PLOS ONE

RESEARCH ARTICLE

An analysis and metric of reusable data licensing practices for biomedical resources

Seth Carbon[1]*, Robin Champieux[2], Julie A. McMurry[3], Lilly Winfree[4], Letisha R. Wyatt[2], Melissa A. Haendel[3,5]

1 Berkeley Bioinformatics Open-source Projects, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **2** OHSU Library, Oregon Health & Science University, Portland, Oregon, United States of America, **3** Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon, United States of America, **4** Open Knowledge International, London, United Kingdom, **5** Oregon Clinical & Translational Research Institute, Oregon Health & Science University, Portland, Oregon, United States of America

* sjcarbon@lbl.gov

OPEN ACCESS

Citation: Carbon S, Champieux R, McMurry JA, Winfree L, Wyatt LR, Haendel MA (2019) An analysis and metric of reusable data licensing practices for biomedical resources. PLoS ONE 14 (3): e0213090. https://doi.org/10.1371/journal.pone.0213090

Editor: Rashid Mehmood, King Abdulaziz University, SAUDI ARABIA

Received: July 26, 2018

Accepted: February 14, 2019

Published: March 27, 2019

Data Availability Statement: All data are within the paper or are available at our GitHub Repository (https://github.com/reusabledata/reusabledata) and Zenodo (https://zenodo.org/record/1247562).

Funding: This work was support by the National Institutes of Health awards OT3TR002019, U24TR002306, and R24OD011883, awarded to Dr. Haendel. Seth Carbon was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The funders

Abstract

Data are the foundation of science, and there is an increasing focus on how data can be reused and enhanced to drive scientific discoveries. However, most seemingly "open data" do not provide legal permissions for reuse and redistribution. The inability to integrate and redistribute our collective data resources blocks innovation and stymies the creation of life-improving diagnostic and drug selection tools. To help the biomedical research and research support communities (e.g. libraries, funders, repositories, etc.) understand and navigate the data licensing landscape, the (Re)usable Data Project (RDP) (http://reusabledata.org) assesses the licensing characteristics of data resources and how licensing behaviors impact reuse. We have created a ruleset to determine the reusability of data resources and have applied it to 56 scientific data resources (e.g. databases) to date. The results show significant reuse and interoperability barriers. Inspired by game-changing projects like Creative Commons, the Wikipedia Foundation, and the Free Software movement, we hope to engage the scientific community in the discussion regarding the legal use and reuse of scientific data, including the balance of openness and how to create sustainable data resources in an increasingly competitive environment.

Introduction

In order for biomedical discoveries to be translated into human health improvements, the underlying data must be thoroughly reusable: one should be able to access and recombine data in new ways and make these recombinations available to others. Significant resources and influence have been invested and leveraged to make biomedical data publicly available and scientifically useful [1–3]. Projects such as the NIH NCATS Translator, Data Commons, Illuminating the Druggable Genome, Bgee, and the Monarch Initiative demonstrate that efforts to aggregate and integrate data are seen as a worthwhile undertaking. However, despite these efforts, technical, logistical, descriptive, and legal barriers continue to impede data interoperability and reusability. We are specifically concerned with the ways in which data licensing practices have created widespread legal and financial barriers across the biomedical domain.

had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

While a great number and variety of publicly-funded biomedical data are ostensibly "open", and some are accessible via aggregated databases, complex licensing issues hinder them from being put to their best use [4–10]. A lack of licensing rigor and standardization forces data users to manually seek, often repeatedly and from multiple data providers, essential reuse and redistribution permissions. Issues include missing licenses, nonstandard licenses, and license provisions that are restrictive or incompatible. The legal interpretation of, and compliance with, database license and reuse agreements has become a significant burden and expense for many fields in the scientific community [11], where a complex and lengthy set of legal negotiations may be required for a data integration project to legally and freely redistribute all of its relevant data. Ironically, few data resources have the capacity to pursue policy violations and, in our experience, most researchers who restrictively license their data do so because they want to be credited for their work and are unaware of the downstream reuse implications. Thus, it is not uncommon for researchers to ignore license restrictions. This landscape does not benefit data providers, users, or scientific progress, especially from the perspective of reuse [12].

