Sharing Bioinformatic Data for Machine Learning: Maximizing Interoperability through License Selection

Alexander Bernier and Adrian Thorogood
Centre of Genomics and Policy, McGill University Faculty of Medicine, Dr. Penfield, Montréal, Canada

Keywords: Big Data, Bioinformatics, Data Commons, Data Licensing, Intellectual Property, Interoperability, License Standardization, Machine Learning, Open Science, Software Licensing, Technology Law.

Abstract: Efficient machine learning in bioinformatics requires a large volume of data from different sources. Bioinformatics is shifting from a paradigm of siloed analysis of individual datasets by researchers to the aggregation and analysis of disparate sets of health and biomedical data across from academic, healthcare and commercial settings. Data generating organizations must give thought to selecting legal terms for dataset release that will promote compatibility with other datasets. In releasing bioinformatic data for open use, care must be taken to ensure that the terms of the licenses selected ensure maximum interoperability. The following technical elements should inform the choice of license: License hybridity; waivers of liability, warranties and guarantees; commercial/non-commercial use; attribution and copyleft; granular permission and bilateral or multilateral licensing. Licenses are compared to inform optimal license selection and enable data integration and analysis; consideration is given to an eventual standard license for open sharing of bioinformatic data.

1 INTRODUCTION

Machine learning needs 'big data' to be most effective. According to the 'Four Vs' model, machine learning requires a great "volume, velocity, variety, and veracity" (Chiang, Grover, Liang & Zhang 2018, p. 384) of data to create usable output, namely "patterns with added value [that] … can be exploited for the creation of wealth, the improvement of human lives, and the advancement of knowledge" (Floridi, 2014, p. 16). Models are hungry for data. With the exception of the data "nouveau riche" – tech giants "such as Facebook [and] Google" – who have access to immense proprietary data troves, many ML researchers and companies must integrate data from varied public and private sources (Floridi, 2014, p. 16). Open-access data sources are often accompanied by terms of use that can be especially restrictive for commercial users, or for the development of commercial algorithms or prediction services. These terms of use are often ambiguous or non-standard, threatening compatibility across databases and effectively restricting integration practices (Carbon et al., 2019).

Bioinformatics stands at the precipice of a sea change. Presently, bioinformatic research is principally performed on isolated datasets. Going forward, single-institution efforts by academic researchers applying traditional bioinformatics methodologies to datasets generated for their particular purposes will be complemented by large-scale big data efforts performed by commercial and institutional entities integrating significant volumes of data from academia, healthcare and industrial research. In actualizing this vision, data generators will need to ensure that the legal rights attached to disparate datasets are standardized or at least compatible. Failing to do so could impose significant costs to understand and comply with the legal limitations on using each dataset and could lead to valuable datasets being impossible to combine lawfully with others, frustrating big data bioinformatics. Consequently, we contend that the bioinformatics community must become adept at understanding license terminology and reading standard licenses so that it can license its data in ways that promote interoperability. We canvass a number of the common elements of data and IP licenses and illustrate how these can affect interoperability. Existing standard licenses are presented in a table that compares how they address the elements discussed (Appendix, Figure 1).

Bernier, A. and Thorogood, A.
Sharing Bioinformatic Data for Machine Learning: Maximizing Interoperability through License Selection.
DOI: 10.5220/0009179502260232
In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 3: BIOINFORMATICS*, pages 226-232
ISBN: 978-989-758-398-8; ISSN: 2184-4305

We consider the potential for standard data licenses to create or to alleviate barriers to bioinformatic data aggregation for machine learning. The following technical elements of licenses are addressed: License hybridity; liability, warranties and guarantees; license duration and conditions of termination; commercial and non-commercial use; attribution and copyleft; permissions granted; and standardization. Finally, we compare how existing standard licenses address the issues discussed to guide data holders in selecting an appropriate data license. The licenses are : Creative Commons Zero (CC0), Creative Commons Attribution 4.0 (CC-BY), the Montreal Data License (MT-DL), the Microsoft Open Use Data Agreement (O-UDA), The Microsoft Computational Use Data Agreement (C-UDA) the Microsoft Data Use Agreement for Open AI Model Development (DUA-OAI), the Linux Community Data License Agreement – Sharing (CDLA-Sharing), the Linux Community Data License Agreement – Permissive (CDLA-Permissive), the Open Data Commons Open Database License (ODC-ODL), the Open Data Commons Attribution License (ODC-BY), and the Open Data Commons Public Domain Dedication and License (ODC-PDDL).

