

Data Sharing, Licenses, and Agreements

Research Report

Katy McKinney-Bock, revised November 20, 2023

1 Introduction

To continue research into best practices and current beliefs/opinions about data collection, data governance, data sharing, and informed consent, this October report turns to mechanisms and reasons for data sharing, asking the question: *What conditions the environment for sharing data, and what are some examples of mechanisms for data sharing?* More specifically, the research questions for October explore *what are the requirements and common elements present in data sharing agreements? What licenses are commonly used to share data?*

This second investigation seeks to understand how data is being shared in 2023, both in terms of specific mechanisms available like licenses or data sharing agreements, but also how data sharing policies discuss and plan for data sharing, which is one component of the data governance models that will be the topic of research in November. This report bridges the September participatory data stewardship report, which explores what it means to steward data in a participatory way, and the next report, which will turn to data institutions and governance models such as data trusts to explore ways to establish formal participation around data. The role it should play is to give an idea of the specifics of how data is being shared in 2023 within some of these broader frameworks, and to highlight some of the present challenges.

The structure of the report first looks at how data can be openly shared, and what are some of the mechanisms that are currently being used to give permission to use data. I look at copyright licenses like [Creative Commons](#), which, in practice, are being used to open data despite the challenges to covering data by copyright, and the [Open Data Commons Open Database Licenses](#). Both of these licenses are listed as requirements for opening a digital public good through the [Digital Public Goods Alliance](#), and are considered approved licenses by the Open Data Commons in addition to being widely used.

When data cannot be opened, there are a variety of different approaches that can be taken to sharing data, such as granting permission to the end user of the data via general agreement terms and conditions, via a code of conduct, or via a specific data sharing agreement. In the *How to Decide How to Share* section, I turn to *data sharing policies* that encourage openness of data and provide a framework for researchers to make decisions around how and when they will share data. In cases where an open license cannot be used, a data sharing agreement is a possible mechanism to share data. In the *How to Share* section, I explore checklists for data sharing and common components of a data sharing agreement. Finally, in the *How to Consider Sharing* section, I turn to frameworks that present principles and parameters around data sharing and decision-making around data.

All three of these *How-to* sections are limited in scope, and there is much work that has been done around data sharing policies, principles, and legal agreements. I did not, and could not, cover everything. Instead, I present some examples of each to illustrate considerations that establish both the environment for sharing and the components that are present when creating agreements.

2 Open Sharing and Open Licenses

In 2022, the [Open Data Institute](#) (ODI) published a Data Governance Playbook for health data, geared toward “non-technical leaders.” In it, they synthesize some of the basics of data sharing. In play ten, they discuss the four (primary) methods for sharing data: licenses, data sharing agreements, terms and conditions, and codes of conduct. In this, they point out that licenses are best used for open data. Non-open data, especially sensitive data, is shared via another mechanism, like a data sharing agreement (formal or informal), or via a terms of service (Open Data Institute, 2022).

Making a decision between licensing data (more open) and requiring a more detailed agreement for sensitive data (less open), requires both an understanding of license and data sharing agreement (DSA) mechanisms, as well as a basic understanding of data classification and data rights. This section starts with a brief discussion of data classification to establish an idea of what types of data are being shared and how, and then a brief discussion of data rights specific to research data. I then move to examples of open licenses and a growing debate around openness in 2023 related to emerging AI technologies.

2.1 Data Classification

The ODI developed a set of graphical descriptions of data classification, the Data Spectrum, which defines a continuum of data from Open to Closed and sets out the basic conditions that would allow data access to happen, with examples.

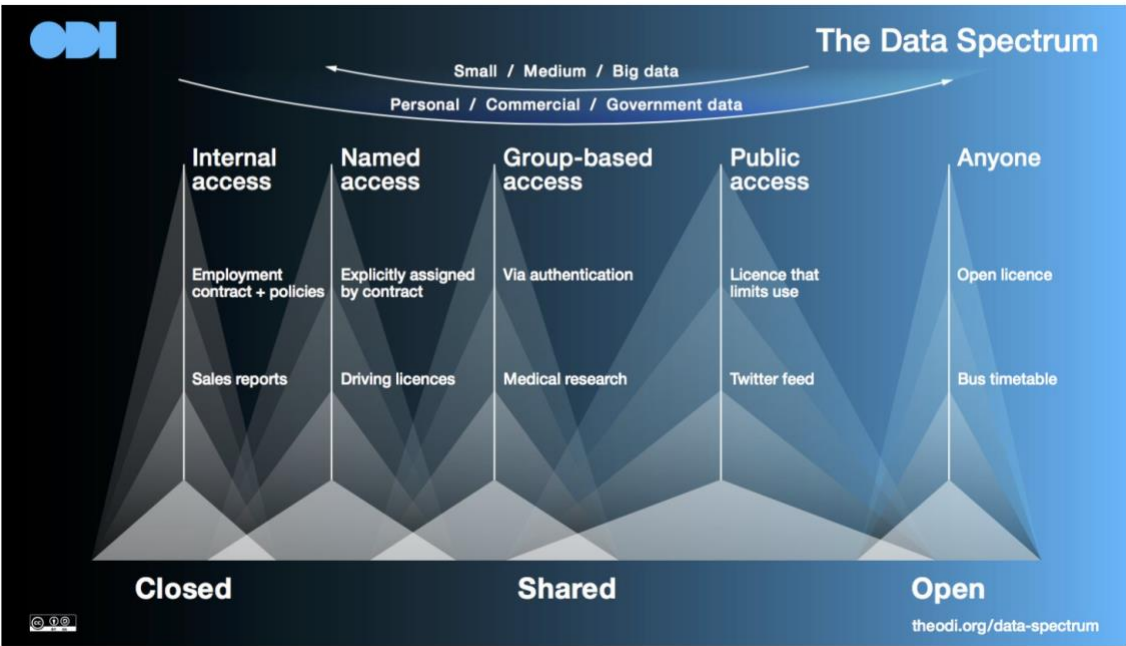


Figure 1: [The Data Spectrum](#) (Open Data Institute, 2020), modified (resized) and used under [CC BY](#).

While it doesn't discuss data sharing agreements, it does provide a sense that non-open data requires some sort of authentication or contract to access it, and open data requires a license granting permission for use.

There are some key terms that are used to describe data, which would place it along the open-to-closed spectrum.

- Sensitive/Non-sensitive
- Personal/Not Personal
- Open/Not Open

To define *open data*, the ODI as well as the Digital Public Goods Alliance, whose work I summarize and report on below, adopt the Open Definition, published by the Open Knowledge Foundation (OKFN). This meaning matches the Open Source Definition of open with respect to software, as well as free/libre from the Free Software Definition and Definition of Free Cultural Works.

The Open Definition, v2.1, from the Open Knowledge Foundation was published in 2015, though the OKFN is working on a current initiative to redefine open based on ambiguities that are present, in particular for AI technologies. (Ávila, 2023; Open Knowledge Foundation, 2015). The summary of the definition is:

“Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.” (Open Knowledge Foundation, 2015)

2.2 Rights to Data

In a primer on research data, Carroll (2015) describes the kinds of rights to data that someone may have: trade secret (proprietary/confidential information), copyright, *sui generis* database rights in the EU and South Korea, and patent rights. I focus here on copyright and database rights.

