Open.Data.Policy.Lab GOVLAB

A BLUEPRINT TO UNLOCK NEW DATA COMMONS FOR AI

By Hannah Chafetz, Andrew J. Zahuranec, and Stefaan G. Verhulst

March 2025

ABOUT THE OPEN DATA POLICY LAB

The Open Data Policy Lab is a collaboration between The Governance Lab and Microsoft that seeks to support decision-makers as they work toward accelerating the responsible re-use and sharing of open data for the benefit of society. Our mission is to make data collaboration more systematic, sustainable, and responsible. Central to our mission has been advancing the Fourth Wave of Open Data: making open data AI-ready—and conversational—and moving towards a data commons approach.

ABOUT THE GOVERNANCE LAB

The Governance Lab (The GovLab) is a global action research center that studies how institutions and individuals can use technology and science to make more effective, equitable, and legitimate decisions to solve public problems. Our Data Program is pioneering innovative ways that researchers, businesses, government, and international organizations can provide access and re-use data to address the world's most pressing policy challenges, including migration, climate change, pandemics, and systemic inequality. We look at how data and technology are tools that can allow people to build a better world.

Cover Photo by Daniel McCullough on Unsplash

ACKNOWLEDGEMENTS

We are very grateful to our peer reviewers: Alek Tarkowski, André Golliez, Andrew Schroeder, Anna Colom, Derek Slater, Gretchen Deo, Inna Kouper, Jordan Gimbel, Jordan Usdan, and Joe Massey.

Thank you to the Open Data Action Labs participants—those who provided reflections on the intersections of data commons and generative AI in a structured workshop format—for sharing your insights. Thanks as well to Adrienne Schmoeker and Abby Zahuranec from The GovLab Data Program team for your support.

Finally, we are grateful for the financial support and guidance received from Microsoft that allowed us to complete this work.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Today's rapidly evolving AI landscape means that it is critical to ensure data remains accessible for those using it for the public's benefit. Reliable and equitable access to data can accelerate the development of public-interest AI use cases. It can ensure AI addresses real stakeholder needs and produces high-quality outputs. Unfortunately, AI-ready data often isn't made available for public-interest applications.

Data commons offer a solution. These collaboratively governed data ecosystems pool and provide responsible access to diverse, high-quality datasets. They can group data assets from one or multiple sectors to enable generative AI applications for the public good.

This Blueprint guides organizations that have or steward data on how to create new data commons to enable public-interest AI use cases. This includes knowledge institutions (such as libraries, universities, and research centers) and other data holders (such as governments, non-profits).

It includes four modules that data commons organizers can use to launch public-interest data commons. Module A, Mapping the Demand and Supply, focuses on understanding why AI systems need data, what data can be made available to train, adapt, or augment AI, and what a viable data commons prototype might look like that incorporates stakeholder needs and values. Module B, Unlocking Participatory Governance, discusses how to co-design key aspects of the data commons with key stakeholders and document these aspects within a formal agreement. Module C, Building the Data Commons, aims to support setting up the data commons from a practical perspective and ensuring all stakeholders are incentivized to implement it. Module D, Assessing and Iterating, focuses on evaluating how the commons is working and iterating as needed.

These modules are supported by two supplementary taxonomies. First, the Taxonomy of Data Types provides a list of data types that can be valuable for public-interest generative AI use cases. Second, the Taxonomy of Use Cases outlines several public-interest generative AI applications that can be developed using a data commons approach along with the potential outcomes and stakeholders involved. These resources are not all inclusive. We recognize they may be harnessed in several ways throughout the data commons planning process.

CONTEXT

The AI ecosystem is rapidly evolving. Yet for most organizations that serve the public interest, data—the lifeblood of AI innovation—is increasingly scarce. They face a potential "data winter" where access to diverse, high-quality and representative data becomes severely restricted.

This does not need to be the case. By guaranteeing reliable access to data, we can ensure that our AI systems address the contexts that communities face. We can better find and act on opportunities to create AI systems that promote public health, science, linguistic diversity, and more. We can do this without compromising the accuracy or reliability of results. We can build AI systems that accommodate diverse public-interest applications.

Open data commons offer a solution. They hold valuable datasets, making them interoperable and accessible for ethical and effective use. Commons can enable public-interest AI use cases while also making AI systems more robust and equitable.

For instance, The Lacuna Fund is providing access to datasets to fuel machine learning applications in low- and middle-income countries, Mozilla's Common Voice is broadening access to voice data to support voice applications in underrepresented languages, and most recently Harvard Law School Library (with the support of Microsoft and OpenAI) has begun working to make data from knowledge institutions ready for AI through its Institutional Data Initiative. All these examples show the potential that open data commons offer.

WHY DATA COMMONS FOR GENERATIVE AI?

In the context of AI development, data commons provide critical infrastructure for data access and re-use. This value is recognized in the United Nations' Global Digital Compact, which suggests that "urgent investments are needed in 'data commons.'" It is also reflected in a recent report by Open Future (in collaboration with the Open Source Initiative) which recommends accelerating data commons to address current data governance challenges across the AI development landscape.

They serve this purpose because they enable select practitioners to engage with large datasets in new and interesting ways. Our review of existing data commons and discussions with industry experts surfaced several important roles that data commons can fill for generative AI development.

- Data commons can bring together disparate and varied datasets needed for AI development and deployment. Increasing access to high quality training data can help improve the quality of the generative AI output and generate more reliable results.

- Data commons can provide the infrastructure needed to standardize data in AI-ready formats. They can help ensure datasets have the necessary metadata and documentation for constructive uses across the AI lifecycle.
- Data commons can have financial benefits. Working with shared data can lower standardization costs and help avoid the duplication of efforts.
- Data commons can increase the cultural diversity of data included in AI systems–helping to reduce bias and distribute the benefits of AI more equally.
- Data commons can let communities decide what data they want to be visible for AI development and training. They can help assess whether it is appropriate to provide access to specific datasets for AI systems and ensure the data is managed responsibly.
- Data commons can operationalize new data re-use models for the common good. Data commons can set up mechanisms to guide data contributions to AI systems and increase the use of open data from official sources.
- Data commons can establish data provenance mechanisms that improve transparency and traceability within AI systems, which can help understand how specific datasets were used in AI applications.
- The developer community can use data commons to build AI applications that address actual needs—from enhancing education and learning in underserved geographies to developing scenarios for humanitarian response efforts.

