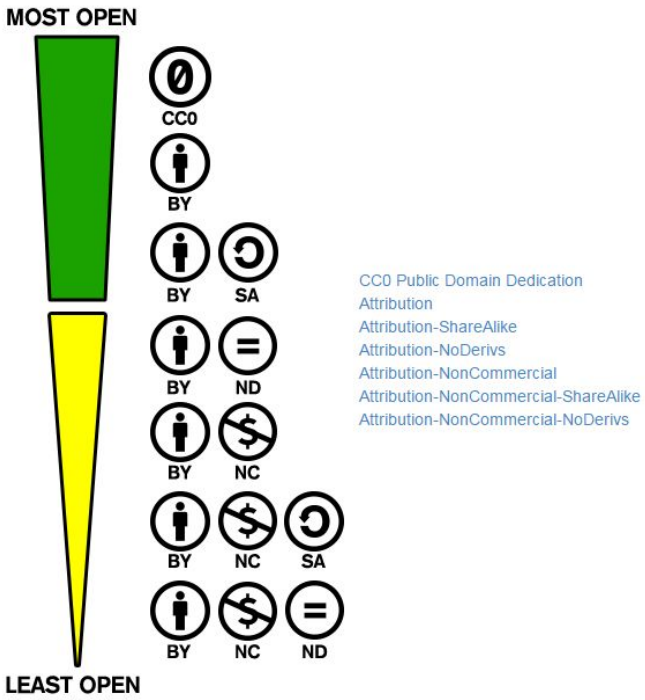
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Open Data

Open data, as indicated in the previous section, is any data that satisfies the three basic requirements: Open Data License, Open Access and Open Data Format. But how can we certify that some data set is "Open"? In the days of a faster and ever more accessible internet, whether data can be freely downloaded can be checked easily by anyone, of course. Also, an Open Data Format typically means that it can be loaded and imported into commonly available software used by most end users (>90%), and exported into most simple tabular formats (without a need for any special software license), which is often relatively easy to check. So usually the only remaining issue is about whether the data is "Open" or not is the Openness of the Data license. Whether a Data license is open or not, can be best checked via the two standard authorities for Open Data:

- **Creative Commons** (https://creativecommons.org/) a nonprofit organization,
- **Open Knowledge Foundation** (http://opendatacommons.org/),

It is, for example, a common misconception that ANY Creative Commons license guarantees freedom and openness. The scheme on the right[16] shows that CC0, CC-BY (Attribution) and CC-BY-SA (Attribution-ShareAlike) are licenses which can be considered as licenses that fully support **Free Cultural Works**[17], while the least open license is the NC-ND. CC-BY-SA can be considered equivalent to the Open Database license (ODbL), which is maintained by the Open Knowledge Foundation.



In summary, only a few data licenses can be considered **truly Open Data licenses** and this should definitely be checked before you claim a product to be "Open Data". The same is true for Open Source Software — what makes software "Open" is primarily defined by the license used.

[16] Image source: http://creativecommons.org/examples
[17] http://creativecommons.org/freeworks

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Why is Open Data important?

Open Data is important not only at a level of organizations or countries, it can also be considered as a fundamental driver of global growth and democratization of society. Consider the following reasoning. The opposite of Open Data is copyrighted data and, in extreme cases, monopolized data (controlled access purely for the benefit of an elite). The big scientific publishers have, for decades, been collecting copyrights for publications written, edited and reviewed by scholars (hence primarily financed from public money). Big scientific publishers, such as Elsevier, Wiley and similar are now considered to be one of the biggest monopolizers of knowledge[18]: *"openness has been severely compromised by the monopolization of knowledge by scientific publishers that has occurred during a great part of the 20th century"*, with many initiatives now emerging to boycott big scientific publishers. In the modern world it is difficult to imagine that paying for a single download of a PDF copy of a research article, even a 30+ years old article, can exceed the daily global GDP (in 2014 estimated at 11,000 USD per year). We do not want to imply here that any publishing business should release their products freely (this would obviously be unsustainable), but transferring the majority of copyrights for articles, books and multimedia funded by public money to big commercial publishers was likely a historic mistake of research societies worldwide — it now limits the global growth and makes science and knowledge more of a luxury, rather than a human right.

So in summary, Open Data is important because:

1. It helps grow the global economy by enabling a wider range of science-driven applications,
2. It helps improve the quality of life and the quality of decision making,
3. It protects from monopolization of public knowledge and corruption.

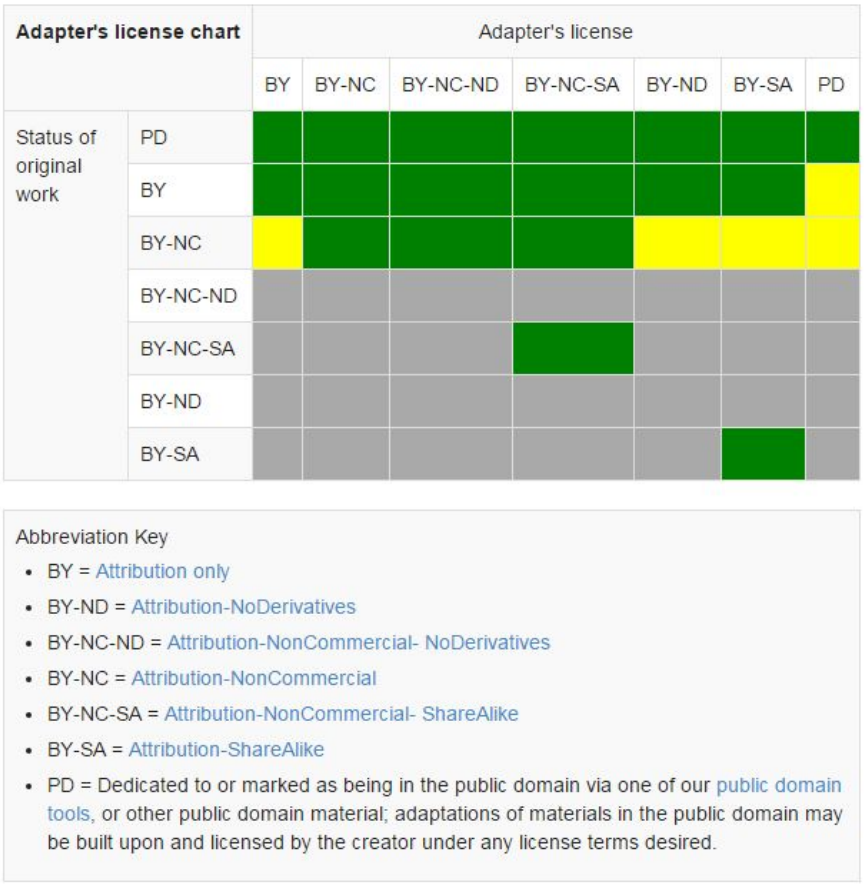
CC-NC

Although very popular among international organizations, the (Non-Commercial) CC-NC license has proven to be over-bureaucratic, inefficient and often confusing to a majority of small and medium size organizations (see also: "reasons not to use a Creative Commons — NC license"[19]). It is really unfortunate that this license has been adopted by many governments and educational or scientific institutions (under the assumption that it will protect the economic interests against the large corporations and business). A seemingly simple choice of forbidding commercial use is not so simple

[18] http://access.okfn.org/2012/05/02/the-access-principle-revisited-open-access-and-the-knowledge-commons/
[19] http://freedomdefined.org/Licenses/NC

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

at all. It is also somewhat confusing that a vendor that releases data under the CC-NC license actually implies that it has full or partial intentions of commercializing the data.