Carroll's insight into the complexities of copyright for research data is worth discussing before licenses. Copyright is challenging to understand in a research context, and copyright does not impose any restrictions on facts or knowledge, i.e. these cannot be copyrighted. Carroll discusses two complexities with copyright and research data: first, there are “layers of copyright in databases” such that not all components in a database may be copyright, but some may be. For example, the data may be facts (not copyrightable), but the database structure may be original enough to be copyrighted. Or, a chart/graph of the dataset produced by the researcher may be subject to copyright, while the data and database are not. Second, the extent to which reuses are permitted varies by nation in ways that impact researchers. One example of this is content mining; the UK created an exception to copyright in 2014 that allowed researchers to mine the research literature. This is one area that is currently a hot topic in the use of text-mined datasets used to train AI.

In cases of databases that are not subject to copyright, certain nationalities grant *sui generis* database rights that recognize the effort made in creating a database even if it does not meet the threshold for creativity that copyright requires. *Sui generis* database rights in the EU/South Korea protect use of large portions of the database.

For the Better Deal for Data, one question that remains is building a solid understanding of what rights exist to different type(s) of data, and which right(s) are relevant to the commitments we make to individuals or organizations that use the technology we have developed. One straightforward example is GDPR's protection of personal data.

2.3 Two Common Open Licenses

The Digital Public Goods Alliance (DPGA), an initiative that supports investment into digital public goods with a focus on achieving SDGs, maintains a list of licenses for software, data, and content that they believe to be conformant to the Open Definition, by SPDX identifiers (Digital Public Goods Alliance, 2022). To note, as of October 2023, the Open Knowledge Foundation is a member of the DPGA (Open Knowledge Foundation, 2023). The DPGA provides this list to assist in determination whether a digital asset counts as a digital public good and can be listed in the DPGA Registry. One of the minimum requirements to enter a digital asset into the DPGA registry is licensing with an approved open license.

For content, they list a set of Creative Commons licenses. For data, they list both a subset of Creative Commons (CC) licenses as well as a set of Open Data Commons Licenses. For software, there is a long list of licenses that are all listed as approved by the Open Source Initiative. Despite listing CC licenses for data, they state the following requirement under the "Use of Approved Open Licenses" in their standard:

"For open content collections the use of a Creative Commons license is required. DPGs are encouraged to use a license that allows for both derivatives and commercial reuse (CC BY and CC BY-SA), or dedicate content to the public domain (CC0); licenses that do not allow for commercial reuse (CC BY-NC and CC BY-NC-SA) are also accepted. For open data, an Open Data Commons approved license is required." (Digital Public Goods Alliance, n.d.)

For this report, I discuss Creative Commons licenses (focusing on 4.0) and the Open Data Commons licenses that the DPGA chooses to list as approved for open data that are considered a digital public good, as they are examples of high use, well-known open license standards. While I am using the DPGA as a starting point, this practice is common for registries; see also [FAIRsharing.org](https://fairsharing.org).

2.3.1 Creative Commons License

Creative Commons is a nonprofit organization that provides a set of simple copyright licenses that, over the past twenty years, has become a standard for giving permission to share and use creative works.

CC licenses grant a set of baseline rights, and then there are four options for a CC license to add or modify additional rights: Attribution, Share-alike, Non-Commercial, and No Derivatives.

- Attribution (BY): "credit must be given to the creator."
- Share-alike (SA): "Adaptations must be shared under the same terms."
- Non-Commercial (NC): "Only noncommercial uses of the work are permitted."
- No Derivatives (ND): "No derivatives or adaptations of the work are permitted."

(Creative Commons, n.d.-a)

The full set of combinations is not generated, as SA and ND cannot co-occur, and the most commonly used licenses all include the BY clause, which means there are six common licenses (listed on the CC License page [here](#)): CC BY, CC BY-SA, CC BY-NC, CC BY-NC-SA, CC BY-ND, CC BY NC-ND. There is also a public domain dedication, CC0, which allows the creator to give up their copyright and dedicate their work to the public domain.

CC BY and CC BY-SA are the only two licenses that fit under the Open Knowledge Foundation's Open Definition, and are the two types of CC licenses that the DPGA lists as acceptable for their standard (in addition to CC0). Version 4.0, released in 2013, allowed for a broader set of jurisdictions to enforce the licenses, as one of CC's goals is for its licenses to work globally; this is often referred to as the "International" license (Creative Commons, n.d.-b).

CC licenses have been used for purposes beyond sharing of creative content, and there has been some discussion about the appropriateness of these types of uses of CC licenses. In particular, use of the CC licenses for both data and databases, especially when the data is highly factual and not necessarily covered under copyright law (facts cannot be copyrighted). We believe that users frequently use a Creative Commons license with the intent to grant permission to use a dataset that they hold some sort of ownership or rights to grant permission, even if the artifact/data is not copyrightable because it does not meet the threshold for creativity. In any case, it is important to know that the presence of a CC license for an open dataset may or may not cover the entirety of the artifact, but only the copyrightable portion.

Creative Commons states that data and databases may be licensed, particularly using the CC0 public domain dedication:

"The CC0 Public Domain Dedication can be particularly important to maximize the re-use of data and databases, since it otherwise may be unclear whether highly factual data and databases are restricted by copyright or other rights. CC0 is intended to cover all copyright and database rights, so that however data and databases are restricted (under copyright or otherwise), those rights are all surrendered." (Creative Commons, n.d.-c)

Creative Commons' wiki also states that Creative Commons 4.0 licenses may be used to license databases and the data contained within the database, as long as the contents are "sufficiently creative" (i.e. copyrightable) and also the database creator has the rights to those contents. The example provided in the CC wiki states that "original poems contained in a database would be protected by copyright, but purely factual data (such as gene names or city populations) would not" (Data - Creative Commons, n.d.). The elements that may be copyrightable by CC licenses are: the database model, data entry and output sheets (or datasheets), field names, and data. Once again, each of these is copyrightable only if it is sufficiently creative - CC points out that a field name such as "address" is unlikely to be copyrightable. Regarding databases, Creative Commons states that their 4.0 licenses can be used for *sui generis* database rights, as per EU legislation, while previous versions, like 3.0. did not consistently grant *sui generis* database rights (Data - Creative Commons, n.d.).

Care should be taken when selecting a license. For example, the CC wiki states that “CC does not recommend use of its NC or ND licenses on databases intended for scholarly and scientific use,” as even privately funded academic research may be barred from using CC NC licensed data because it could be considered commercial.

