Data Commons

Frequently Asked Questions

May 11, 2026

Stefaan Verhulst
Andrew J. Zahuranec
Hannah Chafetz
Leona Verdadero
Jennifer Hansen

THEGOVLAB Open.Data.Policy.Lab Microsoft unesco

TABLE OF CONTENTS

INTRODUCTION

Why Data Commons Matter Now

One of the great paradoxes of our datafied era is that we live amid both unprecedented abundance and scarcity. Even as data grows more central to our ability to promote the public good, so too does it remain deeply—and perhaps increasingly—inaccessible and privately controlled. In response, there have been growing calls for data commons—pools of data that would be (self-)managed by distinctive communities or entities operating in the public's interest. These pools could then be made accessible and reused for the common good.

Data commons offer an alternative to the growing use of privatized data silos or extractive re-use of open datasets. They are a way of organizing data as a shared resource, governed collectively and used to advance public value. While the practice is still emerging and evidence remains limited, data commons could act as useful infrastructure for public interest AI by:

- Enabling equitable access to high-quality data
- Supporting AI systems grounded in diverse, representative information
- Creating institutional pathways for responsible data reuse
- Building trust through participation, transparency, and accountability

Data Commons do not emerge organically. They require intentional governance, stewardship, and social license. This FAQ provides a practical overview of what data commons are, why they matter, and how to design them responsibly.

Questions

What are data commons?

Data commons are collaboratively governed data ecosystems designed to pool and provide responsible (and governed) access to diverse, high-quality datasets from one or multiple sectors to enable the development and deployment of generative AI applications that address public-interest challenges.

This approach is distinct in that it treats data as a shared resource governed by a community rather than controlled solely as individual or institutional property. It further recognizes that open data, open by default and often fully public, does not, on its own, offer a complete solution. In an evolving data landscape, ensuring that openness supports public interest goals requires addressing how recognition and benefit flow, particularly where publicly available data can be widely reused without clear benefit returning to the communities that steward it.

In this reinvented data commons model, distinctive communities operating in the public interest define the terms by which data is collected, shared, and (re-)used. Access is shared, but only under the specific conditions set through governance structures that reflect collective priorities and constraints. Use is often purpose-bound, aligned with the needs and expectations of the contributing communities. Many data commons also incorporate benefit-sharing mechanisms to ensure that those who provide access to knowledge or materials share in the benefits and that participation does not reinforce existing inequities.

At its core, data commons are not just a dataset or a data platform. They are institutional arrangements that clearly define who has authority over data, can access data, under what conditions, for what purposes, with what safeguards and oversight, and for what expected outputs. This includes clear rules on access, mechanisms for oversight and enforcement, and structures for participation that extend beyond data contribution to include governance. They aim to shift the focus from data as an extractive resource toward data as a collectively stewarded asset for the public good.

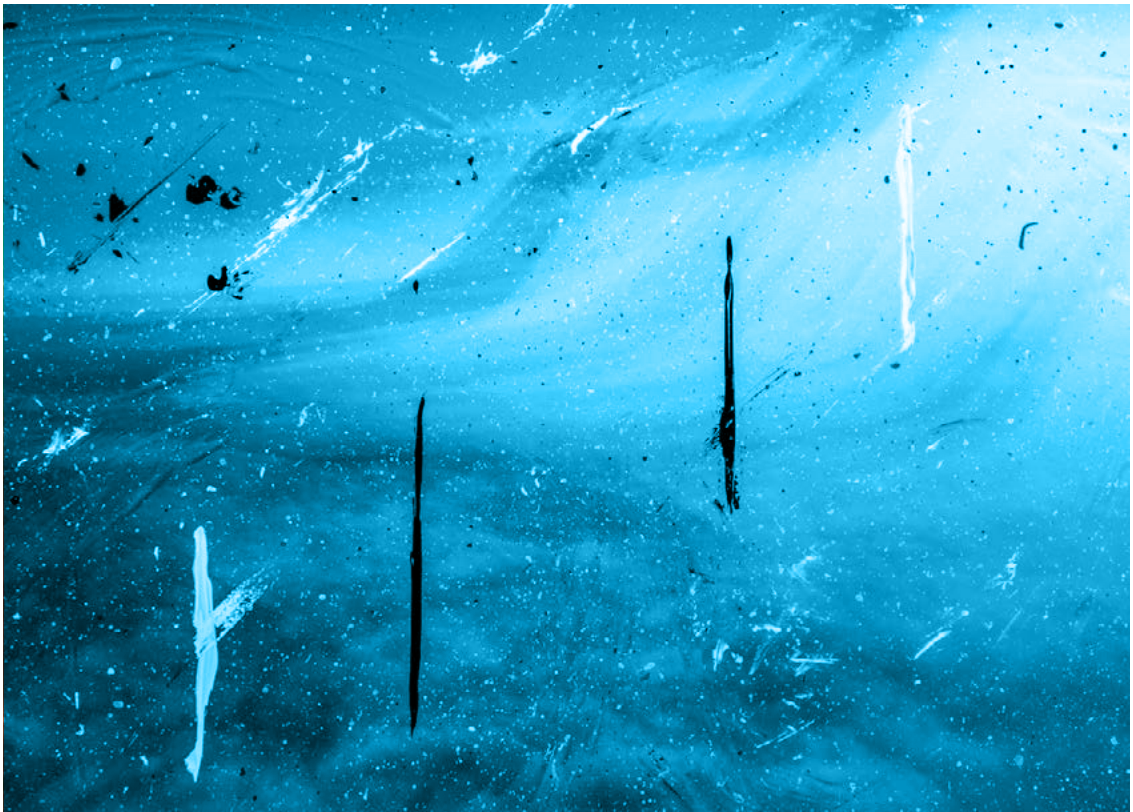
Where does the idea for a commons come from?

