VertNet
Quick(ish) Guide to
Copyright and Licenses for Dataset Publication

Table of Contents

Introduction

This quick guide is intended to give data publishers and users of data published to the VertNet portal, as well as other biodiversity data portals, an understanding of the uses and limits of copyright, licensing, and waivers for biodiversity data and datasets. ***This is NOT a legal document or contract.*** This IS a well-researched guide intended to provide you with a clear understanding of the issues surrounding data sets, copyright, licenses, and the public domain. We've also conversed extensively with our friends at Canadensys, legal professionals (for legal accuracy where appropriate), and others who have conducted a sizable amount of research on these topics on their own.

The descriptions and recommendations in this document are relevant to any data publisher or data user, but the recognition of rights, licenses, waivers for data, and what content is considered to be in the public domain may vary depending upon country, state, province, or institution. ***It is the responsibility of each publisher to consult with local legal counsel for legal opinion in keeping with their home institutions.*** Data users may also need to seek answers to questions of use with local legal professionals, although we are happy to help in any way possible.

In keeping with VertNet's primary goals and beliefs, this guide is intended to make things as

simple and clear as possible, but dealing with these issues can be complicated. The more caveats we/you add, the more complicated things can become. So, buckle up and drive on in.

What We Believe

1. Biodiversity data should be as complete, discoverable, and accessible as possible via portals, public URLs, and other means.
2. Biodiversity data should be standardized so it can be aggregated, shared, and used as easily as possible.
3. Factual data (see, What Can be Protected by Copyright and Licensing) cannot be protected by copyright and should be committed to the Public Domain.
4. Any licenses used to protect compilations or datasets (when applicable) should be licensed using standardized, machine readable licenses, such as licenses from the Creative Commons. Ideally, we would like to see all datasets using the Creative Commons Zero (CC0) waiver.
5. Data publishers have the right to protect any creative content, including, but not limited to images, sound files, videos, and descriptive/speculative text, where applicable.
6. This process should be as simple as possible for both data publishers and for users (results may vary from collection to collection).
7. All data publishers should get the credit they deserve when the data they curate are used by others for broader benefit.
8. It is worth the time and effort to achieve these goals to the greatest extent possible.

Why Are Rights and Licenses Needed (or are they)?

Times have changed. No longer is it necessary for a researcher or scientist to visit a natural history collection in person or to use the phone or parcel post to request and review data that reside within a local database. As the availability of biodiversity data and related content (e.g., imagery and other digital media) from natural history collections continues to grow online and via digital formats, and the ability to share datasets between and among data publishers and users grows easier, many institutions want to be certain that they:

1. receive appropriate recognition and attribution for providing content, and,
2. protect any individual or institutional interests (financial or otherwise) that may arise from shared content.

Ascerting copyrights and licenses on datasets and related materials are two of the tools that data publishers can use to retain some control over their content and exercise applicable rights. Unfortunately, not everything in a publisher’s database can be protected by copyright or licensing. In most cases, the dataset itself may be exempt from protections, but we’ll get into that below.

What Can be Protected by Copyright and Licensing

Perhaps it's better to establish up front what *cannot* be protected by copyright and licensing: facts are not copyrightable. The majority of the data in the datasets published by VertNet and available in other portals such as GBIF, Canadensys, and iDigBio, is factual and, therefore, cannot be protected. Generally, that includes most taxonomy, geography, morphometrics, sex, preparations, and nearly all of the content within Darwin Core fields.

So, what can be protected by copyright or licensing?

Images, media, and some textual descriptions that are clearly and demonstrably the creative work of the collector or observer, or other individual who created or edited the occurrence record (including, but not limited to, the unique expression of ideas, concepts or beliefs).

Compilations and Copyright

Compilations, such as datasets, also can be protected under copyright and licensing, but there are limits to the types of compilations that may be covered. For example, a simple listing of facts in alphabetical, numerical, or other order is not likely to qualify as a compilation that can be protected by copyright. This limitation was tested by the US Supreme Court in the case Feist v. Rural (1991), in which the Court ruled that the Rural Telephone Service Co. could not protect their whitepages because (a) the contents were fact, and therefore, not subject to copyright, and (b) that an alphabetical listing of facts did not demonstrate sufficient creative expression to warrant protection. Other court cases have upheld this via similar rulings, including Assessment Technologies v. Wiredata (2003) and Publications International v. Meredith Corp (1996).

