PRIMER ON DATA SHARING

John Ure

August 2023

he

Commented [MM1]: You want this to be highlighted in green?

Primer on Data Governance for Data Sharing

Preamble

The motivation for this Primer arose from the experience of the data sharing project in Hong Kong 2020-2021, the Inter-Modal Transport Data Sharing Programme that has subsequently become known as Data Trust 1.0 or simply DT1. DT1 was a collaborative initiative between the 'HK Team'[1] and Dr Jiangping Zhou and his colleagues at the University of Hong Kong, whereby the HKU acted as a Trusted Third Party to manage a Data Trust, funded by the Innovation and Technology Fund of Hong Kong's Innovation and Technology Commission.

DT1 was in effect a Proof of Concept that data sharing between public transport companies as data controllers, either privately-owned (bus companies) or government-owned but run as a commercial corporate entity (the metro) could work. 'Could work' meaning that (i) the companies involved were prepared to share certain anonymised and encrypted operational data with a Trusted Third Party, which aggregated the data prior to data analysis – with the assistance of a special Transport Data Analytics Service Provider (TDASP) who was Arup; (ii) the choice of the Central Business District of HK as a manageable location. All data was destroyed once the objectives of the analysis were completed, and curated presentations were made to the data controllers to ensure no confidential data was shared. The aim was to test the possibility of evidence-based public transport planning that went beyond traditional data silos in light of existing inadequacies in the sharing of public transport data.

The Vision of the Data Trust

DT1 was a PoC. It has always been envisaged that it would be succeeded by a DT2 which would have two dimensions to it. First, DT2 would be a scaled-up version of DT1 both geographically embracing representative areas of the territory of Hong Kong, such as urban and rural, long-distance and short, hill terrain and flat, etc. Second, it would become a catalyst for the creation of a more permanent Data Hub into which aggregated data sets could be placed by different stakeholders, private, NGO and public, and these would be available to other stakeholders for further development as applications, information platforms, models and new data sets to be further shared with others through the Data Hub. This would be a major innovation for Hong Kong, comparable to the London Data Store. It is a vision shared with HKPolyU who are partners in a proposal for DT2.

The major task in the establishment of the Data Trust was the creation of a Data Sharing Framework (DSF) which because it was starting from scratch and required universal stakeholder buy-in, took as long as the project itself. A redacted copy of the DSF is provided in the Annex to assist others who may wish to share data. Two outcomes followed. First, a DSF now exists for future adoption. Second, the experience gained from DT1 has been shared internationally with several relevant organizations and agencies, as a result of which it seems appropriate to review the state of data sharing frameworks and projects on an international basis. This is greatly helped by a plethora of available studies and references, so the point of the Primer is partly auto-didactic and partly to provide a hopefully lucid understanding of a complex set of approaches to encourage more researchers, government agencies, commercial organizations as well as civil society such as NGOs to share data

[1] Dr John Ure, at the time director of the Technology Research Project (TRP) at the University of HK, TRP research associates Dr Jenny Wan and Terry Graham, independent transport data analytics specialist Andrew Pickford, and independent smart city analyst Waltraut Ritter.

for both the common good and to help organisations make the most productive use of the data they have by combining their data sets to provide better insights.

Acknowledgments: To achieve these ends, the Primer is published under the Collective Commons for anyone to use or reproduce as they see fit. The only requirement is full acknowledgement. To start that process I wish to give special thanks to Sriganesh Lokanathan for his expertise and insightful comments on a draft of this Primer. I also wish to thank the Global Partnership for Sustainable Development Data (GPSDD) and their participants in webinars on development data sharing, and the National Academies of Sciences, Engineering and Medicine (NASEM) and their participants in webinars on the sharing of data on health and bio-genetics.

Disclaimer: John Ure is a Director of Global Governance Advisory at Access Partnership Pte Ltd. He is solely responsible for any materials used, errors made or views expressed in this Primer.

Primer on Data Governance for Data Sharing

Introduction

This Primer is about **data sharing**. It draws upon a wide range of literature to provide a summary of key issues and practices, especially those that come under the heading **data governance**.

In a connected digital society data *as information,* defined in a digital world as stimuli (called data when it can be stored in a computer) "that has meaning in some context for its receiver " is everywhere, although extracting that information often arises only when the data is shared and/or joined with other data to reveal trends and insights. The process of joining different data or data sets does not need to be an external activity. As AWS points out, it can take place within the confines of a laboratory or an enterprise or a government agency without being shared beyond those boundaries – data warehousing is designed to assist this type of data sharing. However, this Primer is focused rather on data sharing between sperate and independently managed entities, even if ultimate ownership may be similar, such as sharing between different government agencies. In other words, one where the **access to data** in one form or another, such as anonymised or as aggregated data**,** is shared. Reference will also be made to 'federated learning' (see below) which is a hybrid form of data sharing in which the data itself is not shared, it remains internal, but refinements to the analytics model to which the data is exposed is shared.

An oft-cited hyperbole that data is the 'new oil' is *helpfully misleading*. Misleading because oil, unlike data, is not non-rivalrous, it cannot be re-used, and it cannot be shared except in the sense of being consumed by multiple parties simultaneously, such as for the conveyance of passengers in a vehicle driven by an internal combustion engine. The analogy is helpful in the sense that it highlights the value of *not sharing data* where it can offer a commercial advantage in the marketplace, exactly what the EU's 2022 proposed *Data Act* is designed to discourage as an unfair anti-competitive practice that threatens to block innovation among rival businesses. As a December 2022, OECD paper puts it:

> "Data are non-rival but excludable. They exhibit economies of scale, which, coupled with information asymmetries and weak ownership regimes, hinder the emergence of multilateral data markets. Hence, the value of data will be affected by the governance framework that determines how they can be created, shared and used."

By comparison with oil, data as pointed out by the Bennett Institute can simultaneously appear in many different places, and its replication is often at zero marginal cost. However, the **value of data** is often undetected in the sense that the benefits of having the information conveyed by the data through analysis is not fully understood until it has happened or may only become evident later in time. Those who have the data, such as those who generate the data, may have a better understanding of the benefits than others, what is known as **asymmetric information**. For these reasons, the non-rivalrous nature of data which seems to deny its ownership value, or/and asymmetric information which means that other parties are unable to appreciate the benefits of having access to the data, can undermine data sharing by precluding the incentives of data sharing.

Definitions

Nevertheless, data sharing becomes more widespread as its benefits are better recognised, both at the private level and as a public good. One of the drivers is the trend towards 'smart cities' as pointed out by the World Economic Forum, the OECD and the GSMA among others. However, what data is shared, by whom and with whom, and under what conditions, are key questions. These are

not only practical questions in terms of the expected costs and benefits, but also legal and regulatory and the lack of universally agreed definitions can be an obstacle.

As a preamble, what is data sharing? For the purposes of this Primer, it implies the willingness of the party owning or with legitimate authority over the data (such as a Letter of Attorney) to share it with the receiving party and the willingness of that party to have it shared with them, and for a legitimate purpose. It will exclude data sharing designed to cause harm to a third party. By contrast, the data sharing envisaged in this Primer is data to advance a legitimate cause, such as research, policy development, commercial success, information to the public benefit, etc. It will also avoid complex issues such as the sharing of data that may have national security implications. A data sharing framework may itself identify such sensitive areas of data for sharing among limited number of authorised personnel, such as biological data that could be weaponised, or such injunctions can be imposed by a higher authority. However, it needs to be openly recognised that such limitations, however necessary, are *in principle* antithetical to scientific research where unrestricted peer review by others not directly associated with the research is fundamental to the assessment process.

Real difficulties arise when different legal interpretations are attributed to the substantive issues involved, such as what constitutes 'data ownership' or 'who is the data subject' – for example, individual images in a public video. What are judged to be legitimate workarounds to problems can differ across jurisdictions, such as gaining 'subject consent' when the subject is physically or mentally unable to provide consent, or when the data is to be reused for multiple different research projects. These challenges will be especially difficult when data is to be shared across jurisdictional borders.

The following table cites a variety of sources, and they are not exclusive. Different sources focus upon different applications, such as the portability of data between devices versus the portability of personal data between competing service providers. This is an example of a substantial (albeit relatively trivial) difference in meaning, but for the most part the differences in definitions are more semantic than substantial. In addition to the sources cited, OECD (2023) *Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies* is a good reference.

Table 1 - Definitions

Data Issues	Definitions differ from various sources, sometimes substantially
Data and its Management	
Data OECD 2022	Refers to recorded information in structured or unstructured formats, including text, images, sound, and video.
Big Data Oracle	Data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Volumes too large to be processed without software apps.
Data Cleaning Tableau	Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.
Data Standards World bank	Are the rules for structuring information collected by the ID systems which facilitate semantic interoperability. A set of agreed-upon data standards ensures that the data entered into a system can be reliably read, sorted, indexed, retrieved, and communicated between systems. Data standards are therefore crucial for ensuring interoperability and the accuracy and portability of identity data, helping protect its long-term value.
Data Privacy Data Privacy Manager	The GDPR was … the most comprehensive and groundbreaking data protection law that reflected the new digital era in the way data is created and managed in modern everyday business processes. Nevertheless, neither

	GDPR nor other data protection laws (like the US Health Insurance Portability and Accountability Act (HIPAA), California Consumer Privacy Act (CCPA), or the Children's Online Privacy Protection Act (COPPA), gives a strict definition of what Data Privacy is… a definition in any given particular law, there is none.
Data Anonymisation and Data Pseudonymisation Sariyar et al	Erasing or encrypting identifiers that connect an individual to stored data (anonymisation), or replacing original identifiers with artificial identifiers (pseudonymisation). **Note**: precise regulatory purposes differ across neighbouring countries even within the EU
Data Encryption HashedOut	Encryption is a two-way function where information is scrambled in such a way that it can be unscrambled later. Often used for data transfers.
Data Protection Imperva	Data protection is the process of protecting sensitive information from damage, loss, or corruption.
Data Security IBM	The securing of data from unauthorised or illegal access or usage by using an array of security measures, from individual good practices to online firewalls and other cyber security measures.
Data Transfer Informatica	Data transfer refers to the secure exchange of large files between systems or organizations… data transfer is most often used to share data securely among business partners, suppliers, or government agencies for cooperative purposes.
Data Mobility DevOps.com	1. Mobility involves the use of handheld devices, such as a smartphone or tablet, that provide the freedom to move from space to space, allowing employees to complete a wide variety of tasks while on the go. 2. The shift toward a more flexible storage architecture, even in a hybrid cloud environment, has created the need to migrate data much more frequently, fuelling the demand for data mobility.
Data Portability OECD 2021 and Truce	1. The ability (sometimes described as a right) of a natural or legal person to request that a data holder transfer to the person, or to a specific third party, data concerning that person in a structured, commonly used and machine-readable format on an ad hoc or continuous basis. 2. Can refer to the use of laptops and devices that allows once-tethered employees to work remotely, typically performing the same tasks, in much the same way, just in different spaces.
Data Interoperability OECD 2021	1. Systems can work together or "interoperate" in a way that allows for seamless or real-time exchanges, updates or transfers of information or data – can involve protocol and/or data (API) interoperability. 2. International Standards Organization (ISO) defines "interoperability" as the ability of two or more systems or applications to exchange information and to mutually use the information that has been exchanged. 3. In the supply chain, e.g., of electricity, it can be decomposed into functional, technical and commercial interoperability. 4. It can apply also to consumers' technological choices, changes in component devices and changes in suppliers' value chain.
Data Source Talend	A data source may be the initial location where data is born or where physical information is first digitized, however even the most refined data may serve as a source, as long as another process accesses and utilizes it. **Note:** data may come directly from the data subject or from secondary sources such as research, from 'cookies' tracking website users, from customers, etc. This can make data ownership ambiguous.

Data Discovery Snowplough	Data discovery is the process of extracting meaningful patterns from data. This is achieved by collecting data from a wide variety of sources and then applying advanced analytics to it to identify specific patterns or themes. **Note**: It can also imply the use a of catalogue of data assets listing the conditions upon which they can be accessed using an agreed vocabulary.
Data Markets TechTarget	Data marketplaces (platforms) typically offer various types of data for different markets and from different sources. Common types of data sold include business intelligence, advertising, demographics, personal information, research and market data. **Note:** data catalogues are made available.
Data Storage Mongo DB	**Databases/Repositories** - stores the current data required to power an application. **Data Warehouses** - stores current and historical data from one or more systems in a predefined and fixed schema, which allows business analysts and data scientists to easily analyse the data. **Data Lakes** - stores current and historical data from one or more systems in its raw form, which allows business analysts and data scientists to easily analyse the data.
Data Analytics Investopedia	Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
Data Sharing and Accessibility	
Data Sharing OECD 2022	Refers to the act of providing data access for use by others, subject to applicable technical, financial, legal, or organisational use requirements.
Data Sharing Framework (DSF) OECD 2022	Refers to data access and sharing arrangements that permit data access and sharing subject to terms that may include limitations on the users authorised to access the data (discriminatory arrangements), conditions for data use including the purposes for which the data can be used, and requirements on data access control mechanisms through which data access is granted. **Note**: DSFs may be legal, regulatory, policy or guidance, and data sharing authorisation can be paper or blockchain ('Ricardian contracts') based.
Data Access OECD 2022	Refers to the act of querying or retrieving data for its potential use, subject to applicable technical, financial, legal, or organisational access requirements.
Data Trust Cremeglobal	A data trust is a legal and technical framework for sharing and managing data. A Data Trust promotes and facilitates data sharing amongst organizations by ensuring trust in the rules, data security, confidentiality and privacy. A data trust comprises of two key elements: legal agreements and a technology platform to collect, aggregate, protect and manage the data. **Note:** In research work, once the initial purpose of the data has been achieved the data would normally be destroyed unless by the agreement of data subjects or controllers it enters a data hub for future use.
Data Trusted Third Party CASD	A trusted third party is an independent entity which has no stakes in using either source or resulting data guaranteeing the confidentiality of directly identifying data (data which can be used to identify someone). **Note**: Trusted third-party models as the basis of a data trust can be commercially-based, community-based, or individually-based and can have varying degrees of legal and fiduciary status, often derived from an associate data sharing framework.
Data Hub Altexsoft	A data hub is a central mediation point between various data sources and data consumers. It's not a single technology, but rather an architectural

	approach that unites storages, data integration, and orchestration tools. With a data hub, businesses receive the means to structure and harmonize information collected from various sources. Note: a data hub can serve a community of interests, such as the London Data Store or the government data hub in India's state of Odisha.
Data Connectors IDSA	Provides the necessary API endpoints for other participants to negotiate data contracts.
Data Spaces IDSA	Can refer to the computational capacity devoted to data storage and manipulation in gigabytes, or as defined by Fraunhofer/IDSA to "a distributed network of data end points (instances of the International Data Spaces connector) that enables the secure exchange of data and guarantees data sovereignty."
Data Centralised / Federated Learning Xenonstack	Federated learning is a machine learning technique that involves training an algorithm through several decentralized edge devices or servers that carry local data samples without sharing them. This method differs from conventional centralized machine learning methods, which require all local datasets to be submitted to a single server.
Open Data Open Data Handbook	Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike. Note: in that sense it is data sharing free or at a charge) by *de fault*.
Data Ownership and Stewardship	
Data Subject Thomson Reuters	An individual to whom personal data directly or indirectly relates. A natural person about whom a controller holds personal data and who can be identified, directly or indirectly, by reference to that personal data – see GDPR.
Data Producers OECD 2022	Refers to organisations or individuals that create, co-create, generate, or co-generate data, including as a by-product of their social and economic activities, and can therefore be considered a primary data source. Note: also see Amazon's guide to reference architecture
Data Holders Australian CDR	A business that holds consumer data and must transfer the data to an accredited data recipient at the consumer's request.
Data Controller EU	Determines the purposes for which and the means by which … data is processed.
Data Processor EU	Processes… data only on behalf of the controller. The data processor is usually a third-party … however, in the case of groups of undertakings, one undertaker may act as processor for another undertaking.
Data Intermediary OECD 2022	Refers to service providers that facilitate data access and sharing under commercial or non-commercial agreements between data holders, data producers, and/or users. Data holders and trusted third parties can act as data intermediaries.
Data Stewardship TechTarget	Data stewardship is the management and oversight of an organization's data assets to help provide business users with high-quality data that is easily accessible in a consistent manner. While data governance generally focuses on high-level policies and procedures, data stewardship focuses on tactical coordination and implementation.
Data Sovereignty Payments Blog	1. **Internationally**, data governed by the laws of the country in which it originated even if it is stored in a foreign jurisdiction; applying laws extra-judicially is easier if there is an agreement with the other countries in which the data is being stored or used. 2. **Nationally** it can refer to the rights of individual data subjects to own and control their data identify and what happens to their data.