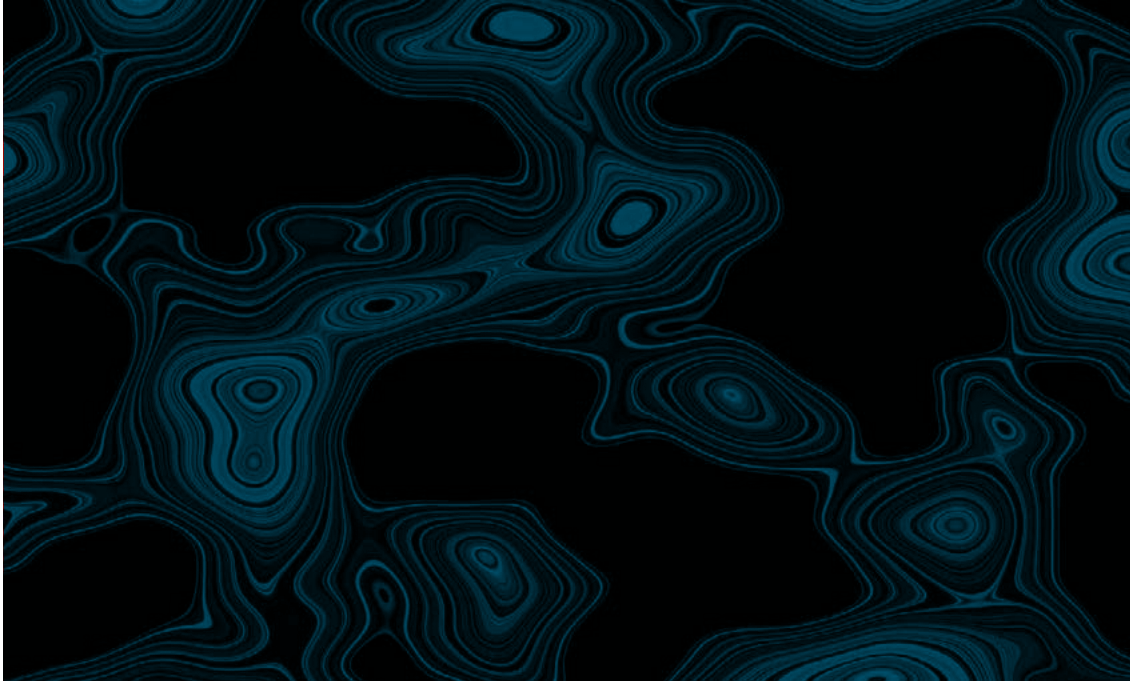
A commons is a shared resource held collectively by a community or user group, governed by rules dictating how it should be used. Commons tend to embody principles of local autonomy, mutual responsibility, and shared benefits. Typical examples of commons can include forests, fisheries, and groundwater.

The concept of the "commons" has existed for centuries but was popularized in the modern day in the context of Garrett Hardin's *The Tragedy of the Commons*. Building on arguments from the 19th century economist William Forester Lloyd, Hardin argued that individuals acting in their own self-interest behave contrary to the common good, that motivated by self-interest they are trapped in behavior that worsens the collective good.

The work of Lloyd and Hardin, which was particularly concerned with overpopulation and limited resources, was extensively refuted by Elinor Ostrom, Nobel Prize-winning economist and political scientist. Her work, particularly her book *Governing the Commons*, demonstrates that human beings are not trapped and helpless, that "tragedy" in managing shared resources is not inevitable but can be maintained through cooperation, monitoring, and enforcement of rules.