Creative Commons license compatibility chart. Source: Wikipedia.

For all those reasons, CC NC license is considered NOT to be an Open License by many researchers and government employees[20, 21]: *"there is a meaningful distinction between attribution and share-alike requirements and others such as non-commercial (NC), and it is a distinction that merits the description of share-alike licenses as being open but non-commercial licenses as NOT being open"*.

Adaptations of NC works to generate freely licensed products is often an area where no legal action happens, unless some ideological or social disputes develop between the original data producer and the adapter. An example from Creative Commons[22]:

[20] http://blog.okfn.org/2010/06/24/why-share-alike-licenses-are-open-but-non-commercial-ones-arent/
[21] http://opendefinition.org/licenses/
[22] https://wiki.creativecommons.org/wiki/Frequently_Asked_Questions

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

"CC does not recommend using a license if the corresponding box is yellow, although doing so is technically permitted by the terms of the license. If you do, you should take additional care to mark the adaptation as involving multiple copyrights under different terms so that downstream users are aware of their obligations to comply with the licenses from all rights holders."

So although it seems that a CC-NC license gives some space also for Open licenses, experience has shown that this license can increase administrative pressure on a data producing organization and can drive away many small and medium-size enterprises that might otherwise want to use the data..

CC-BY-SA (and ODbL)

Unlike CC-NC license, the CC-BY-SA (one of the "copyleft" licenses) is an Open Data license and equally as protective against large scale exploitation as the CC-NC. It is the main license used by OpenStreetMap[23] and the Wikimedia Foundation Inc.[24], i.e. some of the world's largest non-for-profit data organizations (note: from the year 2012, OpenStreetMap uses actually the Open DataBase License which is more suited for informational databases, such as educational or scientific databases). Although the "Share-Alike" (also known as *"pay it forward"*[25]) component implies that the license is partially restrictive, for any organization supporting Free Cultural works CC-BY-SA should be considered as the optimal license for the following two reasons:

- It stimulates the philosophy of Free Cultural Works and Open Knowledge (it emphasizes community norms and *"spreads the word"*),
- It protects from large corporations that could easily exploit and profit on large data generated by public / public agencies (e.g. the OpenStreetMap[26]),

Although it appears to many large corporations as a restrictive license, the SA restriction is in fact minimal — it does not stop companies from using the data for commercial purposes or from selling software or services based on the data, it only prevents them from making profits from data derivatives. There are also other practical reasons why OpenStreetMap and Wikimedia use the SA license: minimum number of law cases against OpenStreetMap and Wikimedia has proven that the ODbL and CC-BY-SA licenses are efficient: claimant's effort would often not be worth a court case because OpenStreetMap makes no profit from the data nor does it stimulate users to generate profit from the data itself.

[23] https://www.openstreetmap.org/about
[24] https://wikimediafoundation.org/wiki/Terms_of_Use#7._Licensing_of_Content
[25] https://en.wikipedia.org/wiki/Pay_it_forward
[26] http://www.citylab.com/design/2015/06/who-owns-the-digital-map-of-the-world/396119/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

The disadvantage of CC-BY-SA license is that large business finds it too restrictive. OpenStreetMap, for example has been criticized for having a SA clause[27]: *"In some places, the legal verbiage is vague on how much and what kinds of data must be contributed back, which makes some businesses nervous that OSM s legally incompatible with their own data (i.e., user data). OSM members see the license as a way of keeping information truly open — and of preventing users from simply profiting from the data without contributing anything back."* This is not a trivial problem and many companies will still be put off. Until they do decide to change their business model and start *giving away* some data themselves.

Open Data and the OpenStreetMap example

An Open Data oriented organization stimulates companies to build up on the basis of data and software products with a minimum of administrative requirements (which under the CC-NC is not possible). One of the best examples of such an organization is the OpenStreetMap foundation. OSM started as a not-for-profit organization, but in 2007 it received venture capital funding of 2.4M euros for CloudMade, a commercial company (source: Wikipedia). Since then, the robustness and potential of OSM data and services have dramatically increased. It is now used by thousands of (commercial) web sites and mobile apps, as the OSM website indicates:

Commercial applications of OpenStreetMap include:

- Web-mapping applications used by Apple, Flickr, MapQuest, MapBox and similar,
- Navigation and vehicle tracking software (offline),
- Geographic analysis and spatial planning ...

So, in summary, Open Data can be an excellent business model. It can help create revenues and increases tax contributions and these can then be used to produce even more Open Data for even more revenue. This has been confirmed by numerous research organizations[28]: *"Open data—public*

[27] http://www.citylab.com/design/2015/06/who-owns-the-digital-map-of-the-world/396119/
[28] http://www.mckinsey.com

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

information and shared data from private sources — can help create $3 trillion a year of value in seven areas of the global economy" (as reported in 2015).

IP extracted through data mining and web-crawling

Something that often slows down environmental data producers and analysts is the unclear terms of use and data license for existing environmental data sets. This is especially true for environmental laboratory data and similar data collected through field work. Can IP extracted through data mining be freely distributed as Open data? The general agreement among many data license specialists is that all point data that have been published in a report or similar publication (with coordinates) can be used by any organization without any need to seek permission. Likewise, if raw tabular data have been, purposively i.e. via an official channel, published in a report / book or as a PDF on the web in plain format (e.g. as an Excel table, CSV file, table PDF or similar), these data can be considered as a collection of facts and added to any public databases along with a record of their lineage (original source). Consider for example geochemical environmental measurements published in a table in an online report or research article. These numbers can be converted to a small database, also through harvesting the Web. Once the tabular data have been extracted and organized, it can be used to run analysis and create new IP (Intellectual Property) without any need to contact the original copyright holders. In other words, there are vast quantities of environmental data waiting for you to use them (as long as you know how to find them, how to interpret them and how to import them into analysis software).