The (Re)usable Data Project (RDP) originated from discussions within the NCATS Biomedical Data Translator project (https://ncats.nih.gov/translator), which aims to integrate and leverage biomedical information across a vast diversity of sources, from fundamental molecule and model organism research all the way to the clinical setting. While licensing issues influenced many of the reuse barriers we discussed, participants could not agree on licensing standards, illustrating the complexity and confusing state of the data licensing landscape. The RDP was created to systematically describe the current data licensing landscape from the perspective of data aggregation, reuse, and redistribution of publicly funded biological and biomedical data resources, starting with the data resources in use within the Translator project, and expanding from there. The RDP's rubric for evaluating data reusability and re-distributability includes a set of criteria and a scoring system that categorizes and weighs licensing and database characteristics, for example the findability and type of licensing terms, negotiation requirements, scope, accessibility, as well as use case and user type restrictions. The RDP aimed to develop a scoring system that is intuitive and comprehensive, and also defensible and agnostic to domain and scientific task. It is important to note that we are not lawyers and the RDP does not provide legal advice. We are a group of scientists, engineers, librarians, and specialists that are concerned about the use and reuse of increasingly interconnected, derived, and reprocessed data. We want to make sure that data-driven scientific endeavors can work with one another in meaningful ways without undue legal burden. We hope the RDP licensing evaluation rubric will help others navigate the legal synthesis and redistribution of public data and enable data providers to choose licensing terms that make it easier for others to use and redistribute their data.

Methods

The RDP's main efforts have been the creation and application of a rubric that defines the licensing characteristics of aggregated data resources (collections of digital biomedical data from multiple contributors) and measures how these licensing behaviors impact reuse. This includes the capture of structured metadata that provide a high-level description of a resource, a working view of its licensing, information to reconstruct the decisions behind our evaluations, and additional notes of interest to others wanting to understand the reusability of a resource's data. The rubric was constructed for the evaluation of and only applied to group and institution size public resources, not to individuals' datasets or contributions.

License categorization

In order to facilitate several points of evaluation in the RDP rubric and illustrate shared qualities among related licenses, the RDP uses an internal categorization of licenses and licensing information, organizing them into six reuse-oriented types—these types are separate from the rubric and are used during analysis. While we acknowledge that the licensing landscape is much more complicated than these categories communicate, classifying licenses via these basic terms was conceptually helpful and provided needed efficiency and simplicity during the evaluation process. The six license types are described, and examples are provided below.

Permissive. Permissive licenses permit reuse, transformation, and redistribution, allowing for attribution. Examples include the Creative Commons Attribution 4.0 International license (CC BY 4.0, the MIT License (MIT), and public domain declarations.

Copyleft. Copyleft licenses allow for reuse, transformation, and redistribution. However, new contributions derived from the original data resource must be distributed under the same license. Examples include the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0) and the GNU General Public License v3.0 (GNU GPL 3.0).

Restrictive. Restrictive licenses provide more permissions compared to data resources wherein all copyrights have been reserved by the provider, but still include terms that may hinder data integration and reuse. Examples include the Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0).

Private pool licenses. A "private pool" license is one where the resource requires data users to add their own data to the pool, or limits the accessibility of derivative data to others that have also joined the pool. Conceptually, this is similar to some copyleft licenses, but without the public "open" component.

Copyright. This category is used both for licensing statements that positively assert a resource provider's exclusive copyrights, often referred to as "all rights reserved", and for when a resource makes no statement about the disposition of its data. Under current US copyright law, creators do not have to explicitly register or copymark their creations to claim their exclusive rights [13].

Unknown. This category captures licensing statements that have conflicting terms, incompatible license references, or are so nonstandard or unclear that a data resource's reuse terms cannot not be reasonably understood or confirmed.

In our view, only data resources within the permissive category facilitate reuse without negotiation, license alignment, or other burdensome tasks. All other categories have issues that hinder reuse.

Evaluation criteria

The RDP's star rubric (http://reusabledata.org/criteria) consists of five criteria that address: the findability and type of licensing terms, the scope and completeness of the licensing, the ability to access the data in a reasonable way, restrictions on how the data may be reused, and restrictions on who may reuse the data. Each of the five criteria (labeled A-E) is quantified by up to a 1.0 star value, so data resource evaluations (i.e., scores) can range from 0 to 5 stars. The rubric is quite extensive, with a branching evaluation workflow, multiple rules and decision points within most parts of the criteria, and bypasses for cases where a particular rule may not apply or make sense. The accumulation of a score on a single scale was a fundamental choice (as opposed to different non-accumulating axes); it has been designed such that a higher score should necessarily equate to a more reusable resource, and a lower score to a less reusable resource. The rubric and workflow used can be found in S1 Appendix.