Though Ostrom had written about knowledge as a commons resource in the early 2000s, the concept of a "data commons" did not emerge until the 2010s. In the biomedical space and research space, data commons were viewed as a way of controlling scholarly data. In 2019, Director of the National Library of Medicine Patti Brennan described how data commons were emerging as an important way of sharing data as well as a "set of principles, governance strategies, and utilities". One recent attempt to apply this concept to data has been undertaken by the Mozilla Foundation.

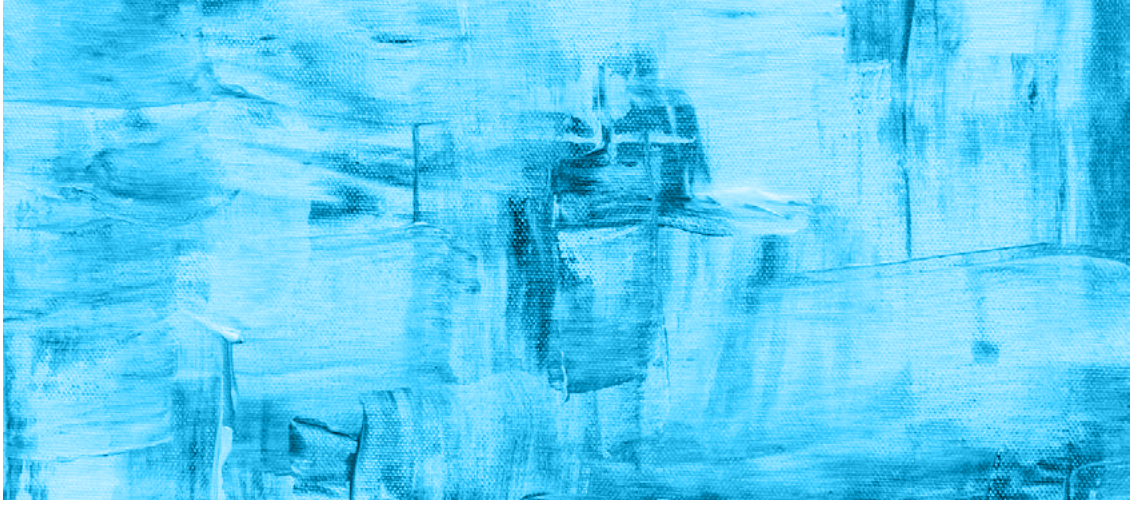




How are commons managed?

Commons can take many forms and be used by communities to manage many different kinds of shared resources. Our research follows Elinor Ostrom's *Eight Principles for Managing a Commons.* These are:

- **Clearly defined boundaries:** The identity of the group and the boundaries of the shared resource are clearly defined.
- **Congruence between local needs and conditions:** Members of the commons negotiate a system that rewards members for their contributions. Disproportionate benefits must be earned. Unfair inequality poisons the effort.
- **Collective decision-making arrangements:** Group members create the rules governing the commons and make decisions by consensus. Even if there are individual disagreements, all are aligned on the overall group goals.
- **Recognized rights to organize:** Groups have their own authority to conduct and oversee their own affairs. They are not constrained by externally imposed rules that cannot be adapted to local circumstances.
- **Effective monitoring:** To avoid free-riding or active exploitation, the commons monitors activities to ensure all members are abiding by the agreed-upon rules. This monitoring is ideally low cost and allows violations of norms to be identified quickly.
- **Graduated sanctions:** Violations of rules result in escalating consequences.
- **Accessible conflict resolution:** The group has agreed on mechanisms that allow disputes to be resolved quickly and in ways that are broadly perceived as fair by the group.
- **Structured nested enterprises:** Responsibility for governing the common resource has been built into nesting tiers. Resource management is handled at the lowest possible level by default but can be linked into a larger interconnected system.



Do all (data) commons fully embody Elinor Ostrom's eight principles?

Ostrom's eight principles can be applied to a variety of sectors. However, they frequently exist as an ideal. In practice, many initiatives fail to fully accomplish each of these standards, particularly when it comes to collective decision-making, community authority, and enforcement mechanisms.

Several commons designed for data operate as data-sharing infrastructures or collaborative platforms maintained by governments and other institutions who claim to be acting on the public's behalf and are not self-governing as understood by scholars. They may enable participation in contributing data and provide structured access but many extant examples stop short of fully granting communities control.

What defines a data commons, distinct from a commons?

Data commons exhibit the following set of commonly implemented characteristics:

Public Purpose: Aims to supply the data needed to solve public problems and develop public-interest AI.

