Augmenting Data and Code Licenses Offered by Deep Blue Data

Author: Jake Gibson
Date: August 8, 2025
Suggested Citation: Gibson, J. (2025). *Augmenting Data and Code Licenses Offered by Deep Blue Data.* University of Michigan - Deep Blue Documents. https://dx.doi.org/10.7302/26733
License: This report is licensed with CC BY 4.0.

Disclaimer: This report is intended to provide guidance on selecting licenses for research data and code. It was not written or reviewed by a lawyer and no aspect of this report constitutes legal advice.

I. Introduction

Central to the goal of making data FAIR (findable, accessible, interoperable, and reusable) is the process of licensing data. A proper license furthers research and scholarship by making it clear how Works can be distributed and modified (i.e. the R in FAIR).

At this time, Deep Blue Data, University of Michigan's institutional data repository, offers three Creative Commons (CC) licenses for depositors to choose from:

- CC0 1.0 Universal (CC0 1.0) Public Domain Dedication
- Attribution 4.0 International (CC BY 4.0)
- Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

In addition, researchers may specify an "Other" license of their choice in a free text field. However, few depositors use this option, so the default choices provided by the deposit form have a significant impact on how researchers license the Works in Deep Blue Data. See Appendix I for a breakdown of current license usage.

The aim of this report is to assess whether the licenses currently offered by Deep Blue Data meet the needs of all types of Works held by the repository. Despite the name, the repository holds not only datasets but also code, which may be standalone software or used to manipulate and analyze data.

The report will begin with an overview of best practices for licensing both data and code, highlighting differences in recommended approaches. This is followed by a brief discussion of research data governance at the University of Michigan, outlining when the University delegates the power to license data to researchers and when the University retains that right. Finally, the report will conclude with recommendations on updates to Deep Blue Data's deposit form and internal/external documentation.

The topline conclusion of the report is that **Deep Blue Data should expand the licenses offered in the deposit form to include the MIT and GNU General Public License (version 3) licenses. In addition, the repository should revise the deposit form to facilitate licensing the data and code components separately.**

A note on terminology:

When referring to Deep Blue Data deposits, this report adopts the terminology used in the repository's Depositor's Guide. "Work" refers to an entire deposit, which includes data files, code files (if relevant), a documentation file (such as a ReadMe.txt), and metadata (see diagram below). The report will largely focus on the distinction between licensing the "data" and "code" components when they are both part of the same "Work."



II. Open Licenses: Overview

Why license research?

While approaches to open scholarship vary between disciplines, funding agencies, journals and publishers, institutional requirements increasingly consider the sharing of the underlying components of research – wherever possible – as a key condition for reproducible scholarship. This includes research data, code, and documentation.

A central part of the open scholarship process is choosing a license under which to release your Work. In the absence of a license, a researcher implicitly claims full copyright over the data, which can limit its access and reuse. **Proper licensing makes it clear how others can (and cannot) reuse research data** (Chinn et al., 2024).

Types of Licenses

To generalize, there are three main types of licenses used for open scholarship: public domain dedication/waiver, permissive, and copyleft. In addition, individuals may choose to apply more than one license on a Work in a dual/multiple licensing regime.

Public Domain Dedication/Waiver

Public domain dedication waives all copyrights over a Work, placing it in the public domain for all to use. This allows the maximum possible discovery, use, adaptation, and citation of the Work, as there are no restrictions placed on the data's access and modification, including for

commercial purposes (“CC0,” n.d.). From a research impact perspective (as well as promotion and tenure), this can have significant benefits.

Permissive License

Permissive licenses allow individuals to adapt, modify, and distribute a Work without seeking permission from the copyright holder. While the terms of the license may vary, it is common for permissive licenses to require subsequent creators to credit the original copyright holder and list any changes made to the Work (Chinn et al., 2024).

Copyleft License

Copyleft licenses require that any subsequent versions of a Work are shared under open licenses. There are both *strong* versions where derivatives must be shared under the same license and *weak* versions where derivatives must be shared under any open license. Like permissive licenses, copyleft licenses also generally require attribution and enumeration of changes made to a Work (Sen et al., 2008). An example of a copyleft license is the Creative Commons Share-Alike 4.0 license (CC BY-SA 4.0), which is a *strong* copyleft license requiring that if individuals “remix, transform, or build upon the” Work, it must be released under CC BY-SA 4.0.

Dual/Multiple Licensing Regime

In instances where universities determine that there may be revenue generating potential from research, a common approach is to use a *dual* or *multiple licensing regime.* In these situations, the data will be openly shared using a license with *non-commercial* provisions and the University will offer individual commercial licenses to for-profit companies for a fee (Ball, 2014).

License Compatibility

When a researcher reuses data or code released under an open license, they must release their new derivative work under a *compatible* license that fulfills any requirements established by the original license(s).

Like many questions around copyright, determining whether two licenses are compatible is not always straightforward and may require the guidance of an intellectual property expert or lawyer. As such, this report cannot provide details on the compatibility of individual licenses, but the following chart attempts to provide general guidance on combining *types* of licenses in a single work (for guidance on individual licenses, see the Licensing Assistant from Interoperable Europe, an initiative from the European Union).

License Combination	Potential Compatibility
Public Domain & ANY license type	Public domain works, since they do not have any restrictions on their distribution or reuse, can be combined with and released under any subsequent license.
Permissive & Permissive	If a researcher is reusing multiple works licensed under permissive licenses, there is a strong likelihood that they are compatible in some form. To release the derivative work, the researcher should choose a license that meets the attribution requirements of the strictest reused material.
Permissive & Copyleft	If a researcher is reusing multiple works that include both permissive & copyleft licenses, they will likely be required to release the entire derivative Work under the same license as material with a copyleft license. Attribution requirements from all licenses must be followed.
Copyleft & Copyleft	If a researcher is reusing multiple works that include two different copyleft licenses, it is possible they will be *incompatible.* Unless one of the licenses has a *weak* copyleft requirement, it may not be possible to uphold the requirements of both licenses. In rare cases, there are some copyleft licenses that have been approved as compatible. Read more about the one-way compatibility between CC BY-SA and GNU GPLv3.