2 LICENSE FEATURES

2.1 License Hybridity

2.1.1 Copyright and *sui generis* Database Rights

Copyright and *sui generis* database rights are intellectual property rights that may or may not apply to datasets or the discrete data points that compose them depending on the jurisdiction. The European Economic Area, Mexico, Russia and South Korea recognize *sui generis* database rights (Doldirina et al., 2018). Copyright applies internationally. Local minimum standards for the application of these rights can vary as to the minimum threshold of human 'creativity,' 'originality,' (Ebrahim, 2019, Stokes 2019) 'investment,' or 'structure' (Gervais, 2019) demonstrated in organizing a dataset. Further, the breadth of exceptions to such rights varies across the world. For instance, the United States permits liberal use even absent necessary permissions, for computational purposes, under the doctrine of transformative 'fair use' (Liu, 2019). Conversely, Europe does not recognize such exceptions or limits them to non-commercial uses (Margoni & Kretschmer, 2018). The uncertain application of copyright and related rights to data complicates licensing selection, as we discuss below. We recommend that bioinformaticians prefer 'hybrid' licenses that act as both intellectual property instruments and contracts, because this will increase their likelihood of being applied by a court even if the existence of underlying rights in the data, and the existence of a valid contract between licensor and recipient are indeterminate.

2.1.2 The Enforceability and Benefits of the Hybrid License

Licensing discussions are often bogged down by uncertainty over underlying IP rights. To avoid this, we recommend selecting hybrid licenses that contain the same content in a contract and in an intellectual property license. The document will bind data users insofar as a valid contract is recognized and third parties insofar as intellectual property interests are recognized in the licensed data. Courts in the United States (Madison, 1998) and Europe (*Ryanair Ltd v PR Aviation BV*, 2015) have recognized such 'hybrid' instruments as valid. A hybrid license can contractually increase the rights of a licensor in the face of overly permissive IP laws, clearly signal to a data user that the IP holder has disclaimed any IP rights in the dataset, and harmonize all the legal regimes applicable to a dataset where it is unclear which IP rights, if any, would apply by default.

We recommend selecting licenses that will address copyright, moral rights, *sui generis* database rights, and clearly state their dual contractual-IP license nature. Failure to do so could lead to the license's effect *not* mirroring the intent of the licensor. The contractual elements of the hybrid license can limit the potential for inconsistent interpretations of IP law to leave ambiguous the rights and obligations of direct parties to the agreement. The IP elements of the hybrid license create some certainty as to the rights and obligations of all third parties, even if ambiguities arise as to the contract's validity. By favoring such licenses, bioinformaticians can create clarity as to the rights granted in their data. This promotes the more widespread use of their data, permits data users to more easily understand if the rights in that data is compatible with the rights in other data, and decreases the risk of license misinterpretation leading to conflict or litigation.

2.2 Commercial/Non-commercial Use

Reserving open data for non-commercial uses may preclude machine learning altogether, as many

applications necessitate resources and expertise only available to sophisticated private-sector entities (Doherty et al., 2016). Further, partnerships across academic institutions and the private sector are the rule rather than the exception in this area, with both parties pooling resources including data, capital, computing power and expertise. Commingling private sector and public sector data and resources benefits public-sector researchers in giving them access to rigorously assembled pools of industry data and permitting them to pursue research goals of academic interest that do not lend themselves to obvious profitability (Perkman & Schildt, 2015). Moreover, the boundaries of 'commercial' and 'non-commercial' use, and 'commercial' and 'non-commercial' actors can be ambiguous. Machine learning also presages the merger of data and open-source software licensing – the latter community considers commercial use restrictions discriminatory. Consequently, we strongly recommend that bioinformaticians avoid licenses that preclude commercial use; such licenses will likely prevent their data from being used for big data applications.