The first three criteria (labeled A, B, C) refer to mechanical aspects of license discovery and resource access. Here, standard licenses (i.e., licenses that are invoked referentially or by template, like Creative Commons licenses, the Open Database License (ODbL), etc.) are preferred since custom language and terms may require negotiations and possible involvement of institutional counsel to clarify and confirm the rights and permissions. Similarly, the ability to access data—to actually act on a license in a reasonable way—is fundamental to the examination of our resources. The last two parts of the criteria (D and E) evaluate the reuse aspects of the licensing terms. Part D considers any restrictions on the *kind* of reuse and part E considers any restrictions on *who* can reuse the data. One star is awarded for each part when all types of reuse are permitted and all audiences can reuse the data without negotiation; however, the rubric does make allowances for some restrictive terms if "research" or "non-commercial" reuse contexts are frictionlessly facilitated. Each part of the criteria can be summarized with the following questions:

Clearly stated (Criteria A). Is the license or terms of use in an easy-to-find location? Is there one, unambiguous license, as opposed to multiple, conflicting versions? Is the license standard?

Comprehensive & Non-negotiated (Criteria B). Does the license clearly define the terms of continuing reuse without need for negotiation with the data creators or resource curators? Does the license have a complete scope that covers all of the data and not just a portion?

Accessible (Criteria C). A license without meaningful access is not an actionable license; does the resource provide its data in a reasonable, good-faith location, and is there a reasonable and transparent method of accessing that data in bulk?

Kinds of reuse (Criteria D). Are all types of reuse (copying, editing, building upon, remixing, distributing) allowable, with or without attribution?

Who may reuse (Criteria E). Can any type of user group reuse the data?

The RDP's rubric emphasizes U.S. based, non-commercial, research requirements for data reuse and redistribution. This perspective reflects our own experience as data resource aggregators and primary data producers, and our frustrations in navigating terms of use that limit certain communities' (e.g., clinical researchers) and kinds of reuse (e.g., new tools) [14]. We also found that this specific and practical point of view was helpful in keeping the rubric and its application logically manageable. When this perspective has limited our evaluations, we have captured the fact that other entities may have different results.

Rubric usage

To date, we have fully evaluated 56 data resources (included in S3 Appendix) with the RDP's star rubric. As the idea for the RDP emerged from an NCATS Biomedical Data Translator meeting, we originally evaluated data resources used by the Translator and the Monarch Initiative, wherein the reuse and free redistribution of publicly available data for disease discovery has been particularly burdensome. We then expanded our scope to evaluate model organism databases (MODs) and data resources that the newly funded NIH Data Commons Pilot Phase will address [15]. We also evaluated several resources, including data aggregators, that were brought to our attention by the community.

Each data resource received a score from 0 to 5 stars according to the rubric (http://reusabledata.org/criteria). The star ratings are designed to give a high-level feeling for the reusability of a resource given our rubric. The ratings are automatically generated from a resource's detailed evaluation file, which includes a detailed list of violations found (by their internal violation code), as well as details that the evaluator found around the violation. These detailed violation notes are available for all resources and public examination on our website at http://

reusabledata.org. The rubric workflow during evaluation was created so that it is a) not possible to have a high-scoring resource that is not "reusable" and b) a "reusable" resource will be high-scoring. Given that the exact weight that any downstream consumer might want to assign to a particular criterion could vary in any particular use case, we felt that it was most important to give a broad feeling, rather than a detailed explanation of license-related reusability rubric violations, which are available within the public evaluation file. Further, the star rating allows for a quick visual understanding of the reusability of a resource in a format that most users are familiar with (similar to how books, restaurants, and services are rated online).

Given the mechanics of the accumulation of stars as the rubric workflow is executed, they can roughly be interpreted as follows:

- 5 stars: The license unambiguously allows the unfettered (re)use and redistribution of the data.
- 4 stars: The license unambiguously allows (re)use and redistribution of the data under some terms.
- 3 stars: The license is clearly stated, unambiguous, and of a standard type, and has clear access, but has terms that may greatly impact the (re)use and redistribution of the data.
- 2.5 or fewer stars: There are likely issues in definitively finding the license, ambiguities in the license that hamper further analysis, issues with clean data access, or terms that require legal advice.

The authors curated the sources directly into the RDP's GitHub repository (https://github.com/reusabledata/reusabledata) as YAML files from a template to help ensure the provenance of statements, also including metadata such as source name, description, source type, license type, data access URL, additional issues uncovered during the evaluation, and any commentary about how the license was evaluated. The evaluations were checked by at least two authors, and comments on the evaluations were made on GitHub pull requests to allow for transparency and continued conversation. Evaluations then went through a battery of syntactic and consistency checks. When necessary, the evaluated resource was contacted for clarification. All data and materials used in the publication of this manuscript are available at Zenodo (https://doi.org/10.5281/zenodo.1247562).

Results

Resource evaluation scores

Complete evaluations can be viewed on the RDP website (http://reusabledata.org) and the RDP GitHub repository (https://github.com/reusabledata/reusabledata).

