The Role of Government as a Provider of Data for Artificial Intelligence

Phase 2: A practical tool for government data sharing for AI

November 2024

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Acknowledgements

This report was developed in the context of the 'The Role of Government as a Provider of Data for Artificial Intelligence' project, with the steering of the Project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Data Governance Expert Working Group. The GPAI Data Governance Expert Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Ching-Yi Liu*, National Taiwan University
Jhalak Kakkar*, Centre for Communication Governance, on leave since June 2024
Avik Sarkar*, Indian School of Business

The report was written by **Annys Rogerson**‡, **Sulamaan Rahim**‡, **Pablo Fuentes Nettel**‡, **Richard Stirling**‡, and **Yaseen Ladak**‡ of Oxford Insights; **Muchiri Nyaggah**†, Local Development Research Institute, **Natalia Carfi**†, Open Data Charter, and **Juan Carlos Lara**†, Derechos Digitales.

GPAI would like to thank the Project Advisory Group Members **Thierry Warin*** (HEC Montréal)**, Seong Oun Hwang*** (Gachon University)**, Andrew Sporle*** (iNZight Analytics Ltd)**, Zümrüt Müftüoğlu*** (Yildiz Technical University), **Jaco Du Toit**** (UNESCO), **Kudakwashe Dandajena*** (AIMS-II), **Zee Kin Yeong*** (IMDA) , **Kim McGrail*** (University of British Columbia), **Mikael Jensen*** (D-Seal), and **Ulises Cortes*** (Universitat Politècnica de Catalunya).

We are also grateful for the review of the text by legal scholars at the Taiwan AICoE, **Chuan-Feng Wu**†, **Tai-Jan Huang**†, **Wen-Tsong Chiou**†, **Ching-Fu Lin**† and **Ting-Chi Liu**†**.**

GPAI would like to acknowledge the tireless efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and GPAI's Data Governance Expert Working Group. We are grateful, in particular, for the support of **Stefan Janusz, Antoine Glory** and **Stephanie King** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Shameek Kundu*** (IMDA) and **Bertrand Monthubert*** (Conseil National de l'Information Géolocalisée).

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2024. The Role of Government as a Provider of Data for Artificial Intelligence, Phase 2: A practical tool for government data sharing for AI, Report, November 2024, Global Partnership on AI.

Table of Contents

Purpose of the Roadmap

This roadmap aims to support government teams to make decisions about when and how to share data for Artificial Intelligence (AI) innovation. It does this by providing a checklist, accompanied by practical guidance and examples, for government teams working on data sharing initiatives for AI. The checklist aims to direct initiatives towards creating public benefits while respecting the rights of affected individuals and communities and contributing to the goals of equality and data justice.

The roadmap is aimed at government teams from any level of government – municipal, local, regional/state or national/federal – who are interested in or already engaged in projects that share data with external AI developers. It can be used by teams to:

- Develop a plan for a new initiative that shares data that can be used in AI development from the beginning.
- Develop a plan for the data sharing aspect of existing, wider projects the team has underway.
- Systematically assess existing data sharing initiatives.

Motivations for sharing data for AI development

Governments hold data that is valuable for AI development. This includes data on society, the economy, and administrative data, generated through service delivery and national statistics exercises. This data could be used by companies, research organisations, and non-profit organisations for the development of AI-driven research, products and services for use inside and outside of government.

Recognising this potential, many government teams are starting to explore how to share data responsibly and effectively with AI developers. Teams have wide-ranging ambitions for sharing data with AI developers. These include:

- Facilitating AI applications to improve government service delivery (Berryhill et al, 2019).
- Commissioning or enabling AI-driven research to better understand complex social issues and inform policy decisions (Ziesche, 2023).
- Supporting innovation of AI products and services outside of government to cultivate the domestic AI sector and support AI adoption (Ceulemans et al, 2021).
- Targeting problems in key sectors and development issues that AI products or services can help to solve (GPAI, 2022b).

The first phase of the GPAI Government as a Data Provider for AI project captured how a number of governments are already sharing data for AI (GPAI, 2023a). The government of Taiwan is facilitating health data sharing between citizens and approved third party AI app developers via its Health Passbook to enable the development of AI-driven health technologies. The government of Colombia is supporting the sharing of government and farmers' data on its Aclímate platform to enable the provision of AI-driven insights into climate and environmental conditions for farmers. Examples of data sharing initiatives for AI are also given throughout the report.

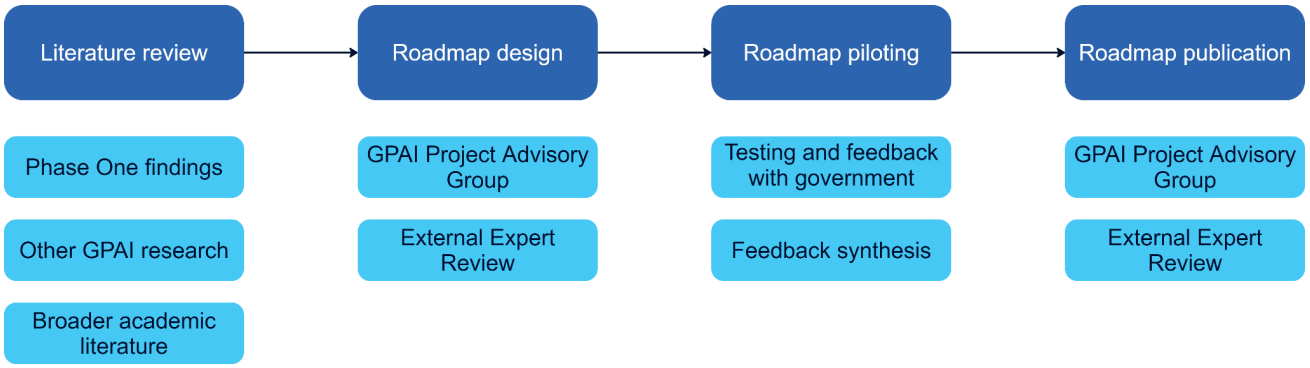
Despite these ambitions, many teams are uncertain about when and how data can be shared with external parties. This exercise must be undertaken in line with principles of human rights, privacy and data protection, inclusion, diversity, innovation and economic growth. To do so, there are a number of challenges that teams often face, including establishing public trust, putting in appropriate privacy measures to comply with data privacy legislation, having regulatory certainty, and building technical capacity (GPAI, 2023a). There are also cultural and commercial barriers to

data sharing, which can exist in government as much as in the private sector (Deloitte, 2017). This roadmap aims to support government teams in navigating their data sharing initiatives and responding to these challenges.

How the roadmap was developed

Oxford Insights was contracted to develop a practical roadmap for governments to share data for AI innovation and test the roadmap with pilot partners.

The roadmap incorporates the findings from the first phase in the GPAI Government as a Data Provider for AI project. This first phase used four case studies of government data sharing for AI to develop a set of recommended principles for governments to follow (GPAI, 2023a). Together with the input of a wider literature review and review from global experts in open data, data sharing, and AI, a first draft of the roadmap was developed.



In order to test the usability of the roadmap, it was subsequently piloted with government teams in three countries across different levels of government:

- Agency for Electronic Government and Information of Uruguay.
- Digital Transformation Office of the Presidency of Turkey.
- Jigawa State Government

We selected these pilot partners after holding an open call for expression of interests from government teams globally. We wanted to ensure that the roadmap is applicable across different contexts. Therefore, we selected pilot partners to represent, among other criteria, a diverse range of geographies, levels of experience with data sharing, and sectors of interest for the pilots.

This final version has been iterated to reflect the valuable insights and lessons learned from these pilots, ensuring it meets the diverse aims and needs of government data sharing initiatives for AI.

Overview of the Roadmap

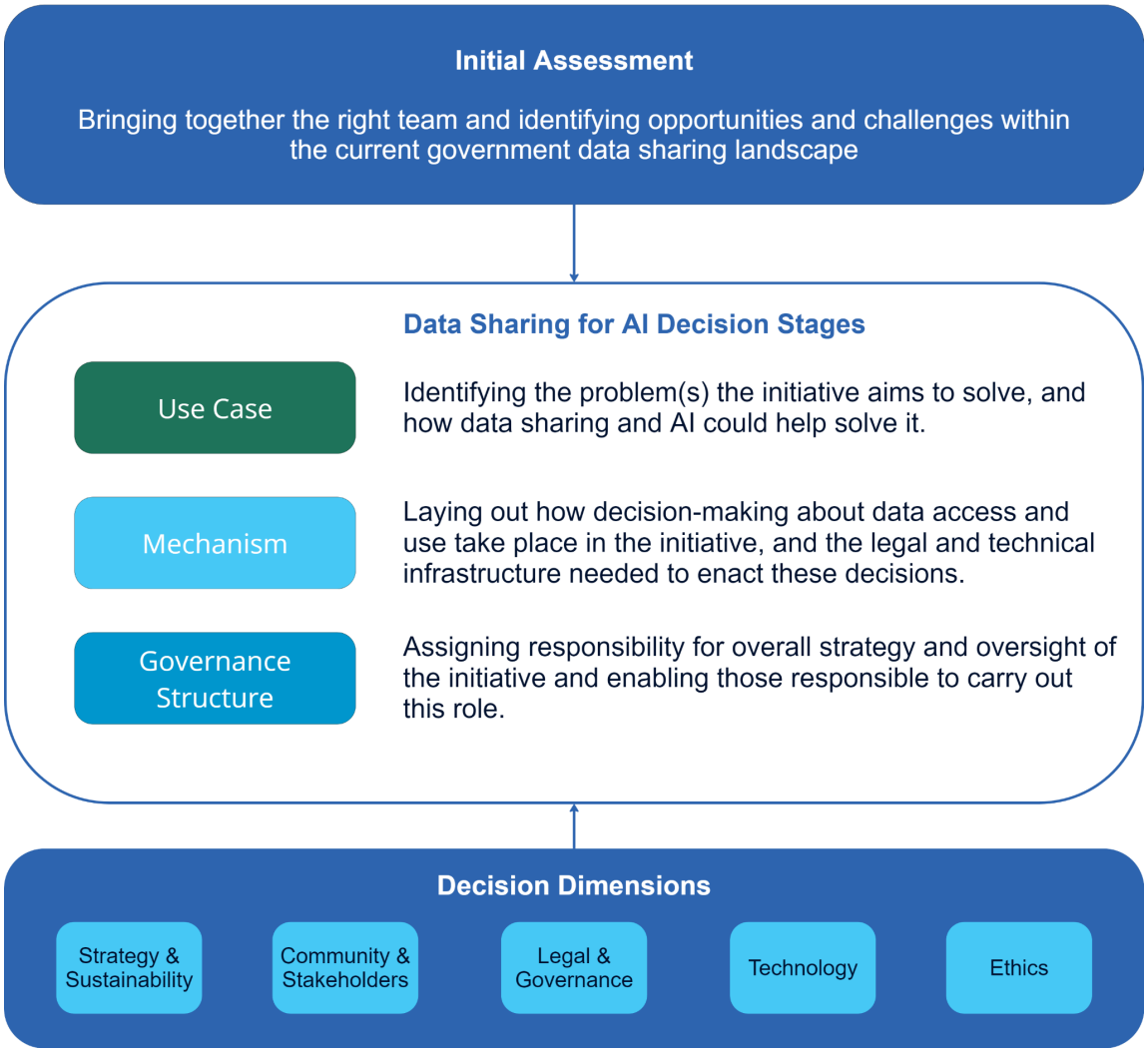
Report Section	Section Description
Framework for Data sharing for AI	Introduces users to the decision-making framework employed within the roadmap.
Using the Roadmap	Advises users on how they can use the roadmap to progress in their data sharing for AI projects.
Initial Assessment	Provides the user with recommended first steps for their project that bring together the right team and assesses the current government data sharing landscape.
Roadmap Checklist	Presents a set of questions for users to consider in order to progress through each stage of the decision-making framework.
Checklist Guidance	Explains why each question in the roadmap checklist is included and a description of what may be required to tick off this question. The guidance also includes international examples, recommended activities, and learnings from the pilot partnerships to provide users with the resources to address the question.
Pilot Case Studies	Presents the learnings from the pilot partnerships as well as a case study of the roadmap implementation journeys and outcomes for each of the three pilot partners.
Supporting Resources	Provides further resources, such as workshop materials, and key term explainers, that were used during the pilot partnerships and could help users address the checklist questions.
Full Methodology	Describes the detailed methodology for the roadmap development, along with key changes made as a result of the learnings from pilot partnerships.

Framework for Data Sharing for AI Roadmap

The roadmap supports government teams through the development of a data sharing for AI initiative by taking teams through an initial assessment, and then beyond this, supports government teams through three decision stages in their projects.

For each decision stage, the roadmap breaks them down into the key considerations for developing effective and responsible initiatives. These considerations are grouped into five decision dimensions and they are presented as questions within a checklist for teams to complete.

Government Data Sharing for AI Roadmap Framework



Initial Assessment

The roadmap begins with an initial assessment of the readiness for undertaking a data sharing for AI initiative. The purpose of the Initial Assessment is firstly to establish that the team and key stakeholders have appropriate capacity to undertake the initiative, and secondly to identify enablers and challenges for the initiative within the government's data-sharing landscape.

Use Cases

Use cases connect the initiative to the problem(s) they aim to solve. Without clear, relevant, and achievable use cases, the initiative will not have real-world impacts.

In the roadmap, a use case for a data sharing for AI initiative specifies the problem the team wants to solve, what AI applications could help to solve it, what data needs to be shared to facilitate these applications, and who the users of data need to be.

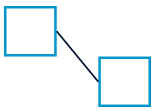
The aims of the Use Case questions in the checklist are to:

1. help the team identify potential use cases;
2. set the legal and ethical boundaries within which data can be shared and for what purposes; and
3. prioritise use cases to move forward with.

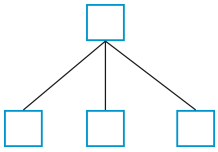
Data Sharing Mechanisms

Data sharing mechanisms lay out how decision-making about data access and data use take place in the initiative, as well as the legal and technical infrastructure needed to enact these decisions. A principal driver for mechanism design is creating trust among participants of the initiative, which is needed for them to take part.

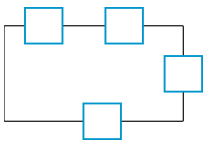
Several mechanisms are emerging for governments to provide data for AI. There are many ways to categorise these mechanisms. In the roadmap, they are broken down into three types of mechanisms based on how data flows between participants of the initiative.



One-to-one data sharing: agreements between two individual parties, who are users and/or controllers of the data. For example, a private contract, or a public-private partnership.



One-to-many data sharing: agreements between one party, who is the data controller or represents the data controllers, and many data users. For example, an open data initiative, a sandbox, or a data stewardship initiative[1].



Many-to-many data sharing: agreements between many parties who are data controllers and many parties who are data users. These may be the same parties. For example, a data commons, or a data marketplace.

[1] Each of these categories are explored in more detail within the Data Sharing Mechanism Taxonomy, found in the Supporting Resources Annex.

Participants in data sharing for AI initiatives

In this roadmap, participants in the initiative are referred to by the following terms (GPAI, 2022b):

- **Data contributors** are those individuals or legal entities who are the subjects of information (data subjects), whose activity results in the generation of new data, those who assemble data, for example in a dataset, and, those who contribute more remotely, such as by providing a device that generates data.
- **Data controllers** are those actors who hold data and, depending on the mechanism, they may also decide the purposes and means of their processing. In this case, they are also **data stewards**. There may be co-controllers who exercise joint decisions about the data together.
- **Data processors** are service providers who process data on behalf of a data controller.
- **Data users** are actors who receive data from data controllers and/or data stewards.
- **Data intermediaries** are data sharing service providers, such as data trustees or data marketplace providers.
- **Service recipients** are actors who use data-based services.
- **Other relevant actors** are regulators and policy makers who set policy, legislative, and regulatory frameworks.

The aim of the Mechanism questions in the checklist is to clarify participant roles and assess necessary data infrastructure and legal infrastructure. This helps tailor the mechanism to the initiative and government setting.

Governance Structures

Governance structures lay out who will be responsible for overall strategy and oversight for the initiative and how they will be able to carry out this responsibility. A principal driver for governance structure design is maintaining trust among participants and the public once the initiative is operational.

The aim of the Governance Structure questions in the checklist is to assign responsibility for strategy and oversight, and ensure that the processes and mechanisms are in place for those responsible to perform this role.

Decision Dimensions

Each of the three decision stages is broken down into considerations relating to five decision dimensions that build on the recommendations from the first phase of this project and wider GPAI Data Governance working group projects (GPAI, 2020; GPAI, 2022a; GPAI, 2022b; GPAI, 2023a; GPAI, 2023b).

Strategy & Sustainability

The use cases target problems that are a strategic or developmental priority for the government, and solving them would have significant public benefits. At the same time, the initiative has funding for an initial pilot phase secured and options for gaining sustainable, long term funding. The financial model should also support a fair distribution of the benefits.

Community & Stakeholders

The needs and perspectives of participants in the initiative, data subjects, and any other affected actors are included from the start of development and reflected in a design that meets the demands of equality and data justice[2]. Transparency is maintained for all of these actors throughout the lifetime of the initiative.

Legal & Governance

The initiative respects legislative and regulatory frameworks, which embed the human rights and data rights of data subjects, and are enforced by capable institutions. Relevant public and private actors have clear obligations and there are mechanisms that enable their accountability.

Technology

The data is shared using safe and secure infrastructure that is continuously monitored. The data meets the quality and accessibility requirements of its users. Safety and security risks are identified and mitigated through impact assessments, auditing, and technical governance techniques. There are fair and robust processes for approving or procuring data users.

Ethics

The initiative does not expose individuals and communities to risk of harm, exclusion, or bias, through either unnecessary data sharing, including breaches of privacy (also covered by legal and governance aspects) or algorithmic decision-making. Impact assessments, human oversight, user centred design, and other built-in processes are employed to ensure that AI systems do not cause harm and are used in ways that promote fairness and bring public benefits.

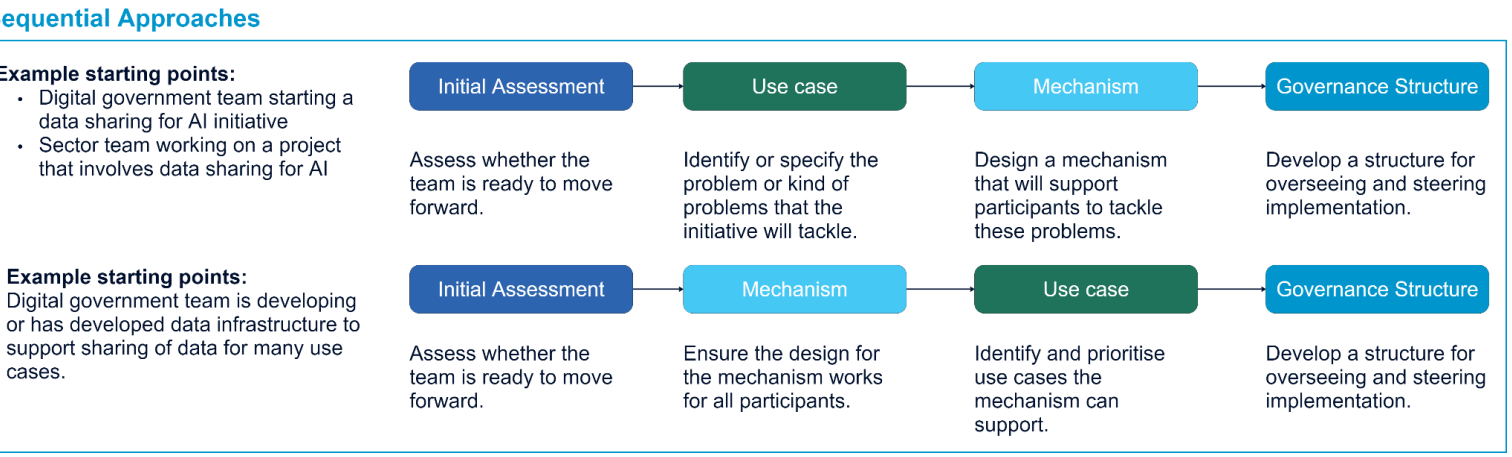
[2] GPAI's 2022 report on data justice asks policymakers to acknowledge and respond to uneven distributions of opportunities and harms that follow historical patterns of social and geographical inequality both within and between countries as a result of data extraction and data-intensive systems (GPAI, 2022c).