In sum, unless there is sufficient and obviously discernible creative expression used to create and organize the compilation, it cannot be protected by copyright. Further, the facts contained within the compilation cannot be protected either.

Copyright places a set of restrictions and limitations on content. Licenses are a means to open up these restrictions and limitations and allow licensees the opportunity to use the content in specific ways as defined by the license type or Terms of Use. The majority of content in datasets published to VertNet, GBIF, or other portals is factual. In addition, the structure of the dataset itself is based on Darwin Core, a community-ratified data sharing standard which is not subject to copyright. As a result the vast majority of the datasets published to biodiversity portals, including nearly all of the content contained within those datasets, are not subject to copyright.

For the few datasets that may contain images, media, or other creative content (e.g., interviews or radio transcriptions, unique descriptive content taken directly from field books), any license used to protect copyright will apply only to the contents that are copyrightable. Thus, once again, all of those facts are still not covered under copyright.

Images, media, and other creative content that can be discovered via hyperlinks within the dataset are not considered to be a part of the dataset and any copyright associated with that creative content remains reserved.

[**NOTE:** for those of you with images and media held within a dataset, we recommend that you consult your local legal counsel for advice about licensing this content - unless, of course, you want to dedicate it to the public domain.]

Sweat of Brow Doctrine

“But what about all of the time and resources we’ve put into curating and maintaining the data?” you may ask. “That must give us the right to protect our datasets.”

Sadly, the answer is very likely “no, it doesn’t.” This argument is called the Sweat of the Brow Doctrine, and it is based on the idea that authors/creators should be awarded copyright for the compilation of factual data. This argument was rejected by the US Supreme Court in Feist v. Rural discussed above.

The Creative Commons

Many folks in our community are looking to the Creative Commons to address the issue of licensing. The Creative Commons “is a nonprofit organization that enables the sharing and use of creativity and knowledge through free legal tools.”[(1)] Their tools are free, easy to use, well documented, and accepted in most countries around the world. The licenses they offer are standardized and meet three important criteria - they are all based firmly in legal standards, they are easy for humans to read and understand (most humans anyway), and they are machine-readable.

The most common licenses being considered for use by the biodiversity community are:

CC-BY (Attribution)
CC-BY-ND (Attribution-NoDerivs)
CC-BY-NC (Attribution-NonCommercial)
CC-BY-SA (Attribution-ShareAlike)
There are additional combinations of these licenses, such as CC-BY-NC-SA as well.

We could describe each of these license types for you, but the Creative Commons no longer recommends that you use any of these licenses to protect your datasets intended for scientific research. This is due largely to the fact that the content is not protected, nor is the compilation. Furthermore, these licenses can add unnecessary and unexpected complications for users (See, Desmet, Rees et al., and the Science Commons now integrated into the Creative Commons).

In brief, Creative Commons has stated, “CC does not recommend use of its NonCommercial (NC) or NoDerivatives (ND) licenses on databases intended for scholarly or scientific use.” [(2)] They also do not recommend the SA for the datasets [(3)] for quite a few reasons, all of which are directly relevant to the kinds of datasets published for biodiversity portals.

The Attribution (BY) license isn’t a bad option, but it creates an unnecessary burden on both data users and publishers in three ways. First, BY can become the cause for unnecessary attribution stacking because users would be obliged to cite the owner for each record used from each data set (a simple example of this is described in this blog post, and in Section 5.3 of this Creative Commons page). Second, the Attribution license places unnecessary burden on the user to decide which data are in the public domain vs those protected by a license. One very likely outcome of this circumstance is that none of the data will be used to avoid any legal confusion or additional work. Finally, it places an additional burden on the data publisher to both police the use of their published data and to invest both human and financial resources in an effort to a) enforce the license, and b) prove that the data in question can be protected.

In all cases, all of these licenses, apply only to any (rare) copyrightable content within the dataset or compilation. The rest of content, including the structure of the database or dataset, will be considered to be in the public domain.

VertNet Recommends...

Based on all that has been described above (and we didn’t even introduce some of the more mind-bending of the issues lurking within an institutional dataset), we are left with a simple choice: CC0 (CCZero).

The CC0 is not a license, it is a waiver that:

1. Commits the contents of the dataset to the public domain (where these facts live already);
2. Waives any and all rights to the data and dataset (which in almost all cases, never existed to begin with);
3. Gives users the freedom to share or use it however they see fit;
4. Exempts the rights holder (if there was one) from any liability for the ways in which the data or dataset are used; and,
5. Is already in use for data shared by a wide-range of scientific institutions.