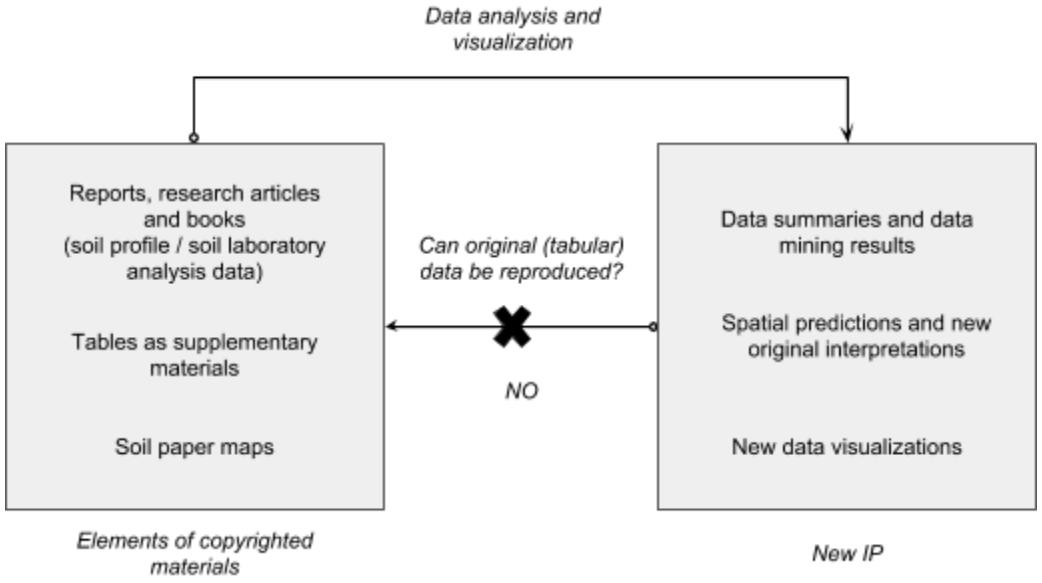
Even restrictive licenses such as CC-NC do not prevent any user from doing data mining and further generation of new IPs. Example from Heliyon[29]:

Creative Commons Attribution-NonCommercial-NoDerivatives (CC-BY-NC-ND 4.0)

"This license allows readers to distribute and copy your article, as long as it is not done for commercial purposes. Under this license, readers cannot change or edit your article for distribution in any way. They are able to distribute and copy your article as long as they give appropriate credit (with the DOI link to the publication), provide a link to the license, and don't claim that you endorse their adaptation of your work. Under this license, readers are also able to mine the text and data."

[29] http://www.heliyon.com/open-access/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018



New IP, i.e. spatial predictions and similar results of meta-analysis can be considered as new IP because the original tabular data can not be reproduced i.e. these can not be used as a replacement for the original point / raster data.

Data mining and meta-analysis of tabular data and facts requires, therefore, no special permission (in most cases, unless there are privacy issues or similar).

Because, for example, spatial predictions can not be used to re-create or reverse-engineer original point data, these data are, effectively, new IP and the license can be set freely. Another example of license specification from the European Commission's' LUCAS point data[30] terms of use:

"Any results can be published once a proper aggregation process that prevents any individual body (private person, private organisation) from being identified has been applied, … graphical representation of individual units on a map is permitted as far as the geographical location of the soil samples is not detectable."

The previous two examples confirm that, it only takes creativity to employ legacy environmental data and create new data products from it that can then be released as Open Data.

Privacy and other data issues

In specific cases, as in the case of the LUCAS points, it is not an issue whether an organization is allowed to use the point samples to generate spatial predictions, but there is another legal obstacle (privacy) that prevents data producers from distributing the data freely. It is really only a privacy law

[30] http://euenvironmentals.jrc.ec.europa.eu/projects/Lucas/Data.html

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

that matters in some cases, but there are effective solutions even when there seem to be serious privacy issues. For example, Google has been initially criticised for driving with their cars and taking photographs of public roads and homes (the Google Street View project). Note that even in the case of strict countries, governments were ultimately OK with Google blurring images rather than removing the Google Street View services all together. Even the biggest critics of Google's intrusion of privacy probably still like to use Google Street View in daily life and travels. As this example from Denmark illustrates[31]:

"The Danish data authorities advised people who are photographed by Google, report Google to the police... Since then, Google hasn't had any legal problems and has continued filming."

So with some creativity and a bit of compromise, there is a practical solution to most privacy / data access conflicts.

Other open license issues that you might check before you release your data in public are: *"how did you obtain the copy of the data?"*, *"are there any privacy concerns considering these data?"*, *"are you respecting the terms of use?"*, *"is the attribution correct?"* Hence, what you really need to check for any source data are the details of data publication, especially terms of use and privacy issues.

Open Data and commercial uses

ODBl has three major consequences:

1. The data will be freely available to companies and others for commercial work e.g. consulting and spatial planning, which was not the case with the previous CC-NC license.
2. With ODbL i.e. CC-BY-SA license your organization can act as a catalyzer of applied environmental research locally and globally. Data registered as Open environmental Data[32] guarantees unlimited: use, transparency, social and commercial value, participation and engagement; which is all within the scope and mission of the OpenGeoHub Foundation.
3. Any remix, transformation and/or adaptation of the ODbL data will stay in the public domain i.e. there will be no additional generation of income using the data and derivative works from these data.

[31] https://en.wikipedia.org/wiki/Google_Street_View_privacy_concerns
[32] https://okfn.org/opendata/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Example of a general decision scheme for data licenses.

Data type	*Input licenses*	*Derivative work*	*Proposed license*
Original data / IP produced by your institute	Variable	NO	CC-BY-SA / CC-BY
Tabular / point data published in literature (raw tables)	(facts hence no license needed)	YES	ODbL / CC-BY-SA
Point data compilations	mainly CC-NC	YES	CC-BY-SA / CC-BY-NC
Spatial predictions based on published point data	various	NO*	ODbL / CC-BY-SA
Spatial predictions based on commercial point data	commercial	YES / NO	*Contractor decides*

*Unless spatial predictions can be used to reproduce the input point data / input raster maps.