Using the Roadmap

User Journeys

Government teams are at different stages in their data-sharing journey—some may have already identified use cases for data-sharing in AI, or developed reusable data sharing infrastructure, while others are just beginning. The roadmap is designed to be used by those at all stages of this journey. To meet these varied needs, the roadmap is therefore designed so that it can be used in different ways. Depending on their aims and starting points, teams can take a number of approaches:

- **A sequential approach**: Teams can follow the roadmap linearly, from the very beginning or starting with a use case or mechanism in mind. We recommend that team's complete the initial assessment and then move onto working through the questions in the checklist. Depending on the team's starting point, the team may begin with the use case or the mechanism sections.



- **A project review approach**: Using the checklist, teams can systematically assess current data sharing projects or existing data-sharing for AI practices and guidance more broadly.
- **A problem-solving approach**: Teams can address specific challenges within a data-sharing for AI project with the guidance included in the checklist.

Initial Assessment

The aim of the initial assessment is to understand the extent of both team and institutional capacity for undertaking a data-sharing for AI initiative. It aims to bring the right team together, support coalition building with other relevant stakeholders, and identify enablers and challenges within the government data-sharing landscape that are relevant to the initiative.

This section provides guidance on the institutional requirements to move ahead with the initiative, the team structure and skills, and resources for conducting a mapping of the government data sharing landscape.

Team and Coalition building

It is important to bring together the right team for the development of the initiative. Developing a data sharing initiative requires collaboration across teams from the beginning.

The core team should **include** members with experience in and responsibility for:

- **Data and AI policy**, including the capacity to conduct impact assessments for data sharing and public sector use of AI systems.
- **Technical expertise**, with capacity to generate use cases for data sharing, discuss the design of data architecture, and lead on its implementation.
- Government **stakeholder management and project management**.
- **Sectoral data expertise** within the sector that the team is hoping to share data within will also be critical. However, the sector may be determined at a later stage in the project and this consideration is covered in question 3 of the checklist.

The core team should **engage** with representatives from:

- **Legal teams** who have experience in applying legal frameworks within a data sharing context and in forming data sharing agreements.
- **Regulatory agencies** responsible for applying the country's data governance frameworks, and have the capacity to engage frequently with the project and the authority to provide independent oversight.
- **Procurement teams** who have experience in software and AI procurement.
- **Communications team** who can support engagement and public engagement exercises.

The core team should have **access** to final decision-makers within the department who can authorise the mandate and budget the team needs to move forward.

Having all of these domains and skills involved from the start facilitates communication at decision-making points and limits conflict further down the line.

Ensuring legal and regulatory capacity

Unlike many of the other challenges that teams can respond to more quickly with sufficient budget and the right mandate, having **a legislative framework and regulatory capacity are prerequisites** for moving ahead with initiative that the team will not be able to resolve themselves.

There must be a legislative framework in place for governing data sharing. This includes data protection and privacy legislation, as well as other legislation that establishes the boundaries and requirements for when data can be shared, by whom, and for what purposes. **These are detailed further in the data sharing landscape mapping on the following page.**

As well as having legislation in place, a further prerequisite is that there is a regulatory agency, responsible for overseeing the implementation of data governance and, if applicable, AI governance frameworks within the country. Beyond having a regulatory agency set up, the agency must have (1) the authority to enforce the regulatory framework so that it cannot be ignored by the government, (2) the know-how for applying legislation in the context of data sharing initiatives, and for use to train or deploy AI systems, and (3) capacity to engage with the team and give guidance during the initiative's development and implementation.

For example, within the country's Data Privacy Act 2012 (Republic Act No 10173, 2012), the Philippines' Privacy Commissioner is given responsibility, among other things, for (1) ensuring compliance among data controllers, including government agencies, (2) publishing guidance to all laws relating to data protection, and (3) providing assistance at the request of government agencies. In other countries, regulators are also developing AI-specific guidance and expertise (OECD, 2024). In France, its data protection authority, CNIL, published recommendations for developers using AI training datasets that include personal data (CNIL, 2024).

Data sharing landscape mapping

A data sharing landscape mapping outlines key enablers to data sharing initiatives in the government (GPAI, 2023a). These are:

- the data infrastructure available to support a data sharing initiative;
- the policy and legislative frameworks that facilitate and set the boundaries of data sharing; and
- the team members, skills, and relationships to facilitate the collaboration.

Recommendation from roadmap pilots

During the pilot implementations of the roadmap, the mapping exercises were conducted as workshops with the core pilot team and any stakeholders with relevant experience or knowledge to create a complete picture of the state of data sharing within the pilot government. The benefits of data sharing landscape mapping were threefold.

1. It supported alignment across team members and other stakeholders about their **motivations for undertaking the initiative and its objectives.**
2. It allowed teams to **identify enablers of data sharing** in their government—e.g., where existing infrastructure and stakeholder relationships can be built on for the initiative—and **potential challenges for data sharing**; e.g., not having experience of applying relevant legislative frameworks within the team.
3. It helped with **coalition building** across teams in government. Bringing together stakeholders from across teams who are involved in data sharing can help to raise awareness of the initiative and build necessary relationships to move forward.

Data infrastructure

A number of components are relevant to mapping data infrastructure (GPAI, 2023a):

- infrastructure used for storing, processing, integrating, managing, accessing, securing, and analysing data;
- servers, either local or cloud-based;
- data access mechanisms such as Application Programming Interfaces (APIs), open government data platforms; and
- policies that aim to establish data standards, interoperability and quality.

Legislative landscape

Current and upcoming legislation that is relevant to data sharing within the jurisdiction are both relevant. While the relevant legislation will differ from country to country, there are six key areas of legislation that are important for ensuring equitable access to data by third-party AI developers (GPAI, 2023a):

- **Data protection and privacy:** protects the rights of data subjects[3], including through requirements for when and how to obtain their consent, and grants enforceable powers to independent oversight institutions to regulate data processing and AI systems.
- **Intellectual Property (IP)**: datasets may be protected under IP laws. Data sharing may therefore have implications on copyright or trade secret laws and such sharing may require prior IP licensing. There are also considerations about IP rights to innovations from shared data use.

[3] Data protection and privacy legislation differs across jurisdictions and therefore so do the protections it provides to data subjects. An explainer on data rights is included in the Supporting Resources to guide teams through what to look out for in legislation and other policies.

- **Antitrust:** access to data can create an economic advantage for companies who are involved in a data sharing initiative.
- **Cross-border data flows:** legislation about maintaining data sovereignty may influence data sharing with international partners. Such legislation may impose certain adequacy requirements prior to international transfer. Or, it may introduce cross-border restrictions for certain types of data in certain sectors.
- **Access to Information:** creates mechanisms to enhance transparency by providing individuals with access to information on government data about them.

Artificial Intelligence specific legislation: imposes specific requirements on development and deployment of AI systems, for example according to their level of risk (European Commission, 2024).

Checklist for Data Sharing

The Government Data-sharing for AI Checklist is presented on the following pages. The aim of this checklist is to identify the considerations that government teams are recommended to take into account when arriving at a decision for the three decision stages in the framework:

- Use Case (U)
- Mechanism (M), and
- Governance Structure (S).

For each decision stage, the decision dimensions shown previously help to frame the question being asked.

Before taking a decision at each stage, it is recommended that **government teams can confidently answer 'yes' to all of the checklist questions**.

The following three **sections** after the checklist (reflecting the decision stages Use Case, Mechanism, and Governance Structure) break down each question and provide teams support to

1. assess whether they can tick off this question in the checklist, and
2. get to this position of being able to tick off this question if not.

The guidance for each question includes:

- **Why it is important:** the rationale for why this question is important for ensuring the data sharing initiative for AI is responsible and effective.
- **What it might look like:** a description of when a team would be able to tick off this question in the checklist. For most questions, this differs for each initiative but illustrative examples are provided where possible.
- **Getting there:** some questions include suggested activities to help the team meet the requirements of the checklist question. Supporting resources are also provided in some cases.
- **Recommendation from roadmap pilots:** learnings about implementing the roadmap from the pilots that were run using the draft version are sometimes included.

Please note: You do not have to read all of the material in the blue boxes if you are satisfied you understand the question, and can answer 'yes' to it.

The guidance provided for each question can be accessed via the links in the right hand column.

A spreadsheet version of the checklist for tracking progress can also be downloaded.

Decision Stage: Use Case (U)		
Decision Dimension	**Question**	**Question reference**

Strategy & Sustainability	Does the initiative try to solve clear problems that are of public value?	U1
Community & Stakeholders	Is there demand for the data among intended data users?	U2
	Is there support for the initiative from the relevant team(s) responsible for the sector or area in which the data is being shared?	U3
	Is there support among the public for sharing data in this sector?	U4
Legal & Governance	Is there a legal framework that clearly sets out conditions for sharing data that apply to this initiative and are complied with by the initiative?	U5
	Does the data to be shared stay within what is necessary for the use case?	U6
Technology	Can the data shared in the initiative be made ready for use by AI developers?	U7
Ethics	Do intended AI applications align with government or other best practice guidance on responsible AI use and development?	U8
	Can the risks to data subjects and other affected groups be mitigated?	U9

Decision Stage: Mechanism (M)		
Decision Dimension	**Question**	**Question reference**
Strategy & Sustainability	Does the initiative's financial model support a fair distribution of its benefits?	M1
Community & Stakeholders	Are data controllers agreed on how data access and use is managed within the initiative?	M2
	Are data controllers able and willing to supply data within the initiative?	M3
Legal & Governance	Are there data sharing agreements between all necessary parties?	M4
Technology	Is there data infrastructure that supports this data sharing mechanism?	M5
	Is it possible to share the data safely and securely?	M6
	Have data management and governance practices been established to ensure continued data quality throughout the initiative?	M7
Ethics	Is the level of openness proportionate to the level of sensitivity of the data?	M8

Decision Stage: Governance Structure (G)		
Decision Dimension	**Question**	**Question reference**
Strategy &	Is responsibility allocated for how the overall strategy of the initiative will	G1

Sustainability	be set, monitored, and steered?	
	Is there a budget secured for this data sharing initiative or are there options for where a budget could be sourced from?	G2
Community & Stakeholder	Are there reporting mechanisms from representatives of all actors affected by the initiative to the people responsible for its oversight?	G3
	Is transparency maintained throughout the development and implementation of the initiative?	G4
	Is there a plan for how a wide range of actors can access the benefits of the initiative?	G5
Legal & Governance	Does the initiative collect consent from any data subjects and facilitate data subjects to remove their consent at any stage?	G6
	Is there a mechanism through which data subjects can seek redress for rights violations?	G7
	Is there a channel for the relevant regulatory agencies to track the initiative and give guidance on its development?	G8
	Are there processes for monitoring compliance with data sharing agreements?	G9
Technology	Are there processes in place to monitor infrastructure and maintain safe and secure sharing of data over the lifetime of the initiative?	G10
	Is the government equipped to conduct a fair procurement process that evaluates AI systems to ensure they meet the requirements of the initiative?	G11
Ethics	Are there processes for monitoring and responding to risks to data subjects and other affected groups?	G12
	Does the initiative require human oversight for any algorithmic decision making?	G13

Decision Stage: Use Case

In the roadmap, use cases of data sharing for AI initiatives are characterised by:

- **A target problem or set of target problems:** the problem(s) that the team wants to tackle through a solution that involves data sharing.
- **Data to be shared:** the data that needs to be shared to tackle this problem.
- **AI applications:** how AI applications can help solve the target problem using the data.
- **Data users:** the types of stakeholders (or specific stakeholders if known) who will use the shared data to develop AI applications.

Why this decision stage is important:

Ensuring that the initiative is driven by solving real-world problems of public value requires identifying and selecting clear use cases. The team requires a framework for identifying use cases based on where there is demand for data, that apply clear legal and governance frameworks setting out the boundaries of data sharing, and that work towards governmental priorities.

How to use this section of the checklist:

These questions help teams identify, assess, and prioritise use cases for the initiative. Prioritisation may be necessary due to resource constraints that require teams to choose between multiple possible use cases.

This section of the checklist can also serve as a quantitative assessment framework for comparing and prioritising potential use cases. A template is included in the Supporting Resources section, along with a prioritisation matrix.

If the team is using the checklist to identify use cases and does not already have a specific sector or challenge area in mind, it is recommended to first narrow down the problem space to a sector, for example based on existing projects or governmental priorities.

Use Case – Strategy & Sustainability

U1. Does the initiative try to solve clear problems that are of public value?

Why it is important

Use cases of the initiative need to identify and target particular problems. These are problems that actors outside of government could address with access to government held, or other privately held, data.

One way of thinking about whether a problem is of public value is to consider governmental priorities within sectors, such as health, or within government, such as public service digitalisation. Concentrating on a problem that is already a government priority is more likely to receive buy-in from leadership and secure funding.

What it might look like

The target problem or problems may be those faced by government, private actors, researchers, or the wider public. For example, Colombia's Aclímate platform is a response to the problems farmers are facing as a result of climate change. In particular, their yields are suffering from changing and increasingly unpredictable weather patterns and data-driven, machine learning insights are now helping them to manage these changes (Young and Verhulst, 2017).

Examples of target problems from data sharing initiatives

Getting there

Identifying use cases will involve working with actors inside and outside of government to understand where a lack of data is hindering advancements in their work. When working with actors to come up with potential use cases, the team can take a bottom-up, top-down, or mixed approach to harness domain knowledge of sectoral actors:

- **A bottom-up approach** invites proposals from multiple actors—both governmental and non-governmental actors—in, for example, open calls to pitch ideas for data sharing. This approach may be less resource intensive and help create partnerships with key actors. However, it requires effective awareness raising and willingness to engage among actors, which may limit the number of potential use cases identified.
- **A top-down approach** involves conducting workshops, surveys, interviews, and desk research into challenges facing actors working in the sector both inside and outside of government to identify where there is data demand, supply, and precedent. This approach can help to identify a range of potential use cases, which allows you to prioritise the most suitable one. However, it may be more resource intensive if there are not existing demands for increased data sharing from actors within the sector.

The processes used to engage actors during the development of the initiative should be open and fair to ensure that any potential future suppliers to the government do not gain an undue advantage. This could include conducting public consultations, or open calls for engagement.

Recommendation from roadmap pilots

To identify actors to engage and explore possible target problems, it is useful to conduct a mapping of the existing data ecosystem that it targets. A data ecosystem map visualises what data assets exist in an ecosystem, and how they are being accessed, used, and shared among actors (ODI, 2022).

Once the map for the data ecosystem is complete, it can be used to explore or validate data sharing opportunities and to visualise current blockers faced by actors and where the initiative could support new data flows and AI system development and deployment within the ecosystem.

The ODI has created a methodology (ODI, 2022) and tools (ODI, 2019) for conducting a data ecosystem mapping. The mapping is best conducted with the involvement of the relevant actors.

Use Case – Community & Stakeholders

U2. Is there demand for the data among intended data users?

Why it is important

Without there being actors who would like to use the data that is shared to work on the target problem, the data would not be used to create public value.

What it might look like

There is clear, demonstrable demand for data related to this problem by actors across industry, academia, civil society, and government. In particular, there is evidence that restrictions on data access are currently blocking potential AI applications and advancements inside or outside of government.

Examples of data demand tackled by data sharing initiatives

Initiative (Country)	Data demand among actors
Aclímate (Colombia)	Researchers/forecasters need accurate, comprehensive, and real-time data to develop analysis around climatic/environmental patterns. Farmers use these insights to guide decisions on farming practices and optimal crop planting strategies for maximising yield. The collaboration between IDEAM, CIAT, and Federroz facilitates secure sharing and integration of government and farmers' data on the Aclímate platform. These are then supplemented with tools and machine learning insights, which allow farmers to access actionable information (Young and Verhulst, 2017). .
The Health Passbook (Taiwan)	Third-party apps which provided personalised healthcare services and those conducting research and development for health AI technologies did not have direct access to National Health Insurance Administration (NHIA) data and patient data. Given the sensitive nature of the data, they could not access servers which made providing their services more difficult, providing a mechanism for patients to consent to their data using NHIA-vetted third-party apps. Patients could directly authorise Third-party AI App developers and researchers, who had been vetted by the NHIA, access to their data to provide additional services (including research for health AI) (GPAI, 2023a).
Copernicus (EU)	In the absence of easily accessible and accurate satellite and geospatial data, government analysts, researchers, and scientists cannot adequately monitor the environment, manage disaster risks, or conduct public health exercises. A paucity of comprehensive and up-to-date data hinders the utilisation of powerful AI tools for climate data analysis. The Copernicus Open Access Hub provides timely, well-structured, and processed data on climate, geography, and the environment, empowering users to conduct in-depth analysis and employ AI tools (Copernicus, n.d.).

All of Us data (USA)	Researchers and public health officials require comprehensive, representative population-wide health data to inform robust health policy decisions. Access to patient-level data is essential for research and personalised medicine development. Particularly, large datasets are vital for leveraging machine learning and AI tools to derive novel insights and inform predictive models on population health, disease management, and treatment outcomes. Previously, accessing such data was challenging due to inconsistencies in processing and cleaning methods, which did not adequately respect privacy rights, alongside limited accessibility. The All of Us platform offers a secure means of accessing this data (All of us, n.d.).
Land Transport Authority (LTA) DataMall (Singapore)	Comprehensive real-time and historical data are required to facilitate good urban transport planning and allow for AI research into improving congestion and traffic flows. App developers also need access to real-time traffic information to provide real-time updates and information for citizens using these services. Such transportation data, however, come from various sources and would be difficult to access. Singapore's DataMall platform aggregates and centralises these data to provide a single point of access to a breadth of datasets. The platform provides clear APIs that enable varied access to datasets upon request. It allows academic researchers, private sector companies, other government departments, and ordinary citizens to access all relevant transportation data in one place (Land Transport Authority, n.d).

U3. Is there support for the initiative from the relevant team(s) responsible for the sector or area in which the data is being shared? *[Community and Stakeholders]*

Why it is important

Support from the team responsible for the sector or area, if this is not the team leading the initiative, is critical to the initiative's success. The sectoral team are likely to be current curators and controllers of data relevant to the sector or problem. They are also likely to have relationships with relevant stakeholders outside of government.

What it might look like

If the initiative is not being driven by a team who is responsible for the sector or area, the initiative should have commitment from the sectoral team(s). For example, if the project is being run by a digital government team that works across government, and would like to support data sharing within the agricultural sector, then the team will need to collaborate closely with the agency or team responsible for agriculture.

U4. Is there support among the public for sharing data in this sector? *[Community and Stakeholders]*

Why it is important

Public support is foundational to establishing trust in the initiative. Establishing whether there is public support is most important where personal data, or other highly sensitive data, such as that relating to health or national security, is being shared.

What it might look like

There are indicators from research – such as surveys, user research,consultations, or previous government initiatives – that the public trusts the government to enable data sharing in this sector.

Use Case – Legal & Governance

U5. Is there a legal framework that clearly sets out conditions for sharing data that apply to this initiative and are complied with by the initiative?

Why it is important

Use cases that the initiative supports must be compliant with legislation and regulatory guidance on data sharing.

What it might look like

Legal frameworks and regulatory guidance will differ between jurisdictions. Across jurisdictions, teams need to identify the frameworks, and accompanying guidance, that apply to their use case(s). In addition to data protection and privacy legislation, there may be more specific legislation that needs to be complied with depending on the nature of the data involved in the use case, for example, additional legal requirements may apply if health data is shared.

Using these frameworks, and with the guidance of team members with legal expertise, teams will need to establish the different permissions and requirements for sharing data within their initiative and apply these to the use cases under consideration. These may include[4]:

- Whether the government body has the legal power to share data
- Whether the initiative needs any permissions from regulatory or other agencies to share data
- Whether there are specific legal bases for sharing the data that must be established
- The purposes for which the body can share data
- Any specific documentation that needs to be concluded
- Any requirements relating to privacy and security measures the team need to take

International examples of different requirements

Legislation (Country)	Example of conditions made on data sharing
Canada's Legal Authority Requirement	**Legal power to share data:** programme or department-specific legislation specifies whether an initiative has the legal authority to collect and use personal data (Government of Canada, 2023).
United Kingdom's Digital Economy Act	**Permissions for data sharing:** a processing body, who is sharing data for research, requires accreditation by the UK's Statistics Authority before it can do so (Digital Economy Act, 2017).
European Union's General Data Protection Regulation (GDPR)	**Legal bases:** processing of personal data is lawful only for a set of specified legal bases, which include receiving the consent of individuals concerned, contractual obligations, and for an organisation's legitimate interest (Regulation 2016/679) .
Turkiye's Statistics Law	**Purposes for sharing data:** Individual level data can be shared for use in scientific studies without any reference to distinct statistical units (Statistics Law of Turkey, 2005).
Indonesia's Personal Data Protection Law	**Specific documentation:** a data protection impact assessment must be completed for processing of personal data that is categorised as high risk (Dharyanto, Murata and Ustriyana, 2023).
China's Population Health Information Measures	**Privacy and security requirements:** requires that population health data is not stored or hosted on servers outside of China and sets requirements on storage in a graded manner based on data sensitivity (Postigo, 2023).