III. Licensing Research Data

Data Copyright

In the United States, individuals hold copyright for *creative* or *original* works. The degree to which copyright applies to data depends on the creative work of building or structuring it. A set of unstructured facts is generally not copyrightable (Chinn et al., 2024). Neither is raw data produced by technical equipment (Ball, 2014). However, when a researcher has made original or creative efforts to create, obtain, arrange, verify, or modify data, copyright may apply (Chinn et al., 2024).

Approaches to Licensing Research Data

Public Domain Dedication/Waiver

Because of potential difficulties in determining whether data constitutes an original or creative work under US copyright law, the most straightforward approach is to dedicate it to the public domain. This approach allows a creator to waive their copyright and opt out of rights and protections automatically granted to creators. This prevents a researcher from accidentally claiming rights on facts that may be un-copyrightable (Fortney, 2016).

One fear that often accompanies the use of public domain dedication is the lack of attribution requirement in the license. However, regardless of licensing requirements, academic and community norms – which don't necessitate legal enforcement – still require attribution and citation. As Katie Fortney (2016) writes, "If professional ethics aren't enough to dissuade someone from plagiarism or other misconduct, why would the restrictions of a CC license dissuade them…?"

The most common way to dedicate a dataset into the public domain (which is already offered by Deep Blue Data) is the Creative Commons CC0 1.0 license.

A less common, but relevant, license is the Open Commons Public Domain Dedication License (PDDL 1.0), which was designed for use with databases (but is also applicable to data). An additional feature of Open Data Commons' PDDL over CC0 is the option to associate "Attribution-Sharealike Community Norms," which are not legally binding but request that future users uphold norms of attribution and keep data open.

Permissive and Copyleft Licenses

Another common approach to sharing data is to apply a license that specifies the conditions under which the data can be reused. These may include *attribution* requirements, *copyleft* conditions that require any subsequent versions to be shared under the same license, restrictions on *derivative* works, and *non-commercial* provisions that prevent the data from being reused in commercial settings.

The most common suite of prepared, open licenses is Creative Commons (CC). CC's six standard licenses, which include both permissive and copyleft options, are well suited to creative works such as text, images, and data. Creative Common's licenses mix various combinations of *attribution, non-derivative, share-alike* (copyleft), and *non-commercial* requirements.

While these licenses provide significant flexibility some versions can cause challenges for reuse. This is particularly true of *share alike*, *non-derivative,* and *non-commercial* restrictions, which might prevent other researchers from incorporating the data into future Works. The term "*attribution stacking*" is sometimes used to describe the challenges of attempting to fulfill the license conditions of multiple works and their derivatives (Chinn et al., 2024).

Note: See Appendix II for descriptions, benefits, and drawbacks of common licenses that could be used for data.

IV. Licensing Code/Software

Do you need a special license to make code open?

In the software development community, there is a long history of creating and sharing Open Source Software (OSS) which can be freely inspected, modified, and distributed. Over time,

software development communities have coalesced around a set of agreed upon Open Source principles. These principles – as outlined by the Open Source Initiative – vary from those incorporated into licenses like Creative Commons, as they are targeted to ensure executable code remains open. Some of these additional principles relate to the distribution of source code, patent rights, and permission for incorporation into future software (including commercial software) (*The Open Source Definition*, 2007).

As an outgrowth of these commitments, over 100 Open Source Software licenses have been developed over the years. Popular licenses include the MIT License, GNU General Public License v3, Apache License v2.0, among others.

Note: See Appendix III for descriptions, benefits, and drawbacks of common code licenses.

Can you use Creative Commons licenses for code?

While there are similarities between the aim of Creative Commons and Open Source Software licenses, there are important differences in their implementation. Specifically, Creative Commons licenses do not provide terms about the distribution of source code and other necessary components of ensuring software is both freely distributable and modifiable. As a result, *Creative Commons does not recommend using their licenses for code* (*Frequently Asked Questions: Can I Apply a Creative Commons License to Software?*, n.d.).

The one exception to this recommendation is the use of the CC0 1.0 public domain dedication for code since it releases all rights on the use and distribution of a work. While Creative Commons allows for the use of CC0 with code, it should be noted it has not been officially approved for use with software by the Open Source Initiative and does not address patent rights (*CC0 FAQ*, 2016).

How to choose a code license?

When code is deposited into a repository like Deep Blue Data, researchers will need to select an appropriate license for their code. If the researcher has created new code entirely from scratch, they have the freedom to select a license that aligns with their own aims for reuse and distribution. If the researcher is using existing open source code, they must ensure that they are fully compliant with any downstream *copyleft* or attribution requirements from the original license(s).

The following chart, excerpted from the Data Curation Network's Primer on Data and Code Licensing (Chinn et al., 2024), breaks down how research may approach different scenarios:

What is the researcher doing?	*What sort of license should be used?*
Creating entirely new code from scratch	*The researcher gets to choose whichever license they want for their work. They should make sure that any collaborators working on the project with them are aware of the selected license.*
Forking an existing project (copying an existing piece of code to develop it separately from existing projects/developers)	*When creating a fork, check to see what license is being used for the existing project. Some licenses will require the use of the same license as the original projects, while others will allow a new license for the code to be used.*
Creating new code that requires other pieces of code or libraries as dependencies	*Some licenses require that any code that uses libraries or other code as dependencies must use the same license, while others allow a new license to be used. Some licenses may be modified from the default version with additional permissions or exceptions that allow them to be used as dependencies.*
Contributing to an existing project	*Those contributing to an existing project should use whichever license the project already uses, otherwise, their code may not be accepted.*

V. Licensing Data and Code (and Documentation) in the same Work

When a Work contains both data and code, researchers should choose a separate license for each component – one designed for use with data and one designed for code. As earlier sections outline, open data and code need different licensing approaches to ensure the content is accessible and reusable.