Data Residency TechTarget	Data residency refers to the physical or geographic location of an organization's data or information. Similar to data sovereignty, data residency also relates to the data laws or regulatory requirements imposed on data based on the data laws that govern a country or region in which it resides. **Note:** Foreign sovereignty laws may apply where data is resident.
Data Custodian Technopedia	Deals with the actual nuts and bolts of transporting and storing data, rather than issues around what data is going into the system and why.

Realms and Limits of Data Sharing

In principle the realms of data sharing possibilities are unbounded, in part because in a digital world the marginal costs of producing data are low, and of reproducing data close to zero. However where the risks of data sharing are perceived to be high, such as highly confidential data, more costly data sharing mechanisms – the implementation and operation of data sharing architectures – may be required. These can be costly in terms of time and effort, and in terms of Capex and Opex. For example, where data sharing involves commercial intermediaries, or trusted third parties, a close adherence to data sharing frameworks is required involving strict regulations. There may also be a shortage of skilled data analysts to handle the process.

But according to the European Commission, commercial data sharing is just not yet happening as it might.

> "Firstly, the value of data as an asset is not yet fully recognised. Secondly, public bodies frequently lack the know-how to identify valuable datasets and the capacity to process them. And, there are currently not enough incentives for businesses to share data with the public sector for the common good. There are a number of other barriers, including a lack of professionals in the field, differences in legalisation between Member States, trust and security issues, ethical questions and the limited interoperability of datasets, amongst others. As a result, B2G data sharing can be a lengthy, uncertain process."

The EU Data Act (see below) is designed to overcome the reluctance.

Access and Sharing

In a digital economy the scope of data sharing is huge. To narrow it down this Primer excludes considerations of how personal private data is used and by whom, except when it is shared with public or research bodies to be added to other data sets for trend analysis, such as aggregated health data or aggregated mobility data, or by commercially bodies authorised to do so. It also distinguishes between **data access** and **data sharing** where *data access refers solely to accessibility without necessarily leading to data sharing* as with third parties or with other data sets to generate new information. Accessibility is necessary but not sufficient to bring about data sharing, whereas sharing is sufficient to ensure accessibility. However, **sharing data access** is different. It can be the basis of a data sharing hub. These may seem subtle plays on words, but their meanings and implications are substantially different.

For clarity, the focus of the Primer is upon data sharing between the public bodies, research bodies and commercial bodies, and the data sharing agreements (DSFs) necessary for their success exist at four levels: **International and national frameworks** initiated by public or private organisations, **private frameworks** initiated by the participating commercial or non-commercial parties, and **research frameworks** initiated by participating research institutions but often in accordance with a national or sector-based DSF, such as for health and biomedical research.

Public Sector - all sectors can be governed by frameworks that have international or national reach, for example, the International Data Transfer Agreement between the UK and the EU for the exchange of trade-related data, and Singapore's Trusted Data Sharing Framework.

Private Sector - private sector agreements are specific to the sector but may be influenced by international or national guidelines, such as banks who need to share data. In the EU for example, agricultural sharing between farms of soil data arising from the use of robots is being promoted. In Hong Kong, as mentioned in the Preface, data sharing between commercial public transport operators, two bus companies and a metro, was run on a Proof-of-Concept basis with the University of Hong Kong acting as a trusted third party.

Research - data sharing among research institutes is often considered absolutely essential in areas such as astronomy and genetics for two reasons. First, because the nature of the data requires multiple sources and efforts, such as multiple telescopes around the globe scanning the night skies for astronomical phenomena. Second, because scientific papers need to be peer reviewed and over time the research findings need to be replicated by independent researchers to verify or challenge them.

Mobility and Portability of Data

Defining mobility as multi-accessibility, that is the ability to access data across many different devices, it is a form of data-sharing in a functional sense, but it can be used to triangulate individuals for better, for example, to solve crimes, or for worse, by hackers and scammers. The portability of data is a more complex issue as outlined the OECD (2021) report *Mapping Data Portability Initiatives, Opportunities and Challenges*. Portability refers to making data of one service supplier, for example customer data or interoperable APIs, available to another, that is the "sharing of data across digital services and platforms", what the OECD refers to as "enhancing access to and sharing of data (EASD)" which is characterised by five key dimensions:

- Sectoral scope – whether data sharing is confined to a specific sector or not.
- Beneficiaries – whether only individuals or businesses have a right to data portability.
- Type of data – limited to personal data or includes volunteered, observed, or derived data.
- Legal obligations – voluntary or mandatory and how enforced.
- Modus operandi – portability 1.0 (ad hoc (one-time) data); portability 2.0 (ad hoc direct transfers to another data holder); portability 3.0 (real time continuous data interoperability).

The report discusses horizontal vs. sector data portability where the former may facilitate data portability in sectors not sufficiently incentivised to it, or where there are sector asymmetries such sectors dominated by companies with significant market power. Providing businesses as well as individuals data portability rights is one way to address this challenge, as in the case of the EU Free Flow of Data Regulation (FFDR) and Australia's Consumer Data Rights (CDR) legislation which distinguishes between consumer data posted online, data created during transactions, purchased data, and other data associated with activities of a digital nature. Sector specific regulations apply, for example, in the case of the banking sector, "volunteered data will be made available, as well as data on financial products such as credit and debit cards, deposit and transaction accounts, and data on mortgages."

The argument for sector specific data portability regulations lies in taking into account specific technical and market conditions of given sector where, for example, data portability may pose a high security risk or a disincentive to investment and innovation, for example of start-ups. Australia also

recognises as ‘accredited data recipients’ (ADRs) third parties able to receive business data securely and manage it within the CDR scheme rules.

Running through the OECD report is the tension between data interoperability enabling data sharing and portability as a means to further competition between platforms and the disincentives to investment and innovation arising from losing control over exclusive data ownership. The report distinguishes between protocol interoperability “without any requirement to allow interoperability with external actors can reduce competition as it increases switching costs” and data or API interoperability. However, data interoperability that results in similar competing products may ultimately reduce the incentive for users to switch. The example is given, where “New digital banks can gain a small market share of 1-5% of the unsecured retail and SME loan market in Singapore. However, it is more natural for small banks to pursue a platform-based business model rather than compete with large banks by providing similar products.”

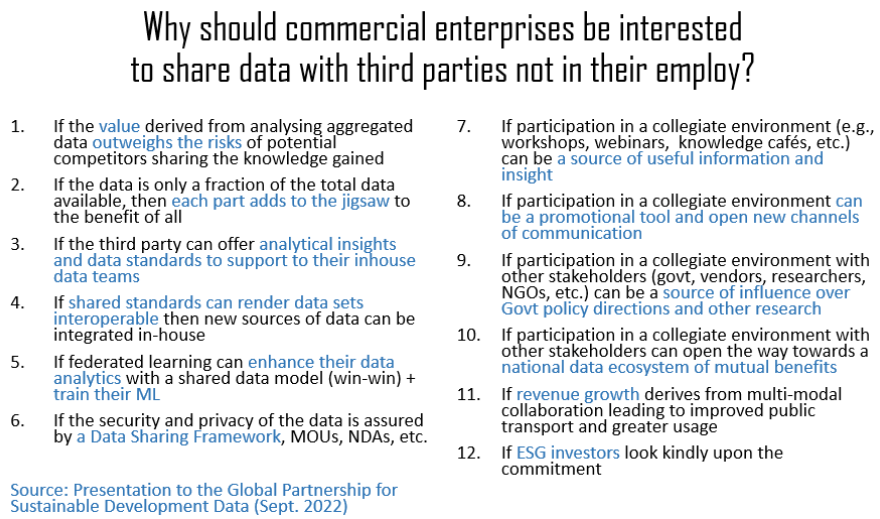
Data portability can therefore be a problematic approach to data sharing. On the downside it can increase the risk of identity fraud, can undo “some of the ‘privacy by design’ efforts of private actors to protect privacy”, and can possibly infringe the privacy rights of third parties when data is ported from one data controller to another. As ever, the OECD report concludes that “countries implementing a data portability regime must have a robust privacy framework.”

Limits and Incentives

Limits to data sharing arise from numerous sources: the ‘data as oil’ metaphor to preserve commercial advantages, company secrets giving competitive advantages, IPRs, etc.; cyber and other national security considerations; lack of harmonious standards; lack of awareness of the value of data sharing or simply lack of agility; dearth of staff trained in data sciences and lack of experience; privacy requirements; etc.

If the motives for *not* wanting to share data are strong then no data sharing framework (DSF) will be effective unless (i) it is mandatory, or (ii) it contains incentives that override the reluctance, such as an agreement to share information that would not normally be available to the data controller, or share the findings of data analysis and perhaps, through federated learning, not to share the data itself, or the right to be consulted and to participate in high-value workshops and events. These benefits might not need to be spelled out in the DSF if prior agreements have been reached with potential data controllers, such as NDAs, IPRs, or access to applications arising from the data analysis, an opportunity to make inputs into policy recommendations, etc.

There is equally the question of what motivates the data processors. In cases of commercial engagement by data controllers the answer is clear. In the case of trusted third parties such as universities and researchers, the answer can be a mix of research interest, a funded public policy engagement and a commitment to the public good as an outcome, mindful this can also embrace a collective good for the private sector. This would be true in cases where plans based upon data outcomes lower the transactions costs to private companies; for example, an integrated public transport system with government backing for a common payments system, or for the development of associated commercial locations. Indeed without combining public and collective industry sector goods, persuading data controllers to share data could be an uphill struggle.



The Value of Data

The Value of Data Sharing – an **OECD** report (2022) *Measuring the Value of Data and Data Flows* makes an attempt to estimate the value of data, but confines its measurement to the cost to an enterprise of producing its own data as fits the Standard National Accounts (SNA) practice. It nevertheless recognises that the "more data are shared the more value can potentially be created from them. At the same time, the cost of producing data might not adequately reflect their contribution to value creation. This would occur if the benefits generated through sharing and re-using data with more firms or individuals are uncompensated."

The ODI's *Understanding the Social and Economic Value of Sharing Data*, uses the ODI's *Theory of Change* to classify data values arising under three headings: stewarding data, creating insights, decision-making; and references the Nuffield-supported *Bennett Institute for Public Policy* at Cambridge University *The Value of Data* and its accompanying literature review. The review cites many different classifications of data, such as 'user provided', 'authored', 'captured', 'master', 'transactional', 'reference data', 'meta', 'structured, semi-structured and unstructured', etc.; and the varying nature and taxonomy of data sets such as 'statistical', 'register', 'personal', 'social graphic', 'climatic', 'sports', 'economic', etc. It also cites examples three valuation methods commonly used: cost, market and income-based, but concludes that none of these methods is convincing and none "sufficient to influence policymakers." A method described as 'novel' and one being experimented by Microsoft among others is to use machine learning "to estimate the marginal effect of new data on predictions." The UITP (Union Internationale des Transports Publics) as the International Association of Public Transport, in a 2020 paper on Sharing Data in Public Transport, suggests the four dimensions to data value creation: cost-based, income-based, market-based and externalities-based, can each offer different perspectives to highlight the areas of greatest value creation.

One conclusion is that data often only gains value when it is combined with other data – noting that data on its own can be asymmetric in terms of accessibility – which rackets up the value of data

sharing. This highlights to distinction between **data sharing and data analytics**, where value is made possible by the former but created by the latter as **data-in-use**.

Finally, the distribution, or *redistribution* of the value of data through data sharing, can be viewed through different lenses: as **data as labour**, compensating individuals for their data, **data as capital**, as an intangible asset, or **data as intellectual property**, licensable for financial gain, and data that is not marketed but is made publicly available by persons and households can be regarded as **owned data** that either reaps a return as a stream of public services, or is handed to public authorities as a **data tax.**

Data Sharing Frameworks

The advantage of a data sharing framework is that it can be comprehensive in its coverage, it can result in clarity, confidence and trust, and thereby it can be re-used for further data sharing projects. A comprehensive introduction to the need for DSFs and data governance for data sharing is the World Bank's *World Development Report 2021* Data for Better Lives. A DSF is not in and of itself a legal document, but if there is a *legislated* DSF, an individual DSF will come under its lawful enforcement and even without a legislated DSF, an individual DSF may have some legal standing. By 2022, research on data sharing by the OECD "revealed the need for more coherent **data governance frameworks** as data access and sharing is increasingly occurring across sectors and jurisdictions."

> In the context of the *OECD Project on Data Governance for Growth and Well-being*, "data governance" refers to diverse arrangements, including technical, policy, regulatory or institutional provisions, that affect data and their creation, collection, storage, use, protection, access, sharing and deletion across policy domains and organisational and national borders. Efforts to govern data take many forms. They often seek to maximise the benefits from data, while addressing related risks and challenges, including to rights and interests. [See also OECD (2023)]

Essentially, a DSF must cover what is to be shared, why it is to be shared, who may be the sharing parties and on what terms and conditions may other parties gain access to the data, who shall have management control over the data sharing process and stewardship over the shared data, what steps need to be taken to ensure personal data privacy and data protection, such as data anonymization, encryption during transfer, data aggregation, data security against hacking, and who has responsibility for the data analysis and what happens to the analysed data, who it can be shared with, whether it can be subsequently entered into a data hub and with who's permission or needs to be destroyed, etc. Such a DSF needs to be agreed with all the relevant parties who may each have their own concerns and their own team of lawyers to vet the DSF, and may want in addition non-disclosure agreements (NDAs) or other types of guarantees. This is likely to be a lengthy and pains-taking process the more parties that are involved, and should entail consultation with responsible regulatory agencies such as Personal Data Protection Commissions. Once accomplished the value of the DSF will be tested by the degree of success of the data sharing project itself, and once proven to be of value is available for re-use with minor adjustments. In cases of purely commercial agreements between enterprises, agencies such as the IDSA can offer to certify DSFs providing them with an industry-based authority – see below.

In the field of academic research data sharing can be either through direct contact or through published papers in journals. One review published in 2021 of data sharing in journals focused upon *Nature* and *Science* concludes that:

"Although data sharing has improved in the last decade and particularly in recent years, data availability and willingness to share data still differ greatly among disciplines. We observed that statements of data availability upon (reasonable) request are inefficient and should not be allowed by journals. To improve data sharing at the time of manuscript acceptance, researchers should be better motivated to release their data with real benefits such as recognition, or bonus points in grant and job applications. We recommend that data management costs should be covered by funding agencies; publicly available research data ought to be included in the evaluation of applications; and surveillance of data sharing should be enforced by both academic publishers and funders. These cross-discipline survey data are available from the PlutoF repository."

The following table provides selected references to data sharing platforms and other related references.

