Community Data License Agreement

cdla.io

Open Source Forum Japan

November 15, 2017

THE LINUX FOUNDATION

COMMUNITY DATA LICENSE AGREEMENT

cdla.io

THE LINUX FOUNDATION


As systems require data to learn and evolve, no one organization can build, maintain and source all data required.


Data communities are forming
• AI and ML use cases
• Autonomous systems
• Connected civil infrastructure

The CDLA license agreements enable sharing data openly, embodying best practices learned over decades sharing source code.

Community Data License Agreement

- On October 23, we announced Version 1.0 of the Community Data License Agreements
- There are two CDLA license agreements:
  - "Sharing" – based on a form of copyleft, designed to encourage recipients to participate in reciprocal sharing of data
  - "Permissive" – an approach similar to permissive open source licenses (e.g. Apache, BSD or MIT) where recipients are not required to share any changes

What are the differences between the two agreements?

- The primary difference relates to your obligations if you decide to publish data that you receive under the Agreement.
- The **Sharing** version of the Agreement requires You to Publish that Data, and any Enhanced Data, under the terms of the Sharing version of the Agreement – similar to a copyleft open source license.
- The **Permissive** version of the Agreement, by contrast, allows Data and Enhanced Data to be Published under different terms, subject to notice and attribution requirements – similar to a permissive open source license.

New advancements are driving interest in "open data"

- Interest in sharing data has grown significantly due to machine learning, AI, blockchain and expansion of open geolocation solutions
- Governments, companies and organizations are interested in sharing data "just like we share source code"
  - They're looking to open source principles and how they may be applied today
  - Open source development is viewed as an ideal model for collaborating on datasets
- Connected civil infrastructure and private systems are starting to intersect (e.g. infrastructure-to-vehicle systems) with data created and shared

Current practices around sharing data vary but generally map to requirements we've dealt with in source code licensing

- Open data publishers are currently using multiple approaches to open licensing data
  - Public Domain, see: https://opendatacommons.org/guide
    - Data.gov "*Additionally, we* ***waive copyright and related rights*** *in the work worldwide through the CC0 1.0 Universal public domain dedication.*"
  - Open Source Licenses, CC BY-SA 2.0
  - Open "Data Licenses", see http://wiki.openstreetmap.org/wiki/Open_Database_License
  - Canadian Government publishes data under the "Open Government Licence", see http://open.canada.ca/en/open-government-licence-canada
- Some communities only ask for attribution…
  - *"The CHIANTI package is freely available. If you use the package, we only ask you to appropriately acknowledge CHIANTI."* (http://www.chiantidatabase.org)

License	Domain	By	SA	Comments
Creative Commons CCZero (CC0)	Content, Data	N	N	Dedicate to the Public Domain (all rights waived)
Open Data Commons Public Domain Dedication and Licence (PDDL)	Data	N	N	Dedicate to the Public Domain (all rights waived)
Creative Commons Attribution 4.0 (CC-BY-4.0)	Content, Data	Y	N	
Open Data Commons Attribution License (ODC-BY)	Data	Y	N	Attribution for data(bases)
Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)	Content, Data	Y	Y	
Open Data Commons Open Database License (ODbL)	Data	Y	Y	Attribution-ShareAlike for data(bases)

http://opendefinition.org/licenses/

Do we need another agreement for data?

- There are current agreements but none has gained traction for a variety of reasons.
  - There is a clear consensus that open source licenses useful for software are not appropriate for data.
  - The Creative Commons licenses have been used – in particular CC0 – but many think that a license specific to data would be preferable.
  - Use restrictions abound and licenses too often enable adding them
  - Many companies hesitate to publish or use data without clear license terms for their use
- The CDLA hopes to avoid a period of license proliferation and aggregations of valuable data under licenses that prevent combinations necessary to optimally exploit the data over time.
  - Tenure of data value is much longer than software – and potentially perpetual.
  - You may never be able to recreate the environment to produce data (e.g. ocean water temperatures over time), and need long term flexibility

Though data is not the same as source code

- In the US and elsewhere, data itself is generally not protectable IP (see Feist Publications, Inc., v. Rural Telephone Service Co.)[1]
- Only the creative expression of the data is protectable by copyright; Facts are not
- Some data provider organizations are trying any means available to lock down access to data, sometimes with direct or ambiguous terms around usage rights
  - "*Intellectual Property Rights means the rights in and to patents, trademarks, service marks, trade and service names, copyrights, database rights and design rights, rights in know-how, moral rights, trade secrets and all rights or forms of protection of a similar nature or having similar or equivalent effect which may subsist anywhere in the world now existing or hereafter arising*."