An existing example of this approach in Deep Blue is *Data for Improvement of Plasma Sheet Neural Network Accuracy with Inclusion of Physical Information [Data set]* (Swiger et al., 2020), which appliances CC BY-NC 4.0 to the data and MIT to the code.

Selecting Compatible Licenses

In most cases, data and code are separate enough to give researchers leeway over the licenses they select for each component. **When licensing data and code separately, it is beneficial for the researcher to select licenses that are broadly compatible with one another, meaning the licenses have similar permissiveness, attribution, or copyleft requirements (e.g., CC BY 4.0 and MIT).** Choosing compatible licenses ensures that data and code can continue to be shared and modified under the same circumstances.

In situations where data has been directly incorporated into code (e.g., data and code are stored in the same files), researchers can still apply separate licenses to data and code but should seek legal guidance to ensure that data and code licenses are fully compatible with one another. If licenses are incompatible, they may significantly hinder the ability for the Work to be distributed or reused.

Questions around license compatibility can be further complicated in scenarios where the researcher is also reusing data or code, as they must ensure that the original licenses are compatible with any new licenses applied to the Work. This can become especially thorny if there are strong *copyleft* provisions.

While the best guidance on license compatibility will ultimately come from a lawyer with intellectual property expertise, there are resources that can help identify potentially compatible licenses. One helpful tool is the Licensing Assistant from Interoperable Europe (an initiative from the European Union), which can check for the interoperability of popular licenses. Additionally, researchers and curators look for guidance directly from the organizations that govern licenses, as they may provide official approval of compatibility between specific licenses, such as the one–way compatibility between CC BY-SA and GNU GPLv3 (*ShareAlike Compatibility: GPLv3*, 2015).

How should code documentation be licensed?

The act of creating documentation (whether for code or data) is considered a creative act and, thus, is also copyrightable. The content of README files, code documentation websites, and metadata can all be released under Creative Commons license.

When documentation is included in the same work as code, the above recommendations about licensing code and content separately may apply. A simple README may not be extensive enough to warrant separate licensing from code, but particularly in-depth or broad documentation, such as wikis, websites, or books, are strong candidates for separate licensing.

When licensing documentation under a separate content license, ensure code samples are exempted from the documentation license and are released under the same license as the rest of the code (Chinn et al., 2024).

When should researchers release data and code under a single license?

In certain instances, it can make sense to release data (and documentation) and code under a single license.

In instances where the researcher would like to dedicate the entire work to the public domain, Creative Common's CC0 dedication is a recommended tool. Creative Commons has approved its use for both data and code.

In instances where data and code are inseparable to the degree of being indistinguishable, it may make sense to release the entire Work under a single code license. If either the data or code are reused, extra care should be taken to ensure any downstream licensing requirements are fulfilled.

VI. Licensing Research Data and Code at the University of Michigan

At the University of Michigan, Research Data is defined as "recorded material commonly accepted in research communities as necessary or useful to validate, reconstruct, evaluate, replicate, and/or generate findings," including supplementary material such as code, metadata, and observational records required to interpret data. (*Research Data Stewardship Policy*, 2023).

Unlike some other forms of scholarly output, the University owns all Research Data produced by faculty, staff, or students with a formal University appointment. However, **the University generally assigns responsibility for managing and disseminating to researchers (Principal Investigators) or their delegates** (*Research Data Stewardship Policy*, 2023). As a part of this process, the University instructs Principal Investigators to "strive to make Research Data publicly available when possible" and make data available following FAIR and CARE principles (*Research Data Stewardship Policy*, 2023). Thus, Deep Blue should continue to encourage researchers to use permissive, open licenses when they deposit Works that include data or code into the repository.

Licensing Commercially Viable Data and Code

In instances where researchers produce research data and code that is intended to generate revenue or was produced under a sponsored activity agreement, the University does not assign them the right to share data. Instead, these Works are managed by Innovation Partnerships – formerly called the Office of Technology Transfer – and must go through a separate licensing process (*Who Holds Copyright at or in Affiliation with the University of Michigan*, 2011). In these situations, the University will generally try to release the Work under a dual license, providing free access for non-commercial academic use and charging a fee for licenses that permit commercial use (*Research Data Stewardship Policy - Procedural Guidelines*, 2023).

Thus, if curators for Deep Blue Data encounter a Work that is potentially revenue generating, they should contact the Innovation Partnerships before releasing the Work under an open license.

VII. Recommendations for Deep Blue Data

Recommended Licenses

In choosing code licenses to recommend for Deep Blue Data, the following characteristics were considered:

- Is there a strong community of use?
- Are the license's requirements clear and understandable?
- Does the license support open scholarship and encourage reuse?
- Is the license compatible with one of the data licenses already offered by Deep Blue Data?

Based on this criteria, the following three licenses are recommended for use with code (see the chart in **Appendix III** for an assessment of these characteristics):

Open Source Software License	Potentially Compatible Data License
CC0 1.0 ● Public Domain Dedication	**CC0 1.0** ● Public Domain Dedication
MIT (Expat) ● Requires Attribution ● No further restrictions on use ● Very strong community (most popular license on GitHub)	**CC BY 4.0** ● Requires Attribution ● No further restrictions on use
GNU GPLv3 ● Copyleft License (requires derivatives to use the same license) ● Permits commercial use, but unattractive to businesses ● Most popular copyleft license *NOTE: Not fully compatible with CC BY-NC*	**CC BY-NC 4.0** ● Restricts commercial use ● Requires attribution

Compatibility issues between GNU GPLv3 and CC BY-NC 4.0

Unfortunately, I am not able to recommend a directly compatible code license for one license already offered by Deep Blue Data: CC BY-NC 4.0. Because the principles of Open Source software prohibit Discrimination Against Fields of Endeavor – meaning approved Open Source Software licenses cannot restrict commercial use – there is not a clearly compatible code license that meets all of the desired characteristics listed above. It is possible to find code licenses that do restrict commercial use, but these licenses won't have a strong community of use and may discourage reuse by future researchers.