Participatory Governance: Possesses a governance structure that offers meaningful avenues for commons members to exercise agency in how their data is used.

Accountability Mechanisms: Monitors the implementation of the commons' rules on an ongoing basis. Has systems in place on how to manage disputes and how to hold those accountable who do not follow the rules.

Contributor Benefits: Offers clear benefits to stakeholders who have contributed to the commons. This might include monetary benefits, early access to AI tools or research developed, attribution, insights for or about them, free membership, access to infrastructure to develop new AI tools, etc.

Access Controls: Includes clearly defined rules governing how, when, and by whom datasets can be accessed and used. This may involve standardized frameworks such as Creative Commons licenses, as well as tiered, restricted, or consent-based access models.

Funding and Support: The data commons is often funded as a public good, often with support of government, philanthropy, or via a membership/cooperative model.

How are data commons different from data collaboratives, trusts, spaces, and other arrangements?

The contemporary data landscape has seen a profusion of governance models. These concepts are often treated as competing alternatives, different pathways toward the same goal. In fact, such a framing obscures the reality that each model addresses distinct governance challenges and includes different models of authority, participation, value distribution, and risk. Seven approaches have gained particular prominence, each addressing a particular governance challenge facing data reuse partnerships:

- **Data intermediaries**, which act as brokers to reduce transaction costs and facilitate trusted exchange between data holders and users. These are particularly useful for addressing **coordination failures** by offering brokerage and transaction-cost reduction through negotiated rules and operational intermediation.
- **Data unions or coalitions**, which aggregate participants to strengthen collective bargaining power in data markets. These are useful for resolving **bargaining asymmetries** by rebalancing power and articulating terms of data use.
- **Data trusts**, which delegate decision-making authority to independent fiduciaries acting on behalf of defined beneficiaries. These trusts address **legitimacy deficits** by creating independent, fiduciary stewardship that ensures neutrality and credible commitments.
- **Data cooperatives**, which embed democratic ownership and member control directly into data governance arrangements. Cooperatives address **ownership inequality** by allowing for democratic ownership and giving members agency to redistribute control and value.
- **Data sandboxes**, which create controlled experimental environments to test governance rules and data uses before scaling. Data sandboxes help address **system uncertainty** by offering an environment to test governance, rules, safeguards, and institutional viability.
- Data spaces, which provide federated infrastructure and shared standards to enable interoperability across sectors or jurisdictions. Data spaces help address scaling complexity by offering systems for federated coordination and fostering interoperability across actors and jurisdictions.
- Data commons, which establish participatory governance structures for shared stewardship of pooled datasets. Data commons address collective governance needs by offering environments for participatory rule-making and shared stewardship for collective benefit.

A data commons may exist alongside or emerge from one of these other models (for example, a data commons may evolve from a sandbox, be facilitated by an intermediary, or scale into a data space), but they are distinct from them.

How is a data commons distinct from open data?

The data commons model aims to address the challenge of access asymmetries (detailed below) where organizations make their data widely accessible, resulting in it being scraped and extracted by others without any benefit in return.

While commons ***can*** make data open for general use, the structure enables their organizers to determine what data they make available, for what purpose, under what conditions. Instead of requiring a model of "open by default", data commons combine accessibility with governance.

They provide structured ways to determine and provide access to data—public, private, or hybrid—while specifying the conditions of use, obligations for benefit-sharing, mechanisms for oversight, and systems for attribution and compensation.

They seek to rethink openness in an era where context, social license, and control matter as much as availability.

What problems do data commons aim to solve?

Data commons aim to address several structural failures:

Access asymmetries: Many large organizations seek to use open data to extract value from others while limiting access to their own proprietary data. Data commons can introduce mechanisms through which communities can influence who accesses their data, how, and for what purpose.

Coordination failures: There are often high transaction costs for sharing data with one actor or on a one-off basis (e.g. formulating bespoke data-sharing agreements, standards). Data commons can create more standardized processes for sharing data under consistent rules.

Trust deficits: In many data-sharing arrangements, the interests and expectations of the public or contributing communities are not meaningfully incorporated. This erodes the social license and limits the legitimacy of any data (re-)use. Data commons try to center these concerns through governance, oversight, and access conditions.

Underutilization: A problem plaguing data re-use is that the most valuable datasets are not used for the public good but instead restricted. Data commons can expand access to a broader set of actors provided they commit to upholding the standards outlined by the organizers.

In these ways, data commons can point toward more sustainable and equitable data sharing relationships for the public good. They can address some of the concerns fueling data scarcity and the ongoing "data winter". They can recenter the conversation on AI to focus on public-interest AI and decision-making.