Overall scoring. Of the 56 data resources we evaluated, 22 (39%) received between 4 and 5 stars, indicating that they met our broadest requirements for being reusable, which allowed for some caveats; only 10 (18%) received 5 stars by meeting all parts (A-E) of the criteria (Fig 1). 23 (41%) of the resources received fewer than 3 stars, which is notable because even data resources for which the provider has reserved all copyrights can receive a score of 3 stars if the data covered by the license are easily accessible. 32 (57%) of the resources had 3 stars or fewer, meaning that a majority of resources had significant issues with basic reusability. Overall, average scores by licensing category were: permissive 4.5, restrictive 2.6, copyright 1.4, unknown 0.7, copyleft 3.0, and private pool 1.0.

Criteria violations. When a resource provided inconsistent or no licensing information, only parts A and C of the rubric were used in the evaluation. While one could assume that some of these resources wished to reserve all of their copyrights when no information was found, the ambiguity and lack of clear intent would require clarification and possibly legal

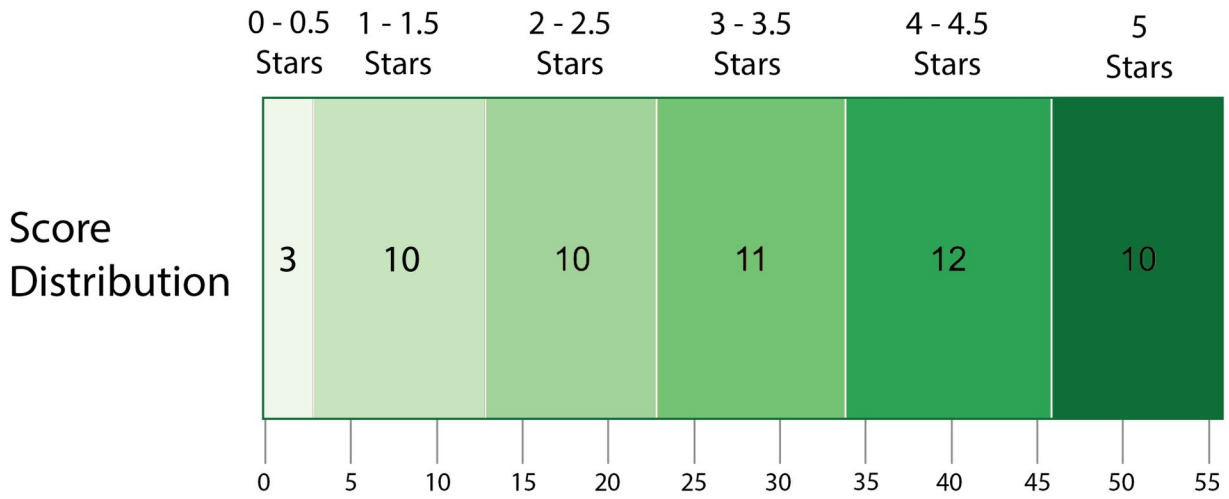


Fig 1. 41% of resources scored fewer than 3 stars. 3 sources scored below 1 star, 10 scored between 1–1.5 stars, 10 resources scored 2–2.5 stars, 11 resources scored 3–3.5 stars, 12 resources scored 4–4.5 stars, and 10 resources scored 5 stars.

https://doi.org/10.1371/journal.pone.0213090.g001

counsel. 11 data resources had such contradictory or missing information; therefore, the summary statistics for parts B, D, and E of the rubric do not include data for these resources (see Fig 2, 'Not Evaluated' category). We have qualified all of the numbers given below to prevent ambiguity.