Table 2 – Examples of Data Sharing Platforms and Other References

Characteristics	Examples
Data sharing platforms (Trusted 3rd Party, Data Markets)	1. **USA:** Open Algorithms (OPAL) Project one of the core projects of MIT Trust: Data Consortium; a P2P data sharing model under the direction of Prof Pentland, enables data analysis by performing algorithm-execution on data at the location of the data repository. Access to raw data is controlled by the repository owner and only aggregate answers are returned. Algorithms are openly published, studied and vetted by experts to comply with privacy requirements. 2. **Singapore:** Dex (start-up) and PwC will launch a blockchain for data sharing; "The data owner can also include access rights (to the data), as well as specific kinds of rights," PwC Singapore's role is to help in the design of a "trusted data framework". 3. **New Zealand**: Data Republic started its Open Data Marketplace in 2017, "built on top of a legal framework that allows people to exchange data, with personal information stripped out. It allows data to move from party A to party B without ever touching the personal information." 4. **China**: The two largest bike-sharing companies, Ofo and Mobike (together account for 90% of the market in China.) both announced their plan to share their big data with the government. Qidian, the world's largest big data platform for bikes developed by Ofo. tracks more than 10 million bikes and 200 million users in 250 cities of 20 countries. Mobike operates 8 million bicycles in more than 200 cities around the world. 5. **China and Malaysia:** Alibaba's 'Tianchi Big Data Program' - A crowd-intelligence platform bring experts together to find solutions to real-world problems. In Malaysia, the program is backed by Malaysian Digital Economy Corporation (MDEC), aims to deliver data-intelligence technology and AI capabilities to 500 data professionals and 300 startups in two years and allow them to engage with about 120,000 developers and 2,700 academic institutes and businesses from 77 countries and regions. 6. **PlutoF Go**: a platform is designed for storing and managing biodiversity data over the web. PlutoF GO app is a tool for gathering biodiversity data - observations, specimens, material samples.

P2P secure data sharing	1. **MaidSafe: (Secure Access for Everyone)** uses "an autonomous network" based upon XOR network architecture that validates chains of data without using blockchain and is based on secure access.
Data sharing for smart traffic management	1. **China, Malaysia and Macau:** Alibaba's 'City Brain' service uses AI and cloud for live traffic predictions, optimizes traffic flow, and detects traffic incidents using open data generated from video footage, traffic bureaus, public transportation systems, and mapping apps. The system can learn traffic patterns overtime and make more sophisticated recommendations to improve traffic efficiency. In **Macau** the system will focus on smart transportation, smart tourism, smart health care, and smart city governance, as well as talent development. In **Malaysia** the platform will initially be used for traffic management, with the potential to help on town planning, incident response, and other emergency services. 2. **Open Transport Partnership:** Easy Taxi, Grab and Le Taxi – three ridesharing companies that, combined, cover more than 30 countries and millions of people – have joined a partnership with the World Bank and other organisations to make traffic data derived from their drivers' GPS streams open to the world.
Open Data	1. **Australia:** Customer Data Rights (CDR) = Open data policies to allow accredited users access to data, e.g., check personal details, and to cover all sectors starting with banking, energy and telecoms at the same time giving customers the choice of "secure automated data technology" to enable their information to be shared. 2. **US Commerce Department**: The Dept is "pushing to make more and more federal data available." Member of the Big Data Working Group. 3. **Ireland**: Enterprise Ireland in December 2022 issued a Public Consultation Paper Proposed Data Sharing Agreement between Enterprise Ireland and all Local Authorities (DSA for LEO Suite of Services Data) based upon the Data Sharing and Governance Act 2019
Sharing apps	1. **USA**: Moovit, a public transit app, seeks to connect riders to public transit as a singular place to plan an entire route across multiple forms of transportation; it can coordinate bus and subway times, ferries, streetcars, cable cars, water taxis, bike shares and ride-sharing apps. 2. **Netherlands**: Marcel Schouwenaar, a Dutch designer, has a plan called Fairbike which uses blockchain for the "smart contracts" that can monitor the fulfilment or breach of any conditions they stipulate--could create self-managing fleets of bikes. "Each bike collects its own money and reinvests these funds back into the network by issuing repairs or if the situation allows it, expand the service by adding a new bicycle to the network." 3. **India/Australia:** Ola mobile app integrates city transportation connecting driver-partners across cabs, auto-rickshaws and taxis. With its real-time tracking, riders can share their travel routes, location and make emergency alerts when needed. Accessible through the Ola app, Ola Play is an in-cab infotainment console which provides entertainment for riders.
Data apps	1. **UK:** 'Data in motion' providing transport real-time transit data feeds and data visualisation design services, including profiles for hyperloops.

Mobility-as-a-Service (MaaS)	1. **EU**: The Mobility as a Service (MaaS) Alliance is a public-private partnership; the main goal is to facilitate a single, open market and full deployment of MaaS services; see projects. 2. **Sweden:** Ubigo (start-up) and Fluidtime (Austrian IT) offering MaaS. 3. **Finland**: 'Whim' is a smartphone apps that will plan journeys by multiple means of transportation**.** 4. **Germany**: Qixxit offers similar to Whim. 5. **UK Parliament**: Transport Committee Inquiry – evidence. **6. Finland: International Transport Forum (ITF)** a shared mobility study on Helsinki in replacement for private car travel with shared vehicles which aims to end congestion, cut emissions and frees public space. The study uses the simulations from mobility data from Lisbon. 7. **World Resources Institute (WRI) - Shared Mobility Principles for Livable Cities:** A consortium of leading city and transport organisations, including WRI developed the Shared Mobility Principles for Liveable Cities to help cities make sense of these changes; with 15 of the world's leading transport and technology companies signing on.
Other Data Schemes	
Big Data Sharing	1. **United Nations**: Global Pulse encourages the use of Big Data as a public good for a sustainable development and humanitarian action through its network of Pulse Labs across the UN.
Data Privacy	1. **USA:** Solid (social linked data), a project led by **Prof. Tim Berners-Lee** developing a set of tools and conventions where internet users can choose where their data resides and who is allowed to access it by decoupling content from the application itself.
GDPR	1. **EU**: The website is a resource to educate the public about the main elements of the General Data Protection Regulation (GDPR).
CCPA	1. **USA California**: California Consumer Privacy Act (CCPA) as amended by the California Privacy Rights Act, 2020 (CPRA) covering areas such as data transfers.
Data Accelerator	1. **Australia**: FUELD is Westpac Group's industry-first data accelerator program to mentor 8 start-ups in anonymised, tokenised data to launch new data-driven products.
Data Trust	1. **New Zealand**: The Data Futures Partnership was tasked by the Government to develop guidelines that public and private organisations can use to develop "social licence" for data use. NZ also added CKAN as a catalogue for sharing government data.
Transport Data Research	1. **Singapore**: The Land Transport Authority (LTA) will set up a new Transport Research Centre with the Singapore University of Technology and Design (SUTD) to foster collaborative research in key areas such as cybersecurity, automation and robotics, data analytics, behavioural studies and user-centric design in transport solutions.
Vehicle Sharing	1. **EU:** Car2go Daimler has bought a stake in Car2go, a JV with Eurocar; China's Geely (owner of Volvo) has become the majority shareholder of Daimler. 2. **London:** 'Smart Ride' operated by Citymapper seats eight people and changes its route on demand. Smart Ride, a bus service using a van that

	operates like a ride-hailing app limited to a specific catchment area. Travellers book a seat in a Smart Ride vehicle at a specific time along a route shown in the Citymapper app. "Think of it as a bus, because it has stops and can be shared, but think of it as a cab, because you can book it as close as possible to you on the network" but works with restrictions imposed by TfL.
Data Research Institutes, Smart City Initiatives and Other Resources	1. **USA - NIST Research Data Framework (RDaF):** will provide the stakeholder community with a structured approach to develop a customizable strategy for the management of research data. The RDaF is a map of the research data space that uses a lifecycle approach with six high-level lifecycle stages to organize key information concerning RDM (Remote Desktop Management) and research data dissemination "anticipation of data sharing". 2. **Germany - The Fraunhofer-Gesellschaft**: is a leading organization for applied research in Europe. Its research activities are conducted by 72 institutes and research units at locations throughout Germany. 3. **Germany** - **Städte und Kommunen müssen die Digitalisierung zur strategischen Aufgabe machen:** smart city digitalisation research. 4. **New Cities:** Canadian-based NGO focusing specifically on housing, environment, mobility, and wellbeing. 5. **USA - ITDP:** The Institute for Transportation & Development Policy began in Nicaragua to promote 'bikes not bombs' and has developed into the promotion of environmentally-friendly transport systems to mitigate climate change. 6. **USA – CM2:** The consortium of Cooperative Mobility for Competitive Megaregions (CM2…aims to advance research, education, and technology transfer initiatives to improve the mobility of people and goods in urban and rural communities of megaregions… partners include the University of Texas at Austin, Louisiana State University, Texas Southern University, and the University of Pennsylvania. 7. **USA - ITS** - **Transportation Sustainability Research Center** (Berkeley) study the technological aspects of sustainable transportation. 8. **UITP – Advancing Public Transport**: Brussels-based mission for sustainable transport community. 9. **ATIS (Alliance for Telecommunications Industry Solutions)**: 'New Framework Helps Promote IoT Data Sharing Among Smart Cities'. 10. **USA: Data Collaboratives Research Network**, The GovLab is a data sharing platform, led by Dr Beth Simone Noveck and Stefaan Verhulst, where participants from different sectors (private companies, research institutions, and government agencies) can exchange data to help solving public problems such as climate change, and share findings with potential collaborators and to publicize their work. 11. **Sweden: Flowminder Foundation**, a registered non-profit organisation, aim to improve public health and welfare of the vulnerable populations in low- and middle-income countries through the collection, aggregation, integration and analyses of anonymous mobile operator data, satellite and household survey data, including Call Record Data from mobile phones.

	12. **C40**: a global network of nearly 100 mayors of the world's leading cities that are united in action to confront the climate crisis. 13. **Chartered Institute of Logistics and Transport** (CILT): globally provide training and webinars on logistics and transport, including digitalization. 14. **Institute for Transportation and Development Policy** (ITDP): to promote high quality public transport … with a focus on transit-oriented development, and inclusive people-centred policies. 15. **International Council on Clean Transportation** (ICCT): a focus on decarbonising transportation 16. **National Association of City Transportation Officials** (NACTO): an association of 96 major North American cities and transit agencies formed to exchange transportation ideas, insights, and practices and cooperatively approach national transportation issues. 17. **International Transport Forum** (OECD-ITF): a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision-makers and civil society. 18. **Intelligent Transport**: newsletter on all things transport 19. **International Journal of Hydrogen Energy**: for the exchange and dissemination of new ideas, technology developments and research results in the field of Hydrogen Energy between scientists and engineers throughout the world. 20. **Data Science Journal**: The CODATA Data Science Journal …an electronic journal, publishing papers on the …use and reuse of research data and databases across… including the CARE Principles for Indigenous Data Governance … inspired by the consortium of rural health journals who will publish "nothing about Indigenous peoples, without Indigenous peoples". 21. **Yale Open Data Access (YODA) Project**: "has iteratively developed a model to make data available to researchers in a sustainable way, in which data sharing becomes a part of the clinical research enterprise of the future." 22. **World Bank's Global Data Facility (GDF)**: The primary mechanism to implement key recommendations and insights from the World Development Report 2021: Data for Better Lives and the Sustainable Development Goals. 23. **Research Data Alliance (RDA)**: a community-driven initiative in 2013 by the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation with the goal of building the social and technical infrastructure to enable open sharing and re-use of data.

Reasonable Data Stewardship

Produced by the Open Data Institute (ODI) *Reasonable Data Stewardship* is a study based upon a literature review and interviews conducted June 2022 – March 2023 of eighteen organisations across Africa, Europe, North and South America on how data should be collected, used and shared.

This was partly motivated by the problematic way data is 'shared' across social media, and what has been termed 'surveillance capitalism'. The scandal of Cambridge Analytica is given as one highly publicised egregious example. In its wake the Nuffield Foundation launched the **Ada Lovelace Institute**.

- **Data stewardship** is the central concept which is variously defined by different organisations. The simplest is the **ODI** working definition of 'the collection, maintenance and sharing of data'. Others include the **Ada Lovelace Institute**: 'responsible, rights-preserving and participatory concept [which] aims to unlock the economic and societal value of data, while upholding the rights of individuals and communities to participate in decisions relating to its collection, management and use'; the **Royal Society**: 'a body mandated to ensure responsible use of data'; the **Mozilla Foundation**: 'the act of empowering agents in relation to their own data and guidance toward a societal goal'; the **Aapti Institute**: 'paradigm which explores how the societal value of data can be unlocked while considering what it takes to empower individuals/communities to better negotiate on their data rights'; and the **GovLab**: a focus on 'agents of change in an organisation, responsible for determining what, when, how and with whom to share private data for public good.'
- **Trust** – central to the success of stewardship is the concept of 'trust', as for example in the public sector the UK's National Trust to keep stewardship of national lands and historic buildings.
- **Ethics** – two broad approaches to ethics have been identified. One is an ***ethics-based approach*** that outlines ethical guidelines, which can also be deemed a ***rules-based approach***. One review in 2019 identified 84 documents on AI ethics, and since 2023 when ChatGPT and similar generative AI apps were launched, many more have appeared. The other approach is ***process-based***, for example how to generate 'data for good'. [See **i4 Institute Partner** *Doing Good with Data*. **Data.org** identified over 600 initiatives that fall into this category. Sub-sets of ethical principles focus on *Connected-by-Data* linking data to the promotion of social justice, and *human-centred data* often associated with accompanying *codes of conduct* and *guiding principles*. Common to these approaches are issues of transparency, agility and diversity closely associated with human rights, and issues of *equity, value and power* closely associated with justice. Interviews further identified concepts of *fairness and inclusiveness* as approaches to justice, the latter especially raised in relation to black and indigenous peoples. For public sector ethics see also OECD *Good Practice Principles for Data Ethics in the Public Sector.*
- **Data Sovereignty** – although often referring to the "laws and governmental policies applicable to data stored in the country where it originated and is geographically located", in the ODI review it is more closely associated with empowering of individuals, with **indigenous data sovereignty** [e.g., see Maori data sets in New Zealand] and with feminism for example, honouring the **FAIR** Guidance Principles for scientific data: *Findable, Accessible, Interoperable, Reusable* complemented by **CARE** principles: *Collective benefits, Authority to control, Responsibility, Ethics* related in Australia to data on and for indigenous people. The seven principles of feminist activism and critical thought in **Data feminism,** as articulated by D'Ignazio and Klein, are: (1) examine power, (2) challenge power, (3) elevate emotion and embodiment, (4) rethink binaries and hierarchies, (5) embrace pluralism, (6) consider context and (7) make labour visible.
- **Data Responsibility** – is closely associated with **data ethics** and **data management**, but no consistent definitions are reported. The **process-based approach,** namely following guidelines, is somewhat abstract if the framework doesn't directly take account of the

specifics of the problems being addressed. Those involved in human rights issues tend to stay close to a **rules-based approach.** In low-income countries with hugely unequal distributions of wealth and representation, a ***data society*** is seen as needing to redress these imbalances, as *data for social justice*. Responsibility is also seen as more than just data protection and data privacy but also needing to encompass safety, ethics and efficacy plus a need to be socially/culturally sensitive. However, this is not so obvious in cases of social media data 'sharing' where the market is often global and straddles multiple cultures. This gives rise to concerns about '*responsibility-washing'*.