Getting there

[4] This is an illustrative list of the kinds of requirements initiatives must identify and apply to the use cases they would like to share within their initiatives. The requirements will differ across jurisdictions. International examples are provided to exemplify the variation in requirements across countries.

The team can engage with representatives from legal and regulatory teams to establish how the legal framework applies to their initiative, including the conditions and boundaries of data sharing that are relevant.

U6. Does the data to be shared stay within what is necessary for the use case? *[Legal and Governance]*

Why it is important

Sharing only the data necessary for specified, legitimate purposes, helps to ensure the data is only being used to solve the specified problem and that the use case respects the principle of proportionality (Black and Stevens, 2013). This is important to help prevent 'function creep', where data is processed for purposes beyond what was originally intended (Koops, 2021). The specification of purposes depends on the type and sensitivity of data being shared. It may be possible for data to be published under an open data licence, in which case, the purposes will be unrestricted. This will be different for personal, or non-personal sensitive data, where the purposes must be defined and the data that is shared restricted in line with these purposes.

What it might look like

In cases where data is not being shared under a open data licence, the data shared does not go beyond what is necessary to achieve the legitimate aim of the use case.

Use Case – Technology

U7. Can the data shared in the initiative be made ready for use by AI developers?

Why it is important

The data needs to be readily available for data users. It should therefore be stored and formatted in a way that meets the needs of potential data users so that they can work on the target problem laid out by the use case.

What it might look like

Strong data management practices and the application of recognised data standards among data controllers are indicators that the data is likely to be useful. An additional initial indicator is if the people who use the data currently believe that it is reliable and trustworthy. However, specific requirements of the data for future users need to be determined and assessed. This may pertain to the format and structure of the data—ensuring, for example, that it is machine-readable and has the required annotations.

Getting there

Understanding and meeting the data needs of users will involve consulting with potential users and assessing the data that will be shared with respect to their requirements.

Data assessment: what makes data AI-ready?

The goal of a data assessment

There are four primary dimensions to be considered when evaluating the appropriateness and usefulness of a dataset for a use case:

- ***Quality and completeness****:* This assesses how accurate and complete the dataset itself is.
- ***Relevance to use case:*** This assesses how well the data conforms to the needs of the particular use case.
- ***Ethics and legal compliance:*** This assesses whether the dataset has been appropriately created and can be made to conform to relevant standards of privacy and security.
- ***(Pre-)processing Needs:*** This assesses how well the data is formatted for the particular use case and whether additional cleaning needs to be conducted.

A full breakdown of these dimensions, including prompt questions for the team to carry out their own assessment, is provided in the Supporting Resources section.

Importantly, the goal of this assessment is to ensure data meets the needs of users. It is separate to a data impact assessment, which may be required in addition as part of establishing the ethics and legal compliance dimension.

Tailoring a data assessment:

The features of the assessment need tailoring to the specific needs of the use case. Different types of data being used in different contexts require different standards of accuracy, entail different regulatory and ethical compliance needs, and require different formats. Here, it is useful to consider:

- the **type (and sensitivity) of the data** used — is it personal data, traffic data, geospatial?;
- the particular **use within AI development** — is it being used as a training set, testing data, or input data?;
- the type of **AI techniques being used** — is it being used for supervised learning, unsupervised learning, generative AI?; and
- the **purpose of the use case** — is it being used to inform policy, make decisions, or generate insights?

Carrying out a data assessment:

Ensuring that the datasets are appropriate for the use case needs to be done with potential participants in the initiative. The data assessment provided acts as a framework for conducting a workshop exercise with those stakeholders relevant to the use case. Actors responsible for collecting, processing, and using the data should be

present (likely to be data teams in the relevant sectoral agency or unit), as well as potential future users of the data provided by the government. Dummy data should be created where appropriate to share and evaluate the properties of the relevant datasets with external parties.

Use Case – Ethics

U8. Do intended AI applications align with government or other best practice guidance on responsible AI use and development?

Why it is important

The initiative should help promote AI development and use that brings public benefits, and that protects the rights and interests of individuals and communities.

What it might look like

If the team is considering a use case in which data is being shared for deployment of AI applications within the public sector, then the team uses either sector-specific or cross-government guidance on public sector AI use, or ethical principles for AI use in the public sector. If there are no government resources relating to this, the priority is that the team is equipped to consider the risks and implications involved in their initiative.

If the team is considering use cases for developing and deploying AI systems outside of government, for example for research or commercial purposes, then the team can consider how to steer the direction of these applications outside of government. For example, Taiwan's NHIA vets third party app developers who can gain access to patient and government health data through its Health Passbook. Similarly, researchers applying to access data from the UK government's Data Service need to submit an application form, detailing how their project aims to bring public benefits (UK Data Service, n.d).

Available Tools

If there aren't established government policies for public sector use of AI, there are existing tools that can be referred to, including UNESCO and OECD's Toolkit for AI in the Public Sector (UNESCO and OECD, 2024), or Oxford Insights' Trustworthy AI Self Assessment (Oxford Insights, 2023).

U9. Can the risks to data subjects and other affected groups be mitigated? *[Ethics]*

Why it is important

Each use case will carry a unique set of risks arising at different points in the sharing and reuse of data, and affecting different groups of actors. Some of the risks that are posed by the use case may not be direct consequences. For example, this includes wider societal impacts on groups being given access to data and those who are not, or how digital inequalities could be amplified.

What it might look like

There is a risk register or other documentation that records and addresses risks that arise at all stages of the initiative. Risks are interpreted broadly and include risks arising from features of the data being shared, the AI systems being used, and the social implications of AI applications.

Getting there

Risks can be identified using techniques such as data impact assessments, AI impact assessments, and consequence scanning. There may already be established practices in government.

> **Available Tools**
>
> If there aren't established government policies for undertaking risk assessments, there are existing tools to support these techniques such as the ODI Data Ethics Canvas (ODI, 2021), Doteveryone's Consequence Scanning tool (Doteveryone, 2021), and UNESCO's AI Ethics Impact Assessment (UNESCO, 2023).

At this stage, it is important to consider whether there are options for responding to a risk effectively. This is because not all risks are manageable and some may prevent the team moving forward with the use case; for example, if a use case risks infringing on fundamental rights. These mitigation options can be developed in more detail in the management and governance structure section.

Decision Stage: Mechanism

The mechanism lays out how decisions will be made about data access and use within the initiative and by whom. In doing so, it determines the legal and data infrastructure required to enact these decisions.

Why this decision stage is important:

Agreeing the mechanism for the initiative is important for establishing trust among participants (Frontier Economics, 2021). It does this by clarifying the:

- **Roles of participants:** it establishes who controls the data at different stages within the initiative and how decisions are made about control over data.
- **Legal agreements:** it identifies the data sharing agreements that are needed to enable the flow of data.
- **Data infrastructure requirements:** it sets the requirements for the data infrastructure needed for providing data access and ensuring privacy and security.

Additionally, the mechanism is important because it has implications for the use cases that are possible within the initiative and the complexity of the governance structure.

How to use this section of the checklist:

The data sharing mechanism taxonomy, provided in the Supporting Resources section, gives examples of how data sharing can be arranged depending on the number of participants involved and the desired flows of data between them.

If the team has not already decided on the kind of data-sharing mechanism it plans to use, then the taxonomy can be reviewed to identify examples that would enable the flows of data they require within the initiative. This part of the checklist can be used to specify participant roles, data infrastructure and legal infrastructure, so that the mechanism is tailored to the initiative and government setting.

If the team has already decided on a broad mechanism, then this part of the checklist can be used to clarify roles and interrogate the required or existing legal and data infrastructure that supports the initiative.

Mechanism – Strategy & Sustainability

M1. Does the initiative's financial model support a fair distribution of its benefits?

Why it is important

The financial model for the initiative lays out who funds the initiative and therefore partly determines how the initiative's economic value is distributed among participants. It is important to consider the fairness of this distribution.

What it might look like

The initiative may be fully or partially publicly funded:

- **Public funding:** funding for the whole initiative secured from within government or from external sources, such as a development bank.
- **Public-private funding:** public funding of the initiative is supported by charging for data use, data stewardship services, or products and services built using the data (Ada Lovelace Institute, 2021a).

Addressing fairness in the financial model of the data sharing mechanism:

(1) Does the financial model reflect the economic value created?

This would mean the model weighs the cost of taking part against the economic value to participants. If the team is charging for actors to take part then it is important to consider how economic value is generated, for whom, and at what scale in order to ensure that participants are charged at a cost they are willing to pay.

For example, data initiatives may create economic value for data users by allowing them to create products or services that they will go on to sell. In this case the scale of economic value depends on the user; some will be better at turning the data into financial gain than others. The scale also depends on the quality of the data being shared. If it is not shared in a readily reusable format then the user may have to invest in preparing it for use (GPAI, 2023a). A fair fee for data access will take these dependencies into account.

(2) How does the model affect the distribution of benefits?

This would mean the model distributes benefits in order to maximise the public value created by the initiative. The financial model partially determines the distribution of benefits by affecting who has access to the data and the products and services built using the data. Any fee might lead to the exclusion of certain actors. It is therefore important to determine which actors might be excluded and what the consequences of this exclusion might be.

The team can encourage broad access to the benefits of products and services built using the data. The team may consider putting conditionalities on their partnership to encourage partners to use the data for public benefit (GPAI, 2023a).

Developing a set of principles that the team would like to reflect in their financial model can help to guide their decision. This approach has been taken by the UK's National Health Service (NHS England, 2023).

Mechanism – Community & Stakeholders

M2. Are data controllers agreed on how data access and use is managed within the initiative?

Why it is important

How data access and use is managed includes establishing who has control over data at different stages across the initiative. It involves establishing how decisions get made and by whom (the individual, the organisation, or the group of organisations responsible). Establishing clear roles and responsibilities for participants throughout the initiative is important for creating trust between them as data is shared with new parties and for new purposes that participants may perceive as a risk.

What it might look like

All groups of actors have agreed upon how decisions are made and by whom. Furthermore, the consequences of these decisions about the control of data are also understood by all groups of actors.

This may be as simple as a government authority with control over all the data needed for a use case to transfer control to one other party. However, in other cases, there may be multiple data controllers, intermediary controllers, or multiple data users. Each of these actors will have different requirements and concerns about how control of data changes throughout the initiative.

Setting the terms of the data sharing initiative

When the team sets up an initiative to enable continued sharing of data between participants, then answering this question also involves setting the terms for who can take part in the initiative. For example, if the team is setting up a data commons, it will need to establish the rules for membership to the commons. Alternatively, if the team is using a data stewardship model then it will need to establish the criteria for accessing the data. Teams may consider the following questions:

Who can become a data provider?

- Data quality— do they meet data quality and formatting standards?
- Data protection and privacy—are they compliant with relevant requirements?

Who can become a data user?

- Purpose of use—is it, for example, for research or specific public interest criteria?
- Conditions of use—how is data accessed; e.g., is it only through a physical location?
- Obligations for monitoring of usage and outputs

How will each application be assessed?

- The process for how these two decisions about taking part in the initiative get made.

Example: INSIGHT (United Kingdom)

INSIGHT is a data trust run by the UK's National Health Service (NHS) to enable responsible data-sharing of highly sensitive eye health data to facilitate research. INSIGHT uses a three-stage approval for data access requests (INSIGHT, n.d.):

- *Initial Screening:* Ensures research adheres to the 'five safes' framework.
- *Ethical Evaluation:* Assesses the research's purpose, value, and benefits. The criteria include evaluating risk mitigation, public consultation, and ensuring the project does not disadvantage any group.
- *Final Approval:* Granted by INSIGHT based on the Data Trust Advisory Board's (DataTAB) recommendations. DataTAB meets every 2 months to discuss requests.

Getting there

It is essential for all participants to agree with how data access is used and managed in order for them to be willing to take part in the initiative. Therefore, agreement should be reached inclusively, and involve the engagement of all participants (Ada Lovelace Institute, 2021b).

Recommendation from roadmap pilots

It can be useful to build on existing responsibilities of organisations and ways of working between organisations. This is because negotiating new responsibilities and establishing new working practices requires significant engagement and they are harder to negotiate.

For example, if decisions about who gains access to the data need to be made collectively, then it may be simplest to take these decisions in existing fora. Alternatively, if decision-making is being delegated to one organisation, then choosing an organisation who already has the trust of participants within to carry out similar functions is appropriate.

M3. Are data controllers able and willing to supply data for the initiative? *[Community and Stakeholders]*

Why it is important

The initiative will make demands on data controllers, which may include preparation and publishing of data, negotiating data sharing agreements, managing data access requests. It is important to establish demands that meet the controllers' expectations in terms of time availability, digital capacity, resources, and willingness to participate.

What it might look like

The roles of current data controllers within the initiative are clear and agreed collectively.

Recommendation from roadmap pilots

The initiative is likely to create new responsibilities for data controllers. By discussing with potential data controllers their capacity to take on these responsibilities, the team can tailor the role of data controllers to their needs. This is particularly important if the initiative is looking to attract data controllers to supply data within the initiative.

Example: Integrated Data Service (United Kingdom)

The UK's Integrated Data Service (IDS), run by the Office for National Statistics (ONS), facilitates the sharing of data on society and the economy from government departments to accredited researchers.

IDS provides linked datasets from across the Government to its users but it offers 3 models for how government departments can provide and link data in order to accommodate their different needs (Office for National Statistics, 2023):

- *Donate model:* ONS data linking team ingest and link data on behalf of the data provider.
- *Deposit model:* There is a staging area inside the IDS platform that provides cloud space for the data provider to do the data architecture and linking.
- *Enable model:* use cloud virtualisation to bring together data but keep within the boundaries of the donating organisations cloud environment.

Mechanism – Legal & Governance

M4. Are there data sharing agreements between all necessary parties?

Why it is important

Data sharing agreements create clarity about control over data at each point in the initiative. This clarity is important for creating trust between participants. Shared understanding of the legal status of the agreements is required for ensuring participants can be held accountable.

What it might look like

There are data sharing agreements that make clear and legally bind the responsibilities and rights at each stage, for each actor. As an example, the European Union's Dataspaces Support Centre provides an overview of the contractual framework (Data Spaces Support Centre, 2024), and provides a summary of what each agreement contains and is intended to do. Similarly, this depository provides examples of full data sharing agreements made by government organisations.

Available Tools

Further resources to support the development of data sharing agreements include designing data sharing agreements: a checklist (Yates et al., 2018) and a database of

published data sharing agreements made by government organisations (C4DC, n.d).

Recommendation from roadmap pilots

It is easier to determine which data sharing agreements will be needed once the control of data along the publication chain has been agreed among participants (question 11 has been answered). Once this is agreed, mapping out flows of data, decision stages, and responsibilities of each actor can help identify where legal agreements are necessary.

Mechanism – Technology

M5. Is there data infrastructure that supports this data sharing mechanism?

Why it is important

Effective data sharing is underpinned by a data infrastructure that is appropriate for the kinds of use case(s) the initiative is supporting and the existing data infrastructure it is being integrated with.

What it might look like

The design of the data infrastructure to support a data sharing initiative will include:

- **Data integration:** the sources of data and how it is ingested into the data storage facility, including transformations that need to happen, and the standards these follow.
- **Data storage:** where the data is stored.
- **Data discovery:** how users find out what data is available and information about the data.
- **Data access:** how user access is managed and enabled.
- **Data services:** other services that the initiative provides to data users such as visualisations, analytics, and applications.

Examples of each of these components are explored further in this Table of data infrastructure components, found in the Supporting Resources section.

Getting there

To design the infrastructure, the team can work with a cross-functional data team (for example, including engineers, owners, users, and compliance staff) to understand current data infrastructure and develop potential future designs. Illustrations of data infrastructure designs that support different data sharing mechanisms are provided as examples in the Supporting Resources section.

What drives the design decisions in data infrastructure?

1. Current data infrastructure

- It integrates with existing systems.

2. The use cases

- It allows the approved users access to the right data through appropriate access controls.
- It supports the governance needed for intended use cases, including security measures, providing visibility of data flows, and facilitating the role of any intermediaries between data controllers and users.
- It enables data sharing at the scale and frequency that users require.

M6. Is it possible to share the data safely and securely? *[Technology]*

Why it is important

Appropriate privacy preserving techniques (for example, those set out in the GPAI PETs+ project[5]) and cybersecurity measures[6] are needed so that participants can trust the safety and robustness of the initiative. Insufficient measures could result in data leakage or privacy breaches that result in harm to data subjects, businesses and organisations, or national security interests.

The legal framework(s) being used to enable the sharing of data within the initiative may include privacy and security requirements for the sharing to be considered lawful. The team should consult legislative and regulatory guidance to ensure any requirements are met.

What it might look like

There are appropriate data security measures in place to protect the data against threats during both storage and transmission. The team has consulted government officials responsible for maintaining data security measures to determine the unique needs of the mechanism and develop a plan for putting them in place. These include firewalls, intrusion detection systems, and secure data transmission protocols to protect the data against security threats.

Technical governance techniques are used where appropriate to preserve privacy of data subjects and preserve other privacy interests, such as national security or sensitive business interests.

[5] *GPAI 2023. Overcoming Data Barriers Trustworthy Privacy-Enhancing Technologies, Report, November 2023, Global Partnership on AI.*
[6] *NIST Cybersecurity Framework (CSF) 2.0*

Getting there

Depending on the data shared, appropriate methods for ensuring data privacy need to be identified with colleagues or partners with technical skill sets.

Technical Methods for Preserving Data Privacy

Privacy-preserving technologies (PPTs) are an important set of tools that enable or facilitate the secure sharing of (potentially sensitive) data.

They ensure data subjects' privacy is protected while allowing for the sharing, storage, and use of such data; this assurance serves to build trust between data subjects, controllers, and users. PPTs can be particularly useful in government data sharing for AI given how they can enable the large-scale sharing and analysis of sensitive data—for example, individual-level data. The relevance and application of different PPTs are explored in more detail in the PPT resource found in the Supporting Resources section.

M7. Have data management and governance practices been established to ensure continued data quality throughout the initiative? *[Technology]*

Why it is important

Ahead of sharing data it will need to be processed so that it is in the format required for use and the formatting requirements depend on the needs of data users. Using common and existing standards can be important for the scalability of the initiative.

What it might look like

There is agreement on the data standards and formatting that will be used within the initiative and the responsibility for meeting these requirements among actors within the initiative.

Recommendation from roadmap pilots

Many data sharing initiatives involve the integration of data currently stored in different systems. In many cases, the current data standards applied will be different across organisations and systems, which creates challenges for data integration. Consequently, teams need to establish data quality and integrity processes and standards that define clear data formats and schemas to ensure consistency and facilitate integration. The implementation of these processes should be negotiated across teams, according to capacity and skill sets.

Data integration processes will include checks prior to movement of data between systems, including profiling, cleansing, and validation processes. These steps can be

automated, semi-automated, or performed manually. For example, checks may be implemented using data validation and cleansing tools that help ensure accuracy and consistency by identifying and correcting errors, duplicates, and inconsistencies.

Data integration is also supported by having centrally managed metadata standards and repositories to help understanding of data lineage, usages, and quality. However, in cases where data originates from diverse sources with varying standards, teams may take a centralised approach after collection of data from sources. There are different, open-source standards that can be used to achieve this by describing the structure and content of datasets.

Mechanism – Ethics

M8. Is the level of openness proportionate to the level of sensitivity of the data?

Why it is important

The appropriate level of data openness is determined by the level of sensitivity. The mechanism needs to ensure that it makes data available only to suitable parties and uses secure access mechanisms and supporting data infrastructure. However, the mechanism should at the same time recognise that government data has been paid for by the public, and therefore, should be made as open as possible. Opening data up to the public is also closely linked to, and sometimes included within, Access to Information laws, which may include proactive transparency obligations (Open Data Charter, 2023). Sensitivity level may be determined by privacy concerns, national security interests, or third party rights to the data.