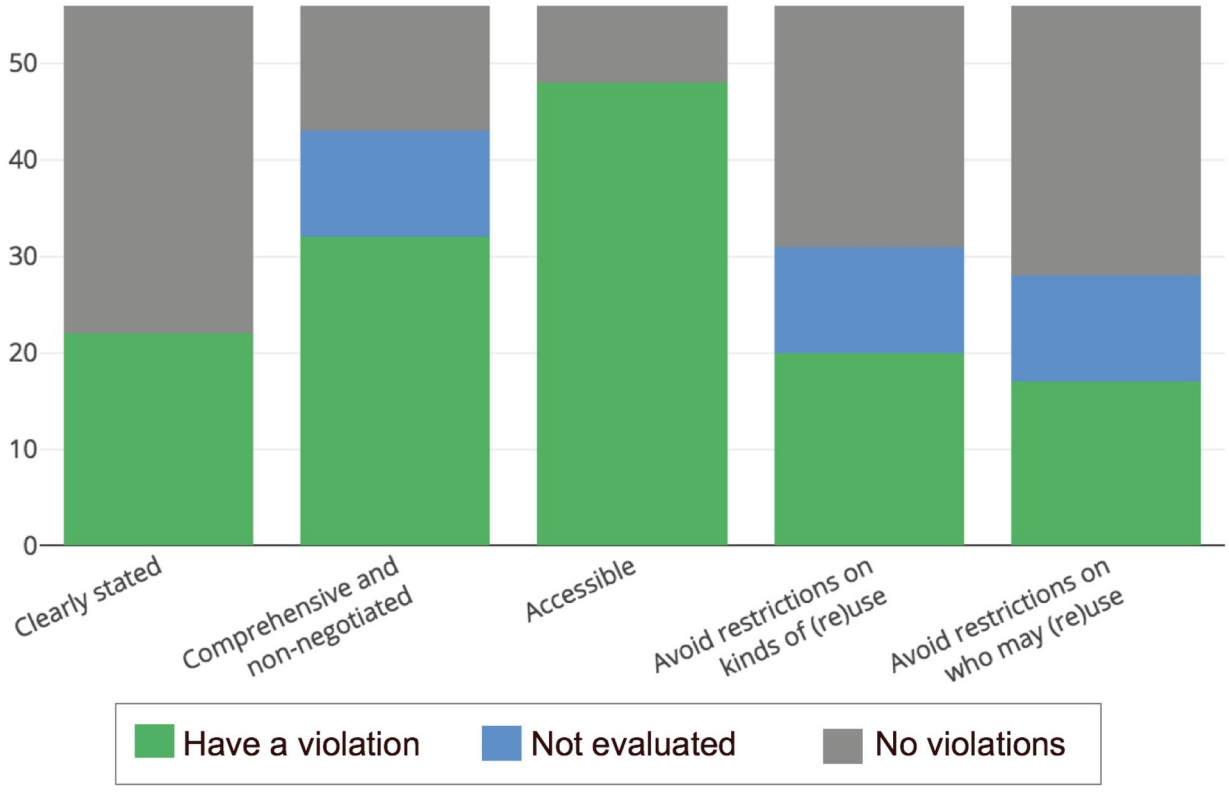


Fig 2. Most sources score poorly on the five categories. 39% of sources have a license that is found & clearly stated. 57% have a license that covers all the data without negotiation. 86% have data that are easily accessible. 36% have no limits on type of reuse. 30% have no limits on who can reuse. Resources that had no findable license were not further evaluated for categories B, D, and E.

https://doi.org/10.1371/journal.pone.0213090.g002

(Clearly Stated, Criteria A) We found that 24 (43%) resources used an explicit standard license and 22 (39%) used custom terms, 5 (9%) had inconsistent licensing information, and 5 (9%) had no licensing information. Table 1 illustrates the count of resources by license type and the associated licensing category. Resources with custom licensing language fell into several licensing categories: 12 were restrictive, 9 permissive, and 1 private pool.

(Comprehensive & Non-negotiated, Criteria B) 32 (57%) of the resources included a license that was explicit, comprehensive, and unambiguous in scope over the data.

(Accessible, Criteria C) We found that 48 (86%) of the resources passed criteria C by making all of their data reasonably accessible at an API endpoint or structured download site.

(Kinds of Reuse, Criteria D) We found that 20 (35%) of the data resources included clear and unambiguous licensing language that provided for unfettered reuse for all purposes

(Who May Reuse, Criteria E) 17 (30%) included clear and unambiguous language that provided for unfettered reuse for all user groups.

The majority of resources failed to receive a full star for parts A, D, or E of the rubric (Fig 2).

The majority of resources (60.7%) had non-permissive licensing, which includes restrictive, copyright, copyleft, private pool, and unknown licenses; only 39.3% of resources had permissive licensing (Fig 3). Furthermore, we found 13 distinct licenses within the 6 license categories, and an additional 22 custom licenses (with 12 of these having custom permissive terms, 9 with custom restrictive terms, and one custom private pool. Fig 4 and Table 1), meaning that for the 56 data resources, we found 35 distinct licenses, with a majority of these being custom, or non-standard.

Counts shows that custom licensing dominates, with 21 out of 56 resources using custom terms. References to most common license terms are available in S3 Appendix.