- **Data Responsible Frameworks** – "Many of our interviewees described responsible data frameworks that set out how to work responsibly with data. These included the Oxfam Responsible Program Data Policy (2015), RD 101: Responsible Data Principles (2018), and The OCHA Data Responsibility Guidelines (2021). This approach is so well established that in 2018 the US-based Center for Democracy and Technology published the report *Responsible Data Frameworks In Their Own Words*, which sought to compare a range of these publications, predominantly in the humanitarian data sector. More practical resources have been developed, such as the IFRC data playbook, The Engine Room RAD tipsheets, CartONG's Information Management Resource portal and Responsible data management toolbox, and practical resources put together by MERL tech."
- **Data Practices –** broadly divided into two groups: **user-centric design** and **participation of data subjects** with examples given of pro-active public services, including Oxfam's moratorium of use of biometric data prior to consultation, and the importance of anonymising data sources. Bad practice cases also exampled. The value of training good data practices is highlighted, including the certification process [see the **International Data Spaces Association**] but auditing is seen as the challenge. Other organizations promoting good data practices include "the **Data Futures Lab at the Mozilla Foundation**, focus on supporting organisations collecting and managing data, while the Responsible Technology focus from the **Omidyar Network** includes supporting a wider ecosystem of actors to realise the positive benefits of data and technology, and the **Engine Room** promoting the use of technology for social justice. The **ODI's** future focus will be on data for 'public benefit', 'reducing harm' and 'readdressing structural inequalities'."

The broader benefits and risks of data sharing and stewardship are also spelt out by the OECD (2022) Responding to Societal Challenges with Data Access, Sharing, Stewardship and Control.

Methods of Data Sharing

The ODI literature review also examines data sharing in terms of **data pools**, such as the *European Data Portal* (2017) as open data, where the highest value datasets as estimated by the *Open Knowledge Foundation* (2019) are: geospatial, earth observation and environment, meteorological, statistics, companies and transport. In addition, as identified by the Government Office for Science (2018) other private data sharing models include **data sharing platforms** providing access to multiple datasets, and **direct data sharing** where a single organisation makes its data available for re-use. Added to these can be the provision of 'clean rooms' as the cloud service provider AWS names them, offering secure and exclusive **data spaces** for data sharing between contracting parties. Specialist platform providers are emerging offering **data rooms** for specific activities, such as M&A and deal-making.

Data Trusts

The role of **data trusts** are reviewed, which come closest to the Data Trust 1.0 of the Hong Kong project. A number of data trust models are identified: legal trust, contractual, corporate, public, and community trust models. The ODI also found different interpretations of 'data trusts', which include:

- A data trust as a repeatable framework of terms and mechanisms.
- A data trust as a mutual organisation.
- A data trust as a legal structure.
- A data trust as a store of data.
- A data trust as public oversight of data access.

Issues of data safety and data equity are frequently promoted by data trusts, but the review also notes that the regulatory framework of such trusts is not always very clear, and references a suggestion that a **Data Trust Law** could exist alongside laws based upon the GDPR.

Government Data Sharing Frameworks

> "Open data has been a priority for all OECD countries and partner economies to support business innovation, social value creation and government transparency. However, fewer countries have adopted broader national data strategies, including in the public sector. In 2018, up to 80% of OECD countries had a strategy for open government data, and 90% had requirements for public sector organisations to publish open data in a machine-readable format. In contrast, only 10% of OECD countries had a dedicated, comprehensive public sector data strategy covering a broader spectrum of data access and sharing arrangements, and their enablers." OECD (2022)

Asia

India's Electronics and Information Technology Ministry (MeitY) as outlined in Data Governance, Asian Alternatives, issued a draft National Data Governance Framework Policy in May 2020 and a revised version was heralded in February 2023. The framework is the culmination of a process begun in 2012 with the issuing of a National Data Sharing and Accessibility Policy (NDSAP). With the advancement of digitalisation, in 2020 a Data Empowerment and Protection Architecture (DEPA) was developed, being the third layer of an India Stack designed to facilitate consented data sharing. This layer includes the Aadhar digital personal ID system and a Unified Payments Interface (UPI). DEPA enables "transfers of a person's data from one data fiduciary to another take place through an encrypted digital workflow that is only triggered after that person's consent has been electronically obtained… and work is underway to implement it in the healthcare system." The author suggests that while public policy has focused on the public good and personal data protection, the "private sector… used to view data governance as a hindrance, but in more recent times companies have come to appreciate that customers view good data governance practices positively."

Non-personal data has been defined as consisting of either data that never refers to an individual or data that has been anonymized, such as health data, and government bodies or private organizations that collect, process, store, or manage data as data businesses. A recommendation in 2020 was that its sharing should be subject to the approval of the Non-Personal Data Authority. Under the NDSAP all non-personal and non-sensitive data generated using public funds across all levels and departments of the government and its authorized agencies shall be open data available through an Open Government Data Platform.

India's position on international data sharing is guarded. The author suggests that compared with the EU's human rights-based approach, and the closely-guarded national security and sovereignty policies of countries such as China, Egypt and Russia, India "declined to sign the Osaka Declaration promoted by Japan at the 2019 Group of 20 (G20) summit out of concerns that the negotiations conflicted with its policy priority for data localization." The conflicts of 'national interest' seem to be stretched between the gains from data sharing and the potential losses of data control to foreign intermediaries.

South Korea

The approach of South Korea to data localisation is characterised in source Data Governance, Asian Alternatives, as 'strategic autonomy' but variously motivated by national security, data sovereignty, data protection, fair taxation and fair competition. This array of drivers arises from different agencies pursuing different goals. Although there is an Open Data Strategy Council, co-chaired by the prime minister and a data expert from the private sector, its focus is upon public open data. An Open Data Mediation Committee handles disputes if public agencies refuse to share data. "Korea's Act on Promotion of the Provision and Use of Public Data controls the data that all public agencies have, but this does not mean the law is applied to each agency in the same manner… while some Korean ministries are mostly focused on data stewardship, others strive first and foremost to facilitate more extensive use of data. Even within the same ministry, different bureaus can have different approaches to opening up data sets."

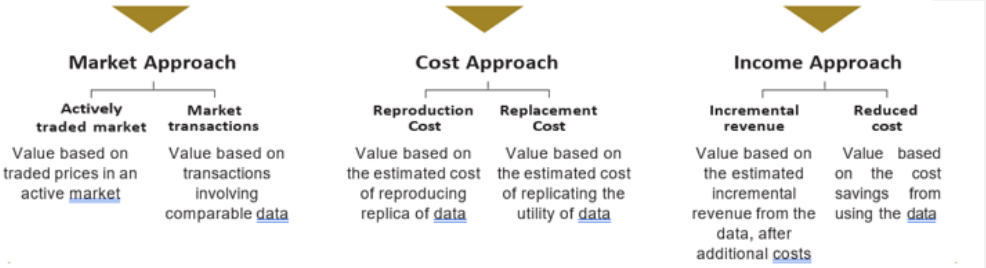
The exercise of data-sharing policy is spread across different ministries: the Ministry of Interior and Safety (MOIS), the Ministry of Science and Information and Communications Technology (MSIT), and Statistics Korea and is different from that for managing open data (the Open Data Strategy Council) and

> "Korea quite simply lacks a unitary national institution for data management and control, which, in turn, makes it difficult to move and share data across sectors, domains, jurisdictions, and organizational boundaries… but the real problem is not a failure of institutional design but a failure of national-level data policy governance: this is because in the Korean government structure, one agency cannot impose policies on multiple ministries."

Ggenerally, "data-related laws in Korea primarily seek to promote data-related industries and economic sectors." But monitoring and regulation of private-sector data-sharing is scant. And according to the source, given the imprecise definitions of what constitutes personal data in the Personal Information Protection Act "tensions between data protection and data sharing arise, a conservative stance commonly prevails."

Singapore

In 2019 the Infocomm Media Development Agency (IMDA) and the Personal Data Protection Commission (PDPC) of Singapore issued a Trusted Data Sharing Framework as a practical guide to the benefits and procedures of data sharing and the relevant legal, technical and business aspects. It lists the PDPC as a reference for basic anonymisation techniques, a guide for data disposal, and a guide for data protection. As a means to value data to motivate data sharing, the Framework offers a summary of the market, the cost and the revenue approaches.



Japan

In April 2022, the Ministry of Economy, Trade, and Industry ('METI') released "a data management framework for collaborative data utilisation and trust to promote the value creation of data." The release, only in Japanese, emphasises a lifecycle approach to data "including generation/acquisition, processing/use, transfer/provision, and storage/disposal." Japan's *Act on the Protection of Personal Information Act* passed in 2003 ("APPI") is the primary legislation that applies to the collection and processing of personal data with guidance published by the *Personal Information Protection Commission*, again in Japanese. DIDOMI provides a guide in English. In tandem with other ministries and the Information Processing Promotion Agency's Digital Architecture Design Center (DADC) and the New Energy and Industrial Technology Development Organization (NEDO), METI has also been working on an interoperable architecture infrastructure Ouranos for the delivery of mechanisms for sharing multiple data managed by various stakeholders.

This is part of Japan's push into "Society 5.0" and "Connected Industries." In the private sector, NTT Communications Corporation (NTT Com) in April 2021 pioneered the sharing of production data via the Smart Data Platform with Trust on a trial basis based on the IDSA standard – and links Japan to the European data infrastructure GAIA-X. Japan has also been advocating within APEC an information-sharing regime similar to GAIA-X along global supply chains. In the research sector a paper in 2019 suggests that while data sharing is widespread, it is often private sharing and varies considerably between disciplines. It also found that, consistent with global surveys, there "were much higher levels of sharing by biological sciences researchers compared with those in other disciplines."

United States

Under the US Constitution, only cross-border transactions, including data, comes directly under Federal laws, so the power of individual *states* to impose data rules is somewhat similar to individual *States* (countries) across the EU. Traditionally this has led to a reliance upon the private commercial and non-government sectors to police themselves in accordance with US policy prescriptions. The USGS issued the following warning about the sharing of data within the US:

> "It is not wise to enter into a data sharing agreement where privacy information may be disclosed since non-Federal organizations are not subject to the Privacy Act. Similarly, the non-Federal organization needs to be alerted that the Federal agencies may be compelled to release information under the [Freedom of Information Act] FOIA."

This dichotomy has led to two successful legal challenges in the EU to the sharing of personal data that may be transferred to the US. The first was to the U.S-EU Safe Harbor agreement and then to the Privacy Shield agreement that replaced it. That in turn was replaced by a new EU-U.S. Data Privacy Framework officially titled the *EU-US Data Protection Umbrella Agreement*.

In March 2023 the US Government fast-tracked a National Strategy to Advance Privacy-Preserving Data Sharing and Analytics (PPDSA).

> "PPDSA solutions include methodological, technical, and sociotechnical approaches that employ privacy-enhancing technologies to derive value from, and enable an analysis of, data to drive innovation while also providing privacy and security. However, adoption of PPDSA technologies has been slow because of challenges related to inadequate understanding of privacy risks and harms, limited access to technical expertise, trust, transparency among participants with regard to data collection and use, uncertainty about legal compliance, financial cost, and the usability and technical maturity of solutions."

For 'privacy-enhancing technologies' (PETs) see below. PPDSA clearly has two dimensions, domestic to the USA and international, notably between the US and the EU. The acceptance of the latter will be largely determined by judicial rulings arising from any future complaints by EU citizens to the activities and practices of US companies, especially social media companies, with regard to EU-based data. Therefore it remains to be seen how successful the PPDSA becomes.

European Union

In 2020 the European Commission (EC) published the European Data Strategy, to create a data economy across the EU by harmonising legislation across Member States and as far as possible comparable with the USA and China. Once proposed, legislation has to pass through the European Parliament and then be adopted by the Council of Ministers. The Strategy consists of five essential Acts, to which could be added the Cyber Resilience Act (CRA):

- Data Act (DA)
- Data Governance Act (DGA)
- Data Markets Act (DMA)
- Data Services Act (DSA)
- Artificial Intelligence Act (AIA)

The Finland-based research company SITRA, provides a useful critical guide to these proposed Acts, but also notes that the NIS - Network and Information Security Directive and Cybersecurity Act - and *NIS2 Directives* are essential to protect data sharing, processing and increase trust in the data economy as envisaged by the first 'Big Five' Acts. In addition, there are common factors in the GDPR and the PSD2 - *Payment Service Directive 2* - that stress the importance of individual choice for any data sharing, thereby giving consumers control over their own personal data.

A research report from the *European Union Institute (EUI)* published in 2023, as a 'Feedback' on the EC's 2022 proposed *European Media Freedom Act* and the amending Directive, refers to all six Acts, where the CRA introduces concepts such as 'security by design' for manufacturers of digital hardware and software products and the design of platforms for data access. One important point the 'Feedback' stresses is the need for regulation to expedite "**transparency of media ownership**", not least because the authenticity and legal purpose of data access and usage is not guaranteed if the ownership of data intermediaries is concealed.

Summary of the EU's Big Six

Acts	How it impacts upon data access and data sharing
Data Act • SITRA calls for a clearer definition of 'data' and especially of 'data ownership'. • The International Road Association (IRU) suggests data types (raw, meta and processed) should be more clearly spelt out. • Bruegel sees data a co-product of user and machine-generated so challenges the distinction; also the distinction between machines and 'other computing devices' (such as smartphones) arguing data links with all tangible products should be included.	1. Sets common framework of practices and rules of non-personal data access across an economy relating to B2B, B2C and B2G; 2. Data to be made available on FRAND (fair, reasonable and non-discriminatory) principles; 3. Data mobility to be specified; 4. Private-sector data is to be *access by design (*or by default) in line with the DGA; 5. During emergencies of "exceptional need" public authorities have the right to free access to privately-held data; 6. Machine-generated data to be excluded from the Act and the Database Directive to be amended accordingly; 7. Data generated by connected products and related services are covered, e.g., IoT data, data from connected vehicles, etc., which is sent back to the manufacturers.
Data Governance Act (DGA) • Note: does not specify what data it applies to, rather safeguards access to and the sharing of data	1. Provides a legal framework to protect existing data rights, such as personal data privacy, IPRs, trade secrets, etc.; 2. Recognises the role of data intermediaries as independent of either data owners/controllers or users, e.g., trusted third parties; 3. Unlike the GDPR does not specify a regulatory authority which is left up to Member States; 4. Promotes re-use of public data, so health data, GPS data, etc., could be used for commercial or non-commercial purposes; 5. Stress upon the public or common good; 6. Key obligations include confidentiality and one-stop shop mechanisms (*access by design*)
Data Markets Act (DMA) • Data sharing in this context is mostly sharing data access; is more about ownership, control and usage than about stewardship. • Issues of data sovereignty and mobility, for example between platforms, are subject to the Data Act.	1. Recognises 'gatekeepers' of platforms who have entrenched market power through strong network effects and control access to data. 2. Gatekeepers must refrain from using in competition with other data users data generated by those users but not publicly available to them, with hefty fines up 20% annual turnover. 3. To be enforced at EC level not national levels. 4. DMA focused only upon gatekeepers, not wider social issues unlike the DSA
Data Services Act (DSA) • SITRA notes different terminologies and focus between DMA and DSA • Data sharing of illicit materials would come under the DSA and other laws.	1. Focus is upon online moderation and content, primarily concerned with online intermediaries. 2. Obligations are graduated from *only* intermediaries to very large online platforms (VLOPs) 3. Regulation through national Digital Services Coordinators
Artificial Intelligence Act (AIA) • SITRA point a connection to the DSA as AI can generate targeted ads. • AI learning uses Big Data that raises issues of personal data and IPRs	1. Based upon a risk system: unacceptable, high, limited and minimal. 2. Create a European AI Board of nationally-determined regulators and members of the EC
Cybersecurity Resilience Act (CRA)	1. Ensure that manufacturers improve the security of products with digital elements since the design and development phase and throughout the whole life cycle;

• See *Principles and Approaches for Security-by-Design and Default* by the Cybersecurity and Infrastructure Security Agency of Australia	2. Ensure a coherent cybersecurity framework, facilitating compliance for hardware and software producers; 3. Enhance the transparency of security properties of products with digital elements, and 4. Enable businesses and consumers to use products with digital elements securely.

European Data Governance Act

It should be noted that the *European Data Governance Act (2022)* which comes into force September 2023, is the European Commission's strategy to "facilitate data sharing across sectors and Member States." But it is not without pushback as the headline in the *Financial Times* 9th May 2023, makes clear 'Tech Chief call for rethink of EU plans on data sharing' as five CEOs complained that "the new rules would force them to give up trade secrets and hand a competitive advantage to China."