ABOUT THE BLUEPRINT

Since Fall 2023, The Open Data Policy Lab (a collaboration between The GovLab and Microsoft) has researched the potential of new data commons for AI. This work builds on The GovLab's prior work to foster models to reuse data across sectors for the public good, most notably data collaboratives.

Drawing on what we've learned, this Blueprint provides guidance and resources to support organizations seeking to create data commons for AI in the public's interest.

Box 1. What we mean by "data commons for AI"

Data commons are collaboratively governed data ecosystems designed to pool and provide responsible 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.

While many types of organizations can set up a data commons for the public good, this Blueprint is most suited for organizations that have or steward data.[1] This includes, for instance, knowledge institutions (such as libraries, universities, research centers) and others that hold unique and high-quality information assets (such as governments, non-profits). As AI applications require vast high-quality data, having this supply is critical.

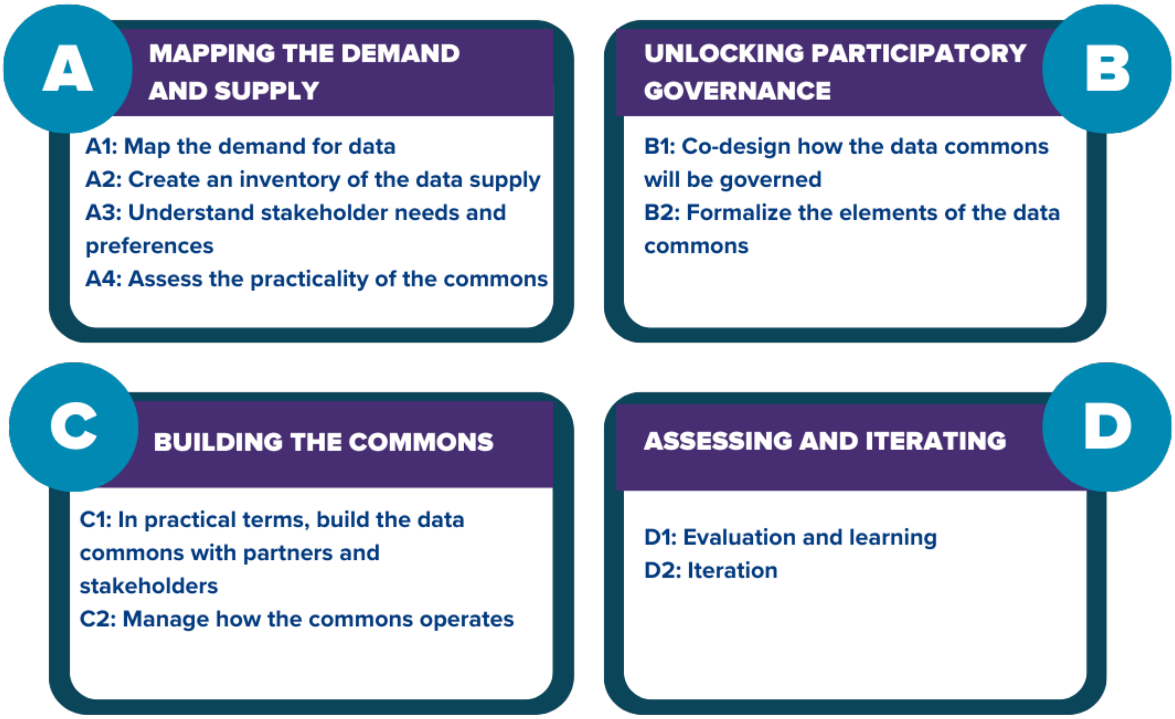
HOW TO USE THIS BLUEPRINT

The Blueprint is divided into four modules:

A. **Mapping the Demand and Supply:** Understanding why AI systems need data, what data can be made available to train, adapt, or augment AI, and what a viable data commons prototype might look like that incorporates stakeholder needs and values;
B. **Unlocking Participatory Governance:** Co-designing key aspects of the data commons with key stakeholders and documenting these aspects within a formal agreement**;**
C. **Building the Commons:** Establishing the data commons from a practical perspective and ensure all stakeholders are incentivized to implement it; and
D. **Assessing and Iterating:** Evaluating how the commons is working and iterating as needed.

Each of these modules includes guidance, resources, and questions that organizations can use when co-designing the main components of the data commons. These modules are not all-inclusive and may be harnessed in different ways throughout the process of creating a data commons. We hope this Blueprint inspires the development of accessible, equitable, and sustainable data commons that unlock AI's full potential.

[1] This includes organizations and institutions that are seeking to create public value through data collaboration or by providing access to datasets to specific groups. More information on the role of data stewards can be found in a blog by Stefaan Verhulst, *Wanted: Data Stewards — Drafting the Job Specs for A Re-imagined Data Stewardship Role*.



A MAPPING THE DEMAND AND SUPPLY

A1: Map the demand for data
A2: Create an inventory of the data supply
A3: Understand stakeholder needs and preferences
A4: Assess the practicality of the commons

B UNLOCKING PARTICIPATORY GOVERNANCE

B1: Co-design how the data commons will be governed
B2: Formalize the elements of the data commons

C BUILDING THE COMMONS

C1: In practical terms, build the data commons with partners and stakeholders
C2: Manage how the commons operates

D ASSESSING AND ITERATING

D1: Evaluation and learning
D2: Iteration

Figure 1. Blueprint Summary

THE BLUEPRINT

MODULE A: MAPPING THE DEMAND AND SUPPLY

Module A focuses on mapping the demand and supply of data. We begin by providing various resources to help data commons organizers understand the need for data and the types of public-interest AI use cases that can be operationalized using a data commons approach. We then discuss what it takes to develop an inventory of data that can be made available to train, adapt, or augment AI. We conclude by providing recommendations for creating a viable data commons prototype that incorporates stakeholder needs and values.

A1: MAP THE DEMAND FOR DATA

Understanding who is looking for data and for what purpose is an important first step in creating a data commons. The data commons organizers (or representatives of the organizations seek to set up a data commons) could start by mapping data holders (within and/or outside of their organization) who could provide access to data. They might also map possible data users (including data scientists, researchers, and public-interest AI developers) looking to access these shared resources to support AI applications. Actors mapping can be done using desk research, but it is also valuable to speak with data users themselves to ensure their needs are well accounted for.