What it might look like

Data openness is prioritised to enable as widespread access to the data as possible. This is to promote fair access and avoid unequal distribution of benefits unnecessarily. Additionally, more open data sharing mechanisms tend to require simpler governance structures, which can be easier to resource and run. However, where the use case(s) the initiative supports involve highly sensitive data, the openness of data should be limited and an appropriate mechanism can be chosen that supports restricted access.

Decision Stage: Governance Structure

The governance structure sets out who will be responsible for overall strategy and oversight for the initiative and how this will be achieved.

Why it's important:

The governance structure is about maintaining trust throughout the lifetime of the initiative among participants and the wider public. The governance structure achieves this through promoting:

- **Inclusivity:** bringing in all participants and affected groups and sharing the benefits of the initiative
- **Accountability:** ensuring independent oversight
- **Transparency:** creating visibility and proactively engaging with affected groups
- **Robustness:** putting in place continued monitoring and maintenance of infrastructure
- **Risk management:** establishing evaluation and escalation practices

How to use this section of the checklist:

This part of the checklist can be completed in order to evaluate an existing governance structure for the team's data sharing initiative or to guide the team in developing one.

Governance Structure – Strategy & Sustainability

G1. Is responsibility allocated for how the overall strategy of the initiative will be set, monitored, and steered?

Why it is important

To support the success of the initiative, responsibility needs to be assigned for making decisions about how it operates and overseeing its delivery. This includes setting the aims of the initiative, developing and enforcing rules for decision-making on data access and use, and developing a sustainable financial model.

What it might look like

Depending on the scale of the initiative and the data being shared, these responsibilities may be given to:

- one individual;
- a single or multiple groups in the form of boards or committees; or
- a trusted third party.

Examples of establishing responsibility for strategy and oversight

Initiative (Country)	Responsibility for strategy and oversight
Mobility Dataspace (Germany)	Germany's Mobility Dataspace (MDS), that facilitates the sharing of data between private and public actors in the mobility sector, is run by a non-profit company, DRM Datenraum Mobilität GmbH. The non-profit company is governed through three groups (Mobility Dataspace, n.d): ● A supervisory board elected by shareholders, which monitors and advises on the management of MDS. ● An advisory board made up of independent experts who advise on strategic orientation and new uses of the MDS. ● Shareholders including private and public sector organisations.
INSIGHT (United Kingdom)	INSIGHT is a data trust run by the UK's National Health Service (NHS) to enable responsible data-sharing of highly sensitive eye health data to facilitate research. The governance of the data trust is coordinated by the NHS, while an advisory board, Data Trust Advisory Board (DataTAB), provides independent review and advisory recommendations on all Data Use Applications received by INSIGHT (INSIGHT, n.d).

G2. Is there a budget secured for this data sharing initiative or options for where a budget could be sourced from?

Why it is important

The initiative needs funding for a pilot, and options for securing long-term funding, to ensure it is deliverable. Without this funding, it will be unable to keep the right people involved, procure the right technology for data infrastructure, or conduct engagement with wider stakeholders.

What it might look like

Funding may be government sourced, sourced from development banks, or privately funded. It may therefore come in the form of a budgetary constraint from central government funders, grants, or a private contract. While there may not be a guaranteed long-term budget, teams will need options for securing a budget for a pilot demonstration of the initiative. Teams should also begin to consider their options for securing the longer term sustainability and scalability of the project (GPAI, 2022b).

Governance Structure – Community & Stakeholders

G3. Are there reporting mechanisms from representatives of all actors affected by the initiative to the people responsible for its oversight?

Why it is important

Those responsible for strategy and oversight may not include representatives from all groups affected by the initiative. The representation and inclusion of these affected groups is important for ensuring the initiative works towards public benefits.

What it might look like

Establishing reporting mechanisms means creating routes for information to flow from affected groups to those responsible for strategy and oversight. This may mean having representatives of groups as part of a board, regular meetings, or through data collection methods such as holding surveys.The following groups are important to consider:

- senior leadership within the government or agency;
- core teams from the initiative;
- contributors to data;
- users of data;
- any data subjects; and
- regulatory agencies.

G4. Is transparency maintained throughout the development and implementation of the initiative? *[Community and Stakeholders]*

Why it is important

Transparency about the initiative is necessary for accountability. It is needed for participants to engage with and contest its management, the public to voice concerns, and for independent authorities to enforce relevant regulations. At the same time as facilitating accountability if there are concerns, transparency also creates trust and supports the team to raise awareness of the benefits of the initiative.

What it might look like

Achieving transparency requires different actions for different actors and use cases. Actions should enable actors to be well informed enough to trust the initiative and what this requires can be determined through engaging with the different groups of actors.

For some actors, this will mean sharing progress and communicating decisions. Depending on the context, examples include publishing governance board meeting minutes, or holding ‘show and tells’ for participants.

For data subjects specifically, transparency requires that they have visibility over who holds their data, for what purpose, and be informed about possible data transfers. This could be achieved through e-consent management systems (Republic of Estonia, 2024), or facilitated by access to information legislation, and implemented through information requests.

G5. Is there a plan to ensure a wide range of actors can access the benefits of the initiative? *[Community and Stakeholders]*

Why it is important

To maximise the initiative's contribution to solving the public value problem(s) it is targeting, teams can consider how the benefits can be spread widely between actors. Benefits here may include access to the data itself, monetary gains from data, AI systems, and products, services, or research developed using AI systems. It is important to interpret benefits broadly so that they can be distributed fairly among affected groups and the wider public.

What it might look like

The team has considered how to make the benefits accessible to as many actors as appropriate. How benefits can be distributed will depend on the data being shared, and what the outputs of the initiative are. However, some example approaches relating to different kinds of initiative are detailed below.

Examples of increasing access to benefits

Approach	Examples
Updating users on data availability	The EU's Copernicus provides news briefings about the publication of new datasets, which users can sign up to for alerts (Copernicus, n.d.). Many open data portals, and data-sharing initiatives with members share updates in a similar way.
Publishing outputs of data sharing	Outputs may include academic research or government studies. In the US, publications and research funded by taxpayers must be made immediately publicly accessible (The White House, 2022). Other forms of outputs may include insights that inform policy making that are for internal government use. Teams could consider how to visualise the data analysis and share findings within government, for example through dashboards.
Designing inclusive services	If data is being used in AI systems within government service delivery, human centred and user centred design techniques can be used to ensure the systems are integrated in a way that meets the needs of all user groups (Chen et al, 2024; Gov.UK, 2024).

Governance Structure – Legal & Governance

G6. Does the initiative collect consent from any data subjects and facilitate data subjects to withdraw their consent at any stage?

Why it is important

Consent is needed in almost all circumstances when personal data is being shared (GPAI, 2023a).

This is important for respecting the rights of data subjects, and in most jurisdictions, for complying with data protection and privacy legislation.

What it might look like

The specific requirements for collecting consent, and what is sufficient for consent, differs across jurisdictions. Requirements will be included in the jurisdiction's data protection and privacy regulation, and any accompanying guidance, as well as any sector specific regulations, for example, relating to health data. The team should engage with data protection regulators to understand how this applies in the specific case of the initiative.

Consent may be requested at the point of data collection or data subjects may later be given the option to either opt-in or opt-out of a data sharing initiative. Teams must ensure that the consent collected covers the scope and purpose of data processing involved in the use case(s) of the initiative. If the team is using an 'opt-out' approach to establishing consent, then this requires active communications with data subjects about the initiative and the option to opt-out.

G7. Is there a mechanism through which data subjects can seek redress for rights violations?

Why it is important

Even though risks should be mitigated and no unacceptable risks should be taken by the initiative, there still needs to be mechanisms for data subjects to raise concerns about their rights and seek redress if their rights are violated. This form of accountability is important for ensuring the initiative stays within legal and ethical boundaries of data sharing and AI use and development. It is also important to sustain trust in the initiative: individuals need to know they have recourse in case something does go wrong.

What it might look like

Mechanisms that enable data subjects to seek redress may be enabled by legislation, and facilitated by the initiative, for example by proactively publishing information about how to raise concerns. Examples include the EU's data protection legislation that restricts automated decision-making with respect to some decisions affecting data subjects (European Commission, n.d).

G8. Is there a channel for the relevant regulatory agencies to track the initiative and give guidance on its development?

Why it is important

The inclusion of regulatory agencies throughout the development and delivery of the initiative is a critical dimension of independent oversight (GPAI, 2023a).

What it might look like

The team has given full disclosure to oversight authorities for their review of the design of the initiative and to receive guidance on its implementation, even when regulatory frameworks do not require such disclosures. Processes are in place for continuous communication and review of the

initiative with regulators throughout its lifetime.

Recommendation from roadmap pilots

Having representatives from the relevant regulatory agency or agencies in the room in key meetings can speed up the delivery of the project. This is because they have experience of applying the relevant legal and regulatory frameworks. This can help make the possibilities and boundaries of the initiative clear from the beginning.

G9. Are there processes in place for monitoring compliance with data sharing agreements?

Why it is important

This is important for managing risks that may arise from parties not complying with the conditions of data use established in data sharing agreements.

What it might look like

There are multiple routes to monitoring compliance.

Examples of mechanisms for monitoring compliance to data sharing agreements

Type of mechanism	Monitoring mechanisms
Business processes	• Destroying data and deactivating access to data sharing environments at the end of the licence period. • Setting up regular reviews of agreements to ensure they continue to meet the needs of involved parties.
Technical features of the data sharing environment	• Automatically recording and logging data access and processing by third parties. • Preventing data downloads to personal devices or copy-pasting data outside of a data sharing environment.
Legislation	• Disclosure control policies to assess outputs of data use.
Regulatory enforcement	• The regulatory agency has power to conduct audits of authorised processors.
Whistleblowing	• Creating routes for actors to raise concerns about possible non-compliance.

Governance Structure – Technology

G10. Are there processes in place to monitor infrastructure and maintain safe and secure sharing of data over the lifetime of the initiative?

Why it is important

For participants to be willing to share data within the initiative, there needs to be robust and reliable data infrastructure. This requires monitoring and maintenance overtime.

What it might look like

The data infrastructure includes servers for data storage, processing, and enabling access for data users. Continuous monitoring may involve the use of performance monitoring tools and conducting security scans. The processes the government takes on itself will depend on the type of server being used and its location.

Processes are in place to ensure that access is given to authorised users and that access is limited to authorised users. Processes will depend on the access mechanism, for example, APIs will require regular health checks.

G11. Is the government equipped to conduct a fair procurement process that evaluates AI systems to ensure they meet the requirements of the initiative?

Why it is important

Many data sharing initiatives will involve the procurement of external partners to use government data to deliver AI-driven products or services. The government needs to be in the position to conduct a fair procurement process that effectively evaluates AI systems to ensure they can deliver against the requirements of the initiative. This includes functional and performance requirements as well as ethical and safety requirements.

What it might look like

For the procurement process to be fair, there should be standardised procurement processes that are visible to, and interpretable by, the public. There may be provisions to encourage small companies to bid.

For the effective evaluation of AI systems, there are evaluation processes in place that are appropriate for whether AI systems are being deployed inside or outside of government and for the tasks the systems are used in. Using a system inside of government will require greater transparency from the system provider. This could include the use of model cards or other model reporting techniques during procurement.

There could be an assessment conducted of suppliers to government to ensure they comply with the Governments own, or internationally recognised, ethics frameworks and service standards.

Conducting evaluations requires that procurement officials have access to expertise in AI technology and its risks. They may have access to other officials with relevant technical expertise, had specialised training, or have guidance to follow.

Available Tools

Resources that can support teams to develop AI system evaluation processes as part of their procurement process include AI Procurement in a Box (World Economic Forum, 2020) , AI Toolkit Procurement Guide (Tony Blair Institute for Global Change, 2022) and we also refer teams to the FAIR Principles (Go Fair, n.d.), Model Cards for Model Reporting (Mitchell et al., 2019),

Governance Structure – Ethics

G12. Are there processes for monitoring and responding to risks to data subjects and other affected groups?

It's important to establish processes for monitoring and responding to risks as the initiative is implemented, which can be reviewed by those responsible for over strategy and oversight. This is ultimately to ensure that risks to the rights of data subjects and other groups are prevented, and in doing so ensure that the initiative achieves its stated public interest benefits.

What it might look like

The processes will be determined by the unique risks identified within a risk register, introduced in question 9, for the initiative. However, the team will have agreed on how the risks will be monitored and will have developed an escalation process for how to respond to any risks if they arise.

Monitoring risk development can be enabled through conditions established in data sharing agreements. For example, agreements can require publication or reporting of outputs from data use back to the initiative.

Collaborators are likely to face a trade-off between operational efficiency and oversight. Responding to this trade-off will require prioritisation of risks so that oversight is built around those considered critical.

G13. Does the initiative require human oversight for any algorithmic decision making?

Why it is important

This question addresses the need for human oversight of how the outputs of algorithms feed into government decision-making, either within public service delivery, back-end operations, or in the form of data analysis that informs policy decisions.

Human oversight of the outputs of AI algorithms and how these outputs are used in decisions is needed to address machine errors, and prevent adverse effects to data subjects, such as being excluded from a public service.

Additionally, delegating authority over a decision that affects an individual to an AI system may be impermissible in contexts where administrative justice frameworks exist (GPAI, 2023a).

What it might look like

Those responsible for strategy and oversight specify how algorithmic outputs will be used in decision-making so that human oversight is maintained. Data may be shared for a specific application of AI in government, in this case, the specification of how to use the outputs can be tailored to the use case. For example, this may mean providing guidance on how public servants should interpret the outputs of AI systems, including their error margins and limitations.

In cases where data is being shared for use in AI development more broadly, for example by researchers and organisations outside of government, the specification may be in the form of guidelines that users agree to before accessing the data.

Available Tools

Further resources that can support teams in determining how to maintain human oversight in their initiative includes Towards meaningful oversight of automated decision-making systems (Digital Future Society, 2022)

Also see the GPAI Report, Algorithmic Transparency in the Public Sector project report, 2024, to be published for the GPAI Summit, December 2024.

Pilot Case Studies

The goal of the second phase of the project was to design a set of guidance materials—a checklist and roadmap—that government organisations could use to facilitate data sharing for public-purpose-driven AI. In order to do this, we identified the need to test any guidance we develop with those involved with data sharing for AI projects on the ground.

We established three pilot partnerships to conduct the testing of the materials. The partnerships provided pilot countries the opportunity to work through and progress existing data sharing projects. They also provided them access to expert advice across governance and technical infrastructure for data sharing, helping them to develop internal capacity.

Alongside this, they allowed us to test the roadmap with those who would, ideally, be using it and identify any areas of improvement. We were able to test both the content and structure of the roadmap, clarifying both whether the actual information we provided was useful and whether it was conveyed in the most appropriate way.

How we chose pilot partners

To ensure that this work had wide reach and relevance—across economic, geographic, government-level, and sectoral lines, for example—we developed a set of criteria for ensuring that we had a representative spread across pilot partners. We released an Expression of Interest and had over 20 responses from federal, national, and city-level government agencies across a number of countries and with a breadth of interests. We scored countries' responses, looking at the capacity they had to engage meaningfully with all stages of the interim roadmap that we had developed. Further information on this process is presented in the Methodology section, Annex 2.

Following this engagement process, we decided upon three pilot partners:

1. Jigawa state government in Nigeria;
2. Türkiye's Digital Transformation Office; and
3. Agency for Electronic Government and the Information and Knowledge Society (AGESIC) in Uruguay.

These pilot partners represented a breadth of national and federal government levels as well as interest in a breadth of sectors and use cases. Each engagement followed roughly the same structure: regular catch-ups with partners; a selection of workshops on sections of the roadmap; and one-to-one sessions with data and policy experts from Oxford Insights.

The rest of this section will give a description of each individual pilot, explaining the:

- **specific country context**: motivations and starting point;
- **output of the roadmap implementation**: the problem the use case is trying to solve, the type of mechanisms discussed, the governance involved; and
- **pilot activities we conducted with countries** and which stakeholders were involved.

Learnings from Pilot Roadmap Implementation

In implementing the roadmap with our pilot partners, we identified a number of key opportunities and challenges for teams undertaking data sharing for AI initiatives, and an understanding of how the roadmap can support teams through their initiatives.

Opportunities

The roadmap is an opportunity for teams to accelerate existing digital initiatives by providing a clear guide for moving forward and for systematically assessing decisions that had already been made in the initiative, if it has already begun. Uruguay and Jigawa both started with sectors they wanted to develop initiatives within. They both wanted to begin by narrowing down from sector to a use case and determine a data sharing mechanism based on this. Türkiye started with a mechanism, and chose use cases based on what data could be shared given the infrastructure and the legal framework that supported this mechanism. We updated the roadmap to accommodate the different entry points partners had to the roadmap and different preferred sequences of decision-making.

Data sharing for AI initiatives requires strong cross-government collaboration, and sometimes the allocation of new responsibilities for stakeholders. The initial assessment in the roadmap can support raising awareness about the opportunities for data sharing for AI with potential cross-government partners. It is useful to keep up this level of engagement throughout the initiative so that roles in the initiative, including data preparation and access management, can be agreed more easily.

There are many emerging models of data sharing for AI and these can be tailored to the needs of the initiative and governmental context. Experimentation within models for data sharing on small and large scales across governments can be useful as examples for how teams arrange the roles of participants and procedures within their own initiatives. However, it is important to tailor these to specific context, including the data infrastructure that exists, what organisations are highly trusted in this domain, and what existing channels of communication there are between parties.

The different stages of the roadmap can be implemented simultaneously. For example, deciding to share sensitive data may raise concerns about whether the team has the capacity to set up the strong technical and operational governance structures required. They may want to address these concerns early on before progressing.

Challenges:

The novelty of data sharing for AI initiatives means that teams have limited experience applying applicable legal frameworks. Even when there is a clear national legal and regulatory framework, there can be limited experience of applying this framework within the team, and to the specific case of a new initiative. This increases the importance of teams seeking advice directly from regulatory agencies, but also points to the need for these agencies to develop applicable and reusable guidance on data sharing within their jurisdictions that can be used as starting points for teams.

There is strong interest from teams in applying emerging technical governance techniques but limited implementation know-how. For example, teams are eager to adopt privacy preserving techniques, such as federating learning or synthetic data, that can enable them to share sensitive data with greater certainty of safety. However, these are new techniques and teams, even if they are aware of them,

do not tend to have experience in implementing them. This challenge calls for the sharing of best practice from governments who are employing these techniques and for greater opportunities for teams to upskill their technical staff in emerging technologies.

Roadmap limitations:

A key limitation of our design process for the pilots was that, because of time constraints, we only engaged with governmental stakeholders. This created a limit on how far we could progress through the design of the initiatives. In all cases, our pilot partners are undertaking external engagement in subsequent phases of their projects.

Another limitation of the roadmap is that, as a tool intended for global use, it cannot address challenges that result from unique government contexts. As a resource intended for use globally, it needed to be, to the best of our knowledge, compatible with all legal, institutional, and technical contexts. The roadmap aims to mitigate the drawbacks of this need by providing practical examples of how governments have responded to challenges within their own contexts.

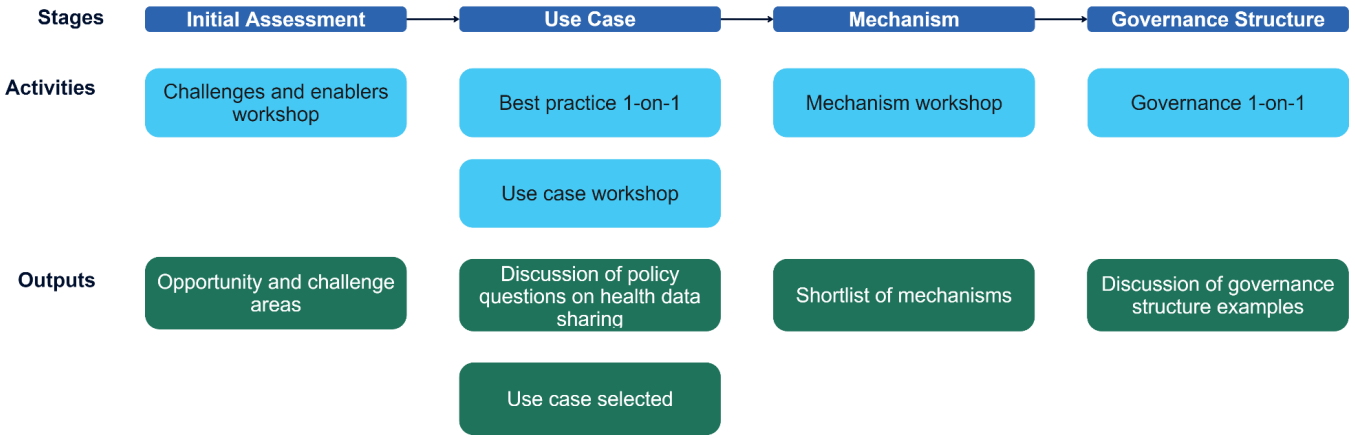
Case Study: Uruguay

Motivations and starting point

Uruguay's Agency for Electronic Government and Knowledge Information Society (AGESIC) responded to the Expression of Interest we issued, expressing interest in data sharing to support the development of AI tools aimed at addressing climate change, as well as data sharing for digital health. They had several publicly accessible climate datasets and were interested in exploring ways to share health-related data with non-government stakeholders (academia, private companies, private health institutions).