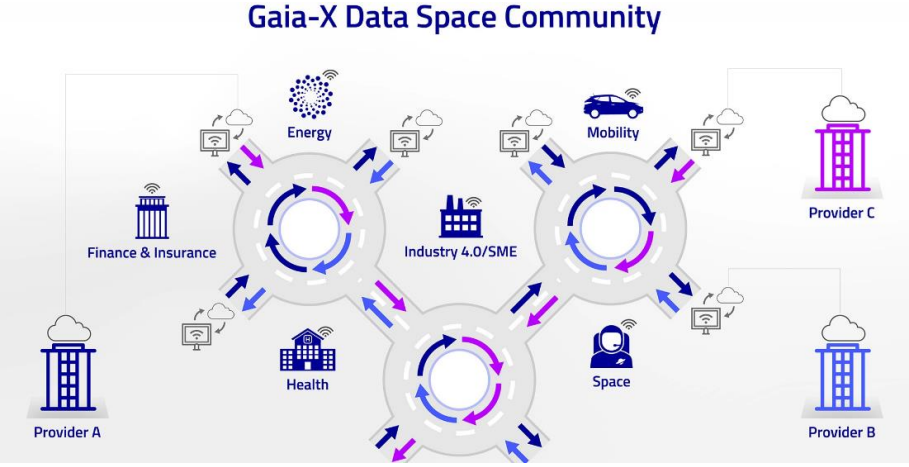
The Act introduces the following:

1. Mechanisms to facilitate the reuse of certain public sector data that cannot be made available as open data; for example, the reuse of health data to advance medical research.
2. Measures to ensure that data intermediaries will function as trustworthy organisers of data sharing or pooling within the common European data spaces.
3. Measures to make it easier for citizens and businesses to make their data available for the benefit of society.
4. Measures to facilitate data sharing, in particular to make it possible for data to be used across sectors and borders, and to enable the right data to be found for the right purpose.

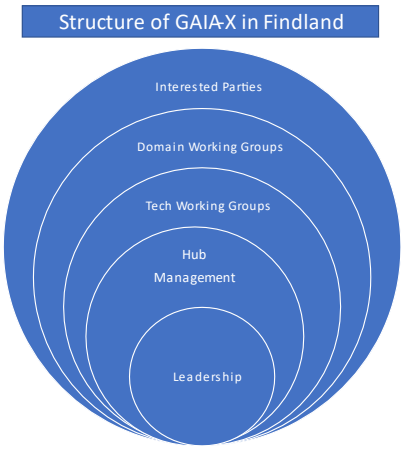
Europe's GAIA-X Project

To promote **data spaces** – that is areas of agreement between data sharing partners – and **data sharing** in general across the EU, the GAIA-X project was initiated in 2019 built upon three pillars: a *Gaia-X Association for Cloud and Infrastructure* (AISBL); *National Gaia-X Hubs* to facilitate and support the creation of **European Data Spaces** that can be specific to the local region or cross-national data spaces; and a *Gaia-X Community*.

> "The goal is a secure and federated data infrastructure that stands for European values, digital sovereignty of the data owners, interoperability of different platforms and open source. Within this ecosystem, it will be possible to provide, share, and use data within a trustworthy environment. Thus, spurring innovation and creating added value for the data economy to all who share data."



An example of a national GAIA-X initiative is **Finland's GAIA-X strategy**, that defines practices for "reliable sharing of decentralised data" to create a **data economy** consisting of **data spaces** that make use of data and services "while maintaining a high degree of reliability and information security." The private sector is seen as the major user of such data spaces, while the public sector "ensures that the operating environment is fair." This is therefore a move towards guaranteeing data sharing frameworks within legislation and regulation, and is largely seen as a way for Europe to catch up with the USA and China as "strong data economy players." The pan-EU potential of GAIA-X is to create similar data economies of scale as in the USA and China.



The structure of Gaia-X in Finland is illustrated by the above diagram, showing a cascade of stakeholders, from the co-ordinator the *Finnish Innovation Fund (Sitra)* which operates under the authority of Finland's parliament. Its objective is to create **data spaces** for **data sharing** projects within Finland and with European partners, and includes working groups of specialists in legal and contractual issues in addition to several of Finland's ministries and Finland's **Technical Research**

Centre (**VTT**). It is Sitra's role to create a GAIA-X **data hub** for the innovative and commercial use of shared data.

The GAIA Ecosystem

A useful guide to the complementary bodies promoting data sharing within the EU, as part of the GAIA-X ecosystem, is the IDSA's March 2023, Data Space Landscape that lists the following:

- IDSA and IDS data space standards certification
- FIWARE provides standards for applications developed on top of IDS that can combine data from different sources and produce information for specific purposes. As such FIWARE builds upon OpenDEI architectures and standards.
- OpenDEI supports the creation of common data platforms based on a unified architecture and an established standard. As part of the Horizon 2020 programme – now superseded by Horizon Europe – the OPEN DEI project focuses on "Platforms and Pilots" to support the implementation of next generation digital platforms in four basic industrial domains: manufacturing, agriculture, energy and healthcare.
- iShare is the European standard for and trust network of international business data sharing in a sovereign way, governed by iSHARE Foundation. Enabling federated trust governance of data spaces.
- Others: Data Spaces Business Alliance (BSBA) is a collaboration between Gaia-X Association for Data and Cloud (AISBL), the Big Data Value Association (BDVA) whose mission is to drive data-driven innovation, FIWARE and the IDSA. The DSBA aims to accelerate business transformation in the data economy by bringing together data providers, users and intermediaries developing a common framework.
- Others: Data Spaces Support Centre (DSSC) acts as a knowledge center and has created a starter kit to help organizations navigate the challenges of creating and maintaining a successful data space.
- Data Space Radar is a circular interactive map of data spaces and use cases produced by IDSA.

Data Intermediaries

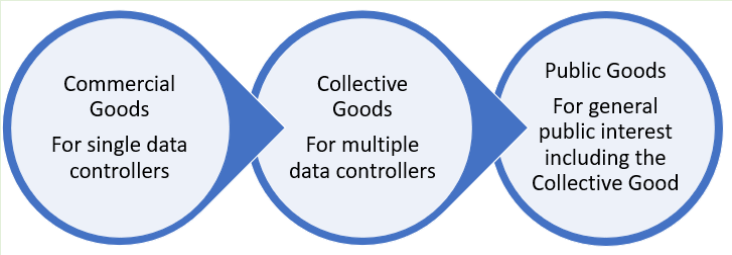
As data sharing *of all kinds* becomes commonplace, the role of data intermediaries grows in importance as trusted third parties. They can be providing services to individual persons or businesses, or to a consortia of businesses or to provide public goods. Data Hubs are a way to institutionalise the latter, and private sector Data Hubs are also appearing, for example Deutsche Telekom's *Telekom Data Intelligence Hub* initiated in 2018. See the following box.

Trust, Intermediaries, Collective and Public Goods, Marketplaces and Hubs

The creation of trust is fundamental to the sharing of data that is considered confidential or of commercial value. In recognition of this there are as many references to 'trust' as there are pages in the World Bank's World Development Report 2021 *Data for Better Lives*, over 300. Establishing trust is a process. Legally-enforceable documents such as NDAs are rarely sufficient as the horse may have bolted by the time the stable door is closed. Even with the best will in the world, honest parties may be subject to data hacking. Trust therefore implies a capacity to be 'trustworthy' in every sense. Reputation and credibility based upon long-time associations and stakeholder networking and partnership is often part of the process. The involvement of independent bodies, such as universities or of UN agencies with track records, is another route.

Data Intermediaries
The role of a trusted third party in this context is broad, but implies data intermediaries. They can be data processors who are independently analysing data entrusted to them by data controllers, for the collective good or for the public good (see below) or data processors who are contracted by data controllers to process data for their commercial interests, or who may be engaged in a wider variety of activities such as, in the words of the UK's Centre for Data Ethics and Innovation, "finding data that is fit-for-purpose, managing transfers and usage rights, and ensuring that the right protections are in place." The data controllers in many of these instances could be private individuals with attorneys acting as data intermediaries. In the case of Singapore, under the Personal Data Protection Act (PDPA) a data intermediary is defined simply as the entity that processes personal data "on behalf of another organisation."

Commercial Goods, Collective Goods and Public Goods
The varying roles of data intermediaries display a range of data sharing activities that, in the commercial rather than the purely private world can be presented in the following spectrum.



There are many data intermediaries providing data sharing facilities and analytics for a commercial good, and for a collective good, for example, using federated models. This Primer is focused on data sharing for the public good, but more often than not this can also serve a collective good. For example, the expansion of a metro line may initially reduce bus services to FMLM feeder services, but with data to identify the public benefits, compensation schemes are possible, such as financial assistance to bus operators, more flexible routings or by helping bus operators expand into adjacent businesses, e.g., on-demand services or the commercial development of bus terminals.

Data Coalitions, Marketplaces and Hubs
There are so many arising permutations of data sharing agreements and frameworks assisted by third parties it is important to make some distinguish between them.

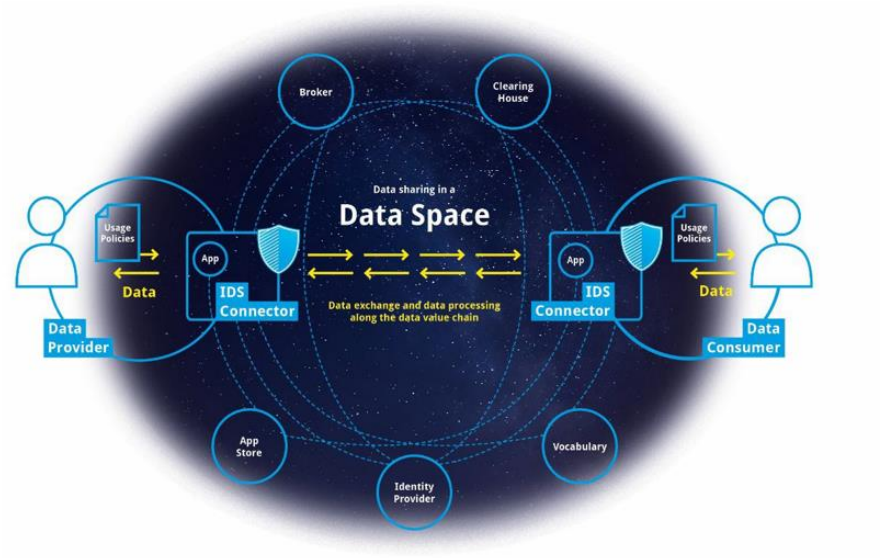
Data Coalitions: is one term that has been adopted to describe commercial arrangements whereby a third party facilitates the sharing of data as commercial data sharing or as data sharing for the collective good. Telecom companies and cloud computing service providers (CSPs) offer 'secure rooms' for such sharing In Europe, IDSA offers data sharing templates for data sharing for both commercial and collective goods, which it also certifies. In Australia, Flusso facilities data coalitions, among companies that regularly have to share data with regulators or overseas partners, and offers Privacy Enhancing Technologies (PETs) to support them.

Data Marketplaces and Data Hubs: the marketplaces provide public or private catalogues of data sets available from different companies and organisations for others to buy or to sell into. Where the commercial fee is zero, they become in effect data hubs serving the public good.

Data Spaces and the IDSA

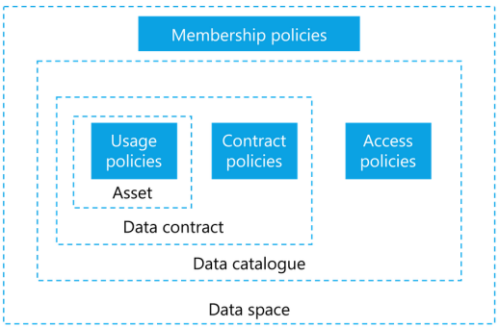
The concept of **data spaces** has been promoted by the International Data Spaces Association (IDSA), a founding member of the Gaia-X Association, to identify a set of conditions and standards that facilitate data sharing between parties. The 'space' consists of data sets linked by software 'connectors' using APIs that ensure standard interoperability and a set of data rules or framework to meet the requirements of those sharing the data as illustrated below. The connectors would, for example, facilitate federated data sharing whereby only the data model, its refinements and findings would be shared, and the data remains in its repository. IDSA has proposed a five-layered reference architecture model (IDS-RAM) using the IDS standard for participants.

IDS architecture overview

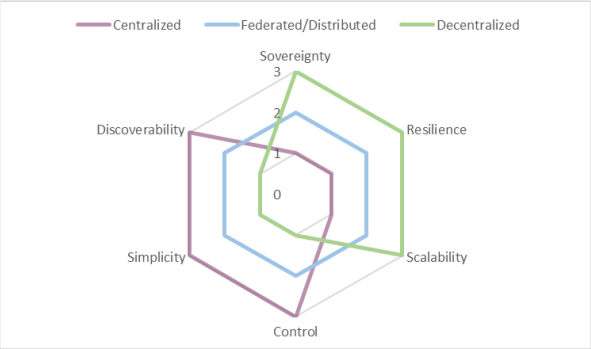


The IDSA portrays the overall governance requirements for creating a data space as composed of the following requirements.

Components of a Data Space

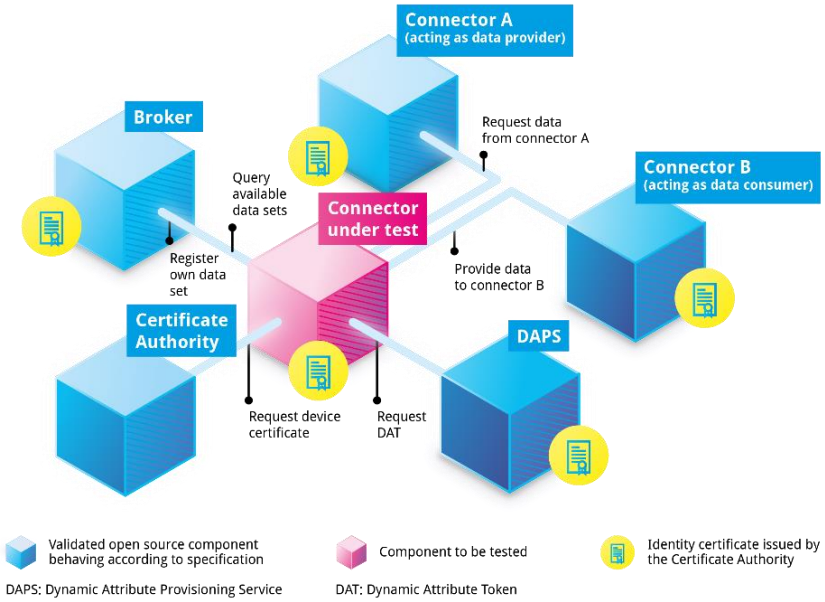


There are three configurations for data spaces: centralised, federated and decentralised. IDSA's assessment of the three options shows the federated model as second-best on all counts, but second-best can also avoid the vulnerabilities of the other two.



IDSA portrays the technical requirements for testing and certifying a data space in the following diagram which illustrates the role of a data broker as well as a data provider and a data user. IDSA states that "has made sovereignty of the data owner its most important design principle." (IDSA Front Matter, 2023.) It may be noted that 'sovereignty' is here being used in the local context of sovereignty of the subject. That places the spotlight upon data privacy concerns and the role of 'PETs' – see below.