The Open Data Policy Lab's *Taxonomy of Data Commons Use Cases for AI* (see Appendix) provides several examples of data points that can be useful for AI and made available through a data commons approach–including public domain books, speech data, and knowledge graphs. The taxonomy also outlines 10 use cases or public-interest AI applications that could be developed using a data commons. For instance, a data commons might support developing education and learning tools for underserved populations, modeling natural disaster scenarios that policymakers can use to develop climate policies, and creating synthetic data for training. When evaluating potential use cases, consider both the specific datasets that data users are looking for and the types of AI use cases that could benefit society.

Considerations:

- Who holds and is looking for data?
- Why are data scientists, researchers, and public-interest AI developers looking for data?
- What public-interest AI applications are needed most?
- What questions are data scientists, researchers, and public-interest AI developers looking to address?

Resources:

- The Open Data Policy Lab's *Taxonomy of Data Commons Use Cases for AI* (see *Appendix*) – Categories of data commons use cases and associated data types that can benefit society
- The Open Data Policy Lab's *The Emergent Landscape of Data Commons: A Brief Survey and Comparison of Existing Initiatives* – Lists several examples of existing data commons initiatives in the context of AI across domains
- The GovLab's *100 Questions Initiative* – A collection of data-actionable questions across 10 domains crowdsourced from data and subject experts
- The Open Data Policy Lab's *Problem Definition Tool* – Guidance and resources to help define problems that can be addressed using data
- The GovLab's *R-Search Methodology* – A methodology to quickly research complex problems
- The GovLab's (in partnership with the Inter-American Development Bank and the support of the French Development Agency) *Toward an Open Data Assessment and Segmentation Methodology* – A guide for organizations on how to engage with an intended audience on a public problem
- The Open Data Institute's *Mapping Data Ecosystem: Methodology* – A tool for mapping data stakeholders
- *The Data Provenance Initiative* – A collection of datasets used to train AI technologies

A2: CREATE AN INVENTORY OF THE DATA SUPPLY

To understand how the data commons could address the demand, first take an inventory of the specific datasets that data holders can make available and to whom. Unstructured data is particularly valuable for generative AI training and adaptation. Microsoft's *Government Data Commons: A Technical Guide* recommends reviewing the unstructured data associated with physical information (e.g. public domain books), records saved digitally that are not available online (e.g. data on government servers), as well as other "text, models, images, audio, and video".

While taking an inventory of the supply, the data commons organizers may also conduct a rapid search to identify if there is a data commons that currently exist for these data assets. This review is important to avoid the duplication of efforts.

After this check, the data commons organizers may start discussing the logistics of making these data sources AI-ready. For example, many unstructured data sources that could be valuable for AI are only available in analog formats (e.g. newspapers) and would need to be digitized and labeled. This labor would require additional time and resources.

One way to compare AI-readiness across data sources is to map ease of access with the potential value (drawing on the needs identified in the demand). Alternatively, organizations can assess the data assets with context-specific criteria.

Considerations:

- What data assets can be made available by whom and to whom?
- From a technical and logistical perspective, what is required to make this data available in the commons?
- Is it feasible to make this data available?
- Does a data commons for these data assets already exist? If so, what's missing?

Resources:

- Microsoft's *Government Data Commons: A Technical Guide* – Detailed guidance on the types of data needed for generative AI training and fine-tuning
- The Responsible Data for Children's *22 Questions* – A tool to interrogate what value a dataset can provide

- The Living Library's *Selected Readings on Crowdsourcing Data* – A curated and annotated collection of recommended works on the topic of crowdsourcing data
- Beeck Center's *Sharing Data for Social Impact: Guidebook to Establishing Responsible Governance Practices: Appendix* – A list of the steps involved in data sharing, considerations for each, and resources that might inform how an organization thinks about them.
- Stefaan Verhulst's *Data to Go: The Value of Data Portability as a Means to Data Liquidity* – An article explaining the importance of data liquidity, the condition in which data can be reused where and when it is needed.
- The Center for the Governance of Change, "*Data Collaboratives: Enabling a Healthy Data Economy Through Partnerships*" – A framework for responsible data sharing and reuse that could increase sharing between the public and private sectors to address some of the most entrenched asymmetries. It presents steps to overcome data asymmetries.

A3: UNDERSTAND STAKEHOLDER NEEDS AND PREFERENCES

Before settling on the data assets for the commons, the organizers should make sure all relevant stakeholder groups—including data subjects and data holders—are comfortable with making these data assets a shared resource. The organizers might identify stakeholders and their concerns by brainstorming stakeholder personas, including their values, priorities, attitudes around specific data assets and groups. These personas might also identify how they would benefit from a data commons for the data in question (beyond monetary benefits). The organizers may engage with stakeholders themselves or representatives of them to help build these personas.

During this step, organizers should assess when data should not be made a shared resource and with whom. It is vital to keep track of the types of actors that should not gain access to the commons and what data assets have the potential to cause harm. In doing so, organizations can improve trust and the well-being of partners.

Organizers can identify "red lines" around sharing by co-developing scenarios for when datasets should not be made available with expert representatives of stakeholder groups. Engaging with expert or technical representatives of stakeholder groups (as opposed to gathering large groups of stakeholders themselves) can help lower capacity/skills gaps and lessen the burden on stakeholders to understand and contribute to technical aspects of the commons.

Considerations:

- What are stakeholders' values, beliefs, expectations and attitudes around making datasets a shared resource? Why do they feel this way?
- Who are stakeholders comfortable with engaging with? Who are they not comfortable engaging with?
- What kinds of data are stakeholders comfortable with providing access and to whom?
- What data should not be made available in the commons?
- How will stakeholders benefit from the commons–beyond monetary benefits?
    - Gain early access to the AI tools developed?
    - Receive a free membership to AI tools that use their data?
    - Receive insights on data about or for them?
    - Gain access to the infrastructure to replicate AI tools for other contexts?