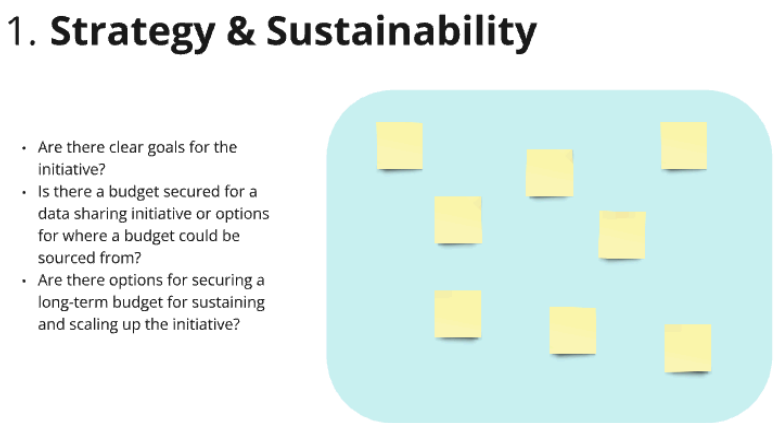
Pilot Activities and outputs

We held three workshops with the AGESIC team alongside two detailed 1-on-1 sessions with experts in Oxford Insights. All sessions were attended by the Digital Health and Technology teams within AGESIC and the Uruguayan Data Protection Authority (URCDP).



Initial Assessment

We began work by conducting a strategic workshop based on the first section of our draft roadmap to determine capacity and strategic priorities within government and AGESIC.



In particular, we looked at challenges and enablers across sectors. We produced a Miro board with opportunity and challenge areas across sectors. Opportunities related to:

- Already having been approached by some interested parties about gaining access to data in some sectors
- Robust legal framework in place for data exchange and personal information
- Experience in forming data sharing agreements
- The government's interoperability platform allows public organisations to exchange data securely.

Meanwhile, challenges centred around:

- Funding depending on the budgets of other departments
- Limited public understanding of the benefits
- Limited experience of applying relevant legal frameworks in the context of AI use.
- Obtaining and managing informed consent

Although climate change was also of interest, following the strategy workshop and independent work on the initial stocktake and sector selection sections of the roadmap, AGESIC ultimately decided to focus on digital health. Uruguay's Ministry of Health holds a great deal of data from a variety of sources including clinics and hospitals. They were aware of its value and the potential that it had to improve a number of different processes/modes of analysis, including public health and disease tracking. They were, however, unclear on how to structure the data sharing programme that would allow the data that was centrally held by the Ministry of Health to be distributed to other parties who would be able to make use of it and in what ways.

Use case

We held a workshop where we worked together to map the relevant stakeholders within the health sector, data flows, and narrow down the specific kinds of problems that might be solved within the sector. This resulted in our narrowing it down to data sharing in health predictive systems; for example, disease prediction and diagnostic medical imaging. Two main factors influenced the use case decision. Firstly, there was a pressing need to enhance disease prediction systems to improve strategic planning and resource allocation. Additionally, the Ministry of Health held relevant data that could be leveraged by AI developers to build these systems. The use case met the checklist requirements of strong data demand, an AI-ready data supply, and addressing strategic priorities.

This use case decision raised a number of governance considerations specific to health data sharing that the team wanted to consider before moving forward. In particular, considerations relating to:

- The factors affecting the scope of how data is used by recipients, including which kinds of data is shared (Personal Identifiable Data (PID), anonymised Data), who the data users are (academia, researchers, private companies), how decision-making takes place (data trustees, cooperative members), what data access controls in place, what data providers have consented to and how (opt-in/opt-out models).

- Ensuring that data that has been shared is handled safely through (1) compliance with National Data Protection Legislation and (2) adherence to Data-Sharing Norms.

Mechanism and Governance

Having determined a use case, we then worked through the next phases of the roadmap: we worked to identify a data sharing mechanism that would be best suited to this kind of data sharing using the roadmap along with real-world examples of health data sharing such as the Swiss data cooperative MIDATA and the UK's NHS data stewardship initiative INSIGHT.

This led to our choosing three potential data sharing mechanisms: data trusts, data stewardships, and data commons based on the suitability of these mechanisms for factors such as the sensitivity of data, the role of participants in decision-making, and the role of government in the initiative.

Dimensión	Data Trust	Data Stewardship	Data Cooperative	Data Commons
Decisiones sobre acceso y uso de datos	Centralizado	Centralizado	Descentralizado	Descentralizado
Confianza entre poseedores y usuarios	Bajo	Bajo	Medio	Medio/Alto
Sensibilidad de los datos	Alta	Media/Alta	Media/Alta	Baja
Nivel de involucramiento de los poseedores	Bajo	Medio	Alto	Alto
Capacidad digital de los poseedores	Baja	Bajo/Medio	Alta	Media/Alta
Rol potencial del gobierno	Proveedor de infraestructura, regulador	Coordinador, regulador	Coordinador, facilitador	Proveedor de infraestructura, facilitador

Finally, we held a workshop working through the practicalities of each of these mechanisms applied to this use case, using international examples and drawing out learnings for the initiative. At this point, AGESIC made the decision to pause work through the roadmap as they felt it would be necessary to return to stakeholders – in particular, the Ministry of Health, because it has the custody of the citizen health data – and conduct more detailed research to determine the most appropriate mechanism, before progressing.

Case Study: Jigawa

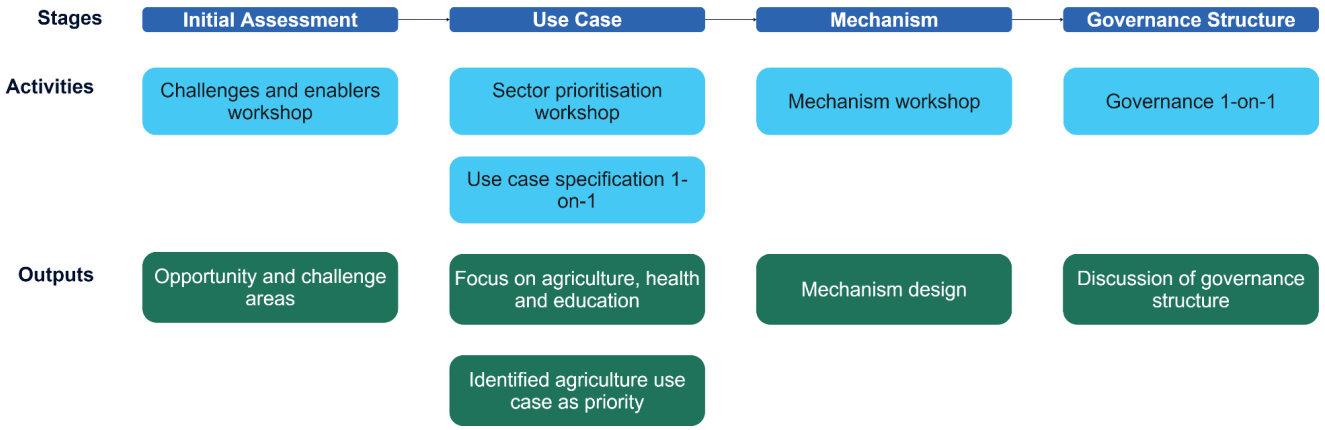
Motivations and starting point

Jigawa approached us with a diverse, multidisciplinary team, with members spanning a number of sectors and expertise. At the beginning of the project, the team had begun leveraging data in collaboration with a number of international organisations, civil society organisations, private sectors, startups and had some data sharing practices established for sharing across teams within the government and out of government. However, they were looking to accelerate general

digitalisation within the state government and use AI to solve developmental challenges. They worked with us to help identify the most promising sectors/use cases for sharing data to develop AI tooling, and create a tailored mechanism for decision-making within the programme.

Pilot Activities and outputs

We held three workshops with the team from Jigawa alongside two detailed 1-on-1 sessions with experts in Oxford Insights. Some sessions were also attended by representatives of agriculture, workforce planning, health, and education units within the state government.



Initial Assessment

We held a challengers and enablers workshop to identify cross-cutting challenges and enablers as well as strategic priorities. These were developed into a summary document for Jigawa to use to inform future decisions about sector prioritisation and use case specification. Areas of opportunity highlighted were:

- The team has established relationships across teams and agencies.
- There is leadership buy-in and the project has significant political support from high level decision makers.
- Data is being collected in digital format that may be helpful to responding to challenges in a number of sectors, including health, education, land administration, judiciary, civil service and human resource administration, financial management and agriculture.
- There are some ad hoc data sharing practices in place between government teams, typically performed through file transfer. Looking ahead, there is an ambition to create central data storage and facilitate access across government using APIs.
- There is momentum around developing policies to support a transition to a more digital government and the government is beginning to introduce data management policies that will support standardisation of government data.
- Agencies are established to regulate and accelerate adoption of ICT and Digital Economy projects in the state (Jigawa State ICT and Digital Economy Agency and JIgawa State Residents Registration Agency)

Meanwhile, areas highlighted as potential blockers were:

- The government has many priorities and does not have available funding to support them.
- Limited access to AI skills means the team would need delivery partners to build and implement AI applications within the government.
- There is still limited awareness within some teams about transitioning to a digital government and the risks and governance practices that will need to come with it.
- Most government operations and services are still 'analog'. This limits the data available, and there is limited existing data sharing infrastructure that can be drawn on to facilitate this new project.
- There is a Nation-wide data protection authority responsible for enforcing the Nigeria Data Protection Act and Jigawa has made contact with the Agency for help in Data Governance. However, it is a new agency with limited experience and a relationship with the team needs to be established.
- The government does not have established policies or practices around data management or data governance, nor experience of conducting data and AI risk assessments, which would be necessary for developing the initiative responsibly.

These challenge and opportunity areas pointed to a number of takeaways for the next steps in Jigawa's data sharing journey. A focus on a small and manageable use case pilot that can be implemented given existing capacity and data availability is appropriate. This focus on undertaking a small pilot also could be used as an example to raise awareness and as a starting point for state-wide policies on data governance and management.

Use case

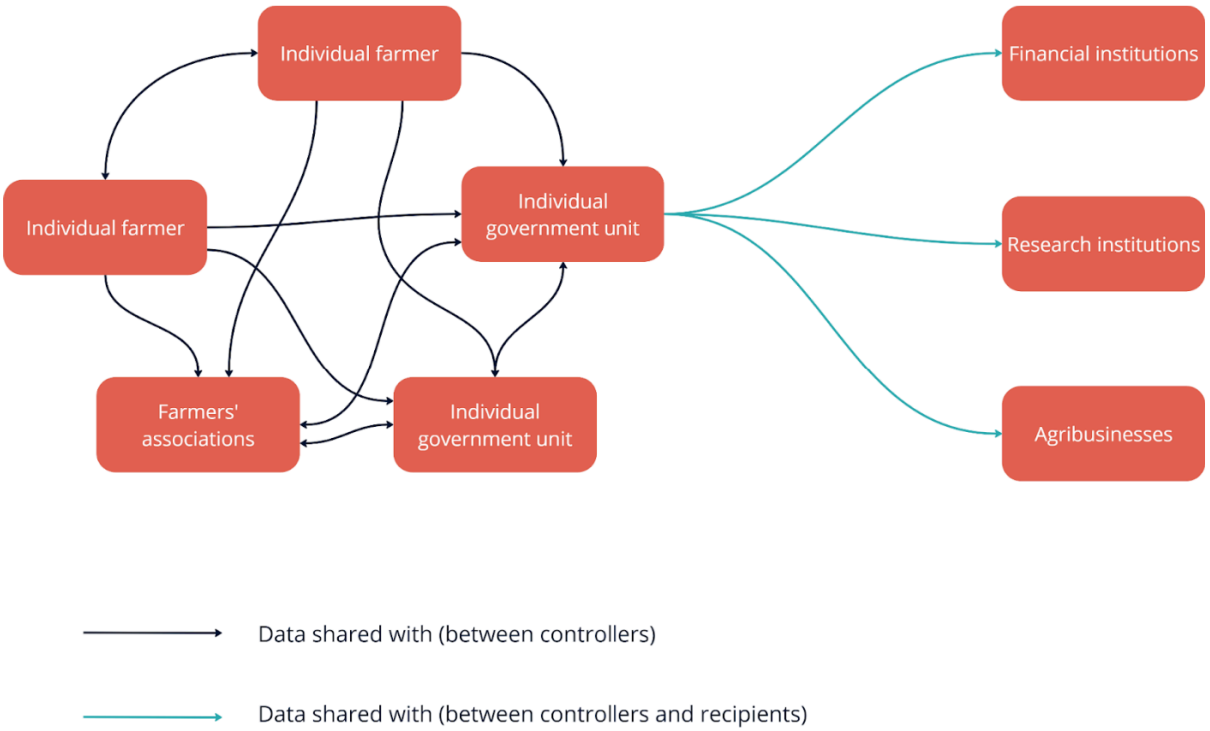
We held a sector prioritisation workshop where we worked through the roadmap to identify the most appropriate sector for a single use case. Here, Jigawa had a number of sectors—agriculture, health, and education—which seemed feasible: they were strategically important and had strong stakeholder engagement. Following a workshop discussing the most important factors for sector choice, Jigawa conducted an assessment of each sector using the checklist. However, each sector showed potential and had strong backing from the respective sectoral teams.

Therefore, we co-produced a document walking through the specific features of potential use cases within each sector, given the existing digital projects underway and available data in each sector. We identified the specific target problems, the relevant data to be shared, possible AI applications, and data recipients. Based on this specification it became clear that the agricultural use case was the furthest developed in terms of data collection, infrastructure development, and having established channels of communication between stakeholders. Given the previously mentioned takeaways from the initial assessment stage of the pilot, the team chose to focus on the agricultural use case for the rest of the pilot.

In the use case workshop, areas of further action were identified before implementing the use case. These included engaging with colleagues from legal teams and the Nigerian data protection agency to receive guidance on how legislation applies to their use cases, considering the risks involved in the agriculture use case, engaging with actors in the agricultural sector to understand what specific data they are interested in and what their access requirements would be.

Mechanism and Governance

We held a workshop with Jigawa to identify relevant data sharing mechanisms. We conducted more detailed data flow mapping and explored issues related to trust, data sensitivity, role of government, and level of involvement of data controllers.



This led the team to decide upon a hybrid model of a data steward and a data cooperative that suited their existing ways of working with external stakeholders as well as other government units. The mechanism involved one government unit coordinating the initiative but involving a number of different government units in decision-making around what data is shared, with whom, and for what purpose, as well as using existing fora to engage with farmers associations on a regular basis.

Finally, we held a governance 1:1 with Jigawa where we discussed what running this mechanism might look like in practice. The takeaways in this final stage were that the next steps included engaging with data users, data controllers, regulators, and the API development team in order to seek agreement on roles within the initiative.

Case Study: Türkiye

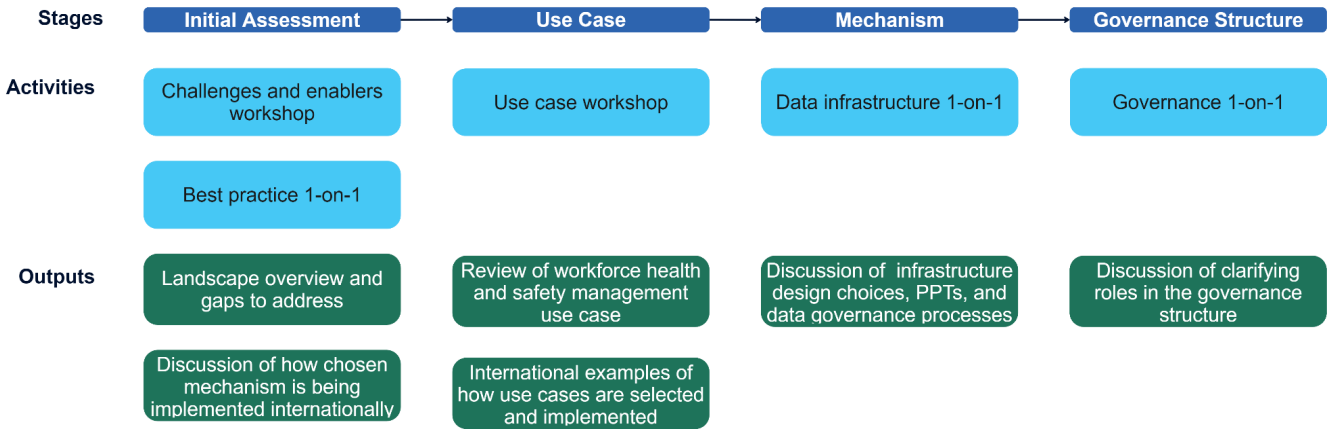
Motivations and starting point

Türkiye's ambition was to progress in their work developing a Public Sector Data Space (PSDS) and the implementation of their National AI Strategy as well as the drafting of the Turkish National Data Strategy .
Their PSDS aimed to improve access to high-quality government data, develop more robust data sharing infrastructure across government, and improve the data governance culture among public institutions.

The Digital Transformation Office of the Presidency of the Republic of Türkiye (DTO) would be coordinating and overseeing the progress. They have an overarching role within data governance across government in coordinating and guiding the other institutions and organisations aligning common strategic goals. The Turkish Statistical Institute (TurkStat) is primarily responsible for the technical implementation of the PSDS, serving as the PSDS Operator. Work had already begun on this project and this pilot sought to address a number of challenges they faced and provide insight into best practice for structure and governance of data spaces.

Pilot Activities and outputs

In this pilot project, the DTO was the main responsible institution and contact point. Alongside TurkStat, the Personal Data Protection Authority, and the Ministry of Industry and Technology they formed the core team participating in the meetings and reviewing the work (including checklists) prepared by Oxford Insights. Some workshops were also attended by wider stakeholders from across several government ministries, including TurkStat, the Personal Data Protection Authority, the Ministry of Health, the Ministry of Environment, Urbanisation and Climate Change, the Social Security Institution , Insurance Information and Monitoring Center , and the Ministry of Labour and Social Security. We held two workshops with the DTO stakeholders alongside three detailed 1-on-1 sessions with experts from Oxford Insights.



Initial Assessment

The pilot began with a challengers and enablers workshop, attended by representatives from several ministries across the Turkish government, with each ministry presenting the data sharing and data governance practices adopted within their departments. Since the PSDS project was

already underway, this exercise was useful for raising awareness among cross-government stakeholders about the project. It produced a landscape overview of data sharing across government, the different levels of progress across departments, and helped identify primary gaps to address. A few challenge and opportunity areas emerged.

Firstly, until the PSDS project, data sharing had largely been coordinated at the departmental level, which meant different data standards had emerged across government. The PSDS project needed to support data providers within the project to make the data usable across government. This work was already being supported by the National Records System project run by TurkStat, that aimed to produce data catalogues and standardising data assets.

Secondly, an advantage of how the PSDS project was being undertaken was that the multidisciplinary core team included legal experts who advocated for a 'privacy by design' and "data protection by default' approach to the data architecture decisions from the beginning. In particular, this resulted in them opting to share microdatasets, in which identifying elements have been removed, preventing the identification of real and legal persons, as well as institutions and organisations about which data is being collected, within the PSDS. However, one challenge that still existed was exploring how the team could use other, new privacy preserving techniques, such as federated learning techniques, which are promoted within the European Union's AI Act while balancing the interests of ensuring efficient and effective data processing with the protection of privacy and security.

Building on this workshop, we held a 1:1 best practice session. Together, we looked specifically at some of the challenges the PSDS team identified, and used the EU's data space standard as an example for how technical architectures aim to address privacy, interoperability, and trust concerns within dataspaces. Introduction to international examples of related projects were useful for the DTO team to see how others were responding to similar challenges.

