REMOVING BARRIERS TO DATA INNOVATION

Empowering people and organizations to share and use data more effectively

Sharing data can help address some of society's biggest challenges and help individuals and organizations be more innovative, efficient, and productive. For example, health researchers can combine data sets to better diagnose and treat cancer. Insights from weather data can be used to promote environmental sustainability. Makers of self-driving cars can better analyze the terabytes of data their cars generate by the hour to make automated driving safe.

And yet, much of the potential that can be realized through data sharing remains untapped – in part because the tools for collaboration are often immature, non-existent, or overly complex. Data sharing agreements can take months to draw up, oftentimes deterring organizations from sharing data at all.

Many people and organizations are doing great work to tackle this problem and we've collected a list of some of these resources at news.microsoft.com/datainnovation. At the same time, we think there's an opportunity to do more. We want to help make it easier for individuals and organizations that want to share data to do so. We're working with companies, academics, and researchers to build better processes and tools. As a first step towards this aim, we're sharing a set of draft data use agreements to govern the sharing of data, particularly in the context of training AI models. The agreements cover the following scenarios:

1. **Open Use of Data Agreement (O-UDA) –** This agreement is intended for use by an individual or organization that owns or has the rights to distribute data for unrestricted uses. This is a "one-to-many" agreement and is intended for use with data for which there is no privacy or confidentiality concern.

    - ***Example scenario*** – A port authority is collecting wind data over time, including the direction of the wind, speed, air temperature, among other related statistics. The port authority could make this data publicly available using the O-UDA for unrestricted uses, as the data has no privacy or confidentiality concern.

2. **Computational Use of Data Agreement (C-UDA) –** This agreement is intended for use with data sets that may include material not owned by the data-providing individual or organization, but where it may have been assembled from lawfully and publicly accessible sources. This agreement allows a provider to make the data publicly available only for "Computational Purposes" (activities necessary to enable the use of data for analysis by a computer, like machine learning). This is a "one-to-many" agreement and is intended for use with data for which there is no privacy or confidentiality concern.

    - ***Example scenario*** – An organization has a database of images that it wants to make publicly available. The database was assembled lawfully from publicly accessible sources, though it may contain images covered by copyright. As a result, the organization could make this data publicly available using the C-UDA for computational use only, to be consistent with copyright laws.

3. **Data Use Agreement for Open AI Model Development (DUA-OAI) –** This agreement provides terms to govern the sharing of data by an organization with another for the purpose of allowing that second organization to use the data to train an AI model, where the trained model is open sourced. This is a "one-to-one" agreement and contemplates the sharing of data for which there may be privacy and/or confidentiality concerns.

    - ***Example scenario*** – A company has a database of information collected from its manufacturing plants. The company would like to share the database with another company in its industry to train an AI model that identifies opportunities for manufacturers to enhance safety measures. The database may contain private or otherwise sensitive information. The companies can share the database using the DUA-OAI, and then open source the trained AI model to allow other companies and governments to use it.

4. **Data Use Agreement for Data Commons (DUA-DC) –** This agreement might be used by multiple parties in possession of large data sets pertaining to a particular subject matter who want to share the data sets through a common, Application Programing Interface (API)-enabled database. This multi-party agreement contemplates that each party will contribute data to a common database through agreed upon APIs and then access data from that database through other agreed upon APIs.

    - ***Example scenario*** – Each of several adjacent municipalities collects data regarding the flow of vehicular traffic within their borders. These municipalities would like to contribute that data to a common database so that the combined database might be accessed by each municipality's traffic signal control system with the goal of improving the flow of traffic across the entire region. These municipalities would like to contribute data to, and access data from, the common database via APIs to ensure proper format of the data and compatibility with their traffic signal control systems. These parties can use the DUA-DC to structure their data sharing arrangement.

Additional agreements will follow covering other scenarios. We look forward to improving these agreements with feedback from the community, and taking additional steps that can help make data more open and accessible for everyone.

Why did you pick these data sharing scenarios?

We started with three common data-sharing scenarios, with the intent to build-out additional data sharing agreement templates, each of which will be designed to address a different data sharing challenge. The templates we create might not address every data sharing scenario and the scenarios we try to address might be improved as we learn more from users. In this process, we hope to listen to your feedback, learn from it, and improve this work product over time.

Why are these files labeled "Data Use Agreements" and not licenses?

Agreements are customarily used when providing access to material, whereas licenses are customarily used when the material being provided is subject to unambiguous intellectual property rights and therefore the rights of use need to be explicitly described. In the case of data, the intellectual property rights (if any exist) applicable to data may vary around the world. Accordingly, we believe that providing data for broad use is more properly accomplished via an agreement, rather than under a license whose legal foundation may be ambiguous.

In the case of the C-UDA, the use of data for computational purposes is permissible under law, and therefore it is more appropriate to define how access and use of data may be accomplished under an agreement rather than via license for the use of such data.

Don't public licenses already exist for licensing data?

Many agreements have been developed for sharing data broadly, including public licenses such as the Open Database License, Open Data Commons license, various Creative Commons licenses, and the Community Data License Agreement. These licenses address data sharing using different approaches but with the goal of encouraging broad use and distribution, and they are useful for a wide variety of data sets.

We developed the O-UDA and C-UDA to be a much shorter and simpler agreement for sharing data. The C-UDA addresses an identified gap among existing public licenses – sharing data sets that include material that the distributor may not own but which was lawfully accessed from publicly available sources and therefore legally permissible for computational uses.

How do the O-UDA and C-UDA differ from each other?

The O-UDA places no restrictions on use of data, which means it is best used with data that the distributor owns or controls. The C-UDA restricts the use to computational uses, which means it is best used with data that the distributor may not own or control specific material in a data set, but wishes to enable uses that are permitted by copyright law in a variety of jurisdictions.

How do the O-UDA and C-UDA differ from other public license examples?

These two agreements are simple, short agreements that cover two scenarios: data created and owned by the distributor, and data assembled by the distributor from lawfully accessed public sources. They are developed to be more "distributor-focused" which means they may not contain warranties or representations present in other public licenses, or they may limit uses to those permissible under law (in the case of data assembled from third party materials).

How do these agreements address the issue of privacy?

Some data sets can be shared without raising privacy concerns. The O-UDA and C-UDA are designed to be used in those scenarios. By contrast, the DUA-OAI allows parties to share data to train AI while accounting for the privacy concerns of the data subjects from which the data was collected. There are varying contractual and technical methods the parties might use to do this which are discussed in the accompanying "READ ME" overview of the DUA-OAI agreement.

Do these agreements allow me to modify or redistribute data?

Yes. Data can be modified, and both the data and any modifications can be redistributed. The O-UDA requires two minor obligations for redistribution. First, the redistributor must pass on all attribution information. Second, the redistributor must pass on the warranty and liability disclaimers from the data provider. The C-UDA requires the redistributor to use the Agreement when redistributing the original or any modified data. These requirements only apply to data, and not to any Outputs.

Will Microsoft distribute its data under these agreements?

Yes, where the agreements are appropriate, we're applying them to Microsoft datasets deemed suitable for use under an open agreement.

Where can I provide feedback?

The O-UDA and C-UDA are hosted on GitHub for comment. If you prefer, and to provide feedback on all four agreements, please email your comments to datainno@microsoft.com and let us know if we can attribute your name to your comments as we resolve them on the master file on GitHub.