2.3 Waivers of Liability, Warranties and Guarantees

2.3.1 General Liability, Warranties and Guarantees

Assuming data generators share data openly on a voluntary basis, licenses must be friendly to them in order to promote data availability. Likewise, if licenses are unfriendly towards users (in terms of being too restrictive or conditional), this will discourage data use.

In selecting a license, licensors should consider the degree of responsibility that best reflects their ability to affirm their rights in, and the quality of, the data, as well as their risk tolerance as regards liability. Waivers of liability, and disclaimers regarding the licensor's rights in the data, and the quality, accuracy and merchantability thereof are common features of licenses. The licensor is better placed than the user to assess the aforementioned features, but open licensing generally provides the licensor little benefit (Wilka, Landry & McKinney, 2018). Consequently, we consider that a license that contains the traditional disclaimers, but also affirms that the licensor "has exercised reasonable care to assure" the disclaimed feature is generally a good compromise position. Nonetheless, bioinformaticians should carefully consider what guarantees to grant data users. If more guarantees are made, data users may feel more comfortable using the data. If less guarantees are made, data licensors may be more protected from the legal risks inherent in making their data available.

2.3.2 Data Protection and Privacy Laws

The General Data Protection Regulation (GDPR) and other data protection legislation modeled on it assesses the right to use data on a subjective 'controller-by-controller' basis; the question is not 'can this data be lawfully used' but 'can *you* lawfully use this data.' Further, data protection is a markedly localized (Custers, Dechesne, Sears, Tani & Hof, 2018) and sector-specific (Archer & Delgadillo, 2016) legal regime. Therefore, it is generally difficult for the data licensor to make any meaningful representation to the recipient regarding data protection.

Bioinformatic and associated health data is often subject to data protection laws, and sometimes to more onerous laws or provisions that hold health information to a higher standard of protection (Kim, Kim & Joly, 2018; Thorogood 2018). Data licensors should consider such legislation before licensing their data, especially if intending to use an open license or public dedication. Presently, most standard data licenses do not address data protection. This could conflict with obligations under some data protection laws (e.g. the GDPR) to distribute responsibilities among data controllers using controllership agreements (Wrigley, 2019). Other data protection laws impose accountability requirements that may favor using licenses and contracts that address data protection (Centre for Information Policy Leadership, 2018). Consequently, we recommend that bioinformaticians use licenses that address data protection, or create data protection annexes is licensing data. In doing so, they should remain mindful of the highly mutable character of data protection obligations across countries and for different entities. We caution them not to rely exclusively on generalized statements about data protection responsibilities in contracts or licenses.

2.4 License Duration and Conditions of Termination

A licensing challenge for machine learning is that it tends to depend on long-term, potentially indefinite, access to a same pool of data in a number of contexts. (Wilka et al., 2018).

Considering that data users will be integrating a large number of datasets to create a single machine learning algorithm, the loss of even a single dataset

could mar their ability to replicate a certain algorithm, or to determine if tuning or retraining on a modified dataset is improving the functioning of an algorithm. (Lehr & Ohm, 2017). For these reasons, it is our recommendation that licensors select licenses that ensure data user breach does not immediately lead to the termination of the license, and that data providers not have the right to unilaterally terminate the license. Further, license terms should be indefinite, renewed automatically on expiration, or renewed at the discretion of the data user rather than the licensor.

2.5 Attribution and Copyleft

2.5.1 Forms of Open Licensing

Open licenses come in many variants. 'Permissive' licenses openly release the work without imposing any limitations on its use. Consequently, recipients are free to create derivative works and commercialize those derivatives, or impose IP protections thereon. 'Copyleft' licenses restrict the recipients from imposing IP protections on the licensed work, and can impose the obligation to license derivative material downstream on equally permissive terms. 'Strong copyleft' requires distribution of the licensed work or a derivate work under the same terms. 'Weak copyleft' requires distribution of the licensed work or a derivative work under the same terms, but permits the combination of the licensed work with other proprietary works (e.g. combining software) under different terms (Hall, 2017). Attribution requirements impose an obligation to attach an attribution to the licensed work, sometimes in a prescribed form.