Resources:

- *Principles for Digital Development* – A compass outlining the obligations that organizations should uphold to promote sustainable and inclusive development in a development context
- Stefaan Verhulst's *Re-use of smart city data: The need to acquire a social license through data assemblies* – Testimony delivered to the New York City Council on the need for public input on data use
- The GovLab and Bertelsmann Foundation's *People-Led Innovation Matrix of People Engagement* – A checklist of stakeholders to identify with and engage in a data collaboration
- The GovLab's (with the support of the Henry Luce Foundation) *The Data Assembly* – An initiative to solicit diverse, actionable public input on data re-use for crisis response in the United States
- International Chamber of Commerce, *Policy primer on non-personal data*: This report gives insights into the different ways that the non-personal data has a positive impact on society, with a particular focus on private sector interests
- Pew Research Center, *How Americans View Data Privacy*: This 2023 poll series provides information on how US citizens feel about technology companies, data rights, and their expectations.
- Lea Skovgaard, Claus Thorn Esktrom, Mette N. Svendsen, and Klaus Hoeyer's *Survey of attitudes in a Danish public towards reuse of health data* – An article exploring the attitudes of Danish residents toward health data reuse.
- Datasphere Initiative's *Datasphere Governance Atlas* – A mapping of organizations working on data governance

A4: ASSESS THE PRACTICALITY OF THE COMMONS

As noted in Open Future's report, establishing equitable data commons requires focusing not only on providing access to data but also how it will be governed and maintained over time. Towards this end, this step focuses on developing a prototype of what the data commons concept might look like and assessing whether this prototype could be sustainable over time. Building on the learnings from the previous steps, data commons organizers might finalize the data assets and groups that will be involved in the commons–incorporating stakeholder needs and preferences.

From there, the organizers could brainstorm the different components of the commons prototype. This includes creating a theory of change, developing a governance strategy for how the data commons will be managed, understanding what funding and resources are needed to maintain the commons (and from whom), assessing the technological infrastructure to ensure secure access to the shared data resources, and the human resources required to manage these tasks.

After developing a prototype, one might conduct a cost-benefit analysis to determine whether this is a viable concept and could be sustained over time. We suggest engaging with partners during this process to ensure they are able and willing to pay the associated costs. If the costs are not feasible, the prototype will need to be revisited.

At this stage, it might be helpful to define how you will measure the success of the commons, such as through key performance indicators (KPIs). KPIs might include the number of data providers or users, inquiries to make data available, generative AI use cases, or model performance. They might also include domain specific indicators.

Considerations:

- What might a data commons concept look like?
  - Governance: How will the data commons be managed? What processes, standards and regulations are needed? Who will manage them?
  - Finances: How will the data commons be funded? What financial resources are needed to build and maintain the data commons? Who needs to supply these resources?

  - ○ Infrastructure: What technologies are needed to build the data commons itself? What technological infrastructure is needed to ensure there is secure access to data? Will this infrastructure be interoperable with other systems?
  - ○ Expertise: What kind of expertise is needed to set up and maintain the data commons? What human resources are needed?
- What are the goals of the data commons and how will you achieve them?
- What is the theory of change?
- How will you measure success?

Resources:

- Stefaan Verhulst and Sarah Telford's *A Framework for Understanding Data Risk* — A four-step approach for evaluating the risk of data
- Open Future's *Data Commons Primer* – A governance framework for data commons
- The Responsible Data for Children's *Opportunity and Risk Diagnostic Tool* – A tool for mapping reasons why one might engage with an organization
- The GovLab's *Leveraging Private Data for Public Good: A Descriptive Analysis and Typology of Existing Practices* – A report outlining the two dimensions of data collaboratives and the six models these collaboratives can take
- The GovLab's *Open Data Impact* — A report and collection of case studies assessing the impact of open data initiatives that indicate the value that data can provide for the public's benefit.
- United Nations Development Group's *Theory of Change: UNDAF Companion Guidance* – A methodology for developing a theory of change that can be applied to a data commons concept
- Dimitrios Sargiotis' book *Data Governance* – The chapter on "Legal and Regulatory Considerations in Data Governance" provides an overview of today's data sharing regulations
- Johan Linåker and Per Runeson, "*Sustaining Open Data as a Digital Common – Design principles for Common Pool Resources applied to Open Data Ecosystems*" – This paper examines how Ostrom's design principles for Common Pool Resources can be applied to open data ecosystems to help ensure their sustainability through effective governance and collaborative management.

MODULE B: UNLOCKING PARTICIPATORY GOVERNANCE

Centered around participatory governance principles, Module B supports data commons organizers in co-designing key aspects of the data commons with outside stakeholders This engagement ensures the data commons structure addresses partner needs and expectations. The latter part of this module outlines various resources to support data commons organizers when documenting these aspects within a formal agreement.

B1: CO-DESIGN HOW THE DATA COMMONS WILL BE GOVERNED

During this step, the data commons organizers work with data contributors and users to establish the specifics of how the data commons prototype will work and how it should be governed. These conversations center on how data users will access the data, how data providers will supply the data, the rules for the re-use of the data, how licensing and data use agreements will be drafted, and how all of the above will be financed. It may be helpful to involve data stewards in these conversations given their role of facilitating collaboration.

A key element of these discussions will be deciding on how to ensure access to data is equitable and secure. There are several models that can be harnessed for secure access. A tiered access model could be valuable in classifying different levels of users and what they can and cannot access. Data sandboxes could provide a secure place for the experimentation with data while minimizing the risk of unwanted access. Federated learning may be harnessed to streamline access. More on each of these models can be found in the resources provided below.

Considerations:

- How can the specifics of the commons be co-developed with stakeholders?
- How will data users and AI developers access the data in a way that is equitable? Should this involve tiered access, federated learning, commercial vs. non commercial access, sandboxes, or trustee model?
- What templated agreements are needed for data supply, data access, and data use?
- How will the data commons be funded? What funding models can be sustained over time (e.g. patronage, micropayments, membership fees, license and re-use fees, public-private partnerships, crowdfunding, or endowments)?
- What technical infrastructure is needed (e.g. cloud-based infrastructure, generative AI applications, automated quality assurance)?

- What are the accountability mechanisms? What happens if someone violates the rules of the commons?
- How long will this operate (e.g. for a set period, indefinitely, until a situation passes)?