Use case

The DTO team had already decided a pilot use case for their PSDS: the sharing of workforce health and safety management data on the PSDS, ensuring that the shared datasets are anonymized and de-identified of personal data. We held a workshop with the Ministry of Labour and Social Security, where they presented on how this use case would facilitate better predictions of workplace accident risk levels using Artificial Intelligence. Using the roadmap checklist, the DTO team identified possible risks that may be encountered in expanding the use case scenario to different public institutions as well as further clarification needed within the team regarding secure and privacy enhanced data sharing and options for anonymization mechanisms.

Since the PSDS will function as reusable data sharing infrastructure for many use cases, we also considered in this stage how other similar projects internationally select and prioritise use cases. In particular, within the UK's data sharing landscape, we looked at the Integrated Data Service (IDS) and the Better Outcomes through Linked Data (BOLD). This session revealed key questions relating to data ownership, how it is defined, and how it changes throughout the chain of data

publication within PSDS. These questions were considered in the next stage of the pilot partnership.

Mechanism and Governance

Since DTO already had established the broad arrangements of their data sharing mechanism, this stage focused on key data infrastructure and governance questions the team had, and which the roadmap could be used to address.

Firstly, the roadmap helped clarify the category of mechanism the PSDS falls within, identifying the PSDS in development as a kind of many-to-one, one-to-many mechanism. This gave us a framework for talking about the specific roles of the different stakeholders working on the PSDS project. In particular, we discussed responsibilities for DTO, Turkstat, data providers and data users. Relatedly, we provided considerations for how control over data changes throughout the PSDS, and how this determines the legal agreements required to establish trust among actors.

Secondly, the roadmap, and additional guidance provided background on privacy-preserving technologies (PPTs) that might be most appropriate and important here, such as the use of differential privacy, anonymisation, pseudonymisation, and synthetic data. Alongside this, we discussed available tools that could be used to support the integration of data from different sources across government, and management of metadata.

These discussions prompted the DTO team to prepare to handle specific challenges they might face moving forward but also provided clarity on how to approach and mitigate risks, supporting them in strategic planning and decision-making capabilities within the PSDS context. The DTO team suggested the roadmap is a tool that can be used not as a ‘one-off’ but for throughout other phases of the project, especially when new participants join or when there are changes in the project's functionality or design.

Appendix 1: Supporting Resources

The supporting resources relate to specific sections or questions within the data-sharing for AI checklist. They are highlighted throughout the checklist guidance. The resources that can be found in this section are:

- Data Rights Explainer
- Template Quantitative Scoring Framework for Use Cases
- AI-Ready Data Assessment
- Data Sharing Mechanism Taxonomy
- Data Infrastructure: Components
- Data Infrastructure Design: Examples
- Privacy Preserving Technologies: Introduction

Data Rights Explainer

It is important to consider the extent to which data rights are established within the team's jurisdiction's data protection legislation and how legislation can be complimented through other policies to guarantee a comprehensive protection of data rights. The explainer acts as an introduction to the concept of data rights and ways in which legislation and policy frameworks can support them.

Explainer on Data Rights

What are data rights?

Emerging out of data protection legislation and case law globally, the concept of data rights for both individuals and collectives has become central to strong data governance and to the ends of achieving data justice (GPAI, 2022a).

Data rights acknowledge that data is often 'co-generated'. That is, it is contributed to by multiple parties: for example, people or groups who are the subject of the data, a company who owns an AI system that played a role in generating the data, or a company who processed the data in a new way.

Co-generation of data gives rise to some rights for contributing parties with respect to the data's use or the value it generates (ALI-ELI, 2021). American Law Institute-European Law Institute (ALI-ELI) have recommended recognition of the following non-exhaustive set of data rights:

- **Access:** right to access data, ranging from reading the data, processing the data, to full portability.
- **Desistance:** right to require that a party desist from particular data activities, including control, processing, and transferral of data.
- **Rectification:** right to require correction of incorrect or incomplete data.
- **Economic share:** an exceptional right to an economic share in profits derived from the use of data.

The applicability of these rights depends on the circumstance. Factors that determine applicability include the nature and scope of a party's contribution, legitimate interests of all parties, public interests, the balance of bargaining power between parties, and the type of data right (Daten Ethik Kommission, 2019).

How can the protection of data rights be evaluated?

When considering whether data rights are protected in a jurisdiction, it is important to consider how they are included in data protection legislation. Legislation may include rights explicitly, such as the right to data portability, but they can also be enabled through requirements about being informed of who holds data and data transfers, consenting to processing by third parties, and opting out of processing (GPAI, 2023a).

Non-legal instruments such as policy frameworks are also important for supporting data rights. For example, the African Union's Data Policy Framework contributes to the concept of the right to an economic share by calling member states to ensure Africans benefit socially and economically from the use of their data, and historical injustices and structural inequities are not perpetuated (GPAI, 2022c). It is important for teams to identify current and upcoming policy frameworks within their government.

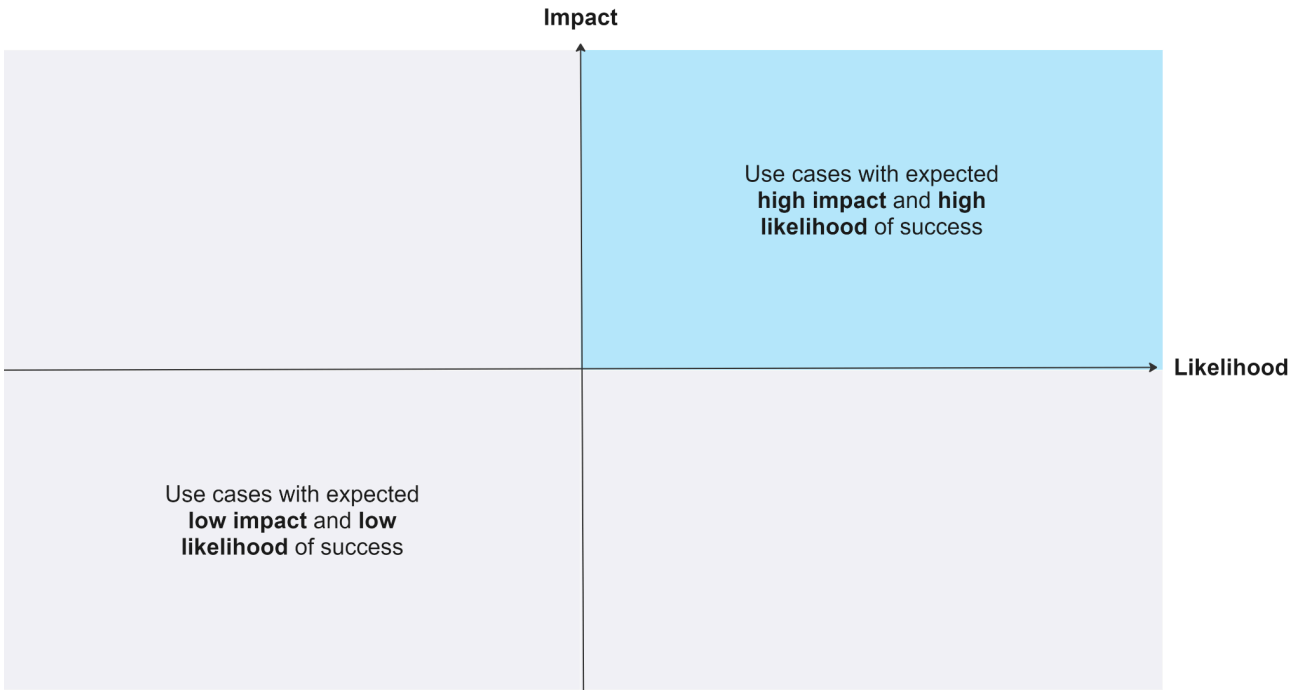
Template:
Quantitative Scoring Framework and Prioritisation Matrix for Use Cases

As an aid for comparing use cases, the checklist can be completed with respect to different use cases and a score can be attached to the outcomes. A template for a simple scoring framework is provided below. Answers can be scored based on the level of confidence the team has that the use case meets the checklist requirement. The total score is calculated by summing up the values attached to each answer. A score provides an overall indication of each use case's potential. However, there will likely be trade offs between the use case(s) that are selected to move forward with. When making a decision, the scale of the public benefits and the likelihood of successfully delivering the project are important dimensions to consider, and are visualised in the below matrix.

Template Quantitative Use Case Scoring Framework

		Confidence Score			
Decision Dimension	**Question**	**0** (No)	**1** (Somew-hat)	**2** (Yes, strongly)	**3** (Yes, very strongly)
Strategy & Sustainability	Does the initiative try to solve clear problems that are of public value?				
Community & Stakeholders	Is there demand for the data among intended data users?				
	Is there support for the initiative from the relevant team(s) responsible for the sector or area in which the data is being shared?				
	Is there support among the public for sharing data in this sector?				
Legal & Governance	Is there a legal framework that clearly sets out conditions for sharing data that apply to this initiative and are complied with by the initiative?				
	Does the data to be shared stay within what is necessary for the use case?				
Technology	Can the data shared in the initiative be made ready for use by AI developers?				
Ethics	Do intended AI applications align with government or other best practice guidance on responsible AI use and development?				
	Can the risks to data subjects and other affected groups be mitigated?				
				Total	

Impact-Likelihood Decision Matrix


Impact
Use cases with expected **high impact** and **high likelihood** of success
Likelihood
Use cases with expected **low impact** and **low likelihood** of success

AI-Ready Data Assessment

The table below breaks down the four dimensions to be considered when evaluating the appropriateness and usefulness of a dataset for a use case (Dama UK, 2013; Jolliffe et al., 2023; Ravi et al., 2022; ESIP Data Readiness Cluster, 2022; World Bank, 2019). As described above, this assessment can be tailored according to the data and use case. The table provides prompt questions to evaluate in a collaborative, or workshop, setting with potential participants in the initiative.

Data Assessment Prompt Questions			
Quality	**Relevance to Use Case**	**Ethical and Legal Compliance**	**(Pre-)processing Needs**
Accuracy: does the data accurately measure the real world concept it is intended to measure? Are its errors known and can they be modelled?	**Coverage:** Does data cover the entire population of interest for the relevant use case? To 'cover' a population of interest, it must simply be representative rather than cover a population in its entirety.	**Privacy**: Can the privacy of data subjects be protected through anonymisation and other privacy preserving techniques?	**Format:** Is the data stored in an appropriate format? Are they machine readable and easy to upload into an AI/ML system?
Granularity: Can the data be broken down by relevant subgroups within the population that it is measuring?	**Frequency:** Are data produced and/or updated at a rate that fits with what is needed for the use case?	**Regulatory compliance:** Does the data adhere to relevant laws and regulations governing data sharing and usage within this sector/pertaining to this use case and its associated data? If the legal basis for data sharing is consent, the consent has been appropriately gathered from data subjects who are supplying the relevant data?	**Labelling:** Does the data have detailed metadata that is human and machine readable? Does it have the annotations required for expected AI techniques?
Integrity: Are data sources, processing,	**Timeliness:** Is data updated soon enough	**Minimisation and proportionality:** Is the	**Interoperability:** Can data be linked to other datasets

and access management tracked and trusted?	after an event occurs for the use case?	data being shared the minimum necessary to fulfil the intended purposes it serves in the use case?	through common and consistent identifiers?
Quantity: Is there sufficient volume of data for the intended use in AI systems in the use case?	**Comparability:** Is data comparable over time and across locations? How well can comparisons be made with past data/similar data from elsewhere?	**Bias:** Is the data free from systematic bias that might lead to disproportionately negative outcomes for certain social groups?	

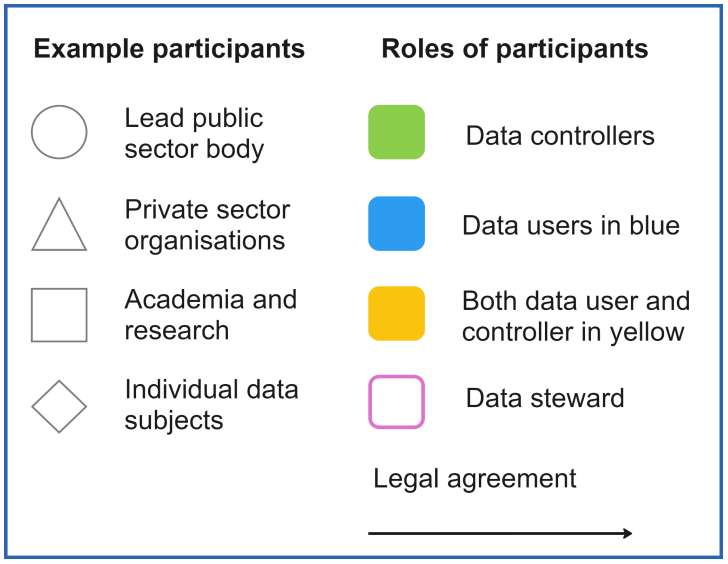
Data Sharing Mechanism Taxonomy

The data sharing mechanism taxonomy provides examples of the three categories of data sharing mechanisms introduced in the framework for the roadmap. The taxonomy breaks each category down into examples, providing a description of the data sharing arrangement, real-world examples, and detail on when each is commonly used.

- **One-to-one data sharing**: Agreements between two individual parties, who are users and/or controllers of the data.
- **One-to-many data sharing:** Agreements between one party, who is the data controller and many data users. The taxonomy also includes many-to-one, one-to-many data sharing. These are a kind of one-to-many data sharing arrangement, where there are agreements between one party, who represents data controllers, and many data users.
- **Many-to-many data sharing:** Agreements between many parties who are data controllers and many parties who are data users. These may be the same parties.

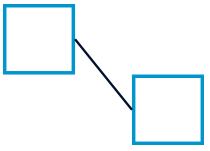
Illustrative diagrams in the taxonomy use the key below. In practice, the type of actors involved in the initiative, and their roles will differ.

Taxonomy Key



One-to-one data sharing

One-to-one data sharing mechanisms establish a connection between the government and another entity through which data can be shared. In the cases relevant to this project, government is always a supplier of data; it can also be a data user. In these cases, there is a multi-directional sharing of data between a government agency and another organisation.



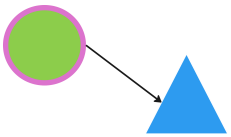
When are they most relevant?

One-to-one mechanisms are useful when a government is procuring research or delivering of a project with a clearly defined scope. It allows a government to engage closely with a particular entity to solve a small, well-defined problem or conduct a piece of research, amongst other things.

What are some examples?

Private contracts

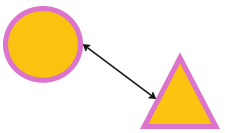
An agreement is made between a government agency and another organisation. This could be for a project that involves data sharing or simply for data itself.



Real world example: The "Rapid Response Register" is a shock response framework designed by the Nigerian National Social Safety-Net Programme (NASSP) in collaboration with the World Bank to provide emergency social interventions to poor and vulnerable households in challenging circumstances.

Public Private Partnerships (PPPs)

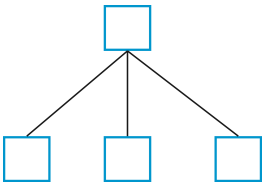
PPPs involve collaboration between government agencies and private entities to conduct what are usually large-scale endeavours. They are usually privately funded with the risk burden on the private entity. In cases like these, data is often shared in both directions, between both parties: government is both a supplier and user of data.



Real world example: The Towards zero-emission road transport (2zero) partnership is an agreement between the European Commission and partners from the automotive industry to accelerate the transition to a carbon-neutral road transport system by 2050.

One-to-many

One-to-many data sharing mechanisms function through multiple legal agreements between a single data supplier and data users. In relevant cases, the data supplier will usually be a government agency or department. Each agreement with the data user can be bespoke, though they may also be identical. In the case where they are not identical, they function as multiple one-to-one data sharing agreements/mechanisms which may be joined through a shared mechanism of access for ease, despite the different agreements with partners.



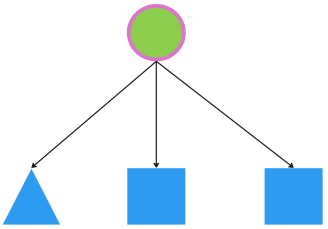
When are they most relevant?

These are useful when government holds data that it wishes to share with multiple, different parties. These may be academic, private, or third sector organisations, all of which may require different agreements and terms of use.

What are some examples?

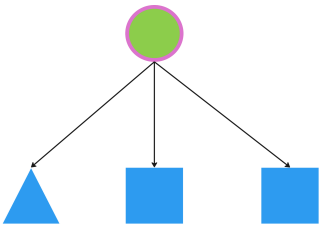
Open data

Data is accessible, exploitable, and shareable to/by anyone under an open data licence. In this case, agreements are all the same despite its use by multiple, different parties.
Real world example: datos.gob.cl, Chile's Open Data Platform, is a data repository where the government bodies publish their data under an open data licence.

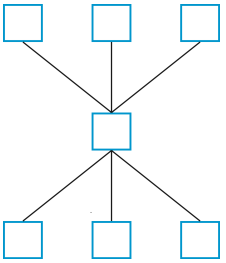


Sandbox

A data sandbox is a secure environment that allows users external to the government to access and use the dataset in a way that does not affect the integrity of the original data set nor compromise the security of the data.
Real world example: The Singaporean Data Regulatory Sandbox provides product developers with a controlled environment in which to develop their data-driven technologies, testing them against existing data protection laws.



Many-to-one, one-to-many



Cases of many-to-one, one-to-many data sharing mechanisms appear to mimic many-to-many relationships but are subtly different and require different governance structures and legal frameworks around them.

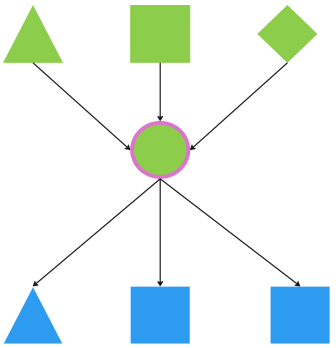
They contain a central node—displayed here as a government entity, though that is not always the case—which collects data from a number of different data suppliers and subsequently distributes it to multiple, different data users. The suppliers and users of the data act through this intermediary node and have legal rights over and responsibilities only to this entity. In many-to-one, one-to-many data sharing mechanisms, the suppliers and users do not engage in a direct agreement.

When are they most relevant?

These are useful when there are complex networks of data users and suppliers and it is difficult for them to engage directly. Many-to-one, one-to-many mechanisms allow an intermediary to act as a conduit, streamlining the interactions. Rather than all data suppliers attempting to interact with all data users, they instead each interact with a designated intermediary. As mentioned, this intermediary does not have to be a government entity. It could just as easily be a third party or special vehicle which is set up to be the middle mode in the above diagram.

What are some examples?

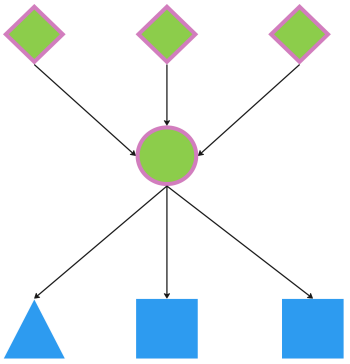
Data trust



A data trust is an entity/ legal structure where data suppliers/subjects grant authority to a designated organisation—sometimes a third party organisation set up specifically for the fact— to make decisions about sharing their data with other entities.

Real world example: The Climate Action Data Trust is a blockchain-based repository of all major carbon credit registry data to enhance transparent accounting in line with Article 6 of the Paris Agreement.

Data steward



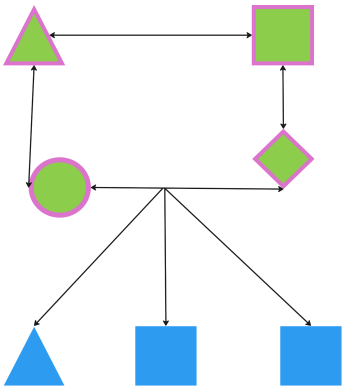
Data stewardship is a participatory concept of data intermediaries where third party actors – governments or NGOs, for example – are given authority by data subjects to facilitate the direct access to data they produce by parties that want to use that data.

Real world example: The Health Passbook in Taiwan where the government helps facilitate access by third parties to patient health data

Data cooperative

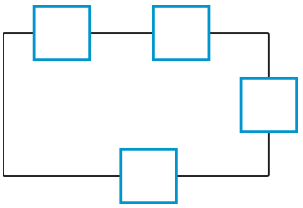
Members of a community jointly contribute to and manage their pooled data. All members maintain their data rights.