A 'public domain dedication' is another popular mechanism for attempting to eliminate all of the licensor's rights in the concerned works. Not all countries recognize the lawfulness thereof. Traditionally, the United States has been permissive in allowing public dedications (Johnson, 2008) and European jurisdictions more reticent (Aishwarya, 2017). Bioinformatics communities should decide what modality of open license best conforms to their values, as ensuring compatibility even between standard licenses is problematic (OpenMinted, n.d.).

2.5.2 Considerations for Machine Learning

In licensing data for data integration, imposing an attribution requirement on a dataset can frustrate the combination of datasets from different sources. The barrier can be resource-based, in that it is insurmountably time-consuming to attribute a large number of datasets used to train a ML algorithm. The barrier can also be rule-based, in that different licenses impose attribution formats that are incompatible such that dual compliance is impossible (Morando, 2013). The most potent algorithmic models ensue from the combination of disparate datasets (Mattioli, 2018).

Copyleft requirements can hamper the interoperability of datasets for ML to varying degrees depending on their formulation. Strong copyleft requirements can pose the same problem as attribution requirements; it could prove impossible to comply with the conflicting copyleft requirements of two datasets. This would preclude the combination thereof.

Copyleft further has the potential to create siloes of 'copyleft' and 'non-copyleft' data. Data recipients hoping to create proprietary technologies from open data are barred from using the 'copyleft' data lest their output become 'infected' and unfit for their private commercial purposes (Thorogood, 2019). From the narrow standpoint of license interoperability, we recommend avoiding strong copyleft or attribution clauses.

2.6 Granular Permissions and Bilateral or Multilateral Licensing

Open licensing generally aspires to the broadest possible permissions, both as regards the parties concerned and the rights in the data. Nonetheless, concerns of data sensitivity, or a desire to foster innovation while safeguarding the right to profit from the data in the future can prohibit totally open licensing (Benjamin et al., 2019). Of the licenses addressed, the Montreal Data License is unique in that it allows for the negotiation of bilateral contracts between parties for particular tiers of rights in data, using language specific to machine learning and algorithmic modelling. Data licensors with an appetite for data release, but who are wary of the privacy violations or commercial opportunities lost in public licensing may want to consider this license.

3 CONCLUSIONS: TOWARD A STANDARD LICENSE

A final consideration in licensing data is that of standardization. The emergence of competing standards does not necessarily reflect the failure of the community to reach consensus, but rather that the culture of openness varies across the academic, bioresources, open patenting, and software

development communities (Liddell, Liddicoat, Jordan & Schovsbo, 2019). License selection does not necessarily mean selecting the most optimal among competing options; it reflects the subjective balancing of differing values. Presently, licensors of bioinformatic data must decide which of the existing options best reflect their objectives. In the future, a standard license for bioinformatic data sharing could benefit the scientific community by ensuring that bioinformaticians have tools for data sharing that enshrine the values of their community. Further, achieving true interoperability may require license standardization, as combining datasets across licenses could create legal ambiguities and inefficient costs. (Morando, 2013). Achieving standardization will require not only appropriate license selection by individuals but successful consensus-building across bioinformatics communities. License literacy will be instrumental in drafting and selecting the licenses needed to make big data bioinformatics a reality and pool data across academia, healthcare and industry.

ACKNOWLEDGEMENTS

The authors graciously thank Genome Canada, Genome Québec and the Canadian Institutes for Health Research for their financial support.

REFERENCES

Aishwarya, S. (2017). The Nature and Enforceability of Open Source License. *11 NUALS Law Journal*, *11*, 53–86.

Archer, J. K., & Delgadillo, C. A. (2016). Key Data Ownership, Privacy and Protection Issues and Strategies for the International Precision Agriculture Industry. *Proceedings of the 13th International Conference on Precision Agriculture*.