Resources:

- The GovLab's *Urgent Need to Reimagine Consent* – Explains the concept of a "social license" and how it might be achieved through different types of community engagements
- Stefaan Verhulst's *Data Stewardship Decoded: Mapping Its Diverse Manifestations and Emerging Relevance at a time of AI* – Explains the main competencies of data stewards in modern data governance
- The Data Tank and French Development Agency's *Responsible Data Re-Use in Developing Countries: Social License Through Public Engagement* – A framework for establishing a social license
- The GovLab and Inter-American Development Bank's *Facilitating Data Flows Through Data Collaboratives* – A series of checklists on the different models and elements of data collaboratives
- Involve's *Methods* – a list of methods to support participatory decision making processes
- Ipsos, the Open Data Institute and Imperial College Health Partners' *NHS AI Lab Public Dialogue on Data Stewardship* – Outcomes from a public engagement on making decisions for data access
- The Living Library's *Selected Readings on Data Governance* – An annotated and curated collection of recommended works on data governance issues
- *Camden's Data Charter* – Example of data stewardship in practice
- *INSIGHT Health Data Research Hub* – Example of data stewardship in practice
- Peter Slattery, Alexander Saeri, and Peter Bragge's "*Research co-design in health: a rapid overview of reviews*": A rapid review of various research co-design approaches in health settings, including focus groups, interviews, and surveys.
- Aapti Institute's *Stewardship Mapper* – An interactive tool comprised of questions and considerations when developing a data governance strategy
- OECD's *Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches* – An overview of how decision makers can harness privacy enhancing technologies

B2: FORMALIZE THE ELEMENTS OF THE DATA COMMONS

Before moving to implementation, the data commons organizers should document each of the decisions made about the data commons governance structure. Organizers may document the conceptual framework through contracts or written agreements. This should only be done once there has been consensus among key stakeholder groups. In some cases, a legal team may be needed to document all decisions and put them into practice. This step is necessary to guarantee legitimacy and accountability around the initiative.

Considerations:

- Is there consensus?
- How can we put the aspects of the data commons into practice?
- Who needs to document these aspects?
- What should be included in the data sharing, data use, and data access agreements?

Resources:

- The Open Data Policy Labs' *Three Resources for Harnessing the Power of Data Sharing Agreements* – Tools and checklists for developing data sharing agreements
- *Joint Development Foundation* – provides the legal infrastructure for new collaborative data initiatives. For instance, its *Overture Maps*, which helps developers and others access open mapping data, used this infrastructure to scale up.
- The New Hanse's "*Governing Urban Data for the Public Interest*" – A report providing recommendations for European cities seeking to better govern and manage data in the public interest.
- Creative Commons' "*Recommended Best Practices for Better Sharing of Climate Data*" – Guidance from Creative Commons to help public climate data-producing institutions to choose the most suitable legal terms and licenses and use metadata.
- Transform Health's "*Health Data Governance Principles*" – A resource meant to help those involved in the collection and use of health data handle it more responsibly.

MODULE C: BUILDING THE COMMONS

Module C includes several considerations and resources that may be harnessed when moving from idea to practice. First we discuss how organizers might go about setting up the data commons from a practical perspective and the associated logistical challenges. We then provide considerations for managing the data commons and making sure all stakeholders are incentivized to participate.

C1: IN PRACTICAL TERMS, BUILD THE DATA COMMONS WITH PARTNERS AND STAKEHOLDERS

Once the lead organizers and partners reach consensus on the specifics of the data commons, it is time to start building the commons with key stakeholder groups. Before beginning implementation, consider defining the roles and responsibilities in managing the commons. This work includes identifying those who will serve as the commons' data stewards, manage participatory governance processes moving forward, handle data user requests, and hold stakeholders accountable if they violate the rules.

This step also includes managing the logistics of making the data available. This might involve setting up the technical infrastructure to host the data assets, digitizing and labeling the data available, establishing data visitation technical guardrails, and making sure the data and technologies are interoperable with other systems.

At this time, organizers can start generating awareness about the initiative with the broader open data community and other interested stakeholders. The goal is to increase excitement about the upcoming initiative as well as transparency around how decisions were made.

Considerations:

- What are the roles and responsibilities of the commons? Who is involved and how?
- What is needed to make the data available?
- What technical infrastructure is being harnessed?
- How will the data be made available?
- Who should know about and participate in the data commons?

Resources:

- The GovLab's *Data Stewards Job Description* – Sets out the specific of the data stewards role

- The Open Data Policy Lab's *A User's Guide to the Periodic Table of Open Data* – A visualization of the enabling conditions and disabling factors that often determine the impact of open data initiatives
- National Archives of Australia's *Data Interoperability Maturity Model* – A framework for evaluating data interoperability
- The United States Department of Commerce's *Generative Artificial Intelligence and Open Data: Guidelines and Best Practices* – Guidance on how to make open data AI-ready
- The Data and Trust Alliance's *Data Provenance Standards* – Metadata standards that can be applied for data across industries
- Timnit Gebru, et al. *Datasheets for Datasets* – An approach for documenting datasets for machine learning applications

C2: MANAGE HOW THE COMMONS OPERATES

Without robust accountability mechanisms, there is a substantial risk that those participating in the commons will lose trust in it, leading to its eventual termination. During this step, data commons representatives (e.g. data stewards, an ethics committee, or the organizers) monitor and enforce the rules of the commons.

What these rules are will vary based on decisions made during commons design. It may include ensuring only specific individuals or groups gain access to the data or monitoring the commons to verify that the data is re-used responsibly. It may also involve ensuring that the right individuals benefit from the commons.

Having a code of ethics can be helpful in navigating situations that may fall within a grey area. At the same time, organizers might look for ways to incentivize stakeholders to participate in the commons responsibly by promoting certain principles and benefits.

Considerations:

- Are the data users and contributors complying with the rules of the commons?
- Are stakeholders receiving the benefits of the commons?
- What does it mean to be ethical within the commons?
- Are stakeholders incentivized to participate in the commons? What other incentives are needed?

Resources:

- The GovLab's *Contracts for Data Collaboration* – A resource that provides a listing of data sharing agreement clauses that outline partner obligations to one another.

MODULE D: ASSESSING AND ITERATING

Module D provides various resources to help data commons organizers reflect on how the commons is working with key stakeholders and co-design actionable solutions. This module emphasizes the need to rapidly iterate to reflect stakeholder needs and the rapidly advancing AI ecosystem.

D1: EVALUATION AND LEARNING

After the organizers launch the data commons, its representatives should continually assess whether the approach is working and document what they learn. This may be done through regular gatherings or an online form.