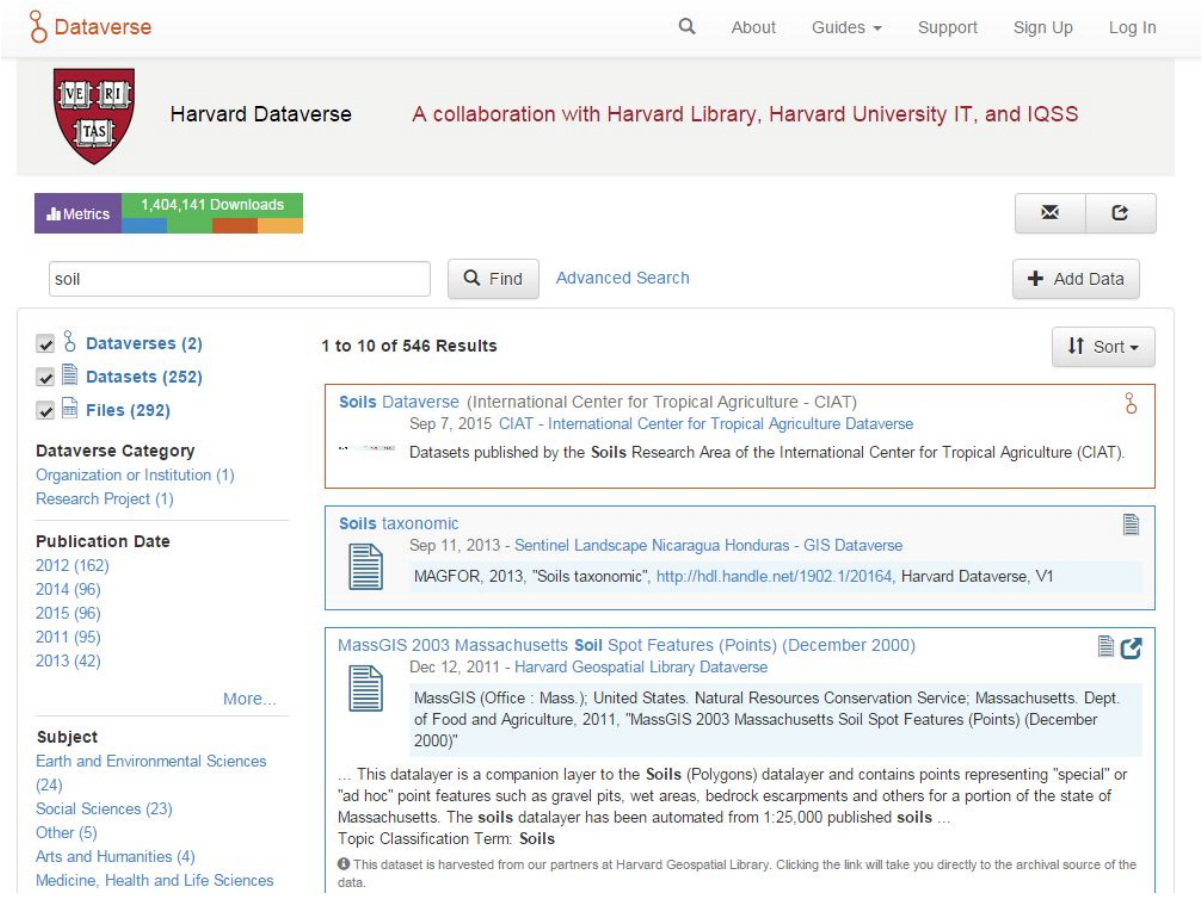
The general business model for Open Data is therefore:

"use the data, no need to ask for a permission but always cite the source; contact us if you need products of even higher accuracy and/or customized solutions"

OpenGeoHub Foundation also aims to stimulate use of its data products (without restrictions) and hence boosting their visibility (free data service is possibly the best marketing strategy) within both the business and research worlds.

Note that it is an important decision to adopt strategic data licenses for the main data products developed in-house. Furthermore, for any new unregistered points, license management can be effectively outsourced to major environmental data registries such as Harvard Dataverse. The example below illustrates some environmental data sets registered on the Harvard Dataverse:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

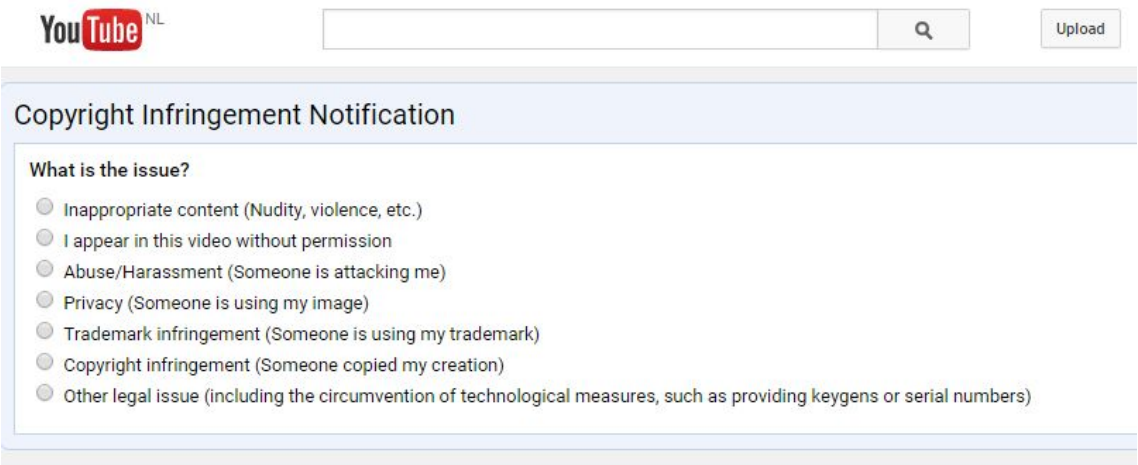


Harvard Dataverse: an example of a free service for registering data sets

Outsourcing environmental data registration to e.g. Harvard Dataverse can speed up clarification of licenses for all public data sets.

For points for which the original producers cannot be located (orphaned points) and/or that do not respond within some reasonable time period, you can opt to use this data assuming that there will be no issues, but then provide a webform *"takedown procedure" / "content reporting procedure"* so that the original data owners can update you about the terms of use and proper attribution or similar efficiently and rapidly. Compare for example with the *Copyright infringement Notification form* used by Youtube.com (in the case of Youtube, the data is removed or modified after a specific request, and not before the upload by users):

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018



To achieve this, your organization can set up a webform with a simple and clear content reporting procedure that should include (at least):

- Report an Abusive User
- Report a Privacy Violation
- Report a Legal Complaint

By setting up a system where the data owners can respond quickly, you aim to create confidence in your services. All this, hopefully, illustrates that there is a lot of opportunity for selecting Open Data licenses for your IP and that you can receive much help by using the functionality of the Creative Commons, Open Knowledge foundation, Harvard Dataverse and similar services.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Open Source Software

As with Open Data, Open Source Software also aims to serve society with functionality and applications. But why is Open Source Software important, and what does it imply? Does switching to Open Source Software imply that only non-commercial / non-for-profit activities are possible? If not, what is the actual business model for organizations and companies using Open Source Software and does this really work? The following sections try to clarify some common misconceptions about Open Source Software and to illustrate, with existing examples, that: yes; Open Source Software is both a viable business model and a platform for doing public good.

Why is FOSS important?

The most common misconception about Open Source Software (in subsequent text **Free and Open Source Software or FOSS**) is that one should use it primarily because it is cheap i.e. costs no money. Although, yes it is free and it typically saves costs, this should definitely not be your primary reason to switch to FOSS. In fact, switching to FOSS also requires an investment because one at least has to invest in migrating all systems and data and training people to use FOSS, so it is definitely not without costs. Also, if you can afford to pay for software, and if this software helps you complete your work more efficiently, you should definitely invest in software. Many commercial companies work hard to produce their software and they deserve to be rewarded for the software design, implementation and support.