Benjamin, M., Gagnon, P., Rostamzadeh, N., Paul, C., Bengio, Y., & Shee, A. (n.d.). Towards Standardization of Data Licenses: The Montreal Data License. eprint arXiv:1903.12262

Carbon, S., Champieux, R., Mcmurry, J. A., Winfree, L., Wyatt, L. R., & Haendel, M. A. (2019). An analysis and metric of reusable data licensing practices for biomedical resources. *Plos One*, *14*(3). doi: 10.1371/journal.pone.0213090

Centre for Information Policy Leadership. (2018). *The Case for Accountability: How it Enables Effective Data Protection and Trust in the Digital Society* (p. 15).

Chiang, R. H. L., Grover, V., Liang, T.-P., & Zhang, D. (2018). Strategic Value of Big Data and Business Analytics. *Journal of Management Information Systems*, *35*(2), 383–387. doi: 10.1080/07421222.2018.1451950

Custers, B., Dechesne, F., Sears, A. M., Tani, T., & Hof, S. V. D. (2018). A Comparison of Data Protection Legislation and Policies Across the EU. *Computer Law and Security Review*, *34*(2), 234–243. doi: 10.2139/ssrn.3091040

Doherty, M., Metcalfe, T., Guardino, E., Peters, E., & Ramage, L. (2016). Precision medicine and oncology: an overview of the opportunities presented by next-generation sequencing and big data and the challenges posed to conventional drug development and regulatory approval pathways. *Annals of Oncology*, *27*(8), 1644–1646. doi: 10.1093/annonc/mdw16

Doldirina, C., Eisenstadt, A., Onsrud, H., & Uhlir, P. (2018). Legal Approaches for Open Access to Research Data. doi: 10.31228/osf.io/n7gfa

Ebrahim, T. Y. (2019). Data-Centric Technologies: Patent and Copyright Doctrinal Disruptions. *Nova Law Review*, *43*(3).

Floridi, L. (2014). Chapter I: Time: Hyperhistory. In *The fourth revolution: how the infosphere is reshaping human reality* (pp. 1–24). Oxford: Oxford University Press

Gervais, D. J. (2019). Exploring the Interfaces Between Big Data and Intellectual Property Law. *Intellectual Property, Information Technology & Electronic Commerce Law*, *22*. doi: 10.2139/ssrn.3360344

Hall, A. J. (2017). Open-Source Licensing and Business Models: Making Money by Giving It Away. *Santa Clara High Technology Law Journal*, *33*(3), 427–437.

Johnson, P. (2008). Dedicating Copyright to the Public Domain. *Modern Law Review*, *71*(4), 587–610. doi: 10.1111/j.1468-2230.2008.00707.x

Kim, H., Kim, S. Y., & Joly, Y. (2018). South Korea: in the midst of a privacy reform centered on data sharing. *Human Genetics*, *137*(8), 627–635. doi: 10.1007/s00439-018-1920-1

Lehr, D., & Ohm, P. (2017). Playing with the Data: What Legal Scholars Should Learn About Machine Learning. *University of California, Davis Law Review*, *51*, 653–717.

Liddell, K., Liddicoat, J., Jordan, M., & Schovsbo, J. (2019). IP policies for Large Bioresources: the Fiction, Fantasy and Future of Openness. In T. Minssen & R. J. Herrmann (Eds.), *Global Genes, Local Concerns: Legal, Ethical and Scientific Challenges in International Biobanking.* (pp. 258–280). Edward Elgar Publishing.

Liu, J. (2019). An Empirical Study of Transformative Use in Copyright Law. *Stanford Technology Law Review*, *22* (Winter), 163–241.