This work is important for two reasons. First, it benefits the larger community of AI developers. The generative AI landscape is rapidly evolving and the intended uses and requirements for the data commons may change based on the latest AI advancements. Second, it benefits the organizers themselves by identifying why one approach works and another does not. During this process, participants may discover that the KPIs they developed in Module 1 are not as effective as they hoped and can make a plan to adjust as needed.

This step may also include evaluating the performance of the AI models developed using the data assets on the commons. Data users and contributors could review how the data supplied impacted model performance and where there is room for improvement.

Considerations:

- How often should partners be engaged?
- How is the data commons performing in terms of its KPIs?
- Has the data on the commons been sufficient to develop public-interest AI applications? What needs to be improved?
- What problems are emerging?

Resources:

- *Impact Assessment and Project Appraisal* – A journal collecting research on how to measure impact over time

- The European Commission's Digital Innovation Lab's *Data Innovation Toolkit* – A framework for managing public-sector data initiatives, including how to evaluate them and measure impact
- Griffith Centre for Systems Innovation's *Now we are all measuring impact — but is anything changing?* – A blog reflecting on the process of impact evaluation and the various approaches it can include.
- The Government of Canada's *Algorithmic Impact Assessment tool* – A risk assessment tool, taking the form of a questionnaire, that determines the impact of an automated decision system.
- data.europa.eu's *Rethinking the impact of open data: A first step towards a European impact assessment for open data* — A report that seeks to establish a methodology for open data impact assessments in a European context.
- *Data-Centric AI* – A series of resources and guidance on how to improve data processes for AI systems
- Sedir Mohammed, et al. *The Effects of Data Quality on Machine Learning Performance* – A paper exploring how several data quality indicators can impact model performance

D2: ITERATION

Lastly, representatives of the data commons should continue monitoring the commons and implement any necessary changes. Continued monitoring is key in making the data commons sustainable over time.

Considerations:

- What changes need to be made in order for the data commons to be sustainable?

Resources:

- *Agile Development* – A methodology to iterate projects based on stakeholder needs
- Sara Marcucci and Stefaan Verhulst, "*Reimagining the Policy Cycle in the Age of Artificial Intelligence*" – An article exploring AI's role across the policy cycle, including in policy implementation and impact assessment, where AI can enhance monitoring, evaluation, and adaptive decision-making.
- The GovLab's "*People-Led Innovation: Toward a Methodology for Solving Urban Problems in the 21st Century*" – A methodology that helps city leaders respond to urban challenges in an iterative manner by placing people, and groups of people at the center of all stages of the problem-solving process.

APPENDIX

APPENDIX A: BLUEPRINT WORKSHEETS

Below we provide a summary of the main decision points from each module of the Blueprint in a worksheet format. These worksheets aim to serve as checklists of elements to consider when setting up a data commons.

MODULE A: MAPPING THE SUPPLY AND DEMAND

Module A seeks to understand why AI systems need data, what data can be made available to train, adapt, or augment AI, and what a viable data commons prototype might look like that incorporates stakeholder needs and values.

A1: MAP THE DEMAND FOR DATA

Considerations	Notes
□ Who holds and is looking for data?	
□ Why are data scientists, researchers, and public-interest AI developers looking for data?	

□ What public-interest AI applications are needed most?	
□ What questions are data scientists, researchers, and public-interest AI developers looking to address?	

A2: CREATE AN INVENTORY OF THE DATA SUPPLY

Considerations	Notes
□ What data assets can be made available by whom and to whom?	

□ From a technical and logistical perspective, what is required to make this data available in the commons?	
□ Is it feasible to make this data available?	
□ Does a data commons for these data assets already exist? If so, what's missing?	

A3: UNDERSTAND STAKEHOLDER NEEDS AND PREFERENCES

Considerations	Notes
□ What are stakeholders' values, beliefs, expectations and attitudes around making datasets a shared resource? Why do they feel this way?	
□ Who are stakeholders comfortable with engaging with? Who are they not comfortable engaging with?	
□ What kinds of data are stakeholders comfortable with providing access and to whom?	

□ What data should not be made available in the commons	
□ How will stakeholders benefit from the commons–beyond monetary benefits? □ Gain early access to the AI tools developed? □ Receive a free membership to AI tools that use their data? □ Receive insights on data about or for them? □ Gain access to the infrastructure to replicate AI tools for other contexts?	

A4: ASSESS THE PRACTICALITY OF THE COMMONS

Considerations	Notes
□ What might a data commons concept look like? □ Governance: How will the data commons be managed? What processes, standards and regulations are needed? Who will manage them? □ Finances: How will the data commons be funded? What financial resources are needed to build and maintain the data commons? Who needs to supply these resources? □ Infrastructure: What technologies are needed to build the data commons itself? What technological infrastructure is needed to ensure there is secure access to data? Will this infrastructure be interoperable with other systems? □ Expertise: What kind of expertise is needed to set up and maintain the data commons? What human resources are needed?	
□ What are the goals of the data commons and how will you achieve them?	

□ What is the theory of change?	
□ How will you measure success?	

MODULE B: UNLOCKING PARTICIPATORY GOVERNANCE

Module B focuses on co-designing key aspects of the data commons with key stakeholders and documenting these aspects within a formal agreement.

B1: CO-DESIGN HOW THE DATA COMMONS WILL BE GOVERNED

Considerations	Notes
□ How can the specifics of the commons be co-developed with stakeholders?	
□ How will data users and AI developers access the data in a way that is equitable? Should this involve tiered access, federated learning, commercial vs. non commercial access, sandboxes, or trustee model?	
□ What templated agreements are needed for data supply, data access, and data use?	

□ How will the data commons be funded? What funding models can be sustained over time (e.g. patronage, micropayments, membership fees, license and re-use fees, public-private partnerships, crowdfunding, or endowments)?	
□ What technical infrastructure is needed (e.g. cloud-based infrastructure, generative AI applications, automated quality assurance)?	
□ What are the accountability mechanisms? What happens if someone violates the rules of the commons?	
□ How long will this operate (e.g. for a set period, indefinitely, until a situation passes)?	

B2: FORMALIZE THE ELEMENTS OF THE DATA COMMONS

Considerations	Notes
□ Is there consensus?	
□ How can we put the aspects of the data commons into practice?	
□ Who needs to document these aspects?	