IDS Testbed Model



Privacy Enhancing Technologies (PETs)

Concerns with data sovereignty applied to data subjects has given rise to a search for *privacy-enhancing technologies* (PETs) as outlined in a 2023 OECD paper Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches. See also the USA's PPDSA above. Besides several clauses in the *General Data Protection Regulation* (GDPR) referring to PETs, a proposal by the European Commission in 2023 with reference to statistics on housing and population, "explicitly supports the adoption of PETs 'to implement data sharing fully in line with the EU's personal data protection legislation' while 'strengthen[ing] the legal basis and encourag[ing] the development of innovative solutions to enable data sharing'." In particular, the proposal recommends "the testing and use of privacy enhancing technologies that implement data minimisation by design…" The paper goes on to say that "PETs are at different stages of development, and will likely need to be part of **data governance frameworks** to ensure they are used properly in line with the associated privacy risks."

As mentioned in the paper, Singapore, for example, funds a multidisciplinary research centre, the Strategic Centre for Research in Privacy-Preserving Technologies & Systems (SRIPTS) to explore PETs that enable "organisations to carry out data mining, analysis and sharing, in compliance with data protection regulation enacted in various jurisdictions."

Making PETS Real

Homomorphic Encryption (HE): uses an encryption scheme that limits the computational functions an external analytics provider can perform: Fully (FHE); Somewhat (SHE); Partial (PHE). Public, private and evaluation keys are issued during which private information remains encrypted 'at rest', in transit and during computation. Attracting the interest of the financial sector, but can be expensive to implement.

Differential Privacy (DP): randomised 'noise' added by computational means to data before or after release to a data processor or aggregator. DP can be interactive-query based or non-query-based; global DP where the real data is shared with a trusted aggregator or local DP where the real data is not shared.

Synthetic Data (SD): techniques that create fully or partially synthetic data sets for AI models without sharing the real data or risking IP; popular in the pharma sector but repeated use can give rise to Model Autophagy Disorder (MAD), a degradation in quality.

Contextual Integrity: uses privacy framework of Prof Helen Nissenbaum to define allowable informational flows based upon cultural, behavioural, regulatory, technical and industrial norms and the context in which those norms are found; source: Flusso - Trustworthy Data Sharing.

Federated Learning and Federated Analysis

The central idea of '**federated**' is the data involved always remains inhouse and it is the model that is shared, updated or improved, and re-iterated across all the organizations providing the data. The reiteration may be many times, for example, where ongoing research from participating laboratories results in frequent updates or new findings which are fed into the model locally and the improved model, but not the data, is then shared accordingly. The most common application to date is

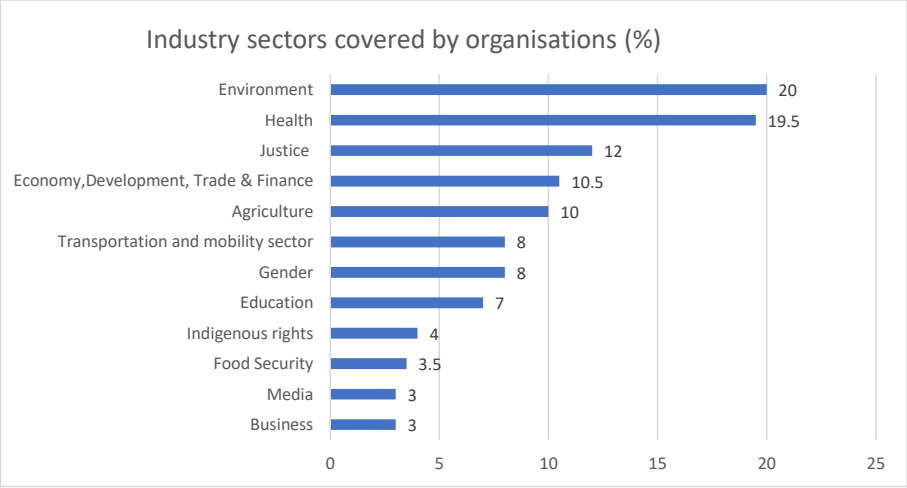
machine learning where an algorithm is taught to learn new words, concepts, identifications, etc. In 2023, ChatGPT and other generative AI apps broke into public consciousness, but the term **'federated learning'** was first **coined by Google in 2016** as they tested a consumer app, which had begun in 2013. Federated learning is often considered as a **PET (privacy-enhancing technology)** yet in terms of training machine learning AI apps the ODI notes that "federated learning lacks concrete privacy and security guarantees… issues of bias and fairness remain with federated learning and require further consideration."

'Federated analysis' is a more recent development, with relatively few publicised examples to date but as a **POC (proof of concept).** It refers to the model not as a learning tool but "is used to generate insights from distributed data, rather than to train models." (ODI) In other words, as an analytical app that researches the data to make discoveries, and yet the model could be learning while simultaneously analysing. One example of a multi-modal transport federated learning model is that of IATA covering land, sea and air.

Data Sharing Frameworks by Sectors: Health and Transport

At international, national and sector levels there are DSFs relating to specific sectors. The Datasphere Atlas provides a "mapping of organizations in the data governance ecosystem." The 2022 Atlas lists 220 government, non-government, private and academic/research organisations, plus 41 intergovernmental organisations. By geography, organisations with either a global (113) or national/domestic (73) focus were dominant (81% of the total), with most focused upon either research (139) and networking (134) which implies that data sharing for other purposes is not so well developed. Most organisations are not sector specific, but of those which are, environment and health dominate.

Table 3: The Data Atlas



To keep the Primer manageable, there follows a focus upon health and transport, the latter being the topic of the Hong Kong project (see above) and the former being an area in which the author has been working with NASEM on health-related data sharing issues.

Health and Biological Sciences Data Sharing Frameworks

Health data, alongside and often overlapping with environmental data, is the jointly most covered sector by industry. As an OECD paper in 2022 points out – see also GSMA 2021 – the pandemic of 2019 added to the urgency not only to collect data, often using mobile digital methods, but to share it between interoperable data repositories, especially across government agencies as a Whole-of-Government approach (WGA).

> The 2019 Coronavirus pandemic (COVID-19) underlined the need for cross sectoral re-use of data and thus for more coherence across sector specific data governance frameworks. For example, anonymised mobile call data records (CDRs) of telecommunications services providers have been re- used to monitor and control the spread of COVID-19 and other pandemics. OECD

Bio-Genomic Frameworks

For this Primer, two examples of data sharing protocols and frameworks are discussed. The first is from the journal BioPreservation and BioBanking (2015) *Shaping and Reuse of Sensitive Data and Samples: Supporting Researchers in identifying Ethical and Legal Requirements*. This article provides numerous references to other articles and policy documents concerned with the sharing of biomedical data, from samples to records. It recognises many trade-offs are involved, including "between maximising data sharing while minimising possible risks to research participants' privacy as a key consideration." The geographical boundaries of data sharing is an issue. For example, if the German Cooperative Health Research in the Region of Augsburg (KORA) project receives a request from beyond Germany, "for kidney cancer-related bio samples from a researcher who is located (for example) in the UK, it has to consider the more general European framework."

The specifics of data regulations do differ across member states of the EU, while Article 8 of the EU Data Protection Directive imposes "an enhanced level of protection on 'special categories' of data, including health data, and although anonymised data is exempt from personal data restrictions, pseudo-anonymous and linked anonymous data "remain personal data, at least for the data controller who has access to the key or the cypher." Sharing such data across the EU is not permitted for linked-anonymous data except where the subject has given specific consent. But specific consent is difficult to apply to a wide variety of research inquiries, especially where a biobank is involved which relies upon the reusability of data sets. To address this practical problem a concept of 'broad consent' has emerged, but local regulations may still restrict its application to specific data sets available in the biobank. A biobank is a form of data trust which is defined by the Open Data Institute as having fiduciary stewardship, such as the UK Biobank established in 2006 to steward genetic data and samples from 0.5m people and takes the form of a charitable company with trustees. Some platforms are available for storing and sharing biological data on the web, such as PlutoF Go.

This 'fragmentation of regulation' can be navigated by researchers and research institutes in two main ways. First, guidance is provided by portals outlining what data is accessible, such as the Genomic Commons Data Portal and the Cancer Genome Atlas (TCGA) provided jointly by the US National Cancer Institute and the National Human Genome Research Institute (NHGRI); and the CORBEL database of the European Marine Biological Resource Centre (EMBRC) which is described as Europe's 'research infrastructure' for marine biological resources. In China, the China Science Technology Cloud (CSTC) offers a cloud-based platform for scientific information sharing.

Second, search tools are available, such as the Human Sample Exchange Research Navigator as a guide to the practical and legal requirements for the sharing of human biological data. As Sariyar et al. argue, "helping data providers to identify relevant ethical and legal requirements and how they might address them is an essential and frequently neglected step in removing possible hurdles to data and sample sharing in the life sciences." The Public Population Project in Genomic and Society (P3G) provides a literature guide to data frameworks, for example, *Framework for responsible sharing of genomic and health-related data* in the Hugo Journal (2014). The 2015 paper provides 43 references, and since then hundreds more have appeared. A series of survey articles of major clinical data sharing platforms is available at YODA.

A second example is the framework of the Canada-based Global Alliance for Genomics and Health (GA4GH.org) available in a paper in 2014 *Framework for Responsible Sharing of Genomic and Heath-Related Data*, published by the *Hugo Journal* and available at the National Library of Medicine. Its ethical foundations, drawn from the UN's 1948 *Universal Declaration of Human Rights*, are stated as: "the right of all people to share in the benefits of scientific progress and its applications as being the duty of data producers and users to engage in responsible scientific inquiry and to access and share genomic and health-related data across the translation continuum, from basic research through practical applications." Of course, the moral stance may not always accord with the commercial interests of, for example, the pharmaceutical industry any more than it did with the tobacco industry. In practice the arguments often revolve as much around how to fund research as to how to realise its benefits.

The response to COVID-19 did produce innovations in the sharing of health-related data alongside its digitalisation. OECD (2022) shows these "included the development of health dataspaces or hubs that could approve access to health data, link health data and provide secure mechanisms for access to data. These innovations are numerous, including a Health Data Hub in France, FinData in Finland, the Health Dataplace in Australia and OpenSAFELY in the United Kingdom." The report adds that "Open government data has proven to be a foundation for the global response to the COVID-19 pandemic …Leading OECD countries in the area of open government data such as Korea were quick to enable creation of citizen services from open data (including those released by private actors)." In many cases these innovations also helped empower patients.

The GA4GH framework develops core elements of responsible data sharing as follows:

Transparency	Risk-Benefit Analysis
Accountability	Recognition and Attribution
Engagement	Sustainability
Data Quality and Security	Education and Training
Privacy, Data Protection and Confidentially	Accessibility and Discrimination

OECD Recommendations on Health Data Governance Frameworks

The OECD (2023) *Emerging Privacy Enhancing Technologies* provides several references to international and country-specific data sharing frameworks and initiatives including those specifically focused upon biological health and medical services. One such is the secure data analytics platform of the National Health Service (NHS) in the UK "enabling researchers to analyse millions of patients' electronic health records" using the industrial OpenSAFETY protocol. OpenSAFETY is described by B&R, a manufacturer of automation technology, as "the first open and only truly bus-independent

safety standard available for all industrial Ethernet and fieldbus solutions." The OECD report further identifies laws in various countries governing health-related data, such as the US (1966) Health Insurance Portability and Accountability Act (HIPPA). Globally, a starting point for all health-related data are the privacy laws of each country.

In 2016 the OECD published a set of recommendations for health governance frameworks and according to OECD (2022) a follow-on study in 2017 examining the use of electronic health records EHR) found "that only a limited sub-group of countries have both strong technical and operational readiness to extract data from these systems for statistics and research, coupled with a health data governance framework and investments supporting data use." The 2022 report makes an assessment of the implementation of the 12 guiding principles set out in the 2016 report.

OECD's 12 principles to encourage greater cross-country harmonisation between health data governance frameworks:

- Engagement and participation of stakeholders in the development of a national health data governance framework;
- Co-ordination within government and co-operation among organisations processing personal health data to encourage common data-related policies and standards;
- Reviews of the capacity of public sector health data systems to serve and protect public interests;
- Clear provision of information to individuals about the processing of their personal health data including notification of any significant data breach or misuse;
- The processing of personal health data by informed consent and appropriate alternatives;
- The implementation of review and approval procedures to process personal health data for research and other health-related public interest purposes;
- Transparency through public information about the purposes for processing of personal health data and approval criteria;
- Maximising the development and use of technology for data processing and data protection;
- Mechanisms to monitor and evaluate the impact of the national health data governance framework, including health data availability, policies and practices to manage privacy, protection of personal health data and digital security risks;
- Training and skills development of personal health data processors;
- Implementation of controls and safeguards within organisations processing personal health data including technological, physical and organisational measures designed to protect privacy and digital security; and
- Requiring that organisations processing personal health data demonstrate that they meet the expectations set out in the national health data governance framework.

Source: OECD (2022) *Health Data Governance for the Digital Age: Implementing the OECD Recommendations on Health Data Governance* covering the period 2016-2022; see OECD (2016) Recommendations on Health Data Governance for the original document.

The potential fragility of the underlying principle of subject privacy which runs through all these framework documents was highlighted in May 2023 by widespread media coverage, such as CNN, that DNA samples from human as well as non-human animals could be plucked freely out of the air. Researchers at the University of Florida could "match genetic information to individual participants who had volunteered to have their DNA recovered as part of the research that published in the

scientific journal *Nature Ecology & Evolution*." Although identifying individuals from a myriad of free floating DNA samples in the open air may be close to impossible, such identification in more enclosed spaces has been demonstrated by a forensic research team from the Oslo University according to the New York Times. To imagine such a technology at work in say Nazi Germany in the 1930s is a horrifying thought, yet it could be only a matter of time before an argument for police detective work to solve crimes might find rational support. The needs for implementable safeguards, laws, accountability and transparency are quite urgent.

Transport Data Sharing Frameworks

Data sharing within and across the public transport sectors seems to have a different set of challenges from those within and across the public health and bio-genetic sectors. There is less laboratory R&D that would call for cross-border collaboration for one thing. Sharing designs for transport technologies, for example, is likely to involve IPRs and national standards issues, unlike the sharing of biological samples. Technological innovations, such as algorithmic systems for autonomous vehicles and standards for electric vehicle recharging points are mainly developed by private manufacturers and vendors who are more comparable to private pharmaceutical companies than to publicly-funded research institutes. Among others IBM's Institute for Business Value promotes the business case for sharing across transport ecosystems. The EU's DATA4PT project 2020-2024 is designed to advance public and private transport data standards and sharing across Europe.

But it is the traffic and other operating data such as routes and frequencies and passenger loadings of different modes of public transport that is most important for public service planning purposes, as well as for plans to shift traffic from private to public modes of transport. The UK Government provides a general overview of transport data sharing. Above all else, such planning needs Big Data, that is data from a variety of sources, such as alternative transport modes, a range of demographic data including the distribution of populations, climatic conditions, emission levels, the availability of refuelling or recharging facilities, licensing laws, etc. Big Data implies a need for data collection and for data sharing between the different sources of the data. In cases, such as Singapore, where government owns much of the transport infrastructure and vehicles and franchises out the services to commercial operators, data collection is written into the agreements. In cases such as Hong Kong where most of the transport infrastructure is privately owned, perceived commercial self-interests inhibits data sharing.

Hong Kong - this was the motivation of the research project initiated by the author of this Primer and a team of collaborators run under the auspices of the University of Hong Kong, 2020-2021, to create a Data Trust which could act as a Trusted Third Party to reassure the metro and bus operators to share a defined and limited amount of data as a Proof-of-Concept that public transport data analysis for planning purposes could be achieved even under privately-owned operating conditions. If there was a defining issue in terms of operator buy-in it came with the acknowledgement that smart city development would rely upon Big Data which implied data sharing for the public good, and success in smart city development was potentially beneficial for *all* operators – but *only if* government used evidence-based planning to structure a risk-reward system accordingly across the entire transport system. This follows in two distinct ways. First, if the result is a redesigned and fully-integrated public transport service system (including MaaS and payments systems for example) that encourages a shift from private cars to public transport, such that even relative losses of market share by some public transport operators result in absolute increases in revenues. Second, feeder

routes may represent relative market share losses if they are substitutes for longer-distance journeys on those modes of transport, such as buses to metro for example. But if there are revenue-sharing arrangements to compensate then both modes of transport can benefit – gross revenue sharing for net revenue gains. As the initial research 2020-2021 was a small sample size as a PoC these arguments were implied rather than developed. A larger scale research project could address them.