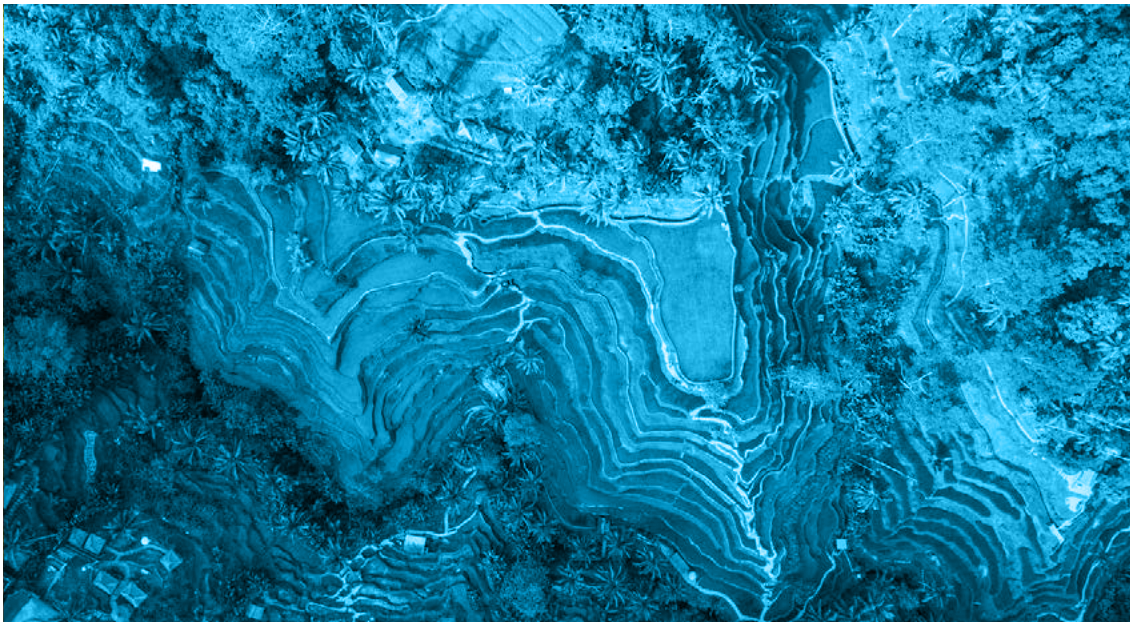
What are some of the critiques of data commons?

Critiques of data commons are important—and necessary. As Tommaso Fia and Gijs van Maanen, among others, argue certain discussions around data commons invoke the language of "community" without adequately specifying what community actually means, who defines it, who benefits, and who has the capacity to participate. They also warn against "commonswashing" and "community-washing," where the rhetoric of collective governance is used primarily for reputational purposes without meaningfully redistributing power or benefits.

These critiques should be taken seriously. There is indeed a risk that some data commons initiatives become overly technocratic, abstract, or solutionist—treating governance as a technical architecture problem while ignoring deeper political, institutional, and economic asymmetries. There is also a tendency in some policy discussions to produce long inventories of governance models (see above)—data trusts, data cooperatives, data commons, data spaces, intermediaries—as if they were interchangeable menu options, without sufficiently interrogating the values, institutional conditions, or power structures underpinning them.

At the same time, there is also a growing tendency among some critics to adopt a kind of governance purism: rejecting experimentation with institutional innovation altogether unless perfect conditions of equity, representation, or democratic legitimacy already exist. While the concerns motivating this caution are understandable, such an approach risks producing paralysis precisely at a moment when existing market and regulatory systems are demonstrably failing to govern data and AI in ways that produce shared societal benefits.

The reality is that many of the current harms associated with extractive AI development, and asymmetrical access to data are themselves the result of insufficient innovation in institutional governance. Simply relying on existing regulatory structures or market incentives will not be enough to address challenges around public-interest data access, benefit sharing, participatory governance, or collective agency in the AI era.



Data commons should therefore not be understood as a silver bullet or replacement for regulation. Nor should they be romanticized as inherently democratic or equitable. Rather, they should be seen as one governance instrument within a broader institutional toolbox; alongside public oversight, fiduciary mechanisms, participatory governance, and public infrastructure.

What matters is not whether an initiative labels itself a "data commons," but whether it meaningfully advances:

- collective agency over data governance;
- more equitable distribution of benefits;
- transparency and accountability;
- participatory decision-making;
- sustainable stewardship of shared resources; and
- mechanisms to counter extractive dynamics.