Real world example: MIDATA, a Swiss data cooperative established in 2015, gives users full control over their health data. Other examples of data cooperatives are the Scottish Agricultural Organisation Society (SAOS) or POSMO, another Swiss cooperative for mobility data.



Many-to-many

Cases of many-to-many data sharing mechanisms appear to mimic those of many-to-one, one-to-many mechanisms illustrated above. They are, however, subtly different and therefore require different governance structures.



Many-to-many data sharing mechanisms establish multi-party, multi-directional data sharing processes, where data often flows with clear processes for access within certain, predetermined actors who may be both suppliers and users. This sharing may be facilitated by a particular platform, but there does not exist a central node through which all data collection and subsequent sharing goes; instead many parties engage with many other parties concurrently.

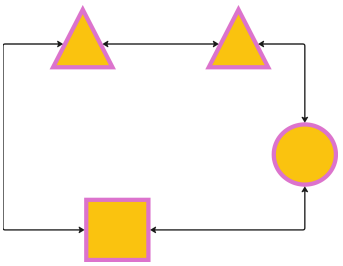
When are they most relevant?

They are useful when there are complex networks of stakeholders where direct engagement can be easily facilitated, as in the case of a many-to-many data marketplace, or where there are many parties within an ecosystem who are already aware of one another. Such mechanisms empower parties within the data sharing structure to engage with one another and exert agency over their particular data holdings/needs.

What are some examples?

Data marketplace

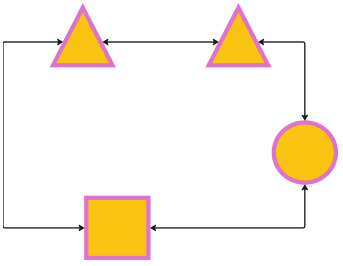
Data marketplace: A many-to-many data marketplace functions slightly differently from the one mentioned above. Instead of an intermediary collecting all of the data from suppliers and then distributing it to users, it simply acts as a platform to facilitate exchange. In this case, suppliers and users engage directly with each other *using* the platform provided.



Real world example: Dawex, a marketplace that allows actors from across the globe to source, exchange, and monetise data.

Data commons

Data commons: Members of a community jointly contribute to, manage, share, and use their pooled data. They may use a cloud platform or other secure environment to facilitate restricted access to relevant members of the community.



Real world example: Aclimate, an agricultural data platform in Colombia.

Data Infrastructure: Components

The table below breaks down the components of a data sharing initiative's data infrastructure and illustrates the options for infrastructure design that the team develops.

Data Infrastructure	
Data Infrastructure Component	**Example Options and Considerations**
Data storage: where the data is stored.	There are three kinds of architecture for data storage that data sharing infrastructure can rely on (either one or a hybrid): • **Centralised**: all data resides in a single repository, often a data warehouse or enterprise data platform. • **Federated**: data remains distributed across multiple systems, each with its own autonomy. • **Decentralised**: data is distributed across a network of nodes without a central authority. Each has different benefits. Centralised architecture may be used when there is, or can be, a single controller of the data. Federated architecture may be suitable when cross-system access is needed but there are multiple data controllers and maintaining autonomy is important. Decentralised architecture may be used when security, privacy, and transparency, are critical within a peer-to-peer data sharing ecosystem. What about hybrid architectures? Hybrid architectures may enable an organisation to leverage the strengths of each type of architecture while addressing specific requirements for analytics, data science, and compliance. For example, an organisation may use a hybrid architecture that combines a centralised data warehouse for structured data, a data lake for unstructured data, and a blockchain for secure transaction records. This can be achieved via a mix of technologies like Hadoop, Snowflake, and blockchain to support the different data types and use cases. There are a few other considerations related to data storage and processing that organisations may want to consider: a) As part of a decentralised approach, organisations can also implement a data mesh. This helps to empower domain-specific teams to manage their own data products. b) Lambda Architecture can be used to manage data processing. This architecture combines batch and real-time processing for analytics and reporting. Kappa Architecture can also be a consideration. It is similar to

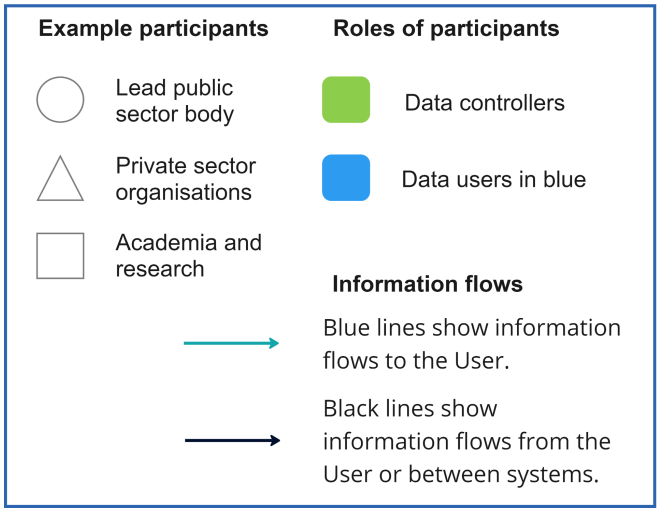
	Lambda, but relies solely on stream processing for real-time data management. The data can be stored on premises, on cloud, or a hybrid. This will depend on security considerations and compliance with applicable legislation and regulations on data hosting.
Data integration: the sources of data and how it is ingested into the data storage facility, including transformations that need to happen, and the standards these follow.	How data integration varies, naturally, by architecture. Centralised architecture depends on ETL (extract, transform, load) pipelines or data warehousing. Data from different sources such as databases, applications, files, APIs, and cloud storage can be ingested in batch mode, by real-time streaming, or both. Transformations are carried out through cleansing, standardisation, enrichment, and aggregation as per the standards: SQL, XML, JSON, CSV. Federated architecture allows access to numerous databases, data warehouses, applications, and cloud services using federated query engines or data fabric platforms. Data is in-place and accessed through virtualisation or replication with minimal changes, though these frequently occur dynamically at query time, using such standards as SQL, REST APIs, or GraphQL. Decentralised data integration uses blockchain platforms and decentralised data marketplaces; the data sources are IoT devices, sensors, social media, and DApps. The ingestion is peer-to-peer sharing through smart contracts, whereas the extent of transformation work done at the source is minimal, where the standards to follow are blockchain-specified data formats and DIDs.
Data discovery: how users find out what data is available and information about the data.	Tools to search for the data. These include data catalogues (Collibra, Alation, IBM watson catalogue).
Data access: how user access is managed and enabled.	The data access method also depends on the data storage choice it this will determine how queries are run: • **Centralised**: data access is enabled through a single point of entry. • **Federated**: unified interface or federated query engine provides access to data across systems • **Decentralised**: peer-to-peer sharing and direct interaction between

	nodes. Data access may be enabled via APIs that bring data securely to users from storage. The use of APIs is common if the data is being accessed externally or running a Machine Learning model to automatically pull the data. Alternatively, batch data transfers can be used if data is updated infrequently. If users are using APIs within a federated architecture, a data fabric may be needed to integrate data from different stores of data. As well as how data access is enabled, data access controls are used to determine which user queries are granted access. Types of controls include Role-Based Access Control, Discretionary Access Control (DAC), Mandatory Access Control (MAC), among others. The choice of control depends on the category of users and the sensitivity of data. More sensitivity data requires stricter controls (e.g., MAC, data-centric security).
Data services: other services that the initiative provides to data users such as visualisations, analytics, and applications.	Data services provide the interface for extracting, transforming, and loading the data within an architectural design that can be done via GUI and non-GUI based interactions. In centralised architectures, BI dashboards offer visual interfaces to explore and analyse data retrieved via APIs from the central repository (e.g., data warehouse), while non-GUI interfaces like SQL clients enable direct querying. In federated architectures, BI dashboards can integrate data from multiple sources through federated query engines, while programmatic access is possible via APIs or data fabric interfaces. In decentralised architectures, BI dashboards might visualise data from smart contracts, while non-GUI interfaces like blockchain explorers or custom scripts interact with the blockchain directly. Any additional services or applications are likely to depend on user needs, aims of the initiative, and resource constraints of the team.

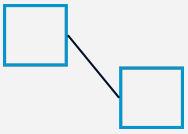
Data Infrastructure Design: Examples

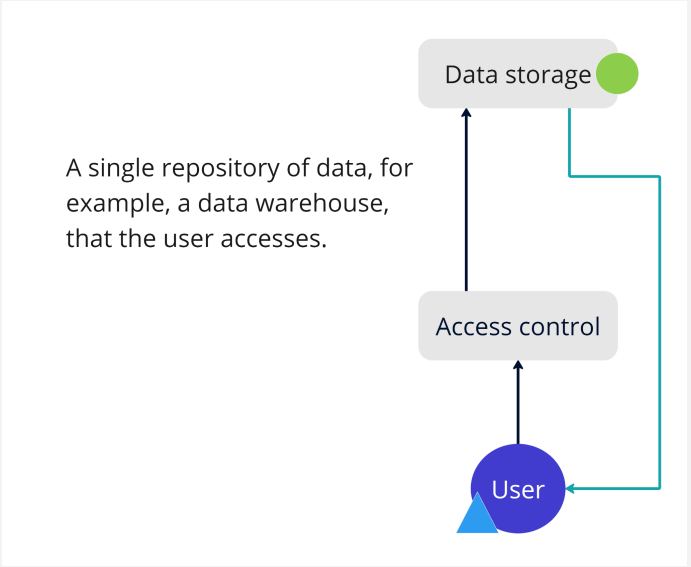
While the data infrastructure will vary across each government and use case, a few broad designs of data infrastructure across different types of data sharing arrangement are illustrated below.

Key for Data Infrastructure Diagrams



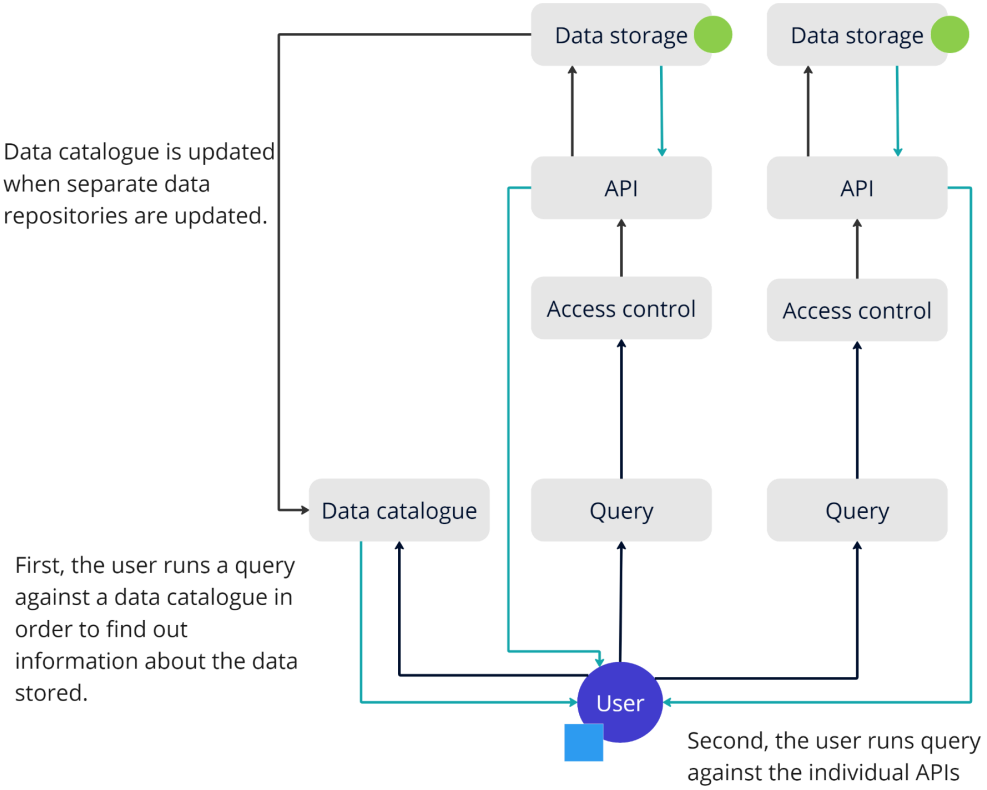
One-to-one data sharing arrangement



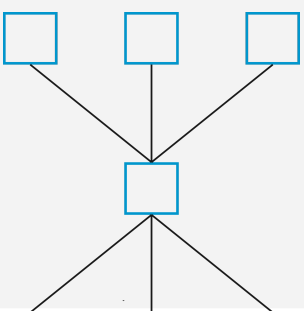


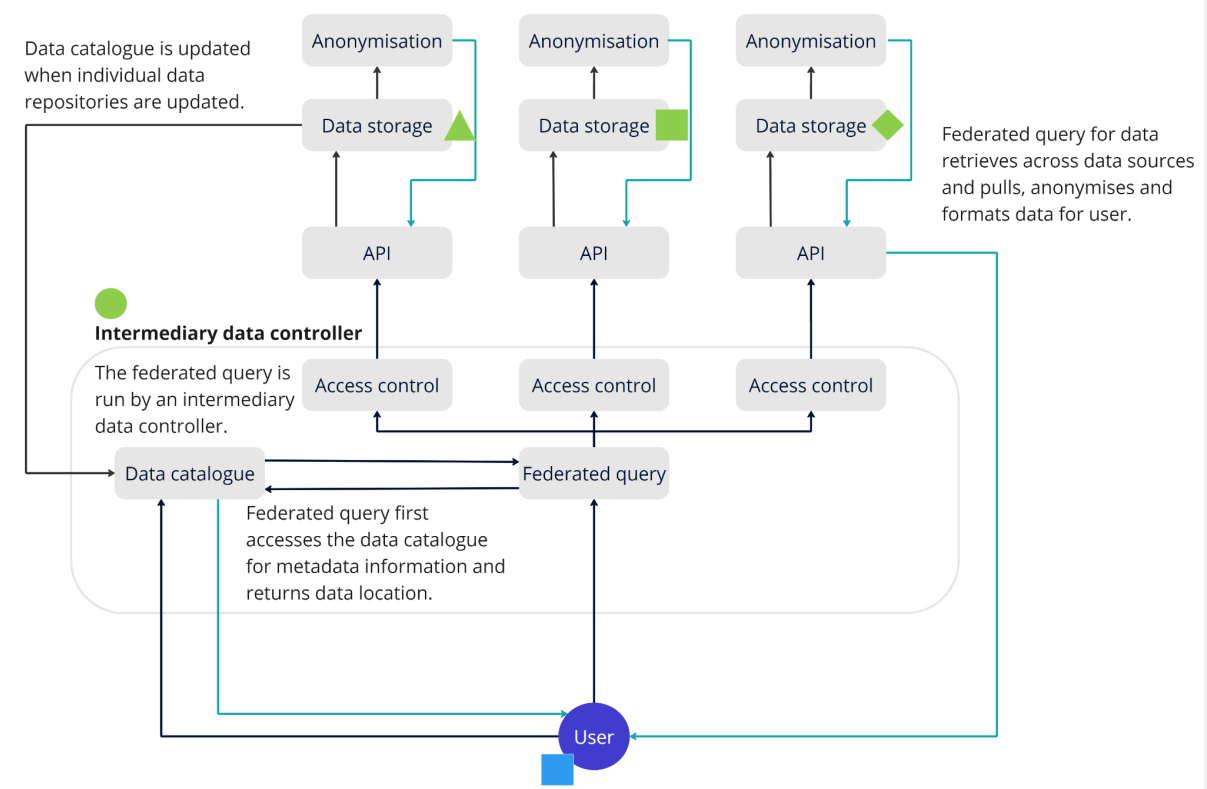
One-to-many data sharing arrangement

A data controller, for example a government agency, makes the data they hold available to data users. In the below case, the agency holds data across different data repositories.

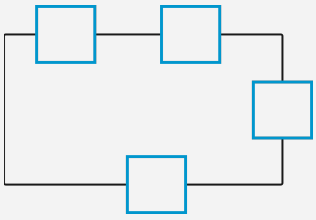


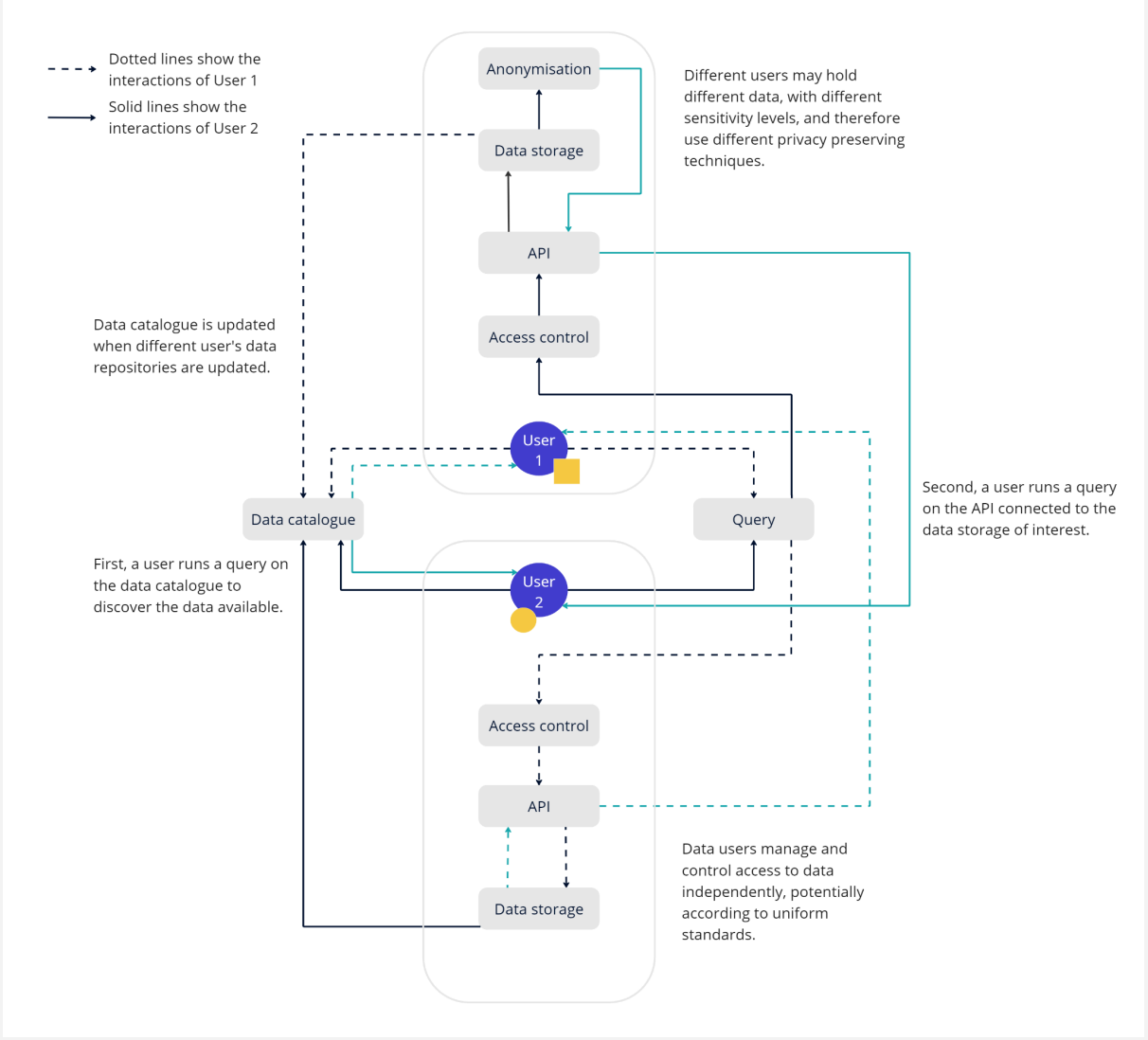
Many-to-one, one-to-many data sharing arrangement





Many-to-many data sharing arrangement




Dotted lines show the interactions of User 1
Solid lines show the interactions of User 2
Anonymisation
Data storage
API
Access control
User 1
Data catalogue
Query
User 2
Access control
API
Data storage
Different users may hold different data, with different sensitivity levels, and therefore use different privacy preserving techniques.
Data catalogue is updated when different user's data repositories are updated.
First, a user runs a query on the data catalogue to discover the data available.
Second, a user runs a query on the API connected to the data storage of interest.
Data users manage and control access to data independently, potentially according to uniform standards.

