Response to Standard licenses for GBIF-mediated data: Proposed options for action

Jonathan Rees (1,2), Karen Cranston (1,2), Hilmar Lapp (2), Todd Vision (2,3)
(1) Open Tree of Life Project, http://opentreeoflife.org
(2) National Evolutionary Synthesis Center (NESCent), Durham, North Carolina, USA
(3) Dept. of Biology, University of North Carolina at Chapel Hill

Summary

As a data aggregator, the goal of GBIF should be to find policies that benefit both its data providers and data reusers. Clearly, a GBIF that has no or few data will have little value, but so will a GBIF full of data that is encumbered with restrictions to an extent that stifles reuse. Our response follows from the proposition that promoting data reuse should be a shared interest of *all* the parties: data providers, data users, and GBIF itself. We feel the consultation document missed the opportunity to recognize this shared interest, and that furthering the goal of data reuse should in fact be a primary yardstick by which different licensing options are measured.

Tracking the reuse of data is a critically important goal, as it provides a means of reward to data providers, allows scrutiny of derived results, and enables discovery of related research. Initiatives such as DataCite have have made considerable progress in recent years in enabling tracking of data reuse by addressing sociotechnical obstacles to tracking data reuse. By contrast, the consultation, in our view, puts undue weight on legal requirements for attribution. Legal instruments such as licenses are unsuitable, not designed for, and of little if any benefit for this purpose. Moreover, in most of the world, there is little to no formally recognized intellectual property protection for data, and it is on such protection that licenses rest.

In short, our recommendations are (1) that all data in GBIF be released under Creative Commons Zero (CC0), which is a public domain dedication that waives copyright rather than asserting it; (2) GBIF should set clear expectations in the form of community norms for how the data that it serves is to be referenced when reused, and (3) GBIF should work with partner organizations in promoting standards and technologies that enable the effective tracking of data reuse.

We note that our analysis is based on our understanding of the law; we are not legal professionals and this is not legal advice.

Copyright law does not apply to data

Copyright licenses are not legally effective for data, including the data served by GBIF, since only creative expression is protected by copyright. This may change when GBIF includes audio-visual media or photographs, but things like the place and time where an organism was collected or spotted, and the determination of what species it belongs to, are not creative expression. Thus, Creative Commons licenses are out of scope.

The only formal instrument fully appropriate for such data is CC0, which is a waiver, not a license. CC0 helps to clarify the legal status of data by signaling that copyright does not apply (or is waived if it does). (The PDDL is similar to CC0 and would have a similar effect.)

Using a legal instrument, such as a copyright notice, in order to compel behavior is useful only if one is prepared to take on the high cost of litigation. The legal case for prosecuting inadequate data attribution would be very weak since copyright does not apply. Thus, a pseudo-legal attribution requirement would be, in the most charitable interpretation, a clumsy way of expressing a community norm and, in the least charitable interpretation, an attempt to coerce behavior through empty legal threat.

Data use agreements stifle reuse and hinder reproducibility

Because copyright does not apply, if one wants to impose legally enforceable conditions on the reuse of data, such as an attribution requirement or non-commercial use restriction, one has to use a contract, usually called a 'Data Use Agreement' (DUA). For a DUA to be in effect, the party receiving the data has to take some action that can be construed as agreement to the DUA. This is the mechanism currently employed by GBIF; before one can proceed to any page with data, the GBIF site requires the user to agree to the GBIF DUA.

However, contract protection for data is weak, since it only governs the actions of the party entering into the DUA. If the data were to "leak" to a third party not bound by the DUA, the third party would be unconstrained in how it copies or uses the data, and the data would become public domain. To prevent this, a DUA must either prohibit further distribution, or be "contagious", i.e. allow distribution only when a new DUA is forged with each receiving party. In general DUA enforcement is difficult because it is hard to attribute any leaked copy to any particular DUA signer.

CC licenses do not work as DUAs because they are not contagious as DUAs. They rely on the statutory nature of copyright protection in order to be binding on *everyone*. Without contagion, the data is public domain for everyone except the parties that have entered into the DUA.

Requiring users to enter into a DUA activates legal departments ('who is authorized to agree to one?', 'might we get sued?') and thus incurs a cost to users. In doing so, it stifles rather than promotes reuse of data. Furthermore, it hinders reproducibility of scientific results. Data under a DUA, and used in obtaining a scientific result, need to be archived so that the result can be reproduced. However, a DUA that prohibits redistribution (and DUAs that don't are ineffective) means that the data cannot be archived.

The non-commercial use condition inhibits scientific use

Even supposing that copyright applies, or that a DUA can be made to be effective and convenient, the above archiving scenario is also inconsistent with the non-commercial-

use restriction. Journal publishers and data repositories engage in commercial transactions that involve articles and their supplemental materials, even when they are open access and/or not-for-profit. Thus, an unintended side effect of such a clause is to exclude the use of data in journals and repositories, effectively putting it outside the reach of scientific discourse.

Legal attribution requirements are inferior to community norms

Common ethical norms in scholarship and journalism already require authors to reference their sources. To be found in violation of this can be professionally damaging, and these norms are effectively enforced by the scholarly communications community.

In contrast, violating a legal requirement under a copyright license (were it to apply) or DUA would require the publisher to take legal action in order for the violation to have negative consequences. Furthermore, the legal attribution requirement may be satisfied in a way that does not satisfy community norms, such as via hidden document metadata in supplementary material.

Data that gets reused and recombined for new research questions may itself have been reused and recombined. In situations like this, attribution requirements can lead to "stacks" of attributions whose management is challenging and error-prone. We suggest that the way to solve this is by participating in initiatives that work on better tools to manage attribution of integrated and recombined datasets, rather than making scientists believe they may be violating the law, or worse yet, that they cannot recombine the data in the way they need to because of technical or legal limitations.

Response to specific questions

Drawing on the above preface and analysis, we respond below to the 3 specific questions posed in the consultation:

1. Do you have any comments on the plan to associate all GBIF-mediated data with a machine readable licence?

> Terms of reuse that are widely used and machine readable are preferable to ones that are custom-written or not machine readable. We favor use of a machine-readable CC0 public domain waiver (not a license).

2. Do you have an opinion on the relative merits of Creative Commons, Open Data Commons or other licence types in the context of the GBIF network?

> In the GBIF context, we feel copyright-based licenses are inappropriate for the reasons given above.

3. Which of the two options described in section 8 of this document should GBIF pursue? If you support “Option 2”, would your position be modified if it resulted in a significant decrease in data published to the GBIF network?

We regret the antagonism between data aggregators and data providers that is implied by the way the question is expressed. Data providers removing or withholding their data from the GBIF network because GBIF chooses terms of reuse that best promote data reuse rather than stifling it are arguably acting against their own best interest. We suggest that rather than raising false hopes about the applicability and effectiveness of licenses for assuring proper attribution of data, GBIF work with its community and with partner organizations to strengthen community norms for data citation and infrastructure for tracking data reuse.