Here are, in our opinion, the real reasons for you to switch to FOSS . These also mirror the experiences of many other developers and businesses[33]:

1. **Transparency** hence **reliability** hence **security**: Open Source implies that all of the code within some workflow is completely visible to the users. There are no hidden processes or black boxes. The same way you would trust some politician or government more if there is more transparency in what and how they do things, you will also place more trust in FOSS if you can actually see and follow, what it does and how it does it.

2. **Quality of code and speed of improvement**: A second advantage of using FOSS is (increasingly) the fact that some of the smartest (and most motivated) people in the world develop solutions for FOSS, and consequently the quality of the code is constantly improved

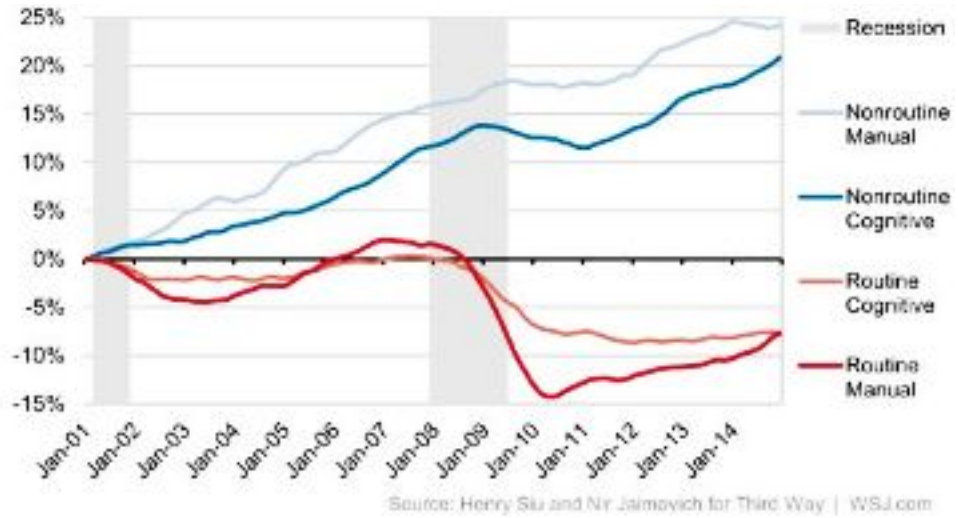
[33] http://www.pcworld.com/article/209891/10_reasons_open_source_is_good_for_business.html

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

and improvements occur at increasingly at faster rates than for proprietary software. Daniel H. Pink in his book "Drive" mentions three universal drivers of the creative process: *autonomy* (or establishment of freedom), *mastery* and *purpose*. FOSS supports all three of these drives, hence it is no wonder that today the majority of servers in the world run Linux, the majority of mobile devices run Android and a majority of researchers use FOSS to run data analysis.

3. **Customizability / extendability**: FOSS puts very few legal or administrative obstacles on users considering possibilities of customizing and extending the existing functionality. The sky is really the limit — you are most welcome to change, extend, simplify or re-design any FOSS code you find.

4. **Quality and efficiency of support**: It is difficult to believe that so many people are eager to help other people (without being paid) through FOSS mailing lists, forums and similar. However, the amount and quality of support available for FOSS users has been constantly increasing and the response time has been decreasing. Likewise, companies that specialize in FOSS are often motivated to try much harder to help you because their revenue is often based purely on their support.



Routine manual (and cognitive!) jobs are rapidly disappearing. There is a going to be an increasing demand for creative jobs that require high-tech software skills and low initial investments, and FOSS is among the most promising platforms for creation.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Ultimately, yes FOSS will often reduce your costs quite dramatically. But here is something that you maybe didn't already know: FOSS usually does not save costs on initial software installation — where it really provides savings for an organization is in:

1) The time required to do upgrades and deal with all administration and payments. For example, the average costs of migrating from one version of Windows to the next can be on the order of 20–60% of the costs of the hardware; the cost of migration for FOSS is often an order of magnitude less).

2) The costs of extending the functionality to fit project needs.

If you are a business interested in FOSS, then you should definitely look at some recent business trends[34]:

- Use of FOSS has increased 2x since 2010,
- Venture investments in FOSS have tripled since 2011,
- In the next 2–5 years, FOSS will likely become the dominant platform for Big Data analytics, cloud computing and server OS. Even the dominant proprietary software development companies, such as Microsoft, are now investing aggressively in FOSS.

As Dries Buytaert, CTO of Acquia, puts it: *"the numbers do not lie — open source is driving innovation for the world's most successful companies"*. So also, if you are a business, you should consider FOSS largely as a business opportunity to rationalize costs, organize a more dynamic and more creative scene... and attract some of the most capable, most productive software developers into your fold.

Open Source Software license

As with Open Data, not all licenses guarantee that your software is truly Open Source. There are probably more software licenses than data licenses in the world today. However, two main OSS data licenses seem to be dominant at present:

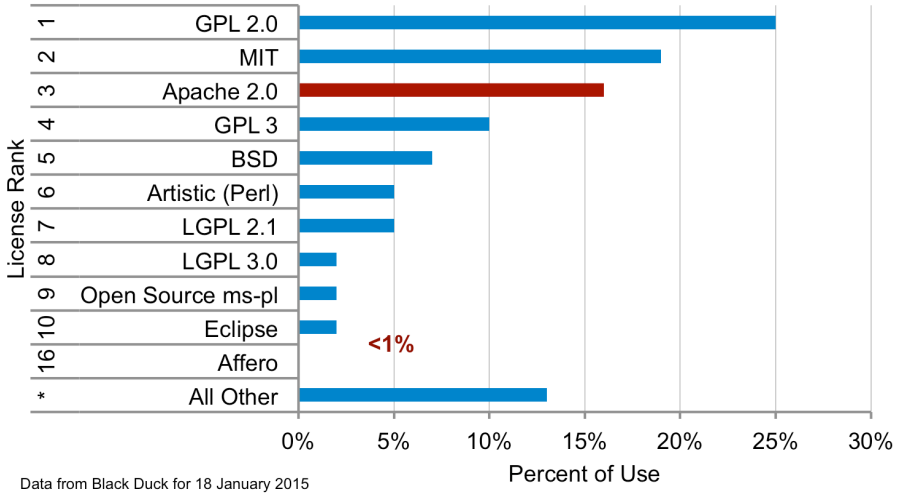
- General Public License or GPL[35], also similar to GNU and LGPL,
- Apache License[36],

There are some subtle differences between the GPL and Apache license. This differences can summarized as follows: the GNU/LGPL is more popular among independent developers and

[34] http://www.slideshare.net/blackducksoftware/2015-future-of-open-source-survey-results
[35] http://www.gnu.org/licenses/
[36] http://www.apache.org/licenses/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

companies which mainly deal with open source software; the Apache License is favored by the big corporations for their open source projects[37]: *"When someone modifies an Apache licensed software and redistributes it, it is not necessary to release the modified version of the source code... If a software under GPL or a modified version of a GPL-ed software is released to the public, the distributor needs to make the source code, including the modified version, if any, available to anyone who wants it. This is very good to promote software freedom but scares away the corporations"*. Note also that software can also be registered using one of the Creative Commons licenses. In this case, GPL seems to be more similar to CC-BY-SA or CC0, and Apache License to CC-BY. A general recommendation, however, is to use CC licenses for publications and multimedia, ODbL for data and GPL, Apache or similar for software.