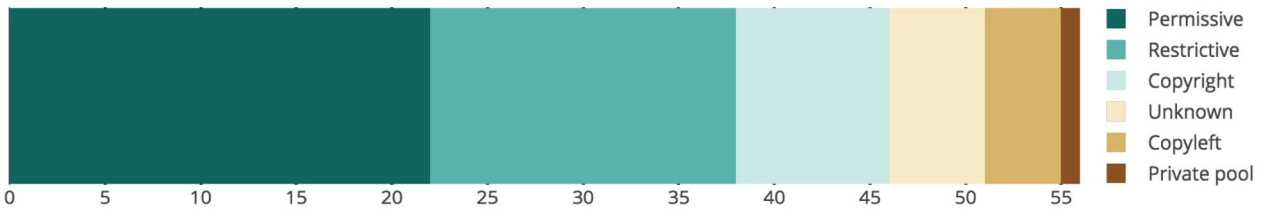
Discussion

As data users and stewards, we have encountered and been frustrated by the ways in which licensing issues hinder data reuse, integration, and redistribution. While 48 (86%) of the resources we evaluated provided easy and actionable data access, only 10 (18%) received a full

Table 1. Licenses used in evaluated resources, with frequency and license categorization.

LICENSE	CATEGORY	COUNT
custom language (restrictive)	custom	12
custom language (permissive)	custom	9
Creative Commons Attribution 4.0 International (CC BY 4.0)	permissive	8
Creative Commons Zero 1.0 (CC0 1.0)	permissive	3
MIT License (MIT)	permissive	1
public domain declaration	permissive	1
no license	copyright	5
all rights reserved	copyright	3
inconsistent or multiple	unknown	5
Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0)	restrictive	1
Creative Commons Attribution-NoDerivatives 3.0 (CC BY-ND 3.0)	restrictive	3
GNU General Public License v3.0 (GPL 3.0)	copyleft	1
Creative Commons Attribution-ShareAlike 3.0 (CC BY-SA 3.0)	copyleft	1
Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0)	copyleft	1
ODC Open Database License v1.0 (ODbL 1.0)	copyleft	1
custom language (private pool)	private pool	1

https://doi.org/10.1371/journal.pone.0213090.t001



License Type	Percent	Count	Description
Permissive	39.3 %	22	Permits reuse, transformation, and redistribution, allowing for attribution; e.g. MIT
Restrictive	28.6 %	16	Includes terms that may hinder data integration & reuse; e.g. CC BY-ND 4.0
Copyright	14.3 %	8	Asserts a resource provider's exclusive copyrights.
Unknown	8.93 %	5	Terms very unclear or conflicting.
Copyleft	7.14 %	4	Allows for reuse, transformation, and redistribution, but derivatives must have the same license; e.g. GPL 3.0
Private pool	1.79 %	1	Requires data users to add their own data to the pool, or limits reuse to those users already in the pool.

Fig 3. Overall, non-permissive licenses are the largest group. The breakdown of evaluated licenses according to their reuse category reveals that 61.8% of resources use non-permissive licenses.

https://doi.org/10.1371/journal.pone.0213090.g003

5-star rating and 32 (57%) of the resources received 3 stars or fewer, indicating that there were serious barriers to reuse. These findings support our experience, in that the data we need are often accessible, but cannot be legally reused or redistributed. Missing licensing information and the variability and potential incompatibility of license types are primary areas needing improvement. For large data integration projects that ingest data from multiple resources to derive new knowledge and provide new tools, this landscape requires costly and lengthy interactions with individual organizations and institutions, that may involve legal considerations.