More strongly, the Open Data Commons (ODC) recommends *against* using CC licenses other than CC0, after discussing four differences between content (covered under copyright) and data:

“These sorts of issues suggest one may should [sic] be cautious in directly applying, say, Creative Commons licenses to data(bases). In fact, Creative Commons have themselves have recommended against using their licenses (other than CC0) for data and databases (which is also a reason, in and of itself, not to use CC licenses (other than CC0) for data, as CC are unlikely to able/willing to fix data-specific issues that arise in relation to those licenses).”

The four issues that ODC mentions are worth following up on. First, the rights in databases are different from content, as there are IP rights to databases, and because copyright applies differently/in a lesser way to databases when compared with typical creative content. Second, ODC points out that it is necessary to distinguish between the database and its contents, which may be licensed separately. Third, further complexity arises when there is a material product generated from both the database and its contents, such as a map that is generated from a geospatial database, which is a product of the database rather than a derivative database. Lastly, databases are much more like ‘code’ than ‘content’ in that they have very high levels of reuse, making licensing of derivative works much more relevant.

2.3.2 Open Data Commons

The Open Data Commons was created in 2007 as a place to host an open database license, the Public Domain Dedication and License. In 2009, the project was transferred to the Open Knowledge Foundation (Open Knowledge Foundation, n.d.-a).

Open Data Commons has a set of three open licenses for data: the Open Data Commons Open Database License, aka “Attribution Share-Alike for data/databases”, the Open Data Commons Attribution License (AL), aka “Attribution for data/databases”, and the Open Data Commons Public Domain Dedication and License (PDDL), aka “Public Domain for data/databases.” (Open Knowledge Foundation, n.d.-c)

The Open Database License (ODbL) allows for someone to *share* (“copy, distribute, and use the database”), *create* (“produce works from the database”), or *adapt* (“modify, transform, and build upon the database”), with the requirements that licensees attribute, share-alike, and keep open the database (which means that while restrictions such as Digital Rights Management/DRM may be used, a non-restricted/non-DRM version must also be distributed alongside it) (Open Knowledge Foundation, n.d.-f). The Attribution License (ODC-By) grants permission for someone to *share*, *create*, or *adapt* the database, with a requirement for attribution ((Open Knowledge Foundation, n.d.-e).

The Public Domain Dedication and License (PDDL) puts the database in the public domain, granting permission for someone to *share*, *create* and *adapt* the database, with no restrictions on use (Open Knowledge Foundation, n.d.-g). Along with the PDDL license, Open Data Commons provides a set of

Community Norms that can be appended. It is very clear that these norms are not a binding contract, but rather, a set of principles that people should adhere to. In these, the norms ask users to share-alike also using the PDDL, to provide attributions, to link to the source, and to use open formats. In summary, the binding parts of the ODbL license are not part of the PDDL license, but are suggested norms to be followed (Open Knowledge Foundation, n.d.-d).

2.3.3 Interaction of common open licenses and key differences

CC0 and PDDL serve the same functionality of waiving copyright and dedicating the work to the public domain. They are also interoperable, in the sense that any content that has a CC0 or PDDL license can be mixed with other content. There is one difference that ODC points out: "Unlike CC0 however, the Open Data Commons system includes a set of Community Norms that can be linked with the license." (Open Knowledge Foundation, n.d.-b)

Both the Open Data Commons and Creative Commons discuss the differences between their licenses, pointing out that CC licenses are intended for a broader set of creative artifacts, while the ODC licenses are specific to databases. Creative Commons states that ODC licenses "apply only to *sui generis* database rights and any copyright in the database structure. These licenses do not apply to the individual contents of the database" while the 4.0 Creative Commons licenses apply to the database and contents (Data - Creative Commons, n.d.).

Additionally, the Creative Commons wiki states that "ODC licenses may create contractual obligations even in jurisdictions where database rights would not otherwise exist and would be necessary only for the license permission. CC has crafted its licenses to ensure that they never impose obligations where permission is not otherwise required to use the licensed material" (Data - Creative Commons, n.d.). Given that CC 4.0 extends to database coverage (not always true in version 3.0), it may make sense to apply Creative Commons licenses where possible, though the benefit of ODC PDDL is that it comes with a set of Community Norms that can accompany the license. ODC states that this is a benefit when applied, in particular, to data: "and one of the strengths of the Open Data Commons approach is the Community Norms document, which makes the Open Data Commons approach particularly applicable for data, especially data produced by the academic sector" (Open Knowledge Foundation, n.d.-b).

Summary table:

CC License	ODC License	Open?	Comment
CC0	PDDL	Yes	Public domain release; PDDL has optional norms
CC BY	ODC-By	Yes	
CC BY-SA	ODbL	Yes	
CC BY-NC CC BY-NC-SA CC BY-NC-ND		No	Commercial use prohibited; problematic for researchers
CC BY-ND		No	No derivatives; not useful for researchers

2.4 Emerging AI licenses

As part of AI growth, new licenses that attempt to manage some of these ambiguities have been developed. I cover three examples: the Montreal Data License Framework (Benjamin et al., 2019), the Responsible AI Licenses (RAIL, Contractor et al., 2022), and the Allen AI ImpACT licenses (Allen Institute for AI, n.d.; Dumas, 2023).

The Montreal Data License Framework was developed to address downstream data use and allow individuals to share data with other parties for use in AI. The published paper was accompanied by a license generator to help people license their data. In Benjamin et al. (2019), the stated goal of the framework is to “build towards a common framework for data licensing akin to the licensing of open source software.” One problem discussed is that existing licenses don’t apply well to how data is used for AI, giving rise to many ambiguities (paraphrased here, see Benjamin et al., 2019, p. 3):

- *Use*: Fair use exemptions aren’t yet tested for AI/ML, leaving users “guessing”
- *Commercial vs Non-Commercial*: The authors take issue with a binary definition of commercial and provide examples such as an academic institution commercializing uses of a dataset in their transfer office
- *Research*: AI/ML has increasing numbers of private/for-profit companies participating in research, and researchers affiliated with both universities and private companies
- *Lack of Uniformity*: The authors point to FOSS’s huge success standardizing; but the authors contrast the definition of “use” for software vs “use” for data in AI/ML and state it is even more challenging to define; current licenses they list lack uniformity or standard terminology.
- *Share-alike Requirements*: Share-alike is challenging in AI/ML because researchers often combine datasets and then SA applies to the whole dataset if it is considered a derivative

While it appears to us that the issues that Benjamin et al. (2019) identify have been present and discussed in non-AI/ML applications (e.g. questions about commercialization in academic institutions, and the challenges of share-alike and downstream uses of software with different licenses), we recognize that there is additional ambiguity around current AI/ML applications; in particular, the lack of uniform terms that apply to AI/ML models. They propose a standard taxonomy for AI/ML as a starting point, and a model license. However, there are several shortcomings: data may still be subject to copyright, there may be personal information, and property rights to data is left unaddressed (which I also don’t explore in this report), and they also don’t address *sui generis* database rights. Their contribution, however, is the creation of a taxonomy and a license that addresses a level of modularity that uses this taxonomy; in particular, the distinction between the use of data and the use of data *with models* (Benjamin et al., 2019).