Madison, M. J. (1998). Legal-Ware: Contract and Copyright in the Digital Age. *Fordham Law Review*, *67*(3), 1025–1143. doi: 10.31228/osf.io/4y2h6

Margoni, T., & Kretschmer, M. (2018). The Text and Data Mining Exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs. *UK Copyright and Creative Economy Centre University of Glasgow Technical Report.*

Mattioli, M. (2018). The Data-Pooling Problem. *Berkeley Technology Law Journal*, *32*(1), 179–235. doi: 10.2139/ssrn.2671939

Morando, F. (2013). Legal Interoperability: Making Open (Government) Data Compatible with Businesses and Communities. *Italian Journal of Library, Archives and Information Science*, *4*(1), 441–452.

OpenMinted. (n.d.). Accessible at: https://openminted.github.io/releases/interop-spec/1.0.0/openminted-interoperability-scenarios/.

Perkmann, M., & Schildt, H. (2015). Open data partnerships between firms and universities: The role of boundary organizations. *Research Policy*, *44*(5), 1133–1143. doi: 10.1016/j.respol.2014.12.006

Stokes, S. (2019). Chapter 2: Digital Copyright, the Basics. In *Digital Copyright: Law and Practice Fifth Edition*. Hart Publishing, Bloomsbury Publishing Plc.

Ryanair Ltd v PR Aviation BV, ECLI:EU:C:2015:10 (E.C.J. 2nd Chamber. 2015).

Thorogood, A. (2018). Canada: will privacy rules continue to favour open science? *Human Genetics*, *137*(8), 595–602. doi: 10.1007/s00439-018-1905-0

Thorogood, A. (2019). Towards Legal Interoperability in International Health Research. *University of Toronto LLM Thesis*.

Wilka, R., Landry, R., & McKinney, S. A. (2018). How Machines Learn: Where Do Companies Get Data for Machine Learning and What Licenses Do They Need. *Washington Journal of Law, Technology and Arts*, *13*(3), 217–244.

Wrigley, S, (2019.). "When People Just Click": Addressing the Difficulties of Controller/Processor Agreements Online. In M. Corrales, M. Fenwick & H. Haapio (Eds.), (pp. 221–252). Kyushu University, Springer.

APPENDIX

License	License Hybridity / Rights Concerned	Liability, Warranties and Guarantees	Duration	Commercial and Non-Commercial Use	Attribution and Copyleft	Parties
Creative Commons Zero (CC0)	Waiver / license of copyright, *sui generis* rights, moral rights	No representations or warranties, guarantees, disclaimer of liability. Disclaimer of responsibility for clearing rights in data.	Indefinite	No limitation	No requirement	One to all.
Creative Commons Attribution 4.0 International (CC-BY)	Hybrid contract (implied) / license of copyright, *sui generis*, waiver of moral rights	No representations or warranties, disclaimer of liability. Disclaimer of responsibility for clearing rights in data.	Indefinite, immediate termination for non-compliance, rectification within 30 days reinstates license, otherwise consent is required.	No limitation.	Specific attribution requirements (prescribed format).	One to all.
Montreal Data License (MT-DL)	Not specified, presumably a pure contract.	Exclusion of warranties, guarantees, disclaimer of liability.	Unspecified duration. Immediate termination on breach.	Licensor's option.	Licensor's option.	One to one.
Microsoft Open Use Data Agreement (O-UDA)	Not specified.	Disclaimer of warranties, limitation of liability for licensor and upstream licensors. No warranty of rights in data. Licensor agrees not to sue recipient and downstream recipient absent breach.	Unspecified duration. No provisions regarding voluntary termination or termination for cause.	No limitation.	Attribution for source and modified data. Must impose warranty disclaimer and limitation of liability for upstream controllers on the downstream recipients. No attribution, warranty, or limitation of liability, requirement for output, so long as the output does not contain more than a 'de minimis' portion of the data.	One to all.
Microsoft Computational Use Data Agreement (C-UDA)	Not specified. Rights limited to computational use.	Disclaimer of warranties, limitation of liability for licensor and upstream licensors. No warranty of rights in data. Licensor agrees not to sue recipient and downstream recipient	Unspecified duration. No provisions regarding voluntary termination or termination for cause.	No limitation.	Attribution for source and modified data. No attribution requirement for output. Copyleft for data (same license must be applied). No copyleft for output or algorithms unless these contain more than a 'de minimis' portion of the data.	One to all.
Microsoft Data Use Agreement for Open AI Model Development (DUA-OAI)	Not specified. Rights in data limited to training the AI model. No right to share or distribute the data or assign license.	Parties each warrant and represent compliance with laws, including data protection laws. Data user warrants rights in the untrained AI model. Data licensor does not warrant rights in or quality of data. Licensor warrants that they are not aware of restrictions that would limit use or distribution. Optional limitation of liability clause, with exception for damages caused by the data recipient's breach of the license.	Duration of one year. Termination with notice after 90 days, termination for breach 30 days after notification of breach, if not cured.	No limitation.	Copyleft in the trained AI model – must publicly release the trained AI model under an open software license that includes a general disclaimer of liability in favor of the data licensor.	One to one.