□ What should be included in the data sharing, data use, and data access agreements?	

MODULE C. BUILDING THE COMMONS

Module C aims to establish the data commons from a practical perspective and ensure all stakeholders are incentivized to implement it.

C1: IN PRACTICAL TERMS, BUILD THE DATA COMMONS WITH PARTNERS AND STAKEHOLDERS

Considerations	Notes
□ What are the roles and responsibilities of the commons? Who is involved and how?	

□ What is needed to make the data available?	
□ What technical infrastructure is being harnessed?	
□ How will the data be made available?	
□ Who should know about and participate in the data commons?	

C2: MANAGE HOW THE COMMONS OPERATES

Considerations	Notes
□ Are the data users and contributors complying with the rules of the commons?	
□ Are stakeholders receiving the benefits of the commons?	
□ What does it mean to be ethical within the commons?	

□ Are stakeholders incentivized to participate in the commons? What other incentives are needed?	

MODULE D: ASSESSING AND ITERATING

Module D involves evaluating how the commons is working and iterating as needed.

D1: EVALUATION AND LEARNING

Considerations	Notes
□ How often should partners be engaged?	
□ How is the data commons performing in terms of its KPIs?	

□ Has the data on the commons been sufficient to develop public-interest AI applications? What needs to be improved?	
□ What problems are emerging?	

D2: ITERATION

Considerations	Notes
□ What changes need to be made in order for the data commons to be sustainable?	

APPENDIX B: TAXONOMY OF DATA COMMONS USE CASES FOR AI

CONTEXT

To organize the energy around data commons in a productive fashion, The Open Data Policy Lab conducted a series of structured analytic exercises to ideate potential use cases.

This effort began with a mapping exercise, identifying sectors where data commons for AI already exist and where there was substantial interest in securing more access to data. We relied on a search on the web (e.g. Google Scholar, news outlets, The Open Data Policy Lab's *Observatory of Generative AI and Open Data Use Cases*, and The GovLab's *Data Collaboratives Explorer*) and our network of experts (e.g. Open Data Action Labs). This review helped us understand current efforts across industries. We summarized this work within the blog: *The Emergent Landscape of Data Commons: A Brief Survey and Comparison of Existing Initiatives*.

We then conducted another mapping of data types that could be useful for different aspects of generative AI, drawing from Microsoft's *Government Data Commons: A Technical Guide*, submissions to the *United States Department of Commerce Request for Information on AI-Ready Open Government Data Assets*, The Data Provenance Initiative's *A*

Large Scale Audit of Dataset Licensing & Attribution in AI, the Open Data Institute's *Data for AI Taxonomy*, *Papers with Code*, and our network of experts.

We focused on data types that were cited as valuable for generative AI or where authors expressed there is a need for data. We chose not to include data types with clear ethical and data responsibility issues (data types such as electronic health records that have been widely criticized by the public for its use within AI systems). We also tried to avoid data types where extensive data commons already exist to avoid the duplication of efforts. Using this information, we developed a taxonomy to organize these data types.

At the same time, we conducted a "What if?" Analysis in which we asked ourselves to consider the types of use cases where new AI data commons could benefit society. Drawing on the information gathered thus far, we synthesized 10 categories of use cases for new data commons in the context of AI.

In what follows we provide a summary of these exercises. While these exercises primarily focus on generative AI, it is important to note that there is still a need for data commons for machine learning and neutral networks as well.

TAXONOMY OF DATA TYPES

Below we provide a list of data types that could be valuable for data commons for AI. These data types are divided into six main categories. Creating data commons that are truly fair and beneficial to all requires sourcing data responsibly. That being said, the data commons should only include data sources with appropriate licensing.

Category	Data Types (Exemplary)
1. **Text**	● Language and linguistics data: Language research, words, definitions, translations, and transcripts ● Legal data: Laws, legal proceedings, and court case decisions ● Research and education data: Exam questions and answers, educational blogs, homework questions and solutions, research reports, peer-reviewed journal articles (focusing on niche scientific subjects), grants, and patents ● News and media data: Press releases, blogs, newspapers and magazines, scripts, books (e.g. textbooks, cultural perspectives, encyclopedias, travel guides, and translations) ● Biomedical data: Notes and insurance policies ● Social media data: Social media posts, customer reviews, online petitions, search terms/keywords
2. **Audio**	● Sounds: Environmental sounds (urban and nature) and physiological sounds (e.g. breathing patterns) ● Music ● Speech Data: Conversations, expert testimonies, phrases, government proceedings (e.g. policy debates, consultations), storytelling data (e.g. audiobooks, recent radio recordings)
3. **Images**	● Knowledge graphs ● Spatial images: GIS/satellite data, maps ● Artwork (including photographs) ● Data visualizations: Charts, tables, infographics

4. Video	● Expert lectures ● Government videos: Policy debates ● Environmental videos: Urban videos (e.g. driving scenarios, traffic camera footage) and nature videos
5. Statistics	● National statistics ● Sensor networks ● Supply chain data: Manufacturing metrics ● Energy consumption data ● Math/physics proofs: 19th century physics proofs
6. Generative AI data	● Synthetically created datasets ● Data about generative AI providers: Models (weights, usage, performance), development processes ● Issues associated with generative AI: Unwanted outcomes, legal cases, other risks ● Best practices when using generative AI: Prompts, policies, strategies, and guidelines

