参考資料6

Community Data License Agreementの概要

○政府、企業、大学等の保有データのオープンデータ化が求められている
○Linux Foundationが2017年10月にオープンデータに関するライセンス契約書を公表
○オープンソースソフトウェア（OSS）の共有スキームをデータに応用
○目的に応じ、２種類の契約書（シェアリング版／パーミッシブ版）を用意

	シェアリング版	パーミッシブ版
許諾	オープンデータの使用、公開	
義務	拡張データ[1]を公開する場合は、シェアリング版で行う義務あり	拡張データを公開する場合には、任意の条件で公開可
メリット	オープンなコミュニティの構築、データの量・質的充実が期待できる	商用利用が容易、利用者を増やしやすい（データの量・質的充実は期待しにくい）
その他	OSSライセンスにおけるGPLライセンスの考え方に近い	OSSライセンスにおけるApache/BSDライセンスの考え方に近い

[1] 受領したデータに対する追加、修正をしたもの

Linux Foundation資料等に基づき内閣府知財事務局にて作成

Community Data License Agreement

cdla.io

Mike Dolan

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.

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

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

Community Data License Agreement

- On October 23, 2017 The Linux Foundation 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.
  - “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.

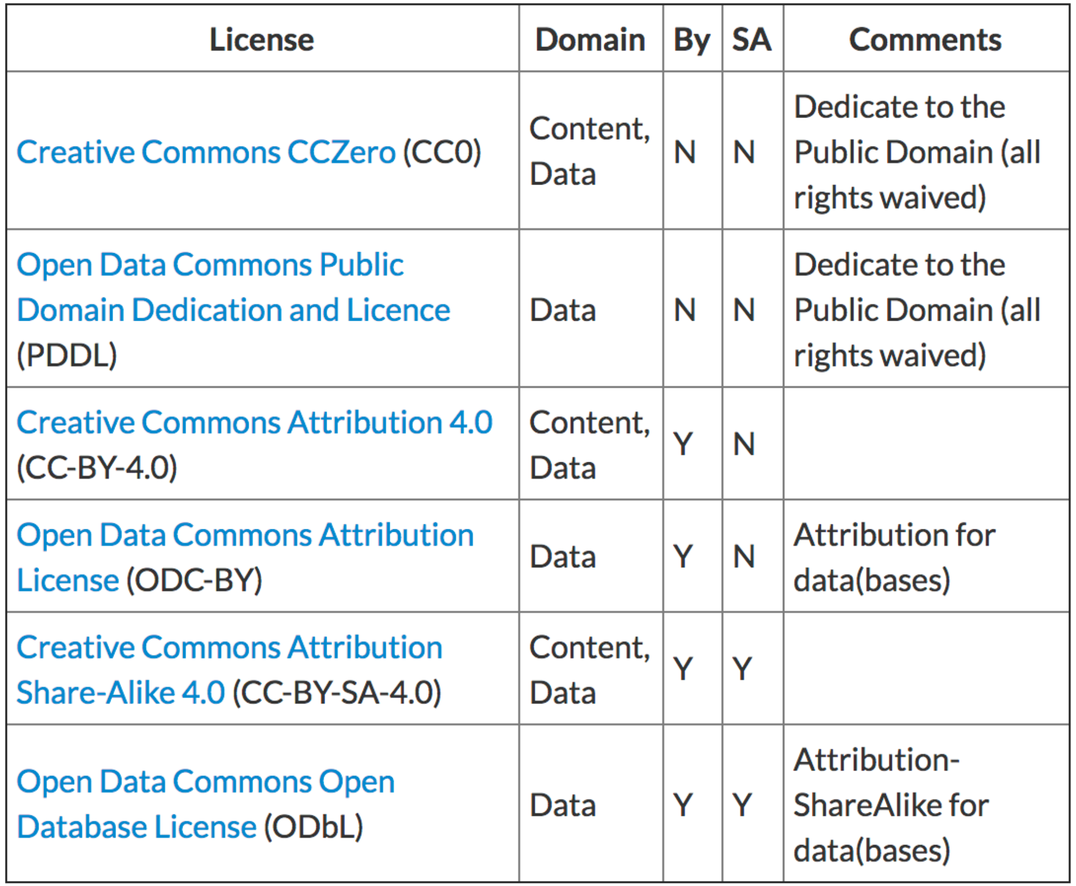
Data is not the same as source code

- In the US and elsewhere, data itself is generally not protectable 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
- 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

Current practices around sharing data vary but generally map to requirements we’ve already 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 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/

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

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

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.”

The CDLA in use – Cisco’s Network Anomaly Telem

NETWORK ANOMALY
TELEMETRY DATASETS

- 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.

cdla.io