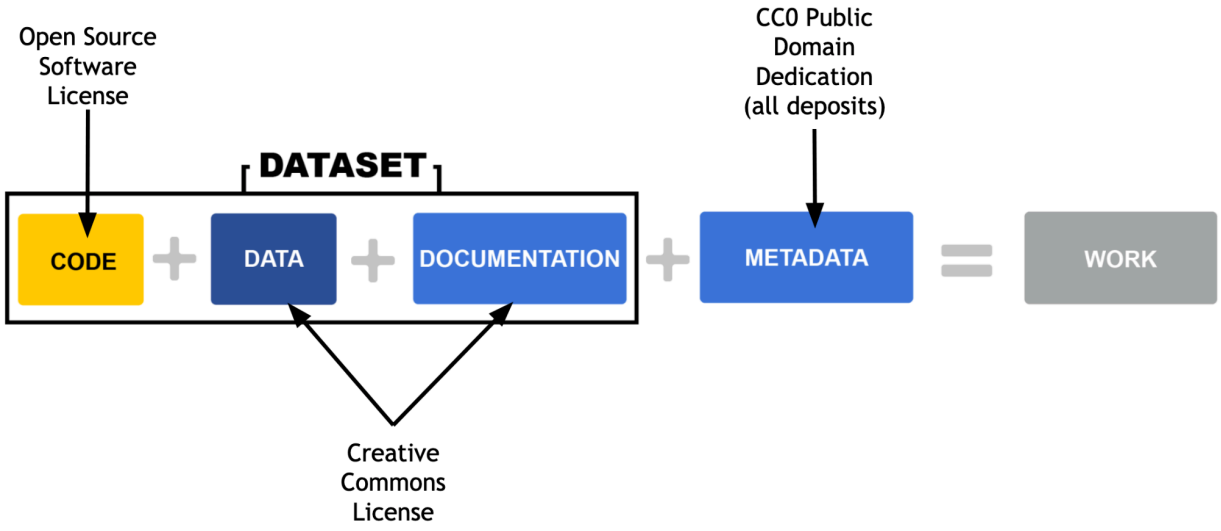
Before allowing research data and code within the same Work to be licensed under both GNU GPLv3 and CC BY-NC 4.0, *the Deep Blue team should consult with U-M's Office of General Counsel to see if this would create significant barriers for reuse and distribution.*

If it is determined that CC BY-NC 4.0 and GPLv3 are not compatible, the following options could be considered:

1. *Offer CC BY-SA 4.0 as the compatible license for use with GNU GPLv3*. The SA license has an approved one-way compatibility with GPLv3, meaning data under the SA license can be incorporated directly into projects licensed under GPLv3. This would allow future researchers to be reassured that they understand the terms of reuse for the Work.
2. Encourage researchers to *deposit data and code as separate works in instances where there are copyleft requirements.* This allows researchers to choose any license for each component, though it may reduce the reusability of their work.

Recommendations for implementation in deposit form

The Deep Blue Data deposit form should incorporate conditional logic to provide different licensing options to researchers depending on whether the Work contains data and/or code. Licenses should be applied as followed:



Currently, researchers can make this distinction manually in the "other" license field – see existing example of this approach in Deep Blue is *Data for Improvement of Plasma Sheet Neural Network Accuracy with Inclusion of Physical Information [Data set]* (Swiger et al., 2020) – but the deposit form should be updated to apply multiple licenses in a machine-readable way.

Preview an example of the conditional logic with Qualtrics or see a the PDF with sample form language included in this the Deep Blue Documents deposit for this report (*data_licensing_sample_form_language.pdf)*.

Curators can use flowcharts in **Appendix IV** to help walk researchers through the process of choosing an appropriate license.

Hyrax Implementation

Hyrax's Metadata Application Profile uses Dublin Core terms for licenses and rights. Since Dublin Core fields can be duplicated, two License fields could be created (one for data and one for code). A canonical URL for each license should be entered into these fields. A Rights note could then be generated to describe how the two licenses are applied to the Work.

This approach would apply two machine-readable licenses to the Work (allowing Hyrax to create RDF predicates to express the licenses). However, the explanation of how the licenses are applied would NOT be machine-readable. It would still require a human to interpret how they are applied to the Work.

Adding License Information to README

In addition to the licenses being incorporated into the repository's metadata, Deep Blue Data should encourage researchers to list the license(s) applied to their Work in the README. This ensures that licensing information stays with the dataset if it is redistributed outside of the repository.

Sample language for Data (only)

> *License:*
> *Data and documentation are licensed under [LICENSE NAME] (license_URL).*

Sample language for Code (only)

> *License:*
> *Code is licensed under [LICENSE NAME] (license_URL). Documentation is licensed under [LICENSE NAME] (license_URL)*

Sample language for Data AND Code

> *License:*
> *Data and documentation are licensed under [LICENSE NAME] (license_URL). Code is licensed under [LICENSE NAME] (license_URL).*

Recommended Next Steps

1. Consult with U-M's Office of General Counsel to determine if code licenses recommended by the report are compatible with existing data licenses offered by the repository.

2. Using information from this report and recommended resources, update internal documentation to reflect curator best practices for guiding researchers through the process of selecting licenses to govern their Works.

3. If additional code licenses are adopted as part of Deep Blue Data, work with the repository's developer to incorporate conditional logic into the deposit form, providing separate license options to Works with data and/or code.

4. Update external documentation (Depositor Guide) to reflect new code licensing options, and provide guidance on selecting compatible licenses for data and code. Consider creating a one-pager (perhaps modeled after the Managing & Organizing Data Informational Bulletin) to widely distribute.

VIII. Recommended Resources

Licensing Assistant (Interoperable Europe)

- Tool developed by the European Union to help "compare, select and combine open-source software licenses, ensuring legal clarity and compliance. It offers content analysis and compatibility check and includes searchable SPDX identifiers."
- Focused on software licenses, but includes some "content" licenses like Creative Commons.