The RAIL licenses primarily address use restrictions when distributing or modifying data, applications, models, or source code for AI; in particular, they aim to prevent specific harmful uses of AI, such as surveillance. The problem space that they are addressing is for creating a license for researchers/engineers to disseminate scientific research in a way that restricts the use of that research. This likely puts RAIL licenses in the general category of “ethical licenses” which are not considered open by OSI. “Releasing algorithms, code, or data as part of the research process, a product, or a service can mean that developers, engineers, and researchers must sacrifice all their control over the application of these artifacts which can lead to a moral dilemma and potentially harm the

dissemination of scientific research" (Responsible AI Licenses (RAIL), n.d.). There is wide use of these licenses. "As of 1 April 2023, there are over 8,000 RAIL-Licensed repositories on the HuggingFace Hub" (Growth and Adoption of RAIL Licenses, n.d.). RAIL credits both the Open Source Initiative as well as Creative Commons for the inspiration for RAIL licenses, with the value of enabling knowledge sharing (FAQ, n.d.).

There are four license types that are part of their framework: RAIL-D, RAIL-A, RAIL-M, RAIL-S, with DAMS (Data, Application/executable, Model, Source Code) being the order to list these licenses. There are different types of RAIL-DAMS licenses for use cases as well: AIPubs OpenRAIL, and AIPubs Research-Use RAIL. OpenRAIL permits free and open access, reuse, and downstream derivatives within the use restrictions in the license, which they refer to as "behavioral-use restrictions". AIPubs OpenRAIL licenses restrict the objects to research use only, which (as they note) allows for less specific behavioral-use restrictions.

The Allen AI ImpACT licenses have been developed as part of the current AI Olmo project, which is an LLM that is being trained on Dolma, both of which will be open, transparent AI products. As part of their transparency initiative, they are releasing logs of how they prepared the dataset and LLM. Unlike RAIL, which focuses on licensing around on the 'artifact' itself, such as data vs. model vs. application, the ImpACT licenses focus on licensing based on risk: "In other words, what you are permitted to do with an artifact depends on its risk category (low, medium, or high), rather than the type of artifact (e.g., data, a model, or something else)" (Allen Institute for AI, n.d.).

The ImpACT licenses allow someone to use the artifact and to create derivatives, with a widely construed interpretation of derivative. There are use-based restrictions that continue 'downstream' to derivatives and further distribution of their artifacts. Attribution is required, and licensees must retain all the notices that came with the artifact. For low and medium-risk models, any derivatives require a "Derivative Impact Report" to be released alongside the derivative (Allen Institute for AI, n.d.).

Without arguing whether the Montreal, RAIL, or ImpACT licenses fit the definition of 'open', they do share some goals in common with the open data licenses they were inspired by, e.g. Creative Commons, and they specifically address use of AI artifacts, which copyright as well as software licensing may not properly cover. Creative Commons is also working to discuss generative AI and copyright recently (Walsh, 2023).

The definition of "open" is under discussion in multiple venues currently. In addition to the initiative undertaken by the Open Knowledge Foundation in 2023, the Open Source Institute is also leading a discussion of 'open' in 2023 to define "Open Source AI" (Open Source Initiative, 2023), and the noticeable lack of an open data license has been under discussion for a while (see e.g. discussion in Arrazolo, 2017).

The OKFN challenges the definition of open beyond the emergence of new technologies:

"At the same time, a new generation of thinkers, academics and activists has emerged, with an agenda that addresses topics not fully discussed or covered during the early days of the open movement, such as data extractivism, digital colonialism, economic, racial and gender-based violence and inequalities, and the effects of an open ecosystem on climate justice. Creating a

bridge between old definitions and new discourses is important to keep the open ecosystem alive and current.” (Ávila, 2023)

The motivation stated here differs from the motivations for the Montreal Data License, which tries to disambiguate ambiguities from *use* that arise from interpretations of data in an AI context vs data in other contexts. However, the RAIL license specifies behavioral use restrictions and tries to limit the harm of AI applications, which could be considered part of the discussion that the OKFN is attempting to foster. The ImpACT licenses also include clearly stated use restrictions, e.g. “military weapons purposes or in the service of nuclear proliferation or nuclear weapons technology” in their licenses.

AI/ML applications have challenged current licensing models because differences in the technology and the utilization of data are not clearly covered by existing open licenses, but they have also reignited discussion about whether “open” can and should include behavioral use restrictions. There are several follow-up research questions here that could be considered as part of the Better Deal for Data project: whether and how behavioral use restrictions play a role in our commitment to data subjects.

3 How to Decide How to Share: Data Sharing Policies

A different approach is required when datasets cannot be made open, such as when they contain personal information or sensitive information, or if there is other risk of harm which would prevent release of the dataset. In these cases, the affected parties must enter into an agreement, whether informal or formal, to share the data for particular use. This is a Data Sharing Agreement. The Health Governance Playbook notes that a Data Sharing Agreement is one mechanism to share data that cannot be openly shared. In this section, I look at three examples of data sharing policies or model agreements that can be compared against the approach of sharing via an open license.

In the following sections, I examine three examples of data sharing policies and principles, which set the context for how data will be shared by an organization, and the expectations and principles that would allow data sharing to happen (or not). I include an examination of a set of principles that could be implemented as a policy, with an example agreement. From there, I examine the components and minimum requirements set out by organizations for best practices in Data Sharing Agreements. Finally, I step back again to look at frameworks for data sharing.

To find examples of data sharing policies, I started with the resource FAIRSharing.org, which is a registry of standards, a registry of databases/digital assets, and a registry of data policies. Running a search for “Data Sharing” over the policy registry, I found several examples of data sharing policies (with 23 results of “maintained” policies). I have selected three policy entries that I found to be recently updated, and/or of particular relevance. The policies contributed to FAIRsharing.org are primarily from publishers (e.g. Wiley, Taylor & Francis) as well as funders (e.g. Wellcome Trust, NIH).

3.1 Policy #1: Wellcome Trust Sanger Institute, March 2023

The Wellcome Trust Sanger Institute is a genomic research institute established in 1936 in the UK. They recognize that open data accelerates research and translation, so they encourage openness in data, in publishing, and in making IP available for the public good. Their Data Sharing Policy begins with four principles, paraphrased here:

- Access: Rapid access to datasets, made public when possible, using data and interoperability standards.
- Ethical Considerations: Privacy of research participants will be protected via access management to datasets.
- Rights of Data Providers: Researchers must be appropriately credited.
- Optimizing Translation: “The Institute recognises that, in specific instances, the use of intellectual property protection and attendant potential delays to data sharing may be necessary to prevent inappropriately exclusive claims by others and to ensure health benefits occur.”

(Wellcome Trust Sanger Institute, 2023)

The last principle addresses the goal in much of genomic research that research results be translated into clinical practice, and the Wellcome data sharing policy states that data sharing “may be delayed to seek IP protection when this is necessary to optimize translation” (Wellcome Trust Sanger Institute, 2023). Following the principles, the document contains a set of Data Sharing Guidelines (v15, March 2023) which provide “practical guidance” for researchers following the Data Sharing Policy. Both open data and managed access data are covered under the policy.