In practice, this means being extremely concrete about several questions:

- Does the community actually exist beyond rhetoric?
- Who defines membership and governance rules?
- What resources and capacities do participants have to govern effectively?
- Are there mechanisms to address existing asymmetries of power and expertise?
- Who captures value from the collaboration?
- What forms of accountability, contestation, and exit exist?
- Do the outcomes materially improve fairness, access, or public value?

Importantly, the answers to these questions will vary depending on context. A data commons built around Indigenous Data Sovereignty, for example, raises fundamentally different governance considerations than a mobility data commons for urban planning or a health data collaborative for research. As Fia and van Maanen emphasize, communities differ in their histories, identities, social ties, governance capacities, and political aspirations.

The goal, therefore, is not to impose a universal governance blueprint, but to develop governance arrangements that are context-sensitive, institutionally grounded, and capable of generating legitimate shared benefits. In many cases, this will require combining commons-based approaches with stronger regulation, public institutions, data stewardship functions, and participatory mechanisms.

Ultimately, the critique of data commons should not lead us to abandon experimentation with collective governance. It should push us to design these systems more rigorously, more transparently, and more politically honestly. The alternative—leaving data governance entirely to either markets or centralized state authority—has already shown its limitations.

Why is a social license critical for data commons?

A social license is an implicit societal acceptance of an activity based on trust, legitimacy, and perceived alignment with community values. It differs from open data licenses as it is not legal in nature but rather is an agreement or plan that is developed through stakeholder engagement and collaboration. In the context of data, it captures collective consent by focusing on the tacit endorsement of individuals and communities for the collection, use, and re-use of their data.

The concept of a social license acknowledges that legal compliance and consent are necessary—but not sufficient. While existing legal regimes seek individual agreement at the point of data collection, a social license seeks to create a process for communities to shape and oversee ongoing data reuse.

A social license ensures that data reflects collective interests. This, in turn, can allow data use to be perceived as legitimate and allow for processes that create ongoing accountability. Instead of one-time consent—a simple opt-in or opt-out process—social license encourages actors to think long-term, to periodically revisit needs and expectations to ensure that all activities are still broadly approved by community stakeholders.

A social license can be critical when dealing with significant power disparities between actors, disparities that create vulnerability or risk of exploitation. They can also be useful in directing efforts not toward short-term needs (e.g. does this actor approve of this work in the moment?) but toward long-term interests (e.g. what is needed to ensure continued involvement?).

What is the role of data stewardship in data commons?

Data stewards are responsible data leaders empowered by their organizations to create public value through cross-sector data exchanges. Data stewards act as the operational backbone of data commons—helping to facilitate and institutionalize the arrangement.

Typically, data stewards would be involved in supporting access and negotiating conditions under which data might be exchanged. They would also ensure compliance and ethical use, help communicate the interests of different data providers and users and work to maintain trust between parties over time.

Are there already data commons?

Data commons are an emerging practice. Relatively few existing initiatives fully meet the definition of collaboratively governed systems where communities exercise authority over data. Instead, current examples span a spectrum of models that incorporate some commons-like features, such as shared access, multi-stakeholder participation, or public-purpose orientation, without always implementing collective governance.

The initiatives below illustrate this range:

CLARIN: CLARIN provides the infrastructure to combine language data from across European institutions for research purposes. Governance is multi-stakeholder at the institutional level, but decision-making authority is exercised by member states and organizations rather than a broader contributing community.

Common Voice: Common Voice is a crowdsourced open voice dataset that can be used to train AI-driven voice applications. The initiative aims to broaden access to voice data for non-English languages and other groups typically underrepresented in voice datasets. The initiative is led by the Mozilla Foundation.

Lacuna Fund: The Lacuna Fund advances the development of labelled datasets about low and middle income countries that can be used for machine learning. The team accepts proposals for the creation of datasets across four domains: agriculture, language, health and climate data.

Language Data Commons of Australia: The Language Data Commons of Australia is a partnership between the Australian Research Data Commons and the School of Languages and Cultures at The University of Queensland and other partners (e.g. First Languages Australia) that seeks to make available Australian language data for both "academic and non-academic uses".

Querido Diário: Open Knowledge Brasil is seeking to develop a data commons for municipal official gazettes. Its Querido Diário platform aims to generate access to these datasets and help analyze information within them. They seek to improve decision making at the local level in Brazil.

Malawi Voice Data Commons: NYU Peace Research and Education Program's Malawi Voice Data Commons, developed in collaboration with Ushahidi, UNDP, and Mozilla Foundation, enables rural Malawians to report emergencies in native languages, creating multilingual, AI-ready datasets for humanitarian response and language preservation. The pilot will take place in Malawi with plans to scale across Sub-Saharan Africa.

What are the funding models used to support data commons?