Public License Selector (Paweł Kamocki, Pavel Straňák, Michal Sedlák)

- Tool developed by researchers at Charles University and Université Paris Descarte to provide guidance to selecting a data or software license.
    - This is the only recommended resource that makes a clear distinction between licensing data and code.
- Takes users through a series of questions to determine potential licenses. Includes questions about reused data/code and meeting requirements of upstream licenses.
- Not the most user-friendly interface (and some of the language is more geared towards experts), but this tool best models the complexity of selecting data and code licenses.
    - Probably most helpful for curators.

A Primer for Applying and Interpreting Licenses for Research Data and Code (Data Curation Network)

- Primer written by Lisa Chinn, Matthew Murray 🦇, and Isaac Wink.
- Provides significantly more detail on the topics discussed in this report.

- Section on “Choosing a license for a newly created dataset” gives particularly valuable recommendations for curators working with researchers to select appropriate licenses.

tl;drLegal
- Knowledge base providing “plain English” summaries of software licenses.
- tl;drLegal is managed by FOSSA and is a community project supported project. License descriptions that have been through a formal review process have a verified badge.
- Focused on software licenses, but includes some “content” licenses like Creative Commons.

Choose a License (GitHub)
- Tool developed by GitHub to help developers choose open source licenses.
- Asks a series of questions to guide users towards licenses that meet their situation/needs (includes guidance on non-software licenses).
    - Includes a page on why creators should choose a license.
- Includes descriptions of what is required, permitted, and forbidden by popular licenses.
- Good resource for sharing with researchers (though geared more towards software).

Open Source Initiative
- Organization that maintains the Open Source Definition and approves Open Source Software (OSS) Licenses.
- Includes a database of approved OSS licenses and their characteristics.

Creative Commons
- Beyond providing full text of licenses, the Creative Commons provides summaries of licenses, information about license compatibility, and an extensive FAQ and wiki.

Best Practices for Sharing Research Software (NIH)
- Answers Frequently Asked Questions regarding NIH’s policies on sharing research software.
- NOTE: NIH does not consider software to be “research data” and is governed by different rules.

Funding Note

This project has been funded with Federal funds from the National Library of Medicine (NLM), National Institutes of Health (NIH), under cooperative agreement number UG4LM01234 with the University of Massachusetts Chan Medical School, Lamar Soutter Library. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

IX. Bibliography

Ball, A. (2014). *How to License Research Data | DCC*. Digital Curation Centre. https://dcc.ac.uk/guidance/how-guides/license-research-data

CC0. (n.d.). *Creative Commons*. Retrieved July 24, 2025, from https://creativecommons.org/public-domain/cc0/

CC0 FAQ. (2016, September 28). [Creative Commons]. https://wiki.creativecommons.org/wiki/CC0_FAQ

Chinn, L., Murray 🦇, M., & Wink, I. (2024). *A Primer for Applying and Interpreting Licenses for Research Data and Code*. Data Curation Network GitHub Repository. https://github.com/DataCurationNetwork/data-primers/blob/main/Data%20and%20Code%20Licensing%20Primer/data-code-licensing.md

Fortney, K. (2016, September 15). CC BY and data: Not always a good fit. *Office of Scholarly Communication*. https://osc.universityofcalifornia.edu/2016/09/cc-by-and-data-not-always-a-good-fit/

Frequently Asked Questions: Can I apply a Creative Commons license to software? (n.d.). Creative Commons. Retrieved July 24, 2025, from https://creativecommons.org/faq/#can-i-apply-a-creative-commons-license-to-software

Research Data Stewardship Policy (Standard Practice Guide (SPG) No. 303.06). (2023). University of Michigan. https://spg.umich.edu/policy/303.06

Research Data Stewardship Policy—Procedural Guidelines. (2023). [Standard Practice Guide (SPG)]. https://research.umich.edu/wp-content/uploads/2023/06/procedures-guide-1.pdf

Sen, R., Subramaniam, C., & Nelson, M. L. (2008). Determinants of the Choice of Open Source Software License. *Journal of Management Information Systems*, *25*(3), 207–240. https://doi.org/10.2753/MIS0742-1222250306

ShareAlike compatibility: GPLv3. (2015, September 29). Creative Commons Wiki. https://wiki.creativecommons.org/wiki/ShareAlike_compatibility:_GPLv3#Considerations_

for_adapters_applying_the_GPLv3

Swiger, B. M., Liemohn, M. W., & Ganushkina, N. Y. (2020). *Data for Improvement of Plasma Sheet Neural Network Accuracy with Inclusion of Physical Information* [Dataset]. University of Michigan - Deep Blue Data. https://doi.org/10.7302/559r-t639

The Open Source Definition (Version 1.9). (2007). https://opensource.org/osd

Who Holds Copyright at or in Affiliation with the University of Michigan (Standard Practice Guide (SPG) No. 601.28). (2011). https://spg.umich.edu/sites/default/files/601x28.pdf