It seems that the primary method of data classification as managed occurs via research participant consent defined at the outset of the study. Managed access data is defined as: “data that may be released to researchers under certain conditions with restrictions on use and re-distribution usually related to the terms of consent given by research participants. Managed access datasets should be submitted to the European Genome-phenome Archive (EGA at EMBL-EBI).”

It is presumed that data will be released according to the policy, and there are certain exceptions listed that would permit no release of the data. The short list of exceptions are:

- Sensitive studies: defined either by socio-political sensitivity or by high risk to individual participants if they are reidentified (even if the risk of reidentification is low).
- Capacity Building Projects: “Collaborations with researchers in low-to-middle income countries may request a delay to data release to give those researchers the opportunity to develop their own skills and expertise in data generation and analysis.”
- Bioterrorism: cases where pathogens or other data could be misused or be a potential security threat.
- PhD Data may be exempt from release until the thesis is published, but it is recognized that the data should then be released according to the policy.
- Seeking IP protection may also delay release of data, “when this is necessary to optimize translation”

The other components of the policy address the types of data that are under the scope of the policy (e.g. sequencing data, genomic data, mass spectrometry, summary statistics) and minimum requirements for submitting metadata. The policy also addresses risk assessment for release of summary statistics and the possibility of reidentification, with examples of low and high-risk data. There

is a release timing table based on data type and a section on anonymization procedures. The policy also requires that all external collaborators adhere to this policy.

Components:

1. Release strategy
2. Data Types
3. Metadata Submission
4. Sharing Summary Statistics
5. Release Timing strategy
6. Exceptions to Data Release
7. Collaborations
8. Publication Moratoria
9. Audit
10. Anonymization
11. Glossary

3.2 Policy #2: NIH Data Management and Sharing Policy, 2023

The United States National Institutes of Health (NIH) Data Sharing Policy in FAIRSharing.org has not yet been updated; it contains the original 2003 statement. However, the NIH has released a new Data Management & Sharing Policy as of this year, 2023, twenty years after the original. The scope of this policy is broader than the Wellcome Trust policy, though the focus is still on scientific research: “The DMS Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. This includes research funded or conducted by extramural grants, contracts, Intramural Research Projects, or other funding agreements regardless of NIH funding level or funding mechanism. The DMS Policy does not apply to research and other activities that do not generate scientific data, including training, infrastructure development, and non-research activities” (National Institute of Health, 2023b).

The components of the policy include:

1. Purpose
2. Definitions
3. Scope
4. Effective Dates
5. Requirements
6. Data Management and Sharing Plans
7. Managing and Sharing Scientific Data
8. Compliance and Enforcement

As the policy applies to all NIH-funded research, the definition of scientific data (in section 2) is much wider: “The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications” (National Institute of Health, 2023b). The definition provides exclusions, such

as lab notebooks and preliminary analyses. The policy establishes a requirement that all projects have a Data Management and Sharing Plan prior to funding release.

The policy takes a broad stance on sharing, stating that “NIH expects that in drafting Plans, researchers will **maximize** the appropriate sharing of scientific data, acknowledging certain factors (i.e., legal, ethical, or technical) that may affect the extent to which scientific data are preserved and shared” (National Institute of Health, 2023b). The policy also addresses human participants and requires compliance with relevant law as well as additional information about how privacy, rights, and confidentiality will be respected. The policy explicitly encourages researchers to plan for how data management and sharing information will be a part of informed consent, “including communicating with prospective participants how their scientific data are expected to be used and shared” and whether the information should be controlled.

In contrast with the exclusions named by the Sanger Institute, the example exclusion reasons named in the NIH FAQ are broad, centering around legal compliance, policy compliance, regulation, as well as informed consent / existing consent, protecting research participants from harm where de-identification is insufficient, and restrictions with other funders, repositories, or HIPAA covered entities. Both policies do address minimizing harm to participants, in particular, human subjects, and delay of release or non-release is discussed (National Institute of Health, 2023a).

Both policies encourage maximal data sharing.

3.3 Policy #3: Chatham House Model Principles for Public Health, 2017

Chatham House, or the Royal Institute of International Affairs, is a UK think tank that has existed since 1920. In 2017, they published a report on data sharing in public health called *A Guide to Sharing the Data and Benefits of Public Health Surveillance*, that has a stated goal of contributing to UN SDG-3, to *ensure healthy lives and promote well-being for all at all ages* (UN Department of Economic and Social Affairs, n.d.). While the project website was not live during the course of this report, the Internet Archive Wayback Machine had an archived snapshot that was used (Jan 9, 2021).

The report developed seven principles of data sharing, and the website provided a model data sharing agreement that could be adopted with additional resource templates for developing data management plans, IP rights, benefit sharing plans, and a form for initiating data sharing.

Principles:

- Building Trust
- Articulating the Value
- Planning for Data Sharing
- Achieving Quality Data
- Understanding the Legal Context
- Creating Data Sharing Agreements
- Monitoring and Evaluation

These seven principles set out a framework “to help create the right environment for data sharing and achieve good practice in data and benefits sharing” (Chatham House, 2017). The principle of building

trust involves two components: transparency towards the community where the data originates, as well as between stakeholders in an environment that has a clear purpose for sharing, where people know each other, have clear expectations, and follow through. The principle of articulating the value means both equitable sharing of benefits as well as upfront discussion about the purpose of sharing and potential value behind it. The principle of planning points out that sharing really is best done when there is a clear need at the outset. These three principles combine to demonstrate the importance of communication, stated purpose, and planning when deciding to share data.

The principle of quality data states that feedback mechanisms with shared data can help improve quality. However, the case study and follow-up questions head a slightly different direction, with using the Google Flu Trends dataset as an illustration of public health data that at first seemed to have predictive power, but over time fell short of other standard measures used by the CDC. The report takes this as an illustration of needing to also ensure that the data is of the quality needed to meet the purpose of sharing. This principle, along with the principle of monitoring and evaluation, prioritize continued understanding of the data sharing context to ensure outcomes are defined, recorded, and whether or not they are accomplished.

The final two principles, understanding the legal context and creating agreements, are about how formal the agreement needs to be given the context of the sharing taking place. An MOU or informal agreement is sometimes sufficient, and other times a formal agreement is needed. It is important to take all legal context into account as well when making plans to share data.

The Model Data Sharing Agreement has several sections:

1. Definitions
2. Purpose and Principles
3. The Agreement and Public Health Surveillance Data Sharing Activities
4. Responsibilities in Sharing Public Health Surveillance Data
5. Intellectual Property Rights Management Plan
6. Benefit Sharing Plan for Activities
7. General Provisions

The agreement defines two parties, the data provider and the data recipient, and the agreement states a general purpose of sharing the data to improve public health. The principles provided include discussion of sharing mechanisms that are compatible with laws and ethical principles, and authorized by both parties; involve timely sharing and interoperable formats; have transparent objectives of the sharing; promote cooperation, protect privacy interests; and are equitable and ensure benefits return to the communities where the data originated.