UITP (Union Internationale des Transports Publics) – in a research paper funded by the Land Transport Authority of Singapore (LTA) in 2020 Sharing of Data in Public Transport focuses upon the business impediments to transport data sharing. Based upon a survey over 100 organisations "from transport and other sectors" it enumerates the 'pros and cons' of data sharing from the business perspective. The 'pros' being listed as ways to achieve greater customer satisfaction, discovering new business opportunities, improving urban mobility and operational excellence and complying with open data policies. The 'cons' or main barriers to data sharing being revealed as privacy concerns, liability risks if something goes array, ambiguity of data ownership and the risk of losing competitive advantage. The report outlines three data sharing models: open access, bilateral restricted access and multilateral restricted access. Removing or lessening the impediments is seen as requiring government initiatives. This accords with the views we received from the private sector in Hong Kong, while the government saw the principle problem as being private sector inertia over data sharing. Risk of data loss, privacy violations, loss of competitive advantage, seems to have a bearing over uncertainty of data ownership, unknown resource commitments involved in data sharing, and of benefits arising from use cases. UITP followed up in 2021 with recommendations in a Policy Brief: A Framework for Sustainable Data Sharing in Public Transport.

World Business Council for Sustainable Development (WBSCD) – together with the International Road Federation (IRF) and the Sustainable Mobility for All Initiative (SuM4All) in 2020 published Enabling data-sharing: Emerging principles for transforming urban mobility, pointing out in the media release the need for "the right policy framework in place to enable data sharing between public and private actors…" The 2020 paper, which has a heavy emphasis upon personal mobility modes of transport, such as ride-sharing, bike-sharing, e-hailing and the use of road space, is very much the product of its period, when the rise of the 'sharing economy' (otherwise known as the P2P economy) was in full swing, founded upon short-term data sharing over the Internet but also too often on a fragile demand which COVID-19 did much to undermine. But the paper goes on to note:

> "While the technology challenges should not be overlooked, arguably the more pressing difficulties stem from the complicated interaction between public policy goals, private citizen rights and expectations and business needs. Those differing priorities manifest in many of the ways common to mobility data-sharing overall: competitive concerns, privacy and security issues and uncertainty over risk and liability."

Concerns over protecting commercial interests, security and liability risk are forever the same issues raised in most of the papers on data sharing. The paper then examines in some detail the data sharing principles that should apply in five use cases: integrated multimodal transport systems, optimisation of real time fleet management, mobility for low-income populations, road infrastructure design to address road bottlenecks, and electric vehicle use. For the integrated multimodal transport use case, the paper espouses the shift towards Mobility-as-a-Service (MaaS) from the integration of different modes to the integration of payment and booking systems and of personalised choice of services. In this context it notes maximum benefits "hinges on a consumer's

willingness to provide personal information about where, when and how they travel." Needless to say, this is a step beyond data sharing between operators and regulators, and in principle calls for the sharing of mobile data *in the public interest*, for example from mobile phones, by telecom operators and service providers subject to forms of 'consent' (specific or broad) with rigorous privacy safeguards.

OECD International Transport Forum (ITF): in 2022 launched a Mobility Innovation Hub with core funding from Korea, publishing Measuring New Mobility Definitions, Indicators, Data Collection as a guide to data collection, including "how governments and private stakeholders can collaborate to improve their understanding of New Mobility and determine if and where policy interventions are needed", but its focus is on "data reporting, not data sharing." As in the case of the WBSCD, the emphasis upon New Mobility leads to a focus on how new shared-mobility operators use data which is shared not between operators but between operators and their customers. This was the likely focus in 2002 of the workshop on Data sharing architecture for Mobility as a Service.

Mobility-as-a-Service(MaaS): the MaaS Alliance defines MaaS as integrating "various forms of transport and transport-related services into a single, comprehensive, and on-demand mobility service… accessing mobility through a single application and a single payment channel (instead of multiple ticketing and payment operations)." It became a business model of many start-ups such as Whim during the dawn of the 'sharing economy' being available through apps on mobile devices and harvesting customer data, but the COVID pandemic which hit all forms of paid transport hard created cash-flow problems for most of these start-ups, including Whim with layoffs and global closures. The data-sharing models involve both customers sharing their data *by default* – although the MaaS companies are bound by privacy laws – and bi-lateral agreements with public transport operators and/or the use of Open Data repositories of routes and timetables. A Parliamentary report in Germany in 2023 maintains the inter-modal vision.

The **National Association of City Transportation Officials** (NACTO) also addresses the role of data sharing and mobility operators in a 2019 policy paper Managing Mobility Data, a joint product with the International Municipal Lawyers Association, which "sets out principles and best practices for city agencies and private sector partners to share, protect, and manage data to meet transportation planning and regulatory goals in a secure and appropriate manner." Mobility data is defined in terms of information:

> "generated by activity, events, or transactions using digitally-enabled mobility devices or services. This data is frequently-recorded as a series of points with latitude and longitude collected at regular intervals by devices such as smartphones, shared micromobility vehicles (shared bikes, e-bikes, scooters etc.), on-board vehicle computers, or app- based navigation systems (e.g. Waze, Google Maps etc.). Mobility data often has a temporal element, assigning time as well as location to each point. Depending on the device used to capture the data, other characteristics, such as the speed of travel, or who is making the trip, can be connected to each individual latitude/longitude point."

And data may be described in terms of:

- GPS Trace/Breadcrumb Trail
- Individual Trip Records
- Location Telemetry
- Data Protection
- Verifiable Data Audit

PII and Aggregation - NACTO is concerned with the challenges of **Personally Identifiable Information** (PII) when data is shared, noting that **Recognizable Travel Patterns** are open to data triangulation, citing a study that found in "a dataset of 1.5 million people over 6 months, and using location points triangulated from cellphone towers, 'four spatio-temporal points are enough to uniquely identify 95% of the individuals'." **Combined With Other Data** sometimes referred to as *indirect or linked PII* is also insufficient to safeguard against a determined researcher with access to the aggregated data. The paper cites the example of "in 2014, a researcher requested anonymized taxi geo-location data from NYC Taxi and Limousine Commission under freedom of information laws, mapped them using MapQuest, and was able identify the home addresses of people hailing taxis in front of the Hustler Club between midnight and 6am. Combining a home address with an address look-up website, Facebook and other sources, the researcher was able to find the property value, ethnicity, relationship status, court records and even a profile picture of an individual patron.'" The paper argues that appropriate "data aggregation is the key tool for managing the balance between access and privacy. However…the exact thresholds (how many data points are needed per hour to ensure anonymity?) are not universally agreed upon."

The paper identifies four principles for managing mobility data with recommendations for each: data for the public good, protected, purposeful and portable. The recommendations for data for the public goods are listed below.

Cities should:

- Require access to data from mobility services operating in the public right-of-way as a default requirement for operating in the public realm. Cities need a wide variety of data in order to make informed decisions and policies.
- Use their authority to issue and enforce contractual agreements to guide private sector actions and protect the public interest. Cities should strive to select vendors who collect, manage, and share data in a manner that aligns with city privacy policies. Where possible, cities should reinforce policy goals through rigorous enforcement of contractual terms. Cities should expand their internal capacity to analyse the data they receive and to confirm data quality.
- Develop or update strategic plans for managing mobility in a digital age to address data management, adequate training, and appropriate insurance coverage and safeguarding procedures.
- Coordinate to create or adopt standardized, open data formats that level the playing field between companies and transportation providers by making expectations about information sharing and management more consistent and predictable across cities. Tools such as the Mobility Data Specification developed by the City of Los Angeles are one step toward a unified standard.

NACTO concludes that city practices need to be continuously updated as the technologies and business models evolve, a point emphasised in NACTO's 2017 paper City Data Sharing Principles: Integrating New Technologies into City Streets. The big question is how many cities have reached this stage of data sharing and data sharing frameworks.

Conclusions

The vision of open data and shared data for the benefit of society is running in parallel with the spread of data in all walks of life. However there are hurdles to be overcome or worked around. The following summarises some of the findings above.

1. Despite many and growing examples of data sharing, they are unevenly spread across industrial sectors and for the achievement of (i) a commercial good among entities with a common ownership structure, (ii) a collective good among coalitions of stakeholders which may be occasional or permanent coalitions, involving trusted third-party intermediaries using either a data trust model or a federated learning model, and (iii) a public good which can also incorporate elements of a collective good.
2. The multiple data sharing frameworks (DSF) each require an elucidated set of data governance principles, and creating these can be as time consuming, or more so, as the actual data sharing and data analytics they are designed to facilitate. Ultimately, they revolve around a concept of trust, a trust in the goodwill of the data intermediaries and the partnering data controllers, and in the enforceability of the DSF itself.
3. There is a plethora of laws and regulations governing different aspects of data sharing, from security to personal data privacy to intellectual property, and while most countries share the principles, such as those enumerated in the EU's GDPR and the OECD's Data Protection Rules, they differ in the way they apply them. This makes cross-border data sharing especially challenging and adds to the costs of compliance.
4. A precondition for successful data sharing are data standards, using compatible hashing algorithms, interoperable APIs and the interconnectivity of data platforms using secure bridges.
5. Industrial and civil sectors differ widely in their contemplation of data sharing. In academia data sharing is widely regarded as the cornerstone of credibility. Across government, a Whole-of-Government Approach (WGA) involving data sharing between agencies and between the public and private sectors is considered necessary for smart city planning and development. In research-focused activities the attitudes towards data sharing range from an academic approach to a public good approach, such as sharing health data during times of a pandemic, to a private commercial approach that stresses IPRs and competitive advantages. This applies to most other private commercial sectors.
6. Sharing the analytical models rather than the data itself is the federated data analytics model . It preserves confidentiality and protects data ownership and is therefore one of the more popular, yet still evolving, methods of data sharing for a commercial good and a collective good.
7. This Primer is focused principally upon data sharing for the public good, whilst recognising that serving a collective good is both desirable as a way to encourage innovation and collaborative planning with public authorities, such as in the case of an integrated transport system or a universal health service, and necessary as an incentive to private commercial businesses to participate in trusted third-party data sharing models.
8. The experience of Data Trust 1.0 (DT1) outlined in the Introduction was a Proof-of-Concept that private and commercially-run transport operators in Hong Kong could share data as data controllers for both a collective and a public good using a trusted third-party model. But many such projects end up as one-offs, the learnings largely lost. This Primer spells out good reasons why this should not be the case and there should be a Data Trust 2.0 (DT2).
9. However parallel lines of their own accord don't meet, and while it should not be difficult to achieve data and information sharing within academia and between research institutes, the

commercial perceptions of the private sector frequently override commitments to the public good, even when the shared benefits arising from certain types of data sharing should be evident. Other reasons include the workloads that could be involved in arranging for safe and secure data sharing. It is not an exercise without costs, but often intermediaries using, for example, a trusted third-party model can reduce these. The shortage of data scientists within the private sector also limits the use and value of the data available, a value that could be realised through data sharing. Again, intermediaries can overcome these limitations, whether through federated or data sharing models.

Annex

[*Location*] Public Transport Interchange (PTI) Proof of Concept:

Information Pack and draft MOU for Participants

The [*location*] represents one of [*the jurisdiction's*] most important public transport hubs. Located in the heart of the Central Business District (CBD), it serves most of [the jurisdiction's] Public Transport Operators (PTOs), including … [details of transport modal types involved]. The coverage, capacity and operational status of each of these transport modes is defined by static and dynamic data as well as [*Metro*] regulations governing the provision of services from various public and private sector organisations.

The parties now wish to research how data sharing can result in better connectivity, improved balance of demand across modes and improved utilisation of the [location] PTI operations, thereby benefiting all user segments, including the people and businesses that depend on the efficient operation of the [location's] PTI ensuring easy access amongst all modes and destinations. Users include transport service providers, travellers to and from the CBD, the tenants of related buildings (and their landlords) and Government. Air quality and other environmental data shall also be integrated into the data-sharing model.

The successful implementation of the [location's] PTI Data-Sharing Proof of Concept ("POC") will depend on selected key stakeholders that will – for a limited time – test the viability of data sharing, linking and using mobility-related time and location data to analyse the potential of improving strategic planning, improve the operational efficiencies of transport assets and demonstrate the potential for closer alignment of transport operations to benefit users. Gaining insight from data collected from independent sources, such as trying to understand the mode choice of users, an understanding of trip purpose, knowledge of the end-to-end total fare, travel time and its variability is required. To fuse data from various sources requires the alignment of one or more common variables such as the identification of the means of payment (e.g. *payment* Card ID), the location of a payment event, and static data that describes transportation infrastructure and the routes applicable to public transport, amongst other factors.

It is noted that several parties have confirmed their interest in enabling the POC, The (*Trusted Third Party*) and one or more sources of transport data:

- The [*Trusted Third Party*] intends to develop a Data Trust and expertise on the technical aspects of data protection and storage. Therefore, [*Trusted Third Party*] offers to define and implement the approach and mechanisms for data hashing, encryption and storage; and
- More than one organisation has agreed with [*Trusted Third Party*] to provide transport data;

The POC aims to establish the principles of data sharing and apply them to historic travel data and static data provided by POC participants, assisted by [*Trusted Third Party*] and other parties subject to agreement. Also, every participant will be invited to submit a limited quantity of research questions to determine if the data provided by all participants collectively could answer such questions.

A draft MOU is attached to invite discussion on the POC and setting out the terms to be refined to the stage where each participant enters into bilateral understanding with [*Trusted Third Party*] (acting in its role of data processor) and a Transport Data Analytics Service Provider (TDASP), on the purposes and responsibility of each party.

This draft MOU between a TDASP and [*Trusted Third Party*] complements MOU v9 that reflects a relationship between a Data Controller and [*Trusted Third Party*]. The attached MOU contains substantially the same terms.

Memorandum of Understanding (MOU) for

the Processing and Use of Data

THIS MEMORANDUM OF UNDERSTANDING ("**MOU**") is made on the ___ day of _____ 2020 ("**Effective Date**") for participation in the Exchange Square PTI data-sharing Proof of Concept Phase 1 ("**POC**").

BETWEEN

(1) [Company]; a limited company/corporation incorporated under the laws of [the jurisdiction] whose registered office is at [address] ("**Company**").

(2) The [*Trusted Third Party*] whose address is at …

collectively known as "**the Parties**" and each of them known as a "**Party**"

WHEREAS

A. Mobility services at and in the proximity of [*the location*] Public Transport Interchange (the "**PTI**") and its environs are primarily mechanised, comprising services provided by [*transport companies and modes involved*];

B. First mile, last mile travel depends on at-grade and grade separated pedestrian facilities including footpaths, subways and elevated walkways;

C. The [*Trusted Third Party*] intends to collect transport-related data from various sources relating to the PTI to be used for preparing statistics and carrying out research into mobility services and the potential for the improvement of such services (the "**Collaboration**"). [*Trusted Third Party*] shall provide IT and related security management resources that is capable of receiving data (as defined below) and the processing of such data (the "**Data Trust**") with the support of one or more service providers (each a "**Transport Data Analytics Service Provider**")

D. There will be several sources of data (each a "**Data Controller***"*)

E. There will be several types of data, including:

   (a) dynamic data, including payment card transaction records and pedestrian flows;

   (b) static data, including the location of bus stops, location of Automatic Fee Collection (AFC) gates at metro station; and

   (c) meta data, including metro exit opening and closing times, pedestrian access restrictions.