Research on this area is still emerging. However, we have identified a few recurring funding structures:

Government-Led Models: Data commons are supported by government agencies, usually in partnership with other entities. These models benefit from long term funding commitments.

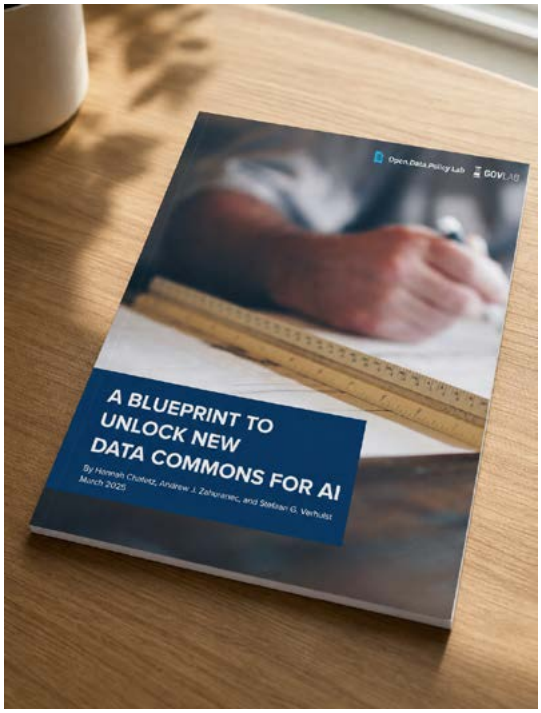
Philanthropy Models: Data commons are supported by donors from the private sector and civil society. Such commons blend donor capital with community participation, aligning with the indirect benefit model identified in the literature: the value lies in mission fulfillment and ecosystem engagement rather than monetization.

Membership and Cooperative Models: Users or contributors (including governments) become shareholders or members, effectively co-owning the commons. These models internalize governance and sustainability through collective ownership, echoing the freemium and razor-blade business archetypes where access is open but added services, tools, or governance privileges require contribution or fees.

Why is it important for artificial intelligence?

Data commons may be valuable in the age of artificial intelligence. They offer critical infrastructure for data access and re-use and can ensure that this use aligns with community interests. We describe this in our *Blueprint*, which notes that data commons can be useful for AI development in that they can:

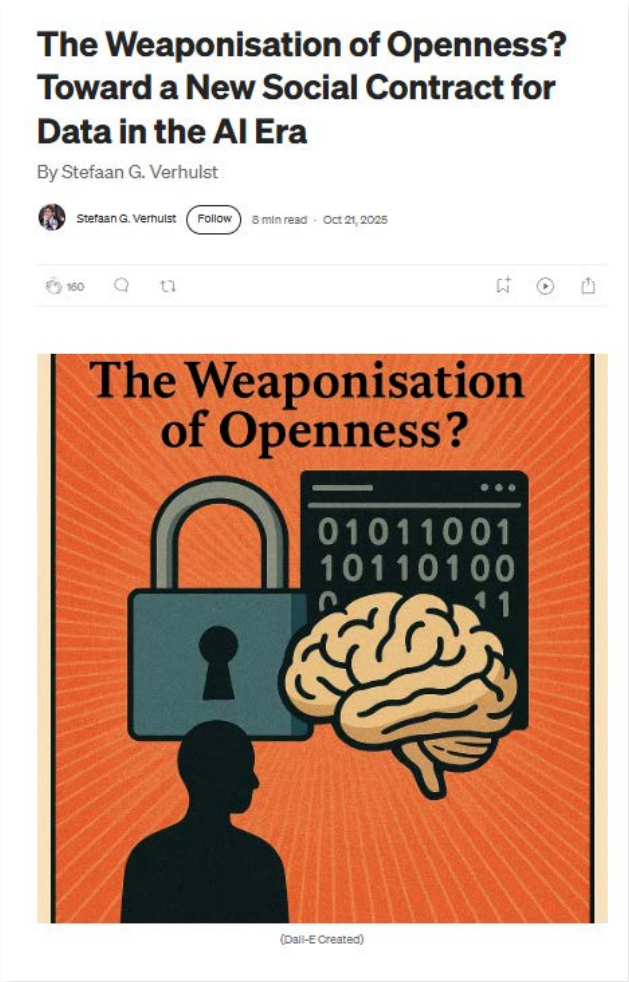
- Bring together disparate and varied datasets needed for AI development and deployment;
- Provide the infrastructure needed to standardize data in AI-ready formats;
- Lower standardization costs and help avoid duplication of effort;
- Increase the cultural diversity of datasets available;
- Operationalize new data re-use models for the common good;
- Establish data provenance mechanisms that improve transparency and traceability within AI systems; and
- Help developer communities build AI applications that address real needs.



