The Community Data License Agreement (”CDLA”)

An Introduction and Overview

Christopher W. O’Neill

Associate General Counsel

Intellectual Property Law

IBM Corporation

cwo@us.ibm.com

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

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

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 for how they may be applied to data problems
  - 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

CDLA Summary

- The Community Data License Agreement (“CDLA”) family is a set of two model agreements, introduced and sponsored by the Linux Foundation
  - Announced on October 23, 2017
  - Modeled on the structure of leading open source agreements
  - Designed for use by independent data communities looking for “open” licenses that reflect the nuances of data licensing (as opposed to software or copyright licensing)
- CDLA promotes free exchange of data
  - Permits data to be freely used, modified, and republished
  - Authorship and source attribution statements must be preserved
  - No use restrictions permitted
  - Broader license coverage than mere copyright
  - Explicit permissions to create separate works and analyses of licensed data
  - Characteristic open-source style warranty disclaimers to minimize the legal exposure of publishing data

CDLA Key Terms

- Two versions of the CDLA:
  - **Sharing Version**: All data re-published must be licensed under the terms of the CDLA Sharing License
  - **Permissive Version**: Data may be republished under any terms not inconsistent with the terms of the CDLA Permissive License
- Both versions of the CDLA license provide for the following:
  - The right to make “Computational Use” of data
    - Broadly defined as enabling the right to analyze data and create analytical works based upon it
    - Analytical works do not need to be relicensed under the terms of the CDLA licenses
    - Analytical works cannot contain more that a de minimis amount of the data itself
  - Minimal reps and no warranties
  - Broad Limitation of Liability
  - No prohibition on commercial use of data

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.
  - “Enhanced Data” means the subset of Data that You Publish and that is composed of (a) Your Additions and/or (b) Modifications to Data You have received under this Agreement.
- 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.

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

- Includes:
  - Modifications to data received, and
  - Additions to data received,
  - That You publish.
- Explicitly 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 GNU GPL licenses

Why is an Open Data Agreement Even Necessary?

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
  - Copyright law is not entirely uniform across national boundaries, particularly with respect to “fair use” rights
- The CDLA is offered to avoid a period of license proliferation and aggregations of valuable data under licenses that prevent combinations necessary to optimally analyze the data over time.
  - Tenure of data value is much longer than software – and potentially perpetual.
  - Many data collections (e.g., temperature readings) grow over time and consistent rights to analyze them (irrespective of changes in the law) is of increasing value over time

Data is not the same as source code

- Traditional open source agreements are primarily based on copyright law
  - Software source and object code are works of authorship protected by copyright
  - Some agreements also contain express or implied licenses to patents that may apply to licensed code
- Data is different from code
  - In the US and elsewhere, data itself is often not protectable by copyright (see Feist Publications, Inc., v. Rural Telephone Service Co.)[1]
  - Only the creative expression of the data is protectable by copyright; Facts are not
  - Patents are typically not applicable to data itself
  - Much of the value of open data (unlike code) is in the ability to study and analyze it, but traditional open source licenses do not explicitly provide for these rights
- In order to protect data usage rights, a data license needs to invoke a broader spectrum of rights to make up for the incomplete coverage of copyright law
  - Some data provider organizations are trying any means available to lock down access to data, sometimes with direct or ambiguous terms around usage rights
  - The CDLA contains explicit terms to prevent access to licensed data from being restricted

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

Current practices around sharing data vary but generally map to requirements addressed 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 Licenses (CC-BY-SA, CC-BY)
  - 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/

Who can use the agreements?

- The CDLA agreements are sponsored by the Linux Foundation for free adoption and use by any data provider or community wishing to use them
- Examples include:
  - 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

Where can you find the CDLA agreements and FAQs?

cdla.io

Questions?