4 How to Share with a Data Sharing Agreement: Components

This section will discuss the required components of a data sharing agreement, building off the third policy example from Chatham House. Then, I turn to some frameworks for data sharing which will lead into next month's report on governance. Data sharing frameworks fit within a broader governance framework.

4.1 ODI Data Sharing Checklist

The ODI drafted a report with a checklist for Data Sharing Agreements. The report defines: “A data sharing agreement is an agreement between two or more organisations about how to share data. It will define what data is being shared and for how long, and any restrictions on its use.” The report also focuses on identifying the need for sharing prior to making any data sharing arrangement.

Finally, the checklist identifies the key components of a “good” Data Sharing Agreement:

- Context: Reasons for sharing & parties involved
- Data to be shared: Description of data
- Sharing process: How it will be accessed, transferred, shared, taking into consideration geographic/cross-border issues, storage, hosting, duration of sharing
- Use: What data can be used for
- Derived data: Who has rights to artifacts that incorporate the data
- Personal data: Set out specific clauses about personal data. Make sure compliance with legislation is there.

4.2 UK ICO Code of Practice

The UK ICO published a Code of Practice for Data Sharing which part of the Data Protection Act of 2018 (DPA). This is intended as a guide for non-open data, or data that is personal, and these practices help ensure that personal data is shared both in compliance with the DPA as well as “fairly and proportionately” (UK Information Commissioner, 2022).

The Code of Practice identifies the key components/questions that should be addressed in a Data Sharing Agreement (UK Information Commissioner, 2022, pp. 27–30):

- Who are the parties to the agreement
- What is the purpose of the data sharing initiative
- Which other organizations will be involved in the data sharing?
- Are we sharing data along with another controller?
- What data items are we going to share?
- What is our lawful basis for sharing?
- Is there any special category data, sensitive data or criminal offense data?
- What about access and individual rights?
- What information governance arrangements should we have?
- What further details should we include?

Here, *controller* is a data controller under the GDPR: “the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data” (see e.g. What Are ‘Controllers’ and ‘Processors’?, 2023).

The components here overlap closely with the ODI’s recommendations to provide information about the context, data, use, rights, and personal/sensitive data. It goes beyond the ODI’s recommendations to include information governance as a requirement.

4.3 ARDC on Data Sharing

The Australian Research Data Consortium (ARDC) also identifies the some “common components” of a successful data sharing agreement in their document, *1, 2, 3 of Data Sharing Agreements*. They acknowledge these are not minimum requirements, nor an exhaustive list (Australian Research Data Commons, 2023).

- Information about the data, parties and context of the agreement
 - Details of the data to be shared
 - Purpose of the data sharing
 - Period of the agreement
 - Parties to the agreement
 - Roles and responsibilities
- Conditions of Sharing
 - Permitted use of the data
 - Allowed or prohibited processing or analyses
 - Sensitivity of the data
 - Existing constraints on the data (legal & policy conditions)
 - Confidentiality
 - Derived data
 - Research outputs
 - Data security
 - Retention and disposal of the data
 - Changes and corrections
 - Reporting of incidental findings
 - Costs associated with the data sharing
 - Method of access
 - Format of the data
 - Timing of the data sharing
 - Support for the data sharing
 - Signoff for the DSA
- Other sections
 - Title
 - License, copyright, IP
 - Reporting requirements
 - Variation in the data
 - Risks
 - Termination

Again, there is heavy overlap in the components with the ARDC’s recommendations for common components and the ODI and UK ICO: parties, context, and purpose are at the forefront, and consideration of sensitive data and use are explicitly mentioned.

4.4 Contracts for Data Collaboration: Framework and Library of DSAs

A collaborative effort between NYU's [GovLab](#) (an NYU lab focused on governance research), [TReNDS](#) (Thematic Research Network on Data and Statistics), the University of Washington's Information Risk Research Initiative (in the UW Applied Physics Laboratory), and the [World Economic Forum](#), the Contracts for Data Collaboration project developed a library of sample DSAs, with a focus on agreements related to the UN SDGs (Contracts for Data Collaboration, n.d.). Their framework consists of six categories that need to be addressed with stakeholders when developing a DSA:

- "WHY is data being shared? What is the context and purpose?"
- WHAT kinds of data are being shared? What are the sources, formats, and other technical requirements?"
- WHO is party to the agreement? Who will be providing and using data resources? Are there any other 'third parties' that are also involved? Who has certain rights and duties?"
- HOW are data being shared? How is the relationship managed? How will issues such as security, privacy, and risk be handled?"
- WHEN will data actions take place? At what point does the agreement start and end?"
- WHERE are data being shared to and from? Are there jurisdictional issues to consider? Are there any international laws that apply?"

(Orrell & Dahmm, 2019)

Additionally, the Contracts for Data Collaboration project developed a library of 43 example agreements that are tagged and can be sorted by shareholder questions, identifying the relevant clause in the agreement text. For example, the question *How are proprietary issues handled?* returns clauses for the set of agreements, including clause 4 from the Agreement in Regards to the Management of Meteorological Networks in the Province of British Columbia:

"4.1 The Parties agree that: 4.1.1 Each Party owns all intellectual property, including copyright, in its own Source Material, other than intellectual property in Source Material which is owned by a third party (who has not entered into this Agreement);"

(Contracts for Data Collaboration, n.d.)

4.5 Synthesis of DSA Components

Across these three descriptions of best practices for data sharing agreements, there is consensus that a DSA should include **who** is involved in the sharing, **why** data is being shared, **what** the data is that will be shared, and involve a set of conditions on the sharing that take into account both the legal and policy context(s), privacy contexts, and sensitivity of the data as well as **what the data will be used for**. Broader compliance with governance, policy, and legal frameworks should all be a part of the agreement.

5 How to Consider Sharing: Frameworks and Principles for Data Sharing

5.1 Five Safes Framework

The Five Safes Framework is a straightforward set of safety considerations when making decisions about sharing data (Australian Bureau of Statistics, 2021; “Five Safes,” 2022; *The Five Safes*, n.d.; UK Data Service, n.d.). There are five dimensions defined, which the Australian Bureau of Statistics calls “a series of adjustable levers or controls to effectively manage risk and maximise the usefulness of a data release” (Australian Bureau of Statistics, 2021). Like the UK ICO’s data sharing agreement components, these are framed as a set of five questions:

- “Safe data: Has appropriate and sufficient protection been applied to the data?”
- Safe projects: Is the data to be used for an appropriate purpose?
- Safe people: Is the researcher appropriately authorised to access and use the data?
- Safe settings: Does the access environment prevent unauthorised use?
- Safe outputs: Are the statistical results non-disclosive?”