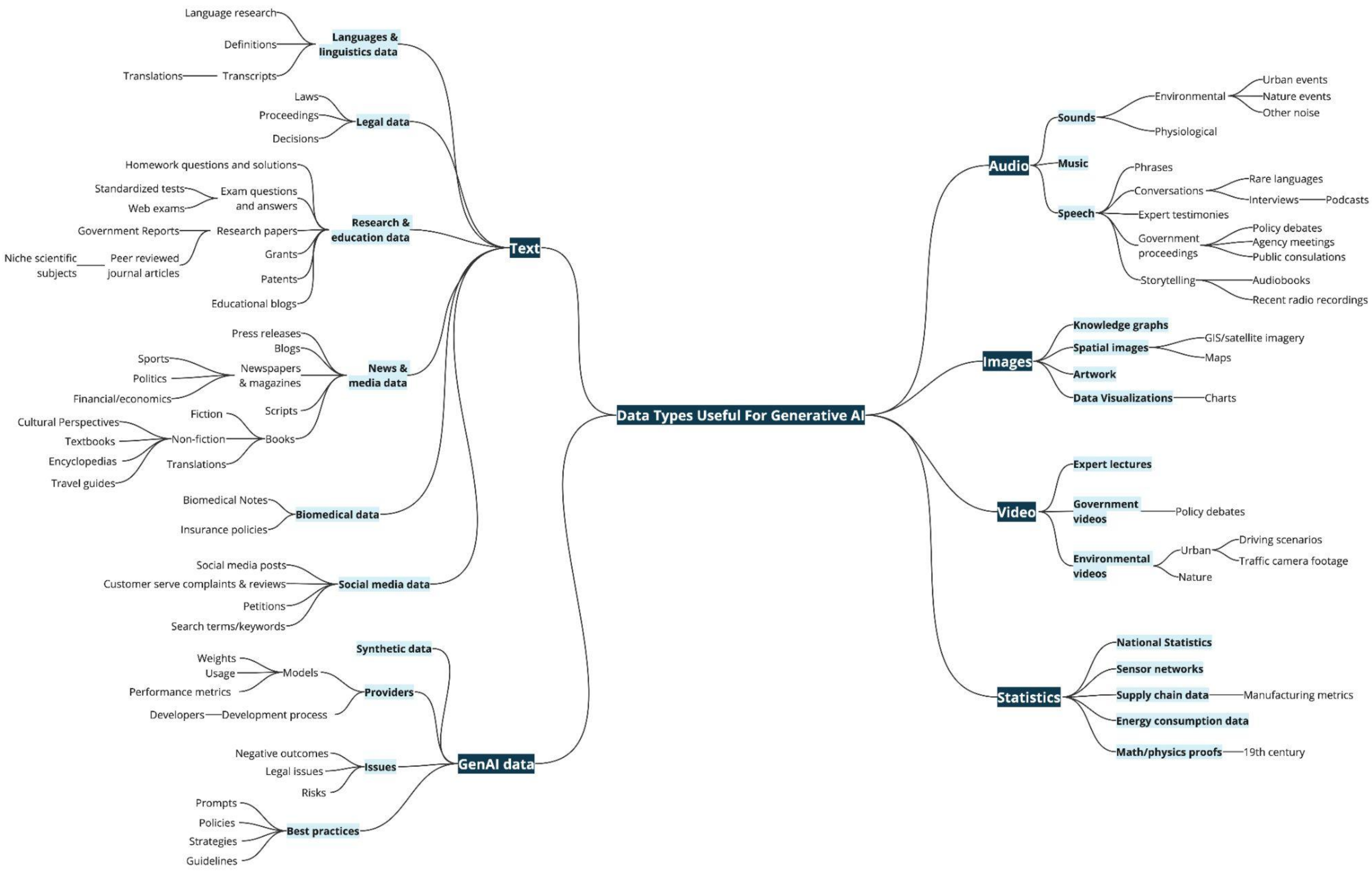


Figure 2. Summary of Data Types

View the full map HERE

TAXONOMY OF USE CASES

The following chart summarizes our taxonomy of data commons use cases that could benefit society. In this taxonomy, we focus on the purpose for developing AI technologies, the types of technologies that could be developed and the scope of the use or who it is affecting most. We also outline potential stakeholders that may be involved. This list of stakeholders is exemplary only and non-exhaustive.

Area of Focus	Purpose	Scope	Potential Outcomes	Stakeholders
1. Research and Scientific Discovery	Enhance scientific exploration and accelerate new discoveries and innovations	Societal	Drug discovery, hypothesis generation, climate modeling, open science collaboration	• Scientists and researchers • Research and academic institutions • Research funders • Publishers/journals • Government research agencies • Multilateral organizations • Representatives of the public
2. Education, Training, and Learning Support	Support education and learning through personalized and interactive tools	Individual	Personalized learning platforms, training simulations, assistive tools, content generation	• Students • Educators • Research and academic institutions • Education technology companies • Government education departments • Multilateral organizations

3. Content Development and Knowledge Preservation	Create and tailor content for education, service delivery, communications, and the preservation of knowledge across cultures	Organizational	Multilingual content, public health campaigns, digital heritage, AI-assisted infographics/visual communications	• Archives, museums, and libraries • Government communications departments • Civil society organizations (NGOs, advocacy groups, and unions) • Multilateral organizations • Representatives of the public
4. Modeling, Simulation, and Anticipatory Decision-Making	Anticipate future outcomes through realistic, data-driven scenario generation	Societal	Climate simulations, policy modeling, urban planning (e.g. digital twins of cities), crisis management, multi-stakeholder decision models, risk analysis	• Foresight and forecasting practitioners • Government strategists and policy makers • Local governments • Philanthropies • Climate agencies • Representatives of the public
5. Pattern Recognition, Inference, and Insight Generation	Analyze and identify trends and anomalies in data for better decision-making	Societal	Traffic optimization, crime pattern simulation, disinformation detection, economic trends	• Research and academic institutions • Government agencies • National Statistics Offices • Local governments • Multilateral organizations • Civil society organizations

6. Personalization and Customization of Services	Tailor services and experiences to individual needs	Individual	Personalized healthcare, adaptive public services, civic engagement tools	• Domain experts • Government agencies • DEI departments • Civic tech • NGOs and multilateral organizations
7. Automation of Repetitive Tasks	Increase efficiency by automating time-consuming processes	Organizational	Automated reporting, legal document drafting, customer support chatbots	• Government agencies • Multilateral organizations • NGOs and non-profits • Philanthropies • Advocacy groups • Businesses
8. Synthetic Data Generation	Preserve privacy while creating synthetic data for model training and research	Organizational	Synthetic datasets for healthcare, research benchmarks, addressing data scarcity	• PII departments • National Statistics Offices • Multilateral organizations
9. Real-Time Insights and Adaptive Systems	Provide feedback in real-time for dynamic system improvement	Societal	Real-time traffic and health monitoring, adaptive disaster response	• Government agencies • Multilateral organizations • Emergency response departments • Local governments
10. Collaborative Intelligence	Enhance collective problem-solving and finding consensus among diverse groups	Organizational	Mediation and conflict resolution, administrative support, public consultation tools	• Government agencies • Multilateral organizations and NGOs • Policy makers • Civil society organizations • Representatives of the public