Appendix I: Current License Usage in Deep Blue Data

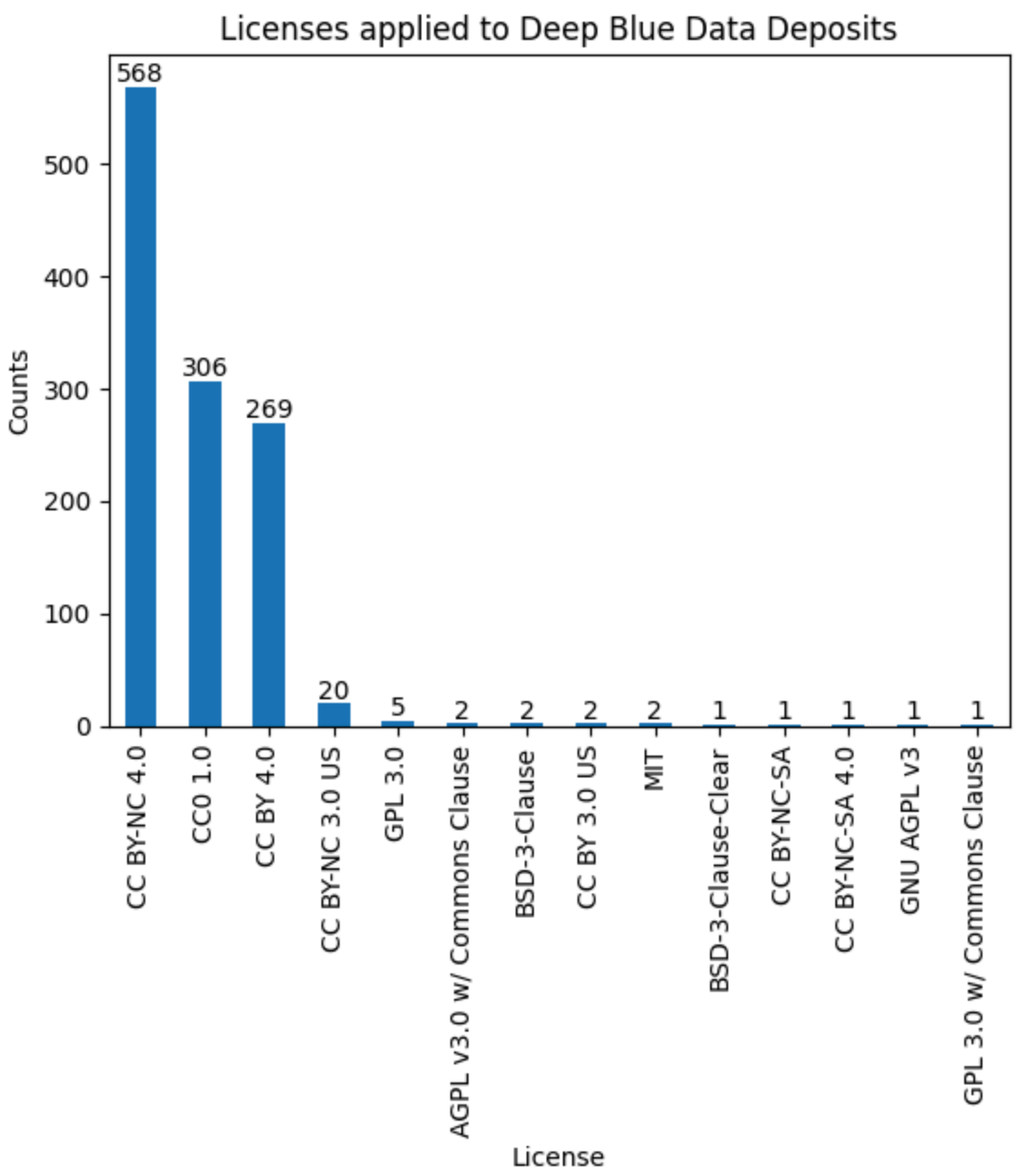
To inform this report, an analysis was conducted of license usage by current Deep Blue Data depositors (excluding the U-M Museums of Zoology and Paleontology, which use bespoke licenses).

A total of 1,178 Works were analyzed, with the vast majority (97%) selecting one of the three default Creative Commons Licenses. CC BY-NC 4.0 was by far the most used license, with 568 applications – slightly less than CC0 1.0 and CC BY 4.0 combined.

Of the works that selected “other” licenses, 14 of those Works used code licenses. While there was not a clearly favored software license by the community, GNU GPLv3 was the most applied license.

Currently, there are three instances where researchers follow practices encouraged in this report by assigning a separate license to data and code:

- Wu, Z., Revzen, S. (2023). Force/Torque BigANT Processed Data [Data set], University of Michigan - Deep Blue Data. https://doi.org/10.7302/vktt-gf88
    - Data: CC BY-NC-SA 4.0
    - Code: GPLv3
- Swiger, B. M., Liemohn, M. W., Ganushkina, N. Y. (2020). Data for Improvement of Plasma Sheet Neural Network Accuracy with Inclusion of Physical Information [Data set], University of Michigan - Deep Blue Data. https://doi.org/10.7302/559r-t639
    - Data: CC BY-NC 4.0
    - Code: MIT
- Wan, Z., Pannala, S., Solbrig, C., Stefanopoulou, A. G., Siegel, J. B., Lee, J. (2025). Dataset for "Degradation and Expansion of Lithium-Ion Batteries with Silicon/Graphite Anodes: Impact of Pretension, Temperature, C-rate and State-of-Charge Window" [Data set], University of Michigan - Deep Blue Data. https://doi.org/10.7302/n5q0-vv61
    - Data: CC BY-NC 4.0
    - Code: MIT


Licenses applied to Deep Blue Data Deposits
Counts
License
0
100
200
300
400
500
568
306
269
20
5
2
2
2
2
1
1
1
1
1
CC BY-NC 4.0
CC0 1.0
CC BY 4.0
CC BY-NC 3.0 US
GPL 3.0
AGPL v3.0 w/ Commons Clause
BSD-3-Clause
CC BY 3.0 US
MIT
BSD-3-Clause-Clear
CC BY-NC-SA
CC BY-NC-SA 4.0
GNU AGPL v3
GPL 3.0 w/ Commons Clause

Appendix II: Common Data Licenses

See TLDRLegal for in-depth descriptions of these licenses in plain English.

* Licenses currently available in Deep Blue Data.

License	Short Description	Benefits	Drawbacks
Creative Commons CC0 1.0 Universal (CC-0) *	Releases data into the **public domain**, waiving copyrights (full description).	• No conditions on adaptation and derivatives, facilitating reuse. • Cannot accidentally claim rights over facts or material already in the public domain. • Can also be used for code	• Attribution not strictly required. • No control over how data is reused.
Creative Commons Attribution 4.0 International (CC BY 4.0) *	**Permissive** license allowing individuals to share, and adapt data as long as proper attribution is given, and any changes are indicated (full description).	• Attribution and enumeration of changes required. • Simple and clear license.	• Potential for attribution stacking where attribution becomes difficult when combined with other reused data, potentially discouraging reuse (Ball, 2014).
Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)	**Copyleft** license allowing Individuals to share and adapt as long as proper attribution is given, any changes are indicated, and adaptations are shared under the same license (full description).	• Ensures that adaptations and reuses continue to have an open license.	• Potential for attribution stacking. • Requiring the same license may restrict who can reuse the data.