[1] Available at: http://caselaw.findlaw.com/us-supreme-court/499/340.html

Does "licensing" data create IPR where none exists?

- This raises questions underlying any consideration of a common license rather than an attempt at a public domain designation.
  - If there is no intellectual property right in the data, can it be a license so that the "remedy" for failure to comply is infringement?
  - Or is it a contract that can only be remedied as a breach?
  - Does it matter?
- Does use of data under something that purports to be a license create an estoppel argument that gives a community participant less rights than an outsider?
  - Does consenting to the terms disadvantage you relative to a data user who does not consent?
- If intellectual property rights in data vary across jurisdictions, does clarity of permissions enable global data aggregation and use?
- Does having clear terms enable broader participation and usage?

Why grant permissions if there is an argument of no IPR existing?

- If it is true in some jurisdictions today, it may not be true tomorrow and it may not be true globally.
- For contributors of data:
  - The disclaimers and limitations of liability are also important to enable sharing
  - Contributors of data want to know that they can use data contributed by others who may be governed by different IP laws
- For consumers of data:
  - They want certainty and do not want to have to seek the pedigree and jurisdiction of all of the data they use
  - They want to have some basis for a good faith belief that the data is suitable for use

If there is a sharing obligation (e.g. copyleft), where does it begin and end?

- Includes:
  - Modifications to data received
  - Additions to data received
- Excludes:
  - The results of any analysis
    - Results may be included voluntarily
    - Contributions will be limited if results have to be shared
    - Similar to internal use exclusion in GPL

But what about Personally Identifiable Information?

- Each Data Provider represents that Publication of the Data that it Publishes does not violate any privacy or confidentiality obligation undertaken by that Data Provider.
- If You choose to Publish Data that You have Received under the Agreement, You are not asked to make a representation that no other Data Provider has included Data that is subject to a privacy or confidentiality obligation that was undertaken by that Data Provider.
- Does that mean that You can pass along Data when You know that someone else has inserted personal or confidential information into that Data?
  - No. Each Data Provider represents that the Data Provider has exercised reasonable care to assure that the Data it Publishes was obtained from others with the right to Publish the Data under this Agreement.
  - Furthermore, although the Agreement may contain no requirement to make representations on behalf of other Data Providers, You are still required to comply with all applicable laws in Publishing and Using Data Received under the Agreement.

Who will use the agreements?

- Communities training AI and ML systems
- Public-private infrastructure (e.g. data on traffic)
- Researchers
- Companies with mutual interests in sharing data
- You?

"Data is replacing concrete as the foundation of 21st century transportation, and knitting this increasingly complex array of public and private data sources together requires new approaches to data licensing and data governance,

The CDLA provides a critical new tool to facilitate collaboration and data sharing between government and private sector innovators."

-- Kevin Webb, Executive Director, Open Transport Partnership.

The CDLA in use – Cisco's Network Anomaly Telemetry data

- https://github.com/cisco-ie/telemetry
- The data sets are based around network anomalies, e.g. port flaps, bgp issues, optic failures, etc.
- The purpose is to allow the development of models to identify the unique signatures of the events as close to the actual time of the event versus identifying it minutes after.
- Cisco is working with a number of universities who are adding the data sets in to their data science and research coursework at both undergrad and graduate levels.
- This level and type of network anomaly data sets have not been available for the data science and machine learning communities let alone the majority of companies to use in developing automation and remediation of network events.
- Like that CDLA is a data-specific agreement as opposed to being just a copyright license, which doesn't really fit data well.

cdla.io

Contact Us

The Linux Foundation

1 Letterman Drive
Building D, Suite D4700
San Francisco CA 94129
Phone/Fax: +1 415 7239709
www.linuxfoundation.org

General Inquiries

info@linuxfoundation.org

Membership

membership@linuxfoundation.org

Corporate Training

training@linuxfoundation.org

Event Sponsorship

sponsorships@linuxfoundation.org

Legal Notices

The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation's Trademark Usage Policy at https://www.linuxfoundation.org/trademark-usage, as may be modified from time to time.

Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute's trademark usage page at https://lmi.linuxfoundation.org for details regarding use of this trademark.

Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.

TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.

Facebook and the "f" logo are trademarks of Facebook or its affiliates.

LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.

YouTube and the YouTube icon are trademarks of YouTube or its affiliates.

All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.

The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at https://www.linuxfoundation.org/privacy and its Antitrust Policy at https://www.linuxfoundation.org/antitrust-policy. each as may be modified from time to time. More information about The Linux Foundation's policies is available at https://www.linuxfoundation.org.

Please email legal@linuxfoundation.org with any questions about The Linux Foundation's policies or the notices set forth on this slide.