(Australian Bureau of Statistics, 2021)

5.2 The Trust Code

The TRUST Code is a code of conduct for research partnerships globally. It recognizes the unethical practice of ‘ethics-dumping’, which involves exporting unethical research practices into less regulated settings. The TRUST Code was developed to counter this practice as well as the practice of ‘helicopter research’, where researchers come into communities, extract data, then leave (TRUST, 2018). The key principles of the TRUST Code are:

- Fairness
- Respect
- Care
- Honesty

Researchers adhering to this code of conduct commit to practices that involve local communities and local researchers throughout the research process in a collaborative manner, and include making research results accessible and free to local communities and local researchers. Data ownership is stated in Article 4: “Local researchers should be included, wherever possible, throughout the research process, including in study design, study implementation, data ownership, intellectual property and authorship of publications.” Article 14 prevents research from being conducted in a lower-income setting when it could not be carried out in a higher-income setting (with exceptions around research on e.g. diseases not prevalent in higher-income settings which is justified and agreed upon by local communities). Utilization of local consent structures and local ethics review is also recognized, as is tailoring of informed consent procedures.

5.3 GovLab - Trust and Collaboration

The GovLab has a framework called the Data Responsibility Journey, which identifies two components of (Data) Sharing: “The Sharing stage involves cross-organizational coordination, trust-building, knowledge transfer, and other partnership considerations” (GovLab, n.d.). The questionnaire tool for

data sharing provides a set of questions for organizations to utilize that identifies both the relevant stakeholders and the topics relevant to trust and coordination. Trust-building questions involve discussions of reputational risk/harm, checking for consensus on policies, checking on data privacy, and asking about communications strategies for reporting back on benefits to the public. Collaboration-building questions involve discussions that compare operational differences and communications/assessments, interoperability, identifying additional needed expertise, and thinking about unintended consequences in advance (GovLab, n.d.).

6 Three Current Moments related to Data Sharing

These three moments illustrate some of the current challenges and global momentum impacting data sharing. First, there is international recognition that guardrails are needed for AI systems. Second, there is a need to balance data sharing in this context, while still recognizing the rights of individuals to be able to manage their data. Third, there is a critical role of an unbiased intermediary, or steward, in safely sharing data.

6.1 AI Executive Order & G7 Code of Conduct

During the writing of this report, the G7 announced on October 30 that they have agreed on a code of conduct for AI companies developing technology, called the Hiroshima Process (started in May 2023 at the G7 summit). Also on October 30, the U.S. President Biden also issued an executive order around AI standards and safety:

- [Hiroshima Process International Code of Conduct for Advanced AI Systems](#)
- [Hiroshima Process International Guiding Principles for Advanced AI system](#)
- [FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence | The White House](#)

6.2 Creative Commons: from Open sharing to Better sharing

Creative Commons' Strategic Plan from 2021-2025 acknowledges that they will expand their focus "beyond copyright licensing" and recognizes that sharing is closely intertwined with economic and ethical concerns: "Indeed, benefits of sharing can be undermined by exploitative practices that threaten financial stability of open endeavors, leading to financial hardship" (Creative Commons, 2020). The example they provide is the use of open content being used to train AI technologies which could cause harm. Their strategy is growing towards "better sharing", beyond "more sharing" and promote sharing "in the public interest." Part of their strategy is to develop a policy agenda that supports sharing in the public interest.

As part of this, CC is advocating for data sharing in four areas:

1. "Neither copyright nor related or *sui generis* rights should be used to raise barriers to the reuse and sharing of data."
2. "People should be able to use, reuse, and share the data they generate when engaging with digital services"
3. "Interoperability can help ensure individuals can access and share their data."

4. “The public sector has a critical role to play in facilitating data uses for public purposes”
(Creative Commons, 2022)

These principles address Creative Commons’s concerns around *better sharing* of data. Their policy advocates for reducing barriers to sharing to prevent legal barriers over sharing facts. They are working to encourage better sharing through data portability and interoperability, to prevent people from being locked into services and creating barriers to new creative uses of data or making data use impossible or impractical. They also advocate for the public sector as playing a key role in opening data.

6.3 Data Economy Lab - Data Sharing and Data Stewardship

The Data Economy Lab at the Aapti Institute recognizes that we are also at a “global flashpoint” around data sharing, and in line with CC recognizes that open data sharing is “insufficient” as not all data can be shared openly. This report, drafted in July 2020, tries to reconcile the numerous methods and models for sharing data with some theoretical principles grounded in models of data stewardship, exploring concepts of data rights and how this can be implemented in models of governance. The report suggests that data sharing and usage be monitored and managed by “the employment of a non-interested party, a steward” (Manohar, 2020).

7 Open Questions and November

This report covers quite a bit of ground, from

- What open licenses are commonly used for data
- Emerging discussions of data licensing for use in AI systems
- What it means to be ‘open’ in 2023, from the Open Source Initiative, the Open Knowledge Foundation, Creative Commons, among others
- Examples of recent Data Sharing Policies and Principles
- Components of Data Sharing Agreements and minimum requirements
- Principles/Frameworks to consider for best practices in sharing: keeping people safe, building trust, and collaborating well

In September, I explored models that would increase participation and empowerment of people around their data, with examples of coop and databank models in practice and an exploration of what kinds of agreements communities enter into when sharing their data. In October, I turned to some of the available mechanisms for data sharing for both open and non-open data, which is currently overshadowed by fast-moving discussions of what it means to be ‘open’ in AI and how to respond to data being used in new AI technologies. In November, I plan to return to data governance models to investigate how new systems are being developed to steward and share data that may increase both participation and empowerment (better stewardship) as well as make sure that data sharing is enabled and maximized while protecting and empowering people who are represented in that data (better sharing - a la Creative Commons).

Research questions include:

- What is data governance?
- What are some newer models of governance, e.g., data coops and data trusts?
- Do/How do data governance models encourage participation and preserve individual rights?
- How is data sharing implemented in these models of data governance?
- How is decision-making and concepts such as *who decides who decides* considered and (to what extent) implemented in these models of data governance?

This work is licensed under CC BY 4.0. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

This work is supported by a subaward from OpenTEAM as an initiative of Wolfe’s Neck Center for Agriculture and the Environment, specifically funded by the U.S. Department of Agriculture under agreement number NR233A750004G032. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of any funder. In addition, any reference to specific brands or types of products or services does not constitute or imply an endorsement.