Overview of licenses used for registering Open Source software.

Open Source Software for environmental data

There are many FOSS of interest to environmental data producers and environmental data users. Here is an overview of some of the mre popular FOSS of interest to environmental producers:

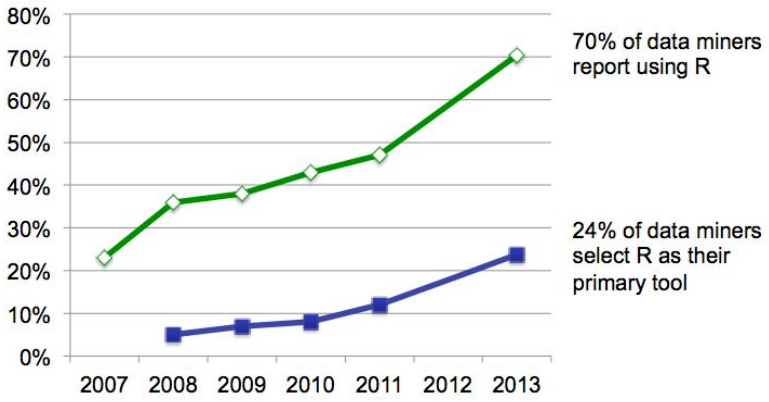
Domain	***OS software***
Data preparation and DBs	PostGIS, PostgreSQL, SciDB, ...
Data and Big Data analytics	R, Python, Java, ...
Geospatial analysis	OSGeo suite (GDAL, GRASS GIS, QGIS, SAGA, etc)

[37] http://digitizor.com/apache-license-vs-the-gpl/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Web and app development	Java, PHP, Javascript, Cordova, …
Data visualization	QGIS, Cesium, …
Data sharing / data distribution	Geoserver, CartoDB, …
Software development	Python, Java, …

One of the data analytics FOSS platforms that has become increasingly popular among environmental scientists is R. Since 2000 R has skyrocketed both considering the number of researchers using it to run data analytics and the number of companies using it as the main platform for data mining, machine learning and statistics.



R usage trends. Source: Rexer Analytics.

Without a doubt, there is a great potential for environmental data producers to utilize FOSS for their work. This does not imply that one should completely abandon proprietary software, but if Microsoft has decided to embrace Open Source Software (e.g. by acquiring recently Revolution Analytics), you should show some interest in it as well.

Keys to making efficient Open Source Software

Richard Barnes[38] suggests the following general qualities of a good software tool:

- It should be a (C++ or similar) library, not a closed program (GUI); this library can then be served through a programming environment such as Python, R or similar,
- It should be ready for High Performance Computing (parallelization),
- It should be Computationally democratic and Communication Avoiding,
- It should have explicit Input / Output functionality (you determine when to write),

[38] http://rbarnes.org/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

- It should be interoperable,
- It should come with an extensive documentation with examples / consists of reproducible Workflows,
- …. it should be Open source,

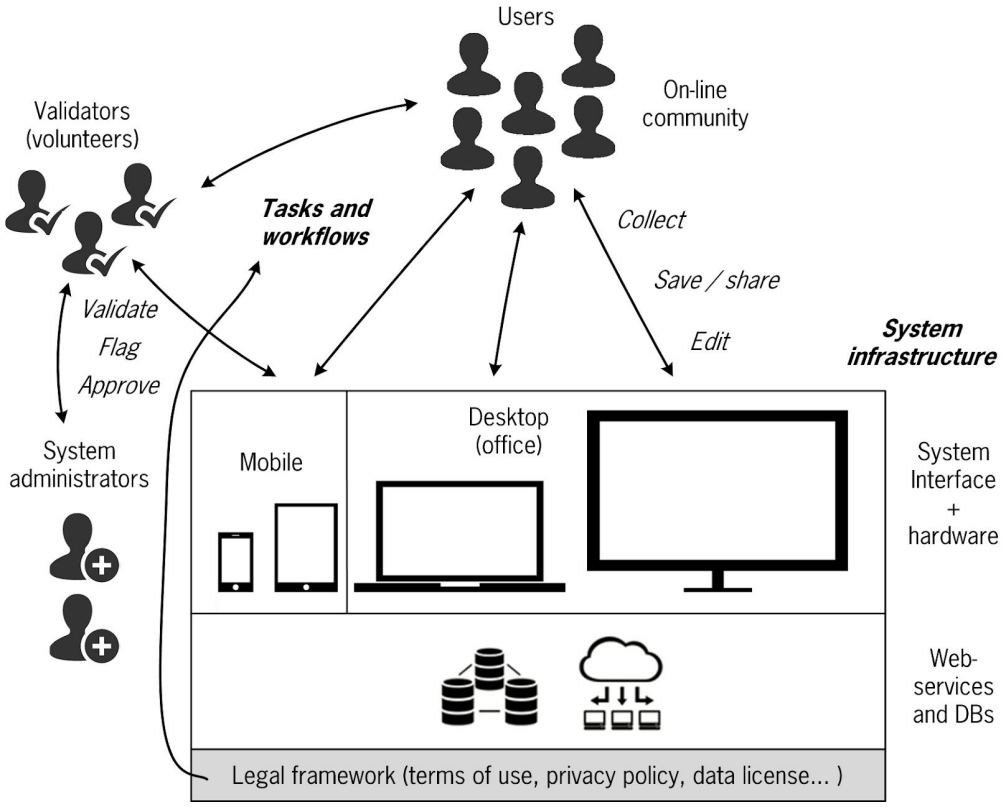
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Crowdsourcing and collective intelligence

Crowdsourcing components and types

Under *"crowdsourcing"* we can categorize all activities geared towards generating data, services, funds and/or ideas from large groups of people and with some minimum staff involvement (self-organized / knowledge communities). In principle, each crowdsourcing system consists of the following four main components (see scheme):

1. **Crowdsourcing moderator** (legal organization or company)
2. **Online community** (the users)
3. **Tasks, workflows and conditions** (rules and responsibilities)
4. **Hardware and software infrastructure** (the system IT)



Typical crowdsourcing system components.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

There are many types of crowdsourcing projects and these can significantly differ in implementation and workflows. Crowdsourcing can be classified as:

1) Based on a business model:
    a) **Commercial** — Where the purpose of the moderator of the crowdsourcing is to commercialize some or all of the outputs of crowdsourcing or profit from the data traffic.
    b) **Non-for-profit** — Where the purpose of the moderator of crowdsourcing is to develop some public good data or service (e.g Wikipedia).
    c) **Combined** (e.g. OpenStreetMap)
2) Based on the manner of contributors' involvement:
    a) **Unconscious, passive or implicit crowdsourcing** — In the case of implicit crowdsourcing, users do not necessarily know they are contributing. For example, one of the biggest crowdsourcing projects at the moment is Facebook: users unconsciously provide information to the Facebook company, which then sells this information to (marketing) companies[39]. Another example of implicit crowdsourcing is the reCAPTCHA project where users unconsciously help translate and digitize various scanned books.
    b) **Active or explicit crowdsourcing (on invitation)** — *"Explicit crowdsourcing lets users work together to evaluate, share and build different specific tasks."* (e.g. Wikipedia).
3) Based on participants level of expertise:
    a) **Crowdsourcing via professional expert groups** — Crowdsourcing focused on professionals, which are often invited through official channels i.e. via their organizations. For example, big commercial publishing companies use scientific societies and scientific networks to quality control scientific publications (e.g. for paper reviews).
    b) **Crowdsourcing via groups of non-professionals** (including children!) or individuals — also known as "Citizen science". Tools developed for non-professionals usually have a more simple interface and have extensive guidelines / data entry wizards.

[39] http://www.theguardian.com/technology/2015/sep/25/facebook-money-advertising-revenue-should-you-be-paid

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

c) **Combined**.

4) Based on functional applications[40]:

a) **OPEN INNOVATION** — uses sources outside of the entity or group to generate, develop and implement ideas.

b) **COMMUNITY BUILDING** — develops new communities through active engagement of individuals who share common passions, beliefs or interests.

c) **COLLECTIVE CREATIVITY** — taps creative talent pools to design and develop original art, media or content.

d) **CIVIC ENGAGEMENT** — fosters collective actions that address issues of public concern.

e) **COLLECTIVE KNOWLEDGE** — develops knowledge assets or information resources from a distributed pool of contributors (e.g. Scientific publishers).

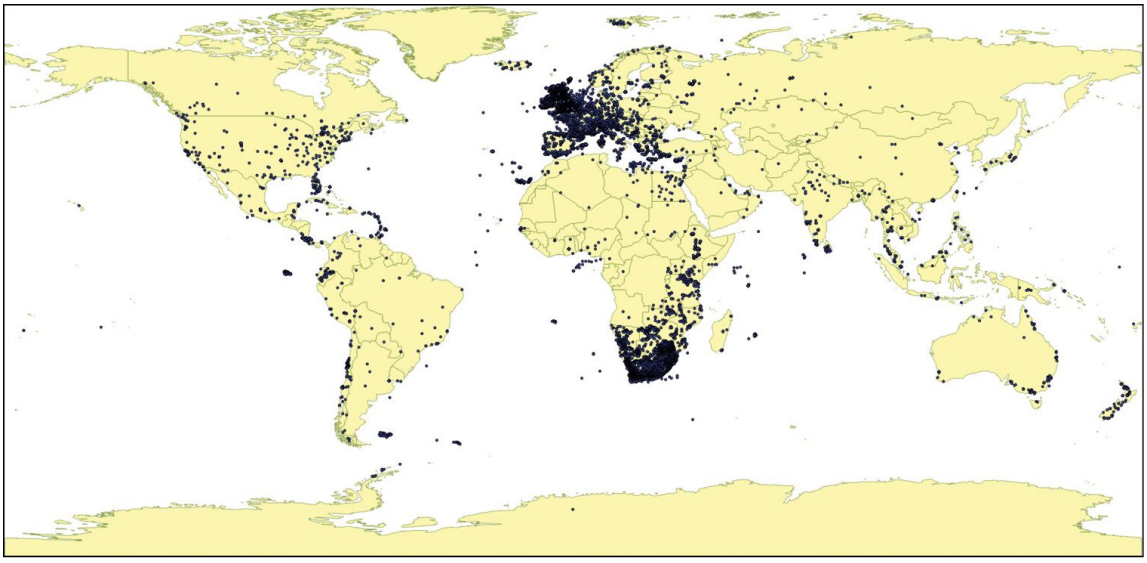
f) **CROWDFUNDING** — solicits financial contributions from online investors, sponsors or donors to fund for-profit or non-for-profit initiatives or enterprises.

g) **CLOUD LABOR** — leverages distributed virtual labor pools, available on-demand, to fulfill a range of tasks from simple to complex (e.g. Amazon Mechanical Turk).

Crowdsourcing, as a data collection mode,l is especially interesting to organizations that perhaps cannot afford the costs of collecting large data: *"Crowdsourcing may produce solutions from amateurs or volunteers, working in their spare time, or from experts or small businesses which were unknown to the initiating organization."* (source: Wikipedia). Crowdsourcing should, however, not be confused with citizen science, which basically implies that only non-professionals are involved in data collection, analysis and interpretation.

[40] Source: http://crowdsourcing.org

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018



iSpot managed to generate thousands of biodiversity observations in only a couple of years since the launch (source: DOI 10.3897/zookeys.480.8803).

Three examples of (global) crowdsourcing systems especially relevant to environmental data producers are for example:

- ➔ iSpot: crowdsourcing collection of biology / biodiversity data,
- ➔ OpenStreetMap: crowdsourcing system for free geographic data, such as street maps,
- ➔ MySoil: crowdsourcing citizen science for soil data collection and for raising awareness,

Other crowdsourcing systems of interest to environmental scientists are:

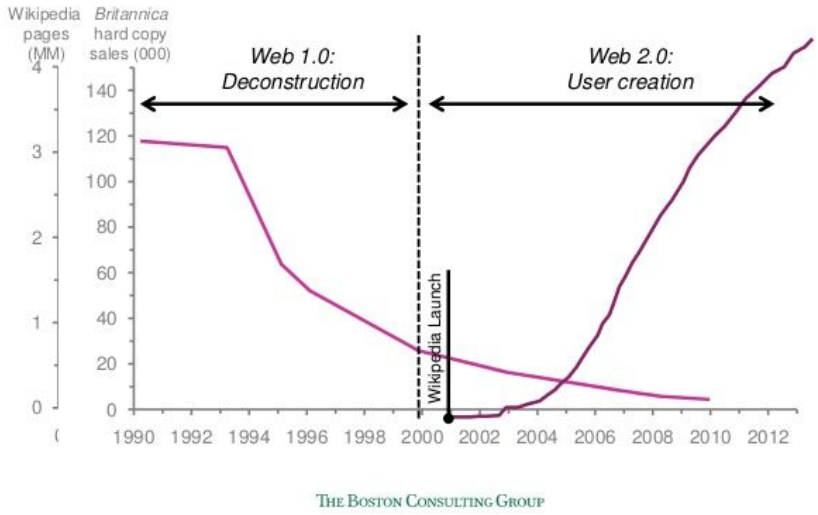
- ➔ Geo-wiki (an app and website) for collecting and sharing land cover observations,
- ➔ Fotoquest (in field validation; comparable to geo-wiki).