Privacy Preserving Technologies: Introduction

Prior to engaging with PPTs, there are a number of general procedures that can be considered to maintain data privacy when sharing data. Firstly, it is important to consider whether or not data sharing is appropriate at all in the relevant use case. Secondly, teams may consider sharing data at a higher level of abstraction that maintains the privacy of data subjects: sharing group-level data only, for example.

If neither of these general procedures are appropriate—for example, if data is needed and individual-level data is required for the particular use case—then teams may consider a number of different PPTs. Below is an illustrative set of examples of techniques that you may use.

Technique	How it works	When is it relevant?	Additional considerations
Anonymisation/ pseudonymisation	Anonymisation removes personally identifiable information (PII) from datasets, making it impossible to trace back to an individual. Pseudonymisation replaces such PII with pseudonyms, which would allow data to be tracked back to the original owner only with additional, confidential information.	Datasets where direct identifiers need to be protected and scenarios requiring compliance with data protection regulations like GDPR.	Anonymisation can be difficult to achieve perfectly, as there's always a risk of re-identification, especially with large datasets or when combined with other information sources. Pseudonymisation is often considered a more practical approach for many use cases.
Differential privacy	Adds statistical noise to data/queries to protect individual data points.	It is relevant for large datasets where individual data points are sensitive and where aggregate insights are needed without revealing specifics.	The level of noise added in differential privacy is a trade-off between privacy and the accuracy of the results: higher privacy levels may result in less accurate or useful aggregate insights.
Homomorphic encryption	Homomorphic encryption allows computations to be carried out on encrypted data without needing to decrypt it. The results, when decrypted, match the results of operations performed on the plaintext.	For sensitive personal data (e.g. health records or financial data) and situations where analysis needs to be conducted without revealing data to third parties.	It can be computationally intensive, especially for complex computations. This can be a limiting factor in some applications.
Synthetic data	Synthetic data is artificially generated data that mimics the (statistical) properties of relevant, real data without containing any actual personal information. Such data is created using models that are trained on real data. It is an increasingly popular option for governments, and one case study from the United Kingdom	Any scenarios where real data is too sensitive to share and applications needing large datasets for testing and development without risking privacy.	The quality and utility of synthetic data depend heavily on the underlying models used to generate it. It may not always be a perfect substitute for real data in all scenarios.

	is explored below.		
Federated learning	ML models are trained across multiple, federated devices/servers holding local data samples, without exchanging them; only model updates are shared to improve a central model.	When there are distributed datasets that cannot be centralised due to privacy concerns and in cases where local data might be used to improve a global model without sharing raw data.	Federated learning can introduce challenges in terms of model convergence and ensuring consistent performance across diverse datasets held by different parties.
Secure Multi-Party Computation (SMPC)	SMPC enables multiple, different parties to jointly compute a function over their inputs while keeping those inputs private. Each party learns only the output of each computation but not about the other parties' inputs.	Cases of multi-stakeholder engagement with sensitive data and joint data analysis tasks where input data from all parties need to remain confidential.	It it can computationally demanding, especially for complex computations involving many parties

Example: NHS England synthetic data pilot

Healthcare data is extremely important: analysis allows researchers and clinicians to better understand existing diseases, treatments, and service provision as well as enabling research into future treatments. That said, it is also extremely sensitive. Robust privacy measures are needed to ensure its security. Access to such data is currently therefore quite a protracted process as only approved users, with approved projects, are granted access to the data. It can also be difficult for prospective users to know how useful the data will be in advance; for example, which datasets will be available, what fields will be present in the data sets etc.

The NHS is piloting a project to help prospective users understand the broad shape of the real patient data through synthetic datasets from approximate values to columns (NHS, n.d). The synthetic data generated mimics the structure, format, and volume of the data but does not, in this case, preserve relationships between fields.

So, while these datasets cannot currently be used for analysis, they will allow potential data users to set up infrastructure, test code, and assess pipelines prior to accessing the real data. This allows users to determine whether the data is appropriate without having to share sensitive data or undergo a lengthy approval process.

Appendix 2: Full Methodology

This roadmap is designed to support governments in making informed decisions about when and how to share data with external actors for AI innovation. To achieve this, we conducted a literature

review, developed an initial draft of the roadmap, which was reviewed by GPAI and external experts, tested the roadmap with three pilot partners, incorporated the lessons learned into a revised version, and then held a second round of external expert review for the updated version.

Reviewing Pertinent Relevant Literature

As a second phase in the GPAI's project, 'The Role of Government as a Provider of Data for AI', a particular focus was placed on GPAI's 2023 report on The Role of Government as a Provider of Data for Artificial Intelligence (GPAI, 2023a). We also reviewed the wider work of GPAI's Working Group on Data Governance (GPAI, 2020), and relevant publications of its other Working Groups. Additionally, we situated our work within the broader research landscape on data-sharing for AI, including academic and grey literature. Key findings from these sources were recorded to inform the design of the initial draft.

Draft Roadmap Design

The roadmap was designed through an iterative process, beginning with internal workshops and discussions, which drew on learnings from previous Oxford Insights projects, like the design of a data-sharing pilot in Colombia (MinTIC, 2022), and the literature review. Based on these resources, we developed a preliminary roadmap which was reviewed by an external group of data governance experts as well as by GPAI's Project Advisory Group (PAG).

Based on the feedback we received from experts and PAG, we iterated the roadmap to create a first draft to be tested practically.

Testing the Draft Roadmap

To ensure that our roadmap is user-centric and addresses the key enablers, challenges, and barriers in the data-sharing journey for AI, we tested it through pilots with three government teams. We invited governments worldwide to express their interest and submit a partner application, assessing each using a scoring framework. This framework evaluated key aspects such as:

- The team's readiness and commitment to the project
- Internal capabilities
- A clear understanding of the project objectives

A critical factor in ensuring the roadmap's effectiveness was its applicability across different contexts. To achieve this, we carefully selected pilot partners to represent a diverse range of:

- Geographic spread.
- Levels of government.
- Economic development.
- Data policy and legislation maturity.
- Experience with data-sharing.
- Previous progress on selecting a sector and/or use case —aiming to have a mix that allows us to test different parts of the roadmap.

After the evaluation process, we established partnerships with the following government teams:

- Agency for Electronic Government and Information (AGESIC) of Uruguay (national level).
- Digital Transformation Office of the Presidency of Turkey (national level).
- Jigawa State Government (state level).

The partnership involved supporting pilot partners in implementing the draft roadmap through a series of workshops and one-on-one problem-solving sessions tailored to the specific needs of

each government team. Additionally, this collaboration allowed us to gather valuable insights into the usability of the roadmap. We achieved this by implementing formal feedback collection forms at the end of each decision stage of the draft roadmap. Moreover, we captured further insights into the roadmap's usability by closely observing the progress of each pilot.

In addition to the pilots, we engaged with other government teams across different regions to gather feedback on the roadmap through one-on-one sessions and a webinar hosted in September 2024.

Iterating the roadmap

The pilots and engagement with government teams provided us with actionable insights to improve the roadmap, leading to the development of an iterated final version. A number of the key learnings we had from the pilot partnerships included:

Structure and user journeys

Our draft roadmap advised a linear path through the project. The roadmap provided an expanded list of decision stages and advised taking each decision stage in turn. In our pilots, we found that partners had different entry points to the roadmap and preferred sequences of decision-making. For example, Türkiye started with a mechanism, and chose use cases based on this, whereas, Uruguay and Jigawa both started with sectors they wanted to develop initiatives within. We updated the roadmap structure to reflect these different user journeys.

Roadmap scope

We learnt the opportunities and limits for how the roadmap can provide genuine support to users. The roadmap can help with structuring decisions, providing considerations, resources, and international examples. The roadmap cannot address challenges that result from unique government contexts. We updated the roadmap to exclude dependencies on specific ways of working or legal or institutional contexts. We included examples where possible to illustrate how governments are working within the constraints of their specific contexts. We provide suggested activities and resources where possible to support teams to overcome challenges themselves.

Decision and Checklist Framing

The framing of decision stages and why they are important became clearer as we developed workshop materials. For example, we found that data sharing mechanisms were best communicated as, at their core, decision making mechanisms adopted by a set of actors for the management of data access and use, with the data infrastructure and legal infrastructure requirements stemming from this agreed mechanism. We updated the framing of the roadmap checklist and decision stages based on this experience.

Responding to challenges users face

Pilot partners encountered and worked through several challenges within the pilots. These include limited experience applying legal frameworks, clarifying and assigning roles across organisations, learning about how to apply new privacy preserving technology techniques, coordinating with external stakeholders. We included materials we developed during the pilots to support partners through these challenges within the roadmap, as well as international examples of how other governments are responding to these challenges.

The new version was reviewed by GPAI's Project Advisory Group as well as our external expert advisors to ensure a high level of quality assurance. These reviews were crucial in refining the

roadmap, ensuring that it met the diverse needs of potential users, and it was both practical and aligned with the latest best practice in data-sharing for AI innovation.

Bibliography

Ada Lovelace Institute, (2021a), Legal mechanisms for data stewardship. www.adalovelaceinstitute.org/wp-content/uploads/2021/03/Legal-mechanisms-for-data-stewardship_report_Ada_AI-Council-2.pdf

Ada Lovelace Institute, (2021b), Participatory data stewardship: A framework for involving people in the use of data. https://www.adalovelaceinstitute.org/wp-content/uploads/2021/11/ADA_Participatory-Data-Stewardship.pdf

ALI-ELI, (2021), Principles for a Data Economy: Data Transactions and Data Rights. www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy_Final_Council_Draft.pdf

All of us, (n.d.), National Institutes of Health, www.researchallofus.org/

DAMA UK, (2013), The Six Primary Dimensions for Data Quality Assessment, https://www.sbctc.edu/resources/documents/colleges-staff/commissions-councils/dgc/data-quality-deminsions.pdf

Deloitte, (2017), New Technologies Case Study: Data Sharing in Infrastructure A final report for the National Infrastructure Commission, nic.org.uk/app/uploads/Data-sharing-in-infrastructure.pdf

Dharyanto, I., Murata, T., Ustriyana, MG., (2023), Indonesian Personal Data Protection Law Update – What to Expect from the Upcoming Implementing Regulation, Data Protection Newsletter, www.nishimura.com/sites/default/files/newsletters/file/asia_data_protection_231002_en.pdf

Berryhill, J., Heang, K., Clogher, R., and Mcbride, K. (2019), Hello, World: Artificial intelligence and its use in the public sector. https://doi.org/10.1787/726fd39d-en

Black, G., Stevens, L., (2013), Enhancing Data Protection and Data Processing in the Public Sector: The Critical Role of Proportionality and the Public Interest, SCRIPTed, vol. 10, no. 1, pp. 93-122. https://doi.org/10.2966/scrip.100113.93

C4DC, (n.d), Library, https://contractsfordatacollaboration.org/library/

Ceulemans, H., Boeckx, T., Dillard, V., Mathieu, G., Oberhuber, M. (2021), From competition to collaboration: How secure data sharing can enable innovation, World Economic Forum, www.weforum.org/agenda/2021/06/collaboration-data-sharing-enable-innovation/

Chen, Y., Clayton, E. W., Novak, L. L., Anders, S., & Malin, B., (2023), Human-Centered Design to Address Biases in Artificial Intelligence, Journal of medical Internet research, 25, e43251. https://doi.org/10.2196/43251

CNIL, (2024), AI how-to sheets, https://www.cnil.fr/fr/ai-how-to-sheets

Copernicus, (n.d.), Copernicus in detail, https://www.copernicus.eu/

Data Spaces Support Centre, (2024), Contractual framework, dssc.eu/space/BVE/357074870/Contractual+framework#Data-space-agreements

Daten Ethik Kommission, (2019), Opinion of the Data Ethics Commission, https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Datenschutz/Data-Ethics-Commission_Opinion.pdf?__blob=publicationFile&v=1.

Digital Economy Act, Section 71, (2017), www.legislation.gov.uk/ukpga/2017/30/contents

Digital Future Society (2022), Towards meaningful oversight of automated decision-making systems, digitalfuturesociety.com/app/uploads/2022/11/Towards_meaningful_oversight_of_automated_decision_making_systems.pdf

Doteveryone (2021), 'Consequence Scanning – an agile practice for responsible innovators', doteveryone.org.uk/project/consequence-scanning/

European Commission, (2024), AI Act, digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Commission, (n.d), Are there restrictions on the use of automated decision-making?, commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/are-there-restrictions-use-automated-decision-making_en

ESIP Data Readiness Cluster, (2022), Checklist to Examine AI-readiness for Open Environmental Datasets. ESIP. Online resource. https://doi.org/10.6084/m9.figshare.19983722.v1

Frontier Economics, (2021), The economic impact of trust in data ecosystems – Frontier Economics for the ODI, theodi.org/insights/reports/the-economic-impact-of-trust-in-data-ecosystems-frontier-economics-for-the-odi-report/

Government of Canada, (2023), Get authority to collect the personal information, The Digital Privacy Playbook, www.canada.ca/en/government/system/digital-government/digital-privacy-playbook/ensure-authority-collect.html

GPAI, (2020), Data Governance Working Group A Framework Paper for GPAI's work on Data Governance, gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf

GPAI, (2022a), Data Justice Policy Brief: Putting data justice into practice, gpai.ai/projects/data-governance/data-justice-policy-brief-putting-data-justice-into-practice.pdf

GPAI, (2022b) Enabling Data Sharing for Social Benefit Through Data Trusts: Data Trusts in Climate, Report, March 2022, Global Partnership on AI, gpai.ai/projects/data-governance/enabling-data-sharing-for-social-benefit-through-data-trusts-in-climate.pdf

GPAI, (2023a), The Role of Government as a Provider of Data for Artificial Intelligence: Phase One Project Report, africa.ai4d.ai/wp-content/uploads/2024/01/DG08-The-Role-of-Government-as-a-Provider-of-Data-for-Artificial-Intelligence-Interim-Report.pdf

GPAI, (2023b), Trustworthy Data Institutional Framework - A practical tool to improve trustworthiness in data ecosystems, https://gpai.ai/projects/data-governance/DG10%20-%20Trustworthy%20Data%20Institutional%20Framework%20%20-%20A%20practical%20tool%20to%20improve%20trustworthiness%20in%20data%20ecosystems.pdf

Hsieh, PJ., Lai, HM., (2020), Exploring people's intentions to use the health passbook in self-management: An extension of the technology acceptance and health behavior theoretical perspectives in health literacy, Technological Forecasting and Social Change, Elsevier, vol. 161(C)., doi.org/10.1016/j.techfore.2020.120328

INSIGHT, (n.d), Data Trust Advisory Board, www.insight.hdrhub.org/data-trust-advisory-board

Jolliffe, D., Mahler, DG., Veerappan, M., Kilic, K., Wollburg, P., What Makes Public Sector Data Valuable for Development?, 2023. The World Bank Research Observer, Volume 38, Issue 2, August 2023, Pages 325–346, doi.org/10.1093/wbro/lkad004

Koops, B. J., (2021), The concept of function creep. Law, Innovation and Technology, 13(1), 29–56. https://doi.org/10.1080/17579961.2021.1898299

Land Transport Authority, (n.d), www.lta.gov.sg/content/ltagov/en.html

MinTIC, (2022), Pasos para pilotear proyectos de intercambios de datos Data Trust, Data Commons y Data Marketplace, https://gobiernodigital.mintic.gov.co/692/articles-238512_recurso_2.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, ID., Gebru, T., (2019), Model Cards for Model Reporting, In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229, https://doi.org/10.1145/3287560.3287596

Mobility Dataspace, (n.d), mobility-dataspace.eu/

NHS England, (2023), Value Sharing Framework for NHS data partnerships. https://transform.england.nhs.uk/key-tools-and-info/centre-improving-data-collaboration/value-sharing-framework-for-nhs-data-partnerships/

NHS, (n.d), Artificial data pilot, digital.nhs.uk/services/artificial-data

OECD, (2024), AI, Data Governance and Privacy: Synergies and Areas of International Co-operation, www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/ai-data-governance-and-privacy_2ac13a42/2476b1a4-en.pdf

Office for National Statistics, (2023), How we're providing better data to power decision making - the technology behind making critical data assets available on IDS, Blog: ONS Digital, Data and Technology, onsdigital.blog.gov.uk/2023/01/18/how-were-providing-better-data-to-power-decision-making-the-technology-behind-making-critical-data-assets-available-on-ids/

Open Data Charter, (2023), Finding The ‘Rights’ Balance: How access to public information and open data complement each other, medium.com/opendatacharter/finding-the-rights-balance-4a1444823e58

Open Data Institute (ODI), (2019), Data Ecosystem Mapping Tool. theodi.org/insights/tools/data-ecosystem-mapping-tool/

Open Data Institute (ODI), (2021), ‘The Data Ethics Canvas’, theodi.org/article/the-data-ethics-canvas-2021/

Open Data Institute (ODI), (2022), Mapping data ecosystems: methodology. theodi.org/insights/reports/mapping-data-ecosystems/

Oxford Insights, (2023), Trustworthy AI Self Assessment, oxfordinsights.com/ai-readiness/trustworthy-ai-self-assessment/

Postigo, A., (2023), Regulation of digital health and health data in the Asia Pacific region, ARTNeT Working Paper Series, No. 230, Asia-Pacific Research and Training Network on Trade (ARTNeT), Bangkok, hdl.handle.net/20.500.12870/6645

Ravi, N., Chaturvedi, P., Huerta, E.A., Liu, Z., Chard, R., Scourtas, A., Schmidt, K.J., Chard, K., Blaiszik, B., Foster, I., (2022), FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. Sci Data 9, 657 (2022). https://doi.org/10.1038/s41597-022-01712-9

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, Article 6, data.europa.eu/eli/reg/2016/679/oj

Republic Act No 10173, (2012), Data Privacy Act of 2012, https://privacy.gov.ph/data-privacy-act/#w7

Republic of Estonia, (2024), Consent Service, Information System Authority, www.ria.ee/en/state-information-system/people-centred-data-exchange/consent-service

Statistics Law of Turkey, (2005), Article 14, unstats.un.org/unsd/dnss/docViewer.aspx?docID=2149#:~:text=ARTICLE%201%2D%20The%20purpose%20of,and%20information%2C%20produce%2C%20publish%20and

Szarfman, A., Levine, J. G., Tonning, J. M., Weichold, F., Bloom, J. C., Soreth, J. M., Geanacopoulos, M., Callahan, L., Spotnitz, M., Ryan, Q., Pease-Fye, M., Brownstein, J. S., Ed Hammond, W., Reich, C., & Altman, R. B., (2022), Recommendations for achieving interoperable and shareable medical data in the USA, Communications medicine, 2, 86. https://doi.org/10.1038/s43856-022-00148-x

Tony Blair Institute for Global Change, (2022), Reaping the Rewards of the Next Technological Revolution: How Africa Can Accelerate AI Adoption Today, www.institute.global/insights/tech-and-digitalisation/reaping-rewards-next-technological-revolution-how-africa-can-accelerate-ai-adoption-today

UK Data Service, (n.d), Apply to access non-ONS controlled data ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/apply-to-access-non-ons-data/

UNESCO, (2023), AI Ethics Impact Assessment, unesdoc.unesco.org/ark:/48223/pf0000386276/PDF/386276eng.pdf.multi

UNESCO, OECD, (2024), G7 toolkit for artificial intelligence in the public sector; report prepared for the 2024 Italian G7 presidency and the G7 digital and tech working group, unesdoc.unesco.org/ark:/48223/pf0000391566

The White House, (2022), OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay, www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/

World Bank, (2019), ID4D Practitioner's Guide: Version 1.0 (October 2019), Washington, DC: World Bank, id4d.worldbank.org/guide/data-protection-and-privacy-laws

World Economic Forum, (2020), AI Procurement in a Box: AI Government Procurement Guidelines, www.nist.gov/system/files/documents/2021/08/23/ai-rmf-rfi-0039-1.pdf

Yates D., Beale T., Marshall S., Parr M., (2018), Designing data sharing agreements: a checklist. In Gates Open Res 2018. https://doi.org/10.21955/gatesopenres.1114886.1

Young, A., Verhulst, S., (2017), Aclímate Colombia: Open Data to Improve Agricultural Resiliency, Open Data's Impact, odimpact.org/case-aclimate-colombia.html

Ziesche, S. (2023), Open data for AI: What now?, UNESCO, doi.org/10.58338/OGYU7382.