License	Short Description	Benefits	Drawbacks
Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) *	**Non-permissive** license allowing individuals to share, adapt, and reuse as long as attribution is given, changes are indicated, and the use is for non-commercial purposes (full description).	• Retains a researcher or institution's right to generate revenue from data.	• Restricts reuse under certain conditions, potentially limiting impact. • Recommendation by the data curation community: only for use in dual licensing regimes (Ball, 2014).
Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0)	**Non-permissive** license where individuals are only free to share data in its original form with attribution (full description).	• Strict control over how data is distributed.	• Severely limits reuse by limiting adaptation and derivatives. (For this reason, this license isn't an option in Deep Blue Data.)
Open Data Commons Public Domain Dedication and License (PDDL)	**Public domain dedication** specifically designed for databases. Differs from CC0 with options to associate suggested community norms for encouraging derivative versions provide attribution and keep data open (full description).	• No conditions on adaptation and derivatives, facilitating reuse of database. • Cannot accidentally claim rights over facts or material already in the public domain. • Ability to associate suggested community norms. • Designed for open data.	• Attribution not strictly required. • No control over how data is reused. • Smaller community of users than Creative Commons licenses.
Open Data Commons Attribution License (ODC-By)	**Permissive** license allowing Individuals to share, and adapt the database as long as proper attribution is given (full description).	• Attribution required (but not an enumerated list of changes). • Simple and clear license. • Designed for open data.	• Smaller community of users than Creative Commons licenses.

Appendix III: Common Software/Code Licenses

See TLDRLegal for in-depth descriptions of these licenses in plain English.

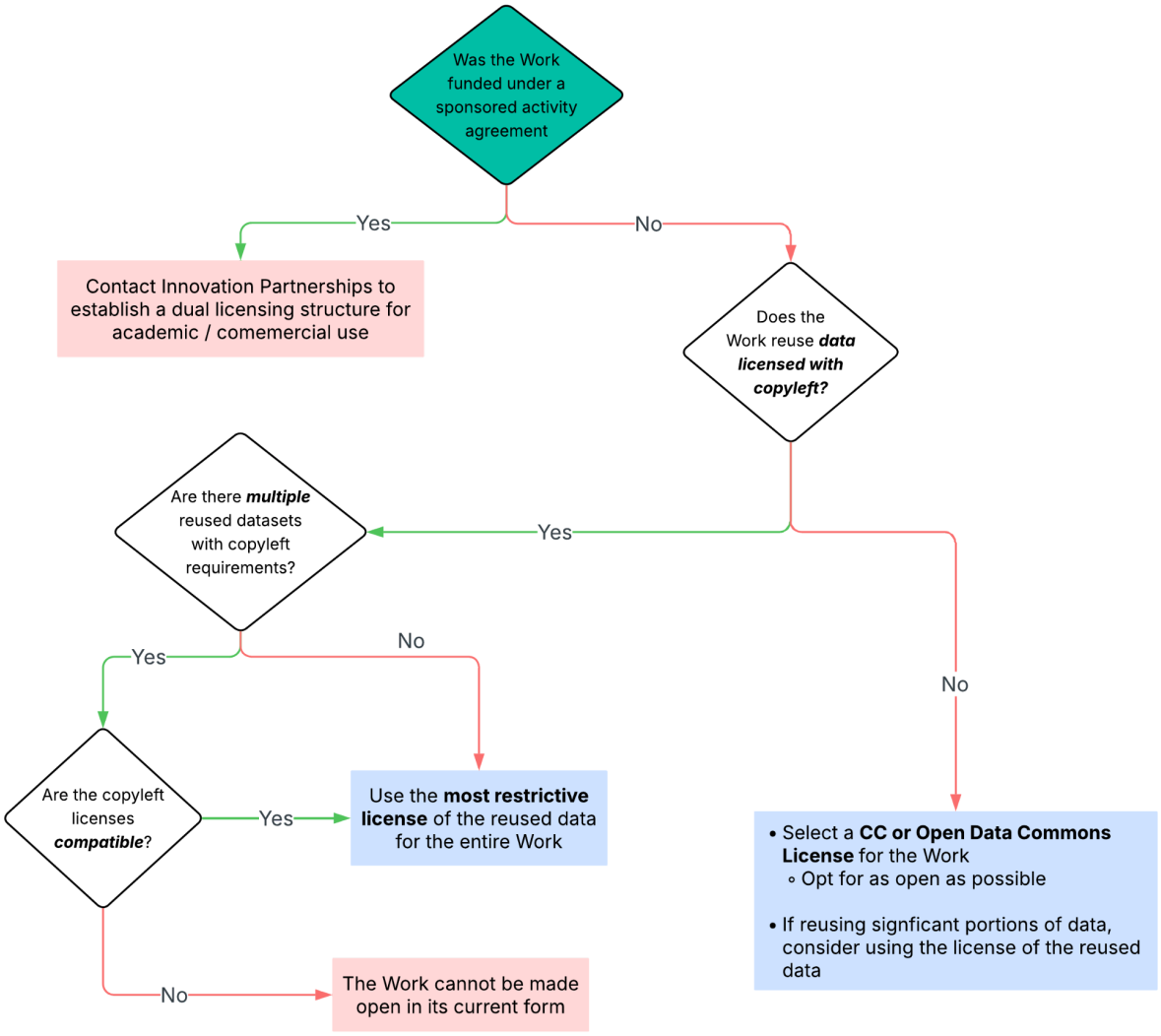
* License currently available in Deep Blue Data.