References

- Allen Institute for AI. (n.d.). AI2 ImpACT Licenses. Retrieved October 30, 2023, from <https://allenai.org/impact-license>
- Arrazolo, S. (2017). Absence of Open Data Licensing in Open Data Efforts—Notes—Dataset by sya. Data.World. <https://data.world/sya/absence-of-open-data-licensing-in-open-data-efforts-notes>
- Australian Bureau of Statistics. (2021, August 11). Five Safes framework. <https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework>
- Australian Research Data Commons. (2023). Data sharing agreement development guidelines. <https://doi.org/10.5281/zenodo.7553198>
- Ávila, R. (2023, March 16). Updating the Open Definition to meet the challenges of today. Open Knowledge Foundation Blog. <https://blog.okfn.org/2023/03/16/updating-the-open-definition-to-meet-the-challenges-of-today/>
- Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., & Shee, A. (2019). Towards Standardization of Data Licenses: The Montreal Data License (arXiv:1903.12262). arXiv. <http://arxiv.org/abs/1903.12262>
- Carroll, M. W. (2015). Sharing Research Data and Intellectual Property Law: A Primer. PLOS Biology, 13(8), e1002235. <https://doi.org/10.1371/journal.pbio.1002235>
- Chatham House. (2017). A Guide to Sharing the Data and Benefits of Public Health Surveillance (p. 40). Chatham House, The Royal Institute of International Affairs. <https://chathamhouse.soutron.net/Portal/Public/en-GB/RecordView/Index/169144>
- Contractor, D., McDuff, D., Haines, J. K., Lee, J., Hines, C., Hecht, B., Vincent, N., & Li, H. (2022). Behavioral Use Licensing for Responsible AI. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 778–788. <https://doi.org/10.1145/3531146.3533143>
- Contracts for Data Collaboration. (n.d.). Library | Contracts for Data Collaboration. Retrieved November 21, 2023, from <https://contractsfordatacollaboration.org/library/>
- Creative Commons. (n.d.-a). About CC Licenses. Creative Commons. Retrieved October 30, 2023, from <https://creativecommons.org/share-your-work/cclicenses/>

Creative Commons. (n.d.-b). Frequently Asked Questions. Retrieved November 6, 2023, from <https://creativecommons.org/faq/>

Creative Commons. (n.d.-c). Open Data. Creative Commons. Retrieved October 10, 2023, from <https://creativecommons.org/about/open-data/>

Creative Commons. (2020). Creative Commons Strategy 2021-2025. https://drive.google.com/file/d/10rQDv5Hzuss38oi1ovGuoxHagmFzqn_f/view?usp=embed_facebook

Creative Commons. (2022, December 9). Our Work in Policy at CC: Data. Creative Commons. <https://creativecommons.org/2022/12/09/our-work-in-policy-at-cc-data/>

Data—Creative Commons. (n.d.). Retrieved October 10, 2023, from <https://wiki.creativecommons.org/wiki/data>

Digital Public Goods Alliance. (n.d.). Digital Public Goods Standard. Retrieved October 10, 2023, from <https://digitalpublicgoods.net/standard/>

Digital Public Goods Alliance. (2022, September 20). Approved Licenses for Digital Public Goods. GitHub. <https://github.com/DPGAlliance/publicgoods-candidates/blob/main/help-center/licenses.md>

Dumas, J. (2023, July 27). The AI2 ImpACT License—A new way to think about AI licensing. Medium. <https://blog.allenai.org/the-ai2-impact-license-a-new-way-to-think-about-ai-licensing-bc90ff26a9ee>

FAQ. (n.d.). Responsible AI Licenses (RAIL). Retrieved October 30, 2023, from <https://www.licenses.ai/faq-2>

Five safes. (2022). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Five_safes&oldid=1117226298

GovLab. (n.d.). Data Responsibility Journey—Sharing. Retrieved October 20, 2023, from <https://dataresponsibilityjourney.org/sharing>

Growth and Adoption of RAIL Licenses. (n.d.). Responsible AI Licenses (RAIL). Retrieved October 30, 2023, from <https://www.licenses.ai/license-adoption>

Manohar, S. (2020, July 31). Data Sharing for Public Good: Theoretical Bases and Policy Tools. The Data Economy Lab. <https://thedataeconomylab.com/2020/07/31/data-sharing-for-public-good-theoretical-bases-and-policy-tools/>

National Institute of Health. (2023a). Data Management and Sharing Policy: Frequently Asked Questions. <https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm?anchor=56545>

National Institute of Health. (2023b). NOT-OD-21-013: Final NIH Policy for Data Management and Sharing. <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>

Open Data Institute. (2022). Health data governance: A playbook for non-technical leaders. <https://open-data-institute.gitbook.io/data-governance-playbook/>

Open Data Institute. (2020, September 26). The Data Spectrum. The ODI. <https://theodi.org/insights/tools/the-data-spectrum/>

Open Knowledge Foundation. (n.d.-a). About. Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/about/>

Open Knowledge Foundation. (n.d.-b). Frequently Asked Questions. Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/faq/>

Open Knowledge Foundation. (n.d.-c). Licenses. Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/licenses/>

Open Knowledge Foundation. (n.d.-d). ODC Attribution-Sharealike Community Norms. Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/norms/odc-by-sa/>

Open Knowledge Foundation. (n.d.-e). Open Data Commons Attribution License (ODC-By) Summary. Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/licenses/by/summary/>

Open Knowledge Foundation. (n.d.-f). Open Data Commons Open Database License (ODbL). Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/licenses/odbl/>

Open Knowledge Foundation. (n.d.-g). Open Data Commons Public Domain Dedication and License (PDDL). Open Data Commons: Legal Tools for Open Data. Retrieved October 30, 2023, from <https://opendatacommons.org/licenses/pddl/>

Open Knowledge Foundation. (2015). Open Definition 2.1. <https://opendefinition.org/od/2.1/en/>

Open Knowledge Foundation. (2023, October 10). Open Knowledge Foundation joins the Digital Public Goods Alliance. <https://blog.okfn.org/2023/10/10/open-knowledge-foundation-joins-the-digital-public-goods-alliance/>

Open Source Initiative. (2023, August 25). Deep Dive: AI Webinar Series (2023). Open Source Initiative. <https://opensource.org/events/deep-dive-ai-webinar-series-2023/>

Orrell, T., & Dahmm, H. (2019, April 24). Partnerships founded on trust: Introducing Contracts for Data Collaboration (C4DC). TReNDS. <https://www.sdsntrends.org/research/2019/4/24/partnerships-trust-c4dc?locale=en>

Responsible AI Licenses (RAIL). (n.d.). About. Responsible AI Licenses (RAIL). Retrieved October 30, 2023, from <https://www.licenses.ai/about>

The Five Safes. (n.d.). Retrieved October 31, 2023, from <http://www.fivesafes.org/>

TRUST. (2018). The TRUST code -A Global Code of Conduct for Equitable Research Partnerships. <https://doi.org/10.48508/GCC/2018.05>

UK Data Service. (n.d.). What is the Five Safes framework? UK Data Service. Retrieved October 20, 2023, from <https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/>

UK Information Commissioner. (2022). Data sharing: A code of practice (v1.0.31). UK Information Commissioner. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/data-sharing-a-code-of-practice/>

UN Department of Economic and Social Affairs. (n.d.). Goal 3. UN Department of Economic and Social Affairs. Retrieved October 31, 2023, from <https://sdgs.un.org/goals/goal3>

Walsh, K. (2023, August 18). Understanding CC Licenses and Generative AI. Creative Commons. <https://creativecommons.org/2023/08/18/understanding-cc-licenses-and-generative-ai/>

Wellcome Trust Sanger Institute. (2023, March). Data Sharing Policy and Guidelines. <https://fairsharing.org/FAIRsharing.cg2j1y>

What are 'controllers' and 'processors'? (2023, September 29). ICO. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/controllers-and-processors/controllers-and-processors/what-are-controllers-and-processors/>