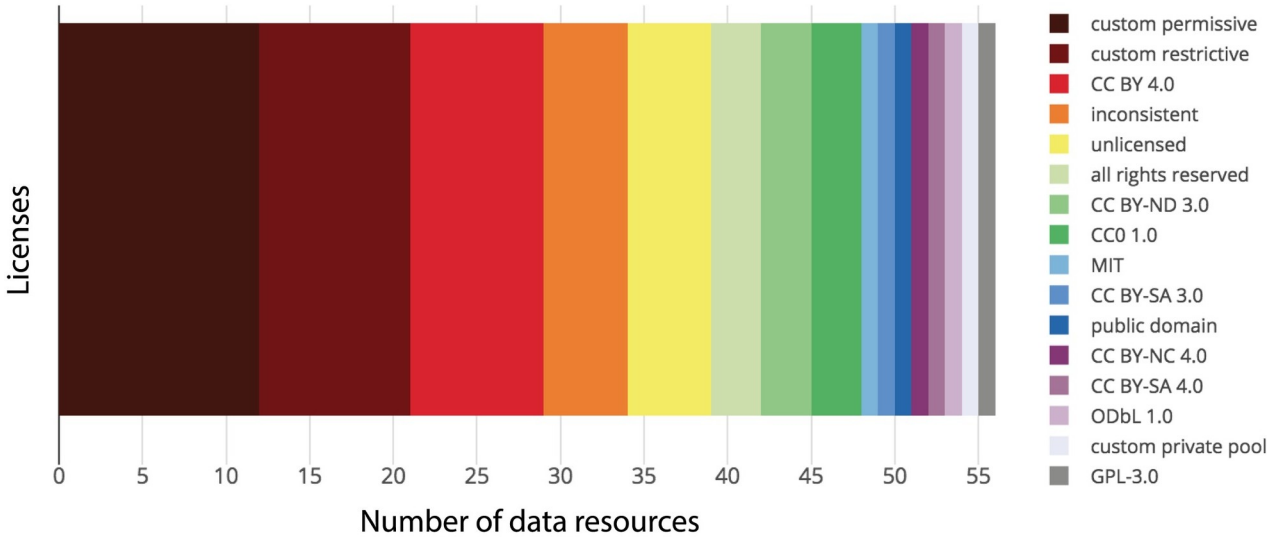


Fig 4. A wide variety of licenses were used within the evaluated resources. In the 56 resources, we found 35 distinct licenses, including large numbers of custom-restrictive and custom-permissive licenses.

https://doi.org/10.1371/journal.pone.0213090.g004

It is noteworthy that the largest single type of licenses were custom licenses, suggesting that resource providers either felt that a standard license did not meet their needs or that they were not knowledgeable about standard licenses. Moreover, while the majority of custom licenses were restrictive, 9 were permissive, leading us to wonder if some needs and intentions are not being met by the existing set of standard permissive licenses. Although it is encouraging that the largest single license category is permissive, the total body of non-permissive license types is larger. In the future, we would like to continue the community discussion to understand if there are any gaps between resource licensing intention and license selection.

Our goal with the (Re)usable Data Project is to draw attention to the licensing issues that are challenging the reuse of valuable biomedical data, not to criticize any specific organization or data resource in the community. Rather, we hope the RDP will encourage the community to work together to improve licensing practices in order to facilitate reusable resources for all, with the option of using the RDP website (http://reusabledata.org) as a possible initial community focal point. Reusing data *en masse* comes with numerous challenges and can be better enabled via the practices articulated in initiatives like the FAIR Data Principles and FAIR-TLC evaluation framework [16–17]. The RDP's focus on licensing issues—narrowly scoped to reusability—is meant to draw attention to the pervasiveness of current practice failures and their effects. The current proliferation of and changes to funder and journal data sharing policies offer an opportunity to provide explicit direction to upstream data contributors about licensing and other practices that are consistent with the FAIR data principles and positively impact reuse. Additionally, repositories and aggregated resources can require that data deposited meet specific licensing terms. For examples, resources available within Wikidata must have a license at least as permissive as the Creative Commons CC0, the Creative Commons "public domain" tool. Ultimately, these kinds of licensing choices can impact the lifetime of a resource: a resource with reusable terms and access can be forked or incorporated into other resources, extending its value, length of use, and audience reach.

As part of our evaluation process, we often contacted data resources with clarifying questions about their licensing information and tracked these conversations on the RDP GitHub repository. These exchanges led to more accurate evaluations and sparked dialogue about how resource curators could improve the clarity of their licenses. Additionally, in response to our outreach on social media, we received requests to evaluate eight additional data resources. We believe this early engagement demonstrates a community interest in enabling reuse, and a desire to contribute to open discussions about how to address licensing problems. Moreover, while the RDP has been focused on biological and biomedical data resources, the goals and problems we have raised are domain agnostic, and we are communicating with other data communities to ensure that our rubric is relevant and applicable across disciplines, such as the Research Data Alliance (RDA) Legal Interoperability Interest Group.

We do not envision the RDP star rubric and evaluation data as only a tool for analyzing past licensing selections; rather, it could be used to test and guide future licensing choices as well. For example, it could be used by groups considering how to plan for the long-term sustainability of data resources, which may include a variety of monetization options. The RDP rubric could be applied to understand the implications of different strategies, including the potential interoperability between resources and as check on continued data reusability.

While the RDP's star rubric and evaluation results provide specific insight into the data resource licensing landscape, enhancements could include developing criteria to define and assess more complicated interactions and compatibility characteristics between data resources. Exploring the license interaction space more deeply would require the creation of a richer internal model for our data, possibly using ontologies and leveraging the use of reasoners to aid in the task. Finally, licensing is only one barrier to data reusability, and we would like to

capture and add to our analysis information about data resource size, data and resource connectivity, and structured funder information. These improvements could provide a more complete and holistic picture of the reusability and impact of publicly funded research data resources.