License	Short Description	Benefits	Drawbacks
Creative Commons CC0 1.0 Universal (CC-0) *	**Public domain dedication** releasing code into the public domain, waiving copyrights (full description).	• No conditions on adaptation and derivatives, facilitating reuse. • Does not waive patent rights. • Can also be used for data.	• Attribution not strictly required. • No control over how code is reused. • Approved by the Creative Commons organization for use with code but not the Open Source Initiative so there's weaker community buy-in in the software space.
MIT (Expat) License	A **permissive** (and short) license that allows any future use of code as long as the original copyright notice/license is included (full description).	• Very few conditions on derivatives (including use in proprietary software). • Strong compatibility with other licenses since it does not have a copyleft provision. • Strong community buy-in (most popular on GitHub).	• Does not require that derivatives are open source. • Attribution not strictly required. • Does not directly address patent rights.

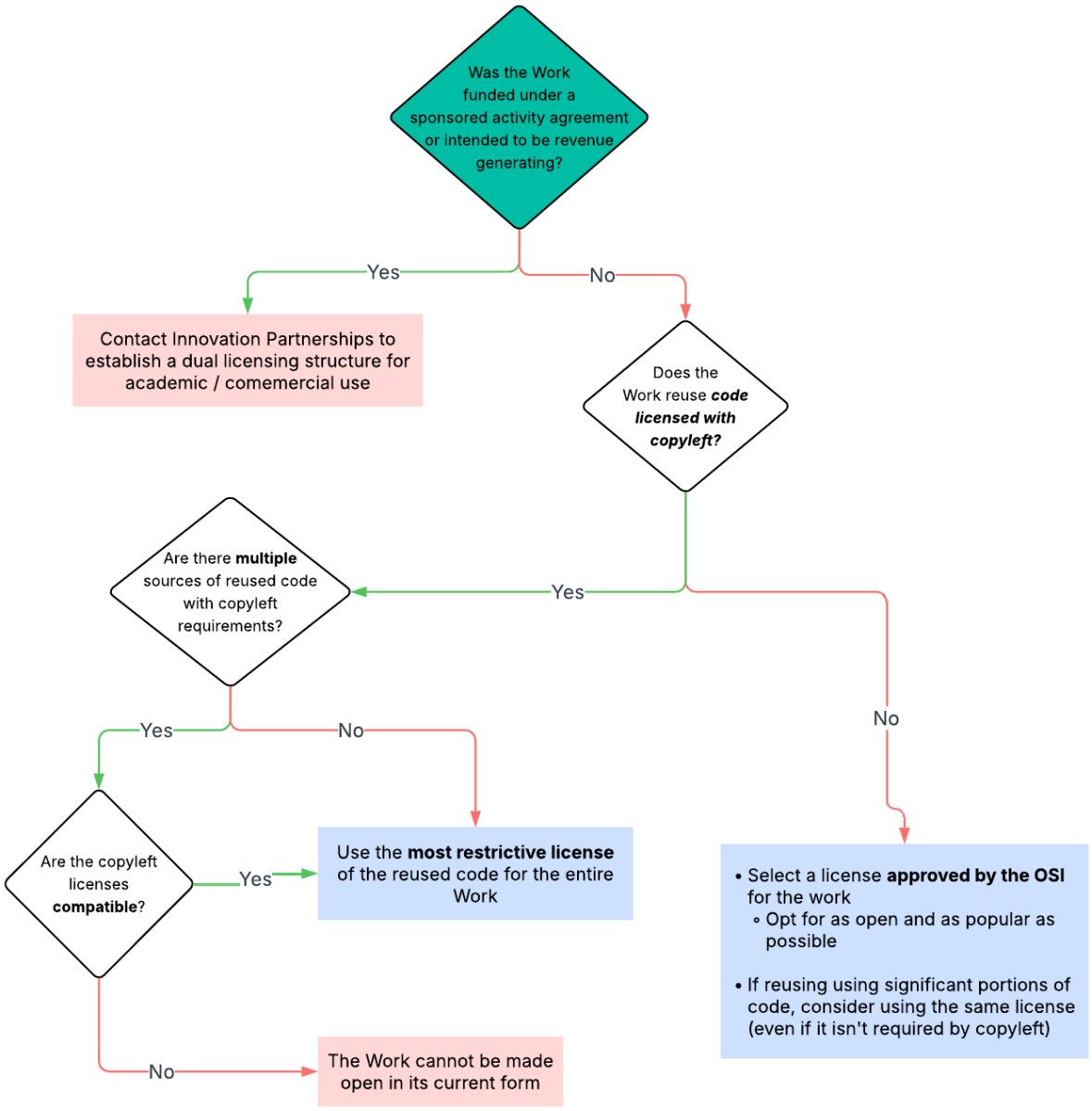
License	Short Description	Benefits	Drawbacks
GNU General Public License v3.0 (GNU GPLv3)	A **copyleft** license requiring derivative code or works to track changes and dates of modification, as well as maintain use of the GPLv3 license and include original copyright notice/license (full description).	• Ensures future derivatives are open source. • Attribution and any changes to derivatives are clear and enumerated. • Directly address patent rights.	• Requiring derivatives to retain the GPLv3 license can cause license compatibility issues and severely reduce potential for reuse.
GNU Affero General Public License v3.0 (GNU AGPL)	Version of GNU public license with **copyleft** requirements designed for networked software (e.g. web apps), adding an additional requirement for source code to be distributed alongside the web application (full description).	• Designed for networked and web applications. • Ensures future derivatives are open source. • Attribution and any changes to derivatives are clear and enumerated. • Directly address patent rights.	• Requiring derivatives to retain the AGPLv3 license can severely prevent reuse (may not be compatible with other code).
Apache License v2.0	A **permissive** license, similar to MIT (Expat) license, that allows any future use of the software as long as the original copyright/notice is included. Directly addresses patent rights (full description).	• Very few conditions on derivatives (including use in proprietary software). • Strong compatibility with other licenses since it does not have a copyleft provision. • Attribution and any changes to derivatives are clear and enumerated. • Directly address patent rights.	• Does not require that derivatives are open source. • Attribution not strictly required.

Appendix IV: Licensing Flowcharts

Research Data Licensing Flowchart



Code Licensing Flowchart



Research Data and Code Licensing Flowchart