In summary, regardless of whether users are non-professionals (citizen science), professional government organizations or NGOs, at the core of the system is a robust on-line system that can host information in a manner that is useful both to the crowdsourcing organization and to the users themselves.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Web 3.0

Any modern crowdsourcing system is heavily based on Web 2.0 and Web 3.0 technologies. To Goodchild (2007)[41], Web 3.0 is a web-technology (actually the complete cyber-infrastructure) that allows users to read, write and execute operations, hence it includes Web Processing Services and various API's in addition to data serving facilities. Web 3.0 enables the widest possible community of users to participate, often independently of their hardware and software requirements.



The power of the crowd: over the last fifteen years Wikipedia has established itself as the world's leading encyclopedia with comparable accuracy to commercial products with budgets often hundreds of times larger. It is clear that Wikipedia and Web 2.0 approaches to content creation have displaced other traditional systems. Image source: The Boston Consulting Group.

Web 3.0 and the future

To Murgante and Garramone (2013)[42] the ultimate purpose of Web 3.0 is to support collective intelligence:

"Collective intelligence occurs wherever there is a huge human interaction and new technologies can easily encourage synergies even among geographically distant people. Someone who lives in a remote part of the world can interact with other people with complementary knowledge, living very distant, continuously communicating with each other, exchanging their experiences, cooperating, etc... Collective intelligence can

[41] DOI: 10.1007/s10708-007-9111-y
[42] DOI: 10.1007/978-3-642-39646-5_44

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

be defined as an information mixed with different points of view producing synergies and developing complementary aspects."

Traditional legacy environmental data has been mainly collected via top-down systems e.g. through big government campaigns. There is now an increasing potential for crowdsourcing of soil data collection, validation, transformation and environmental data exchange. Not only for the purpose of collecting new data, but also for the purpose of enabling merging and harmonization of existing legacy environmental data (e.g. crowdsourcing of data validation). Almost everybody has a mobile phone today. People are also increasingly aware of the importance of environmental and land protection. This is a software/web niche that will likely be filled by various environmental science and non-environmental science institutes and companies in the decades to come.

It is realistic to expect that most of the content produced by OpenGeoHub, and similar organizations, beyond 2016–2017 will be accessed via the mobile phone apps and web services such as REST API. Also many major information businesses in the world (e.g. Microsoft, Apple, Google) have highlighted mobile phones and web-of-things technologies as their main development areas.

Implementing crowdsourcing for environmental data collection

Implementing crowdsourcing projects to the point of self-sufficiency is not trivial. To Rossiter et al. (2015)[43], soils have not been very attractive for large crowdsourcing projects mainly because *"soil is not 'attractive' in the same way as birds, plants, or stars; it is not easily 'visible' in the same way as the atmosphere or biosphere and citizens have little knowledge of soil science, compared to sciences with wide popularity such as medicine, astronomy or even cosmology"*. Rossiter et al. (2015) hence suggest multiple initiatives/campaigns before a citizen science project could reach stability (in this example citizen science for DSM):

1. *"a clear description of what additional information for DSM can be easily and safely provided by non-specialists;*
2. *the identification of citizen groups that would be in a position to provide these;*
3. *a strong publicity campaign, with appropriate materials and perhaps training opportunities;*
4. *protocols for data collection and submission by the citizens; and*
5. *protocols for dealing with citizen-submitted information, to make it useful for DSM."*

[43] DOI: 10.1016/j.geoderma.2015.05.006

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

One of the first crowdsourcing projects for collecting spatial data was the OpenStreetMap. This had fantastic success considering the data usability and should be considered an important milestone and a successful proof of concept. Nevertheless, only once companies, businesses and governments / political elites decide to use this data will it become significant.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Conclusion

There is great potential in establishing crowdsourcing systems and Open Source software, especially for environmental data collection, harmonization, monitoring of land degradation and for spatial planning. Selecting restrictive licenses, such as the Creative Commons NonCommercial (NC) license, for new IP generated from public data or data harvested from the web should be avoided as much as possible because: (a) it increases administrative pressure on an environmental data serving organization, (b) it drives away all small and medium sized enterprises. On the other hand, Creative Commons ShareAlike license and or Open Database Licenses are especially suited for organizations aiming to stimulate the free exchange of data, but they can also be used to support development of commercial applications and commercial uses. Many corporations are put off by the ShareAlike component of these licenses but the interpretation that such licenses prevent these companies from generating profits are likely misunderstood and over-exaggerated.

Using any external point data (public or private) for any purpose without consulting carries inherent risks considering the terms of use, data license and privacy issues. There is still a wide grey area where it is not entirely clear what the data license is and how data can be legally used. An organization is, however, unlikely to be prosecuted across national boundaries for releasing newly generated environmental data under the CC-BY-SA license because the claimant's effort would not be worth it: if nowhere in the chain is any profit being made from the data itself.

Providing environmental data under an Open Data license helps boosts an organization's visibility and broaden its network. In the coming 2–5 years, FOSS will likely become the dominant platform for Big Data analytics, cloud computing and server OS. Using and contributing to Open Source software communities helps all government organizations, NGOs and commercial companies to develop or improve their businesses and to meet some of the smartest geeks on the planet (maybe even hire them to work for you?).

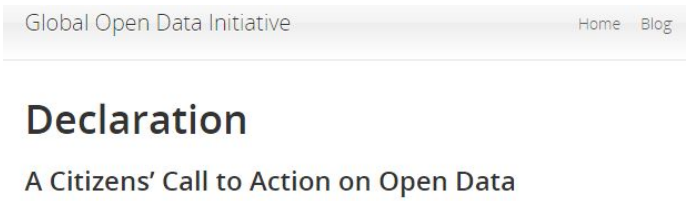
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018

Further reading

Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture ...
Book by Lawrence Lessig

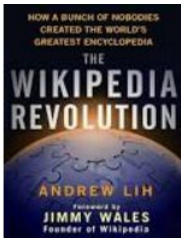
http://www.free-culture.cc/

http://openaccess.mpg.de/Berlin-Declaration



http://globalopendatainitiative.org/

The Wikipedia Revolution: How A Bunch of Nobodies Created The World's ...
Book by Andrew Lih



http://www.wikipediarevolution.com/

Good Faith Collaboration
Book by Joseph M. Reagle Jr.

http://reagle.org/joseph/2010/gfc/

The Open-Source Everything Manifesto: Transparency, Truth, and Trust
Book by Robert David Steele

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27127v2 | CC BY 4.0 Open Access | rec: 19 Oct 2018, publ: 19 Oct 2018