Supporting information

S1 Appendix. Star criteria for the (Re)usable data project.
(DOCX)

S2 Appendix. List of all reviewed resources, their categorization, and their license.
(DOCX)

S3 Appendix. Reference information for commonly use licenses.
(DOCX)

Acknowledgments

We would like to thank Arvin Paranjpe, Senior Technology Development Manager, OHSU, for his thoughtful discussion and expertise on academic data licensing. We would also like to thank Noel Southall and Christine Colvis at NCATS, Andrew Su at Scripps Research Institute, and John Wilbanks at Sage Bionetworks for their ongoing commitment to helping us address the data licensing problem.

Author Contributions

Conceptualization: Seth Carbon, Julie A. McMurry, Melissa A. Haendel.

Data curation: Seth Carbon, Robin Champieux, Julie A. McMurry, Lilly Winfree, Letisha R. Wyatt, Melissa A. Haendel.

Formal analysis: Seth Carbon, Lilly Winfree.

Funding acquisition: Robin Champieux, Melissa A. Haendel.

Investigation: Seth Carbon.

Methodology: Seth Carbon, Robin Champieux, Julie A. McMurry, Lilly Winfree, Letisha R. Wyatt, Melissa A. Haendel.

Project administration: Seth Carbon, Robin Champieux, Lilly Winfree.

Resources: Seth Carbon.

Software: Seth Carbon.

Supervision: Melissa A. Haendel.

Validation: Seth Carbon, Robin Champieux, Lilly Winfree.

Visualization: Seth Carbon, Robin Champieux, Lilly Winfree.

Writing – original draft: Seth Carbon, Robin Champieux, Julie A. McMurry, Lilly Winfree, Letisha R. Wyatt, Melissa A. Haendel.

Writing – review & editing: Seth Carbon, Robin Champieux, Julie A. McMurry, Lilly Winfree, Letisha R. Wyatt, Melissa A. Haendel.

References

1. 2018 Nucleic Acids Research database issue and the online molecular biology database collection | Nucleic Acids Research | Oxford Academic. Available at: https://academic.oup.com/nar/article/46/D1/D1/4781210. (Accessed: 15th March 2018)
2. NIH Data Sharing Information—Main Page. NIH Data Sharing Policy
3. NIH awards to test ways to store, access, share, and compute on biomedical data in the cloud. National Institutes of Health (NIH) (2017). Available at: https://www.nih.gov/news-events/news-releases/nih-awards-test-ways-store-access-share-compute-biomedical-data-cloud. (Accessed: 10th March 2018)
4. Cancer Moonshot. Enhanced Data Sharing Working Group Recommendation: The Cancer Data Ecosystem [Internet]. Zenodo; 2016. https://doi.org/10.5281/zenodo.193064
5. Oxenham S. Legal confusion threatens to slow data science. Nature. 2016; 536: 16–17. https://doi.org/10.1038/536016a PMID: 27488781
6. Wilbanks J, Friend SH. First, design for data sharing. Nat Biotechnol. 2016; 34: 377–379. https://doi.org/10.1038/nbt.3516 PMID: 26939011
7. Gilbert N. Legal tussle delays launch of huge toxicity database. Nature News. 2016; https://doi.org/10.1038/nature.2016.19365
8. Balasegaram M, Kolb P, McKew J, Menon J, Olliaro P, Sablinski T, et al. An open source pharma roadmap. PLoS Med. 2017; 14: e1002276. https://doi.org/10.1371/journal.pmed.1002276 PMID: 28419094
9. Haendel M, Mungall C, Su A, Robinson P, Chute C, B Altman R, et al. Request for Community partnership in data resource licensing planning. figshare. 2017; https://doi.org/10.6084/m9.figshare.4972709.v1
10. Policies and Disclaimers—NCBI. (n.d.). Retrieved March 15, 2018, from https://www.ncbi.nlm.nih.gov/home/about/policies/
11. Analyzing the licenses of all 11,000+ GBIF registered datasets—Peter Desmet. Available at: http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html. (Accessed: 10th March 2018)
12. Hrynaszkiewicz I, Cockerill MJ. Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. BMC Research Notes. 2012 Sep 7; 5 (1):494.
13. Copyright in General (FAQ) | U.S. Copyright Office. Available at: https://www.copyright.gov/help/faq/faq-general.html. (Accessed: 15th March 2018)
14. Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species | Nucleic Acids Research | Oxford Academic. Available at: https://academic.oup.com/nar/article/45/D1/D712/2605791. (Accessed: 15th March 2018)
15. "NIH Awards to Test Ways to Store, Access, Share, and Compute on Biomedical Data in the Cloud." 2017. National Institutes of Health (NIH). https://www.nih.gov/news-events/news-releases/nih-awards-test-ways-store-access-share-compute-biomedical-data-cloud (March 10, 2018).
16. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3: 160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
17. Haendel MA, Su A, McMurry J. Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133 [Internet]. 2016. Available: https://zenodo.org/record/203295