CC0 is really the only option for sharing datasets that makes sense because it doesn’t place any restrictions on data use and it doesn’t assume rights exist where they are absent.

There is one other option known as the PDDL, or Open Data Commons Public Domain Dedication and Licence, created by the Open Knowledge Foundation via the Open Data Commons. The PDDL predates the CC0 and it performs a nearly identical function by waiving all rights and liabilities associated with the data, dataset, and use. Unfortunately, it is not as widely recognized or used as the CC0 around the planet. Thus, the CC0 still makes the most sense, especially for scientific data.

[**NOTE:** The Open Data Commons provides two different licenses for databases, the ODbL and the ODC-BY. These are some excellent licenses, but they apply only to database frameworks and structures, not to the content within a database. Since we, as a community, use a community-ratified standard for sharing data (Darwin Core), the structure of our datasets or databases is not subject to copyright.]

As a good friend and colleague of VertNet once said:

"1. ODbL covers only database
2. Still a license needed for (copyrightable) content
3. Standard licenses that can be used for content: CC licenses or CC0 (or PDDL)
4. Scientific content = 99% facts = no copyright applicable => Use CC0 to clearly mark it as such
5. User happy" [(4)]

We would add only that you may be a data publisher, but you're probably a data user too, so we encourage you not to make sharing data any more complicated than it needs to be, both for you and for your colleagues.

The Real Solution

As the number of datasets grows and access to these datasets expands, the use of data licenses, even when they apply, is not the solution to the problems of giving credit where credit is due. We believe the solution is a robust set of social norms that govern the behavior of people who publish and use biodiversity data. This won't happen overnight, but the process has begun. Canadensys has already published a set of norms for data use and publication. VertNet is in the process of adopting these norms (as of March, 2014), and encourages other aggregators to encourage the adoption of similar norms within their communities.

As with all VertNet resources, if you have any questions or comments, please let us know (dbloom@vertnet.org).

I've Selected a License, Where Does it Go?

Once you have decided on the license or waiver you wish to use, you need to make sure it's

attached to your data set. We recommend the following protocol:

List any and all copyrights, licenses, and waivers in the Darwin Core field ***rights***. To do this, please provide a web link to the rights or license statement, instead of including the license type itself. For example:

Good content for ***dwc:rights***: https://creativecommons.org/publicdomain/zero/1.0/

Not so good content for ***dwc:rights***: CC0

You are welcome to have more than one link in this field, if you find yourself in a circumstance that requires more than one statement of copyright, license, or waiver. We do recognize that licenses and waivers are not rights as defined by Darwin Core, but given the current status of Darwin Core, this is the field in which this data is best presented.

If you have additional terms of use or norms that you require data users to follow, links to these requirements should be posted in the ***dwc:accessRights*** field. Some examples of these additional terms of use include the Norms for Data Publication and Use by Canadensys, and adapted by the University of Kansas. VertNet will have a version of these norms available for public use in the near future.

If you are using the GBIF's Integrated Publishing Toolkit to publish your data, the content from both ***dwc:rights*** and ***dwc:accessrights*** should be in the Additional Meta Data section in the metadata field, ***IP Rights***.

Footnotes and Other Resources

(1) https://creativecommons.org/about
(2) http://wiki.creativecommons.org/Data
(3) http://sciencecommons.org/resources/faq/database-protocol/#why_not_sa
(4) Personal Correspondence between David Bloom and Peter Desmet, 14 March 2014.

Resources

Data - Creative Commons
Why we should publish our data under CC0
CC0
CC0 use for data
Feist v. Rural
Assessment Technologies v. Wiredata
Publications International v. Meredith Corp
Sweat of the Brow Doctrine

Creative Commons
Showing you this map of aggregated bullfrog occurrences would be illegal
Response to GBIF request for consultation on data licenses
Open Knowledge Foundation
Open Data Commons
Darwin Core
Norms for data use and publication
Creative Commons Licenses
Analyzing GBIF data licenses
Attribution Stacking via Creative Commons
Attribution Stacking Example

Version History

David Bloom	*Orig Release, 14Mar2014*
David Bloom: Copy edits, minor legal corrections	Revised, 31Mar2014