Figure 1: Standard License Comparison Table.

Linux Community Data License Agreement – Sharing (CDLA-Sharing)	Hybrid contract (implied) / license of copyright, *sui generis*, waiver of moral rights.	Parties each warrant and represent reasonable care in ensuring use in compliance with rights of others, privacy and confidentiality. Disclaimer of warranties and limitation of liability.	Termination for data recipient's breach if not rectified within a 'reasonable time' of becoming aware. Termination if litigation against data provider or data recipient concerning dispute not related to the data license.	No limitation.	Attribution of source and modified data; flagging of modified data, integration of those notices into the data files. No attribution requirement for output / results unless these contain more than a 'de minimis' portion of the data. Copyleft for data (same license must be applied); no copyleft for output / 'results' unless these contain more than a 'de minimis' portion of the data. Explicit preclusion of restriction using technological measures.	One to all.
Linux Community Data License Agreement – Permissive (CDLA-Permissive)	Hybrid contract (implied) / license of copyright, *sui generis*, waiver of moral rights.	Parties each warrant and represent reasonable care in ensuring use in compliance with rights of others, privacy and confidentiality. Disclaimer of warranties and limitation of liability.	Termination for data recipient's breach if not rectified within a 'reasonable time' of becoming aware. Termination if litigation against data provider or data recipient concerning dispute not related to the data license.	No limitation.	Attribution of source and modified data; flagging of modified data, integration of those notices into the data files. No attribution requirement for output / results unless these contain more than a 'de minimis' portion of the data. Modified data or a combination of original and modified data can be released under a different license. No copyleft for output / 'results' unless these contain more than a 'de minimis' portion of the data.	One to all.
Open Data Commons Open Database License (ODC-ODL)	Explicit hybrid contract / license copyright and *sui generis*, waiver of moral rights.	Disclaimer of warranties and exclusion of liability.	Immediate termination for breach; can be reinstated if first breach and rectifies within 30 days of notice of breach. Otherwise reinstated 60 days after cessation of breach, if licensor does not send notice of permanent termination in that time.	No limitation.	Attribution requirement for the data/base, a derivative data/base, or output. Copyleft, must license the data/base or a derivative database under the same license or "a compatible license." No additional legal "terms or technological measures" can be imposed, excepting a limited right to 'parallel release.' No copyleft for output.	One to all.
Open Data Commons Attribution License (ODC-BY)	Explicit hybrid contract / license of copyright and *sui generis*, waiver of moral rights.	Disclaimer of warranties and exclusion of liability.	Immediate termination for breach; can be reinstated if first breach and rectifies within 30 days of notice of breach. Otherwise reinstated 60 days after cessation of breach, if licensor does not send notice of permanent termination in that time.	No limitation.	Attribution requirement for data/base, modified database, or output. Copyleft for data/base or derivative database (same license must be applied).	One to all.
Open Data Commons Public Domain Dedication and License (ODC-PDDL)	Public domain dedication of copyright, database rights / license of copyright, *sui generis*, and waiver of moral rights.	Disclaimer of warranties and exclusion of liability.	Indefinite.	No limitation.	No attribution or copyleft requirement.	One to all.

Figure 1: Standard License Comparison Table (continued).