Why is this model appropriate for marginalized groups?

A major problem with the development of AI and other data-driven tools is that they can be extractive. Data made open to serve a community is collected at mass scale to train proprietary systems, with little interest in recognizing or compensating those communities. The result is that a few large organizations gain further assets, increasing power asymmetries, and people have little control as information about them is used in ways that undermine their interests or go against their explicit desires.

We call this problem the *weaponization of openness*. It is a problem that is familiar to many Indigenous communities who have seen their traditional knowledge and biodiversity extracted by pharmaceutical firms without recognition or return.



Data commons offer a framework for combining accessibility with governance. They can provide structured ways to access data—whether public, private, or hybrid—while specifying conditions of use, obligations for benefit-sharing, mechanisms for oversight, and systems for attribution and compensation. This can ensure that data is used in ways that comport to a community's expectations, needs, and interests.

However, it is not a cure-all. It needs to exist within an environment where a community is well-defined, has resources that can be brought to bear to protect its interests, and where consensus is possible. Data commons may benefit from larger regulatory environments specifically designed to preserve group interests (e.g. treaty mechanisms for Indigenous communities) against larger, extractive entities.

Additional Readings

Chafetz, Hannah, Andrew J. Zahuranec, and Stefaan Verhulst. 2025a. "10 Data Commons for Cultural Knowledge and Preservation." *10 Data Commons for Cultural Knowledge and Preservation*, November 24. https://opendatapolicylab.org//articles/blog-post-10-data-commons-for-cultural-knowledge-and-preservation/.

Chafetz, Hannah, Andrew J. Zahuranec, and Stefaan Verhulst. 2025b. *A Blueprint to Unlock New Data Commons for AI.* The Open Data Policy Lab. https://incubator.opendatapolicylab.org/files/data-commons-for-ai-blueprint.pdf.

Chafetz, Hannah, Andrew J. Zahuranec, and Stefaan Verhulst. 2025c. "Appendix B: Taxonomy of Data Commons Use Cases for AI." Brooklyn, New York, March. https://incubator.opendatapolicylab.org/files/appendix%20b.pdf.

NYU Tandon. 2025. "The New Commons Challenge Proves the Power of Data Collectives | NYU Tandon School of Engineering." Brooklyn, New York, October 20. https://engineering.nyu.edu/news/new-commons-challenge-proves-power-data-collectives.

Verhulst, Stefaan, Hannah Chafetz, and Andrew Zahuranec. 2025. "Emerging Funding and Business Models for Data Commons: A Comparison." *Open Data Policy Lab*, October 29. https://opendatapolicylab.org//articles/blog-post-emerging-funding-and-business-models-for-data-commons-a-comparison/.

Verhulst, Stefaan, Hannah Chafetz, and Andrew J. Zahuranec. 2024. "Data Commons: Under Threat by or The Solution for a Generative AI Era? Rethinking Data Access and Re-Use." SSRN Scholarly Paper No. 4836354. Social Science Research Network, May 21. https://doi.org/10.2139/ssrn.4836354.

Verhulst, Stefaan, Andrew J. Zahuranec, and Adam Zable. 2025. "Participatory Approaches to Responsible Data Reuse and Establishing a Social License." *In Global Public Goods Communication: Mapping Actors, Policies, and Narratives*, edited by Sónia Pedro Sebastião and Anne-Marie Cotton. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-90667-1_10.

Verhulst, Stefaan, Andrew J. Zahuranec, Adam Zable, and Peter Addo. 2025. "Reimagining Data Governance for AI: Operationalizing a Social License for Data Reuse." SSRN Scholarly Paper No. 5255677. Social Science Research Network, April 28. https://doi.org/10.2139/ssrn.5255677.

Verhulst, Stefaan, Burton Davis, and Andrew Schroeder. 2025. "Data Commons: The Missing Infrastructure for Public Interest Artificial Intelligence." *LinkedIn*, April 29. https://www.linkedin.com/pulse/data-commons-missing-infrastructure-public-interest-verhulst-phd-k8eec/.

Zahuranec, Andrew J., Hannah Chafetz, and Verhulst Stefaan. 2025. "Why Responsible Data Access Will Determine the Future of AI: The Increased Importance of Data Commons." *Open Data Policy Lab*, February 8. https://opendatapolicylab.org//articles/blog-post-why-responsible-data-access-will-determine-the-future-of-ai-the-increased-importance-of-data-commons/.

Data Commons

Frequently Asked Questions

THEGOVLAB Open.Data.Policy.Lab Microsoft unesco