   (collectively referred to as "**Data**").

THE PARTIES HEREBY AGREE

1. For the avoidance of misunderstanding, it is expressly agreed that, this MOU does not create any legal obligations between the *Parties*. Instead, it sets out some intended key terms and conditions of an agreement for the *Collaboration* which will be negotiated in good faith between the *Parties* following the execution of this MOU (the "**Agreement**").

2. To establish the principles and mechanisms of data sharing for transport services in [the jurisdiction], as applied to data that relates to user behaviour and transport infrastructure (the **"Purpose**").

3. The *Company*, as a *Transport Data Analytics Service Provider* ("**TDASP**"), consents to provide one or more data processing algorithms (each an "**Algorithm**") and directly related support services to [*the Trusted Third Party*] as the *Data Processor* for the *Purpose*, in compliance with all relevant laws of the [the jurisdiction] Special Administrative Region, including the Personal Data (Privacy) Ordinance ("**PDPO**").

4. The provision of services by the *TDASP* shall be at the sole cost and expense of the *TDASP*. The hosting of the *Algorithm* by the *Data Processor* and related processing and storage of *Data* by the *Data Processor* shall be at the sole cost and expense of the *Data Processor*.

5. The *Data* will include data collected on every calendar day of May 2019 or otherwise agreed between the *Data Controller* and *Data Processor* ("**Period**") in machine readable format (e.g. CSV, JSON, XML), as illustrated in Appendix B.

6. The *Data* may also include payment events and their location and validated by the *Data Controllers* – for journeys that originate and/or terminate in [*the selected location*] and its environs for the Period.

7. Prior to the transfer of *Data* to the *Data Processor*, each *Data Controller* shall apply a hashing algorithm in accordance with Appendix A to Personal Data included in the *Data* (as defined by the PDPO) and then encrypt any Personal Data included in the *Data* (e.g. *payment* Card ID). See Appendix A.

8. The approach to hashing and encryption shall be defined by agreement between the *Data Processor* and each *Data Controller*, with the aim of ensuring a common approach by all *Data Controllers*. See Appendix A.

9. The *Data Processor* will prepare **Aggregated Data** based on the *Data*. The *Data Processor* will grant the *TDASP* access to the *Aggregated Data* as:

   (a) a small sample ("**Sample Data**");

   (b) an extract for the purposes of training the *Algorithm* ("**Training Data**") by the *TDASP*; and

   (c) the remaining *Aggregated Data*.

10. Throughout the term of the Agreement, [*the Trusted Third Party*] offers the use of its Data Trust, data storage and data analysis resources for the Purpose at no cost to the *TDASP*.

11. Save and except for the publication right under Clause 16, communications with the general public will be limited to an announcement that the *TDASP* and other parties as applicable, have agreed to the *Collaboration* for the benefit of the travelling public and to reduce air pollution ("**The Announcement**"). See Appendix D. Other than the Announcement, the *Parties* shall not use the name "The [*Trusted Third Party*]", the Company's name, or any variation, adaptation, or abbreviation thereof, or the name of any of [*the Trusted Third Party's*] trustees, officers, faculty members, students, employees, or agents, or any trademark owned by [the Trusted Third Party] or the Company, in any promotional material or other public announcement or disclosure without the prior written consent of [the Trusted Third Party] or the *Company* (as the case may be), which consent [the Trusted Third Party] or the *Company* may withhold in its sole discretion.

12. Save and except for the rights provided to the *Company* under Clause 17 related to the *Algorithm* provided by the *TDASP* to the *Data Processor*, the *Agreement* shall include provisions indicating that the sole and beneficial ownership of intellectual property rights ("**IPR**") arising from and in connection with the *Projec*t, including but not limited to the Inventions, results, findings and information of the *Project* shall be vested in [the Trusted Third Party]. Subject to the TDASP being a company incorporated in [the jurisdiction] or otherwise a party approved in writing by the Commissioner for Innovation and Technology, within three months after the expiration of the Agreement, the TDASP may obtain a non-exclusive, non-sublicensable, royalty-free licence for the TDASP to use the *IPR* for non-commercial purposes, the terms of such license shall be negotiated in good faith between the [Trusted Third Party] and the Company.

13. The Agreement shall also include provisions relating to the following:

    (a) purpose(s) of the processing of *Aggregated Data* , substantially as defined by the *Collaboration*;

    (b) permitted use(s) of the *Aggregated Data*;

    (c) requirement to keep records of *Aggregated Data* that is shared (including the parties involved; the permitted uses of personal data; details of the transfer, update or deletion of the data);

    (d) restrictions on further sharing of the *Aggregated Data*;

    (e) data security measures to be implemented by the *Data Processor* to protect the *Algorithms;*

    (f) requirement on the *Data Processor* to give notification to the *TDASP* in case of breach of the data security measures;

    (g) the retention period and deletion arrangement of the *Algorithms*, to prevent any *Algorithm* transferred to the *Data Processor* from being kept longer than is necessary for the *Collaboration*;

    (h) individuals' rights (e.g. data access and data correction rights; right to withdraw consent) as to the Algorithm;

    (i) the arrangement between the *TDASP* and the *Data Processor* for correcting and/or updating *the Algorithm*;

    (j) sanctions for failure to comply with the *Agreement*;

14. The *Algorithm* shall not be not used for any other purpose than as defined by the *Collaboration*.

15. Confidential Information

    15.1 "**Confidential Information**" means all information provided by one Party (the "**Disclosing Party**") to the other Party (the "**Receiving Party**") and clearly identified as confidential by the Disclosing Party at the time of disclosure. 15.2 Specifically excepted from this definition is all information: (a) known by the *Receiving Party* at the time of disclosure; (b) publicly disclosed except by breach of this *Agreement*; (c) rightfully received by the *Receiving Party* from any third party without an express obligation of confidence; (d) independently developed by the employees or agents of the *Receiving Party* without any knowledge of the confidential information provided by the *Disclosing Party*; or (e) is disclosed pursuant to any judicial or government request, requirement or order, provided

that the *Receiving Party* takes reasonable steps to provide the *Disclosing Party* with sufficient prior notice in order to allow the *Disclosing Party* to contest such request, requirement or order.

15.3 The *Receiving Party* agrees to hold the *Confidential Information* in trust and confidence for the *Disclosing Party*, using the same care and discretion that the *Receiving Party* uses with similar information which it considers confidential. The *Receiving Party* shall not use the *Confidential Information* other than for the benefit of the Parties and relating to this Agreement, the *Receiving Party* shall not disclose the *Confidential Information* without authorisation from the *Disclosing Party*.

15.4 This provision shall remain in effect during the term of this *Agreement* and for three (3) years thereafter.

16. Publication

16.1 [*The Trusted Third Party*] shall be entitled to publish the results of the Collaboration and [*Trusted Third Party*] shall submit the manuscript of each proposed publication to the *Company* for its review at least ten (10) days prior to the scheduled publication. The *Company* will discuss in good faith with [the Trusted Third Party] and complete its review within 10 days by a written reply. If [the Trusted Third Party] does not receive any written reply from the *Company* within the 10-day period, the proposed publication shall be deemed to be approved by the *Company*.

16.2 Authorship of each publication that includes content developed by the *Company* or from use of an Algorithm provided by the *Company* will be determined by the principle of fairness and the relative contribution of the *Parties* to the results published in such publication.

17. Intellectual Property Rights

17.1 The *Company* shall retain the **IPR** in each *Algorithm* that it provides to the *Data Processor*, in its sole name at all times. The disclosure of an *Algorithm* by the *Data Processor* shall be subject to the written approval of the *Company*.

17.2 The *Company* shall forthwith upon request of the *Data Processor*, conditionally grant to the *Data Processor* a non-exclusive, royalty free license to use the IPRs in the *Algorithm*, and for a period of 2 years thereafter or until after the Effective Date or on completion of the *Collaboration*, whichever is the shorter.

18. One of more Data Controllers intend to enter into a separate agreement with the Data Processor to participate in the POC Phase 1 on substantially the same terms as set out herein. See Appendix C.

19. Term

21.1 Subject to Clause 21 hereof, the term of this MOU shall commence on the *Effective Date* and last for two (2) years after the *Effective Date* or on completion of the *Collaboration*, whichever is the shorter ("**Term**"). The Parties will negotiate in good faith in order to sign a final and legally binding Agreement in relation to the Collaboration within 30 days after the *Effective Date*, failing which this MOU shall automatically become null and void.

20. Termination

22.1 Either *Party* shall have the right to terminate this MOU forthwith at any time by giving to the other *Party* notice in writing to that effect and claim for damages if the other *Party* is in breach of any of its obligations under this MOU and the defaulting party has failed to remedy that breach within 30 days after receiving written notice from the other *Party* requiring remedy.

22.2 Notwithstanding anything to the contrary in this MOU, the *Company* may at any time give 14 days' prior written notice to [the Trusted Third Party] to terminate this MOU.

22.3 The right to terminate this MOU given by this MOU shall be without prejudice to any other right or remedy of either Party in respect of the breach concerned, if any, or any other breach.

21. Other parties may collaborate closely with [the Trusted Third Party] to offer technical advisory services and liaison to all parties of the *Collaboration*. Each such party will enter into a separate Non-Disclosure Agreement with [the Trusted Third Party] to protect *Confidential Information* relating to the Purpose.

22. The *Parties* agree to jointly collaborate on a non-exclusive basis, and each Party shall be free to carry out activities similar to the ones mentioned herein, including without limitation by developing products, receiving services from others or providing services to others.

23. Each *Party* shall bear and be responsible for its own costs, including reasonable legal costs, incurred in connection with the preparation of this MOU and the preparation and completion of the Agreement and all matters related thereto. For the avoidance of doubt, the *Company* will not be responsible for any costs and expenses incurred by [the Trusted Third Party], *Data Controllers* or any other *TDASPs* or parties involved in the *Collaboration* or *Project*.

24. Any notice requiring to be given under this MOU shall be given in writing and shall be deemed duly served if left at or sent by registered or recorded delivery post or sent by facsimile transmission or e-mail to the relevant party:

(a) in the case of the Company:

Address	:	[*]
Facsimile	:	[*]
Attention	:	[*]
E-mail	:	[*]

(b) in the case of [the Trusted Third Party]:

Address	:	[*]
Facsimile	:	[*]
Attention	:	[*]
E-mail	:	[*]

25. Governing Law

The validity and interpretation of this MOU and the legal relationship of the Parties to it shall be governed by [the jurisdiction] laws. Any dispute, controversy or claim arising out of or relating to this MOU, or breach, termination or invalidity hereof, shall be settled by arbitration in accordance with the UNCITRAL Arbitration Rules as at present in force and as may be amended by the rest of this Clause. The appointing authority shall be the [*official Arbitration Centre*]. The place of

arbitration shall be [*the address of the official Arbitration Centre*]. There shall be only one arbitrator. Any such arbitration shall be administered by the [*official Arbitration Centre*] in accordance with the [*official Arbitration Centre*] Procedures for Arbitration in force at the date of this MOU including such additions to the UNCITRAL Arbitration Rules as are therein contained. The language to be used in the arbitral proceedings shall be English.

Signed for and on behalf of	Signed for and on behalf of
[Company]	[*The Trusted Third Party*]
_____	_____
Name:	Name:
Title:	Title:

Appendix A

Minimum Requirements on Data

Each *Data Controller* shall ensure that the data fields provided include sufficient overlap with the data fields (i.e., common keys) from another source. This not only requires agreement amongst potential *Data Controllers* and *Data Processors* but also common process to protect the Data during its transfer, processing and use in compliance with the PDPO (if applicable).

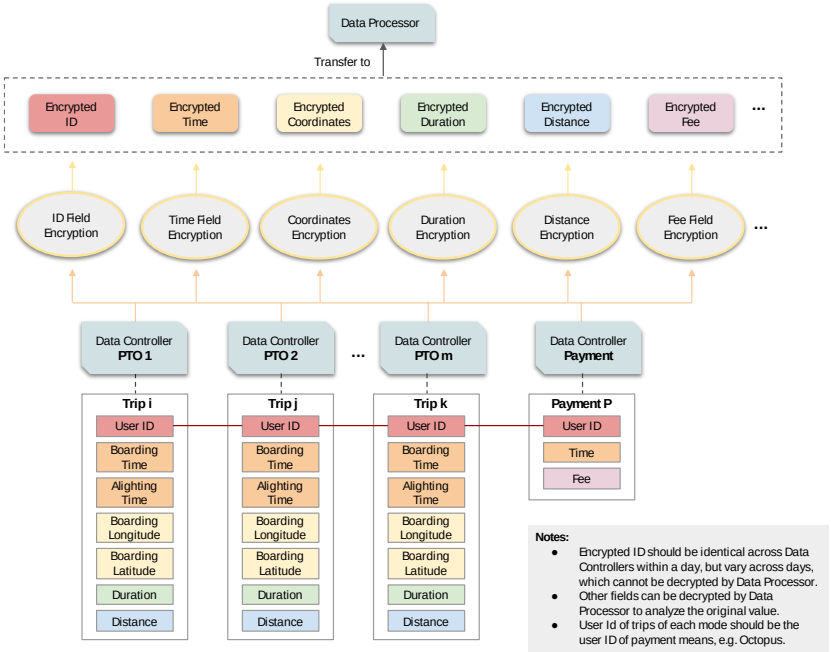
To ensure sufficient overlap, operators need to provide hashed Octopus Card ID (User ID) as a common data field as shown in the figure below.

Each *Data Controller* that participates in the Collaboration shall use a common approach to hashing of its Data that includes a minimum number of data points. The hashed Data provided by each *Data Controller* will be transferred to the *Data Processor*, and recombined with the data sets provided by other *Data Controllers* to generate Aggregated Data. The process will ensure that 'personally identifiable data' cannot be extracted and that individual identity will not be revealed.

The Parties intend to agree on the approach to the processing of *Aggregated Data*.

The Data provided by the *Data Controllers* as well as any *Aggregated Data* will be used by the *Data Processor* to conduct transport-related research, modelling and spatial analysis for the *Collaboration* and not for any other purpose.

The *TDASP* will be provided with access to the results of the research and analysis on the *Aggregated Data*, but not the *Data* provided by any *Data Controllers*.



Appendix B

Statement of Data Management Protocols and Principles relating to the Data Processor and a TDASP

1. An *Algorithm* provided by a *TDASP* shall not be used for any purpose other than the *Purpose*. An *Algorithm* of a particular *TDASP* shall not be distributed or transferred in part or in whole and in whatever forms and media to a *Data Controller* or a third party. The *Data Processor* shall keep such *Algorithm* confidential.
2. An *Algorithm* stored in the computer systems and storage media of the *Data Processor* shall be destroyed when the Purpose has been fulfilled and no later than on completion of the Collaboration.
3. The *Sample Data* and *Training Data* stored in the computer systems and storage media of the *TDASP* shall be destroyed when the purpose for which the *Sample Data* or *Training Data* has been fulfilled and no later than on completion of the *Collaboration*.
4. The expected date of completion is no later than 6 months from receipt of *Sample Data* or *Training Data* from the *Data Processor*. The *Data Processor* shall confirm that it is no longer in possession of an *Algorithm* or any part thereof in any media or in any form.
5. The Parties acknowledge that an *Algorithm* has not been prepared specifically to meet the *Data Processor*'s individual requirements and purpose and that the *Data Processor* will not have any recourse against a *TDASP* for any damage or loss it may suffer in any use or attempted use of the Algorithm.
6. A *TDASP* gives no warranty that an *Algorithm* is error-free and the *TDASP* shall in no way be held for any loss or damage which may be suffered by the *Data Processor* or any other person from the use of an *Algorithm*.
7. Each *TDASP* remains the owner of each *Algorithm* it provides, at all times. [*Trusted Third Party*] shall not copy or otherwise infringe any rights (including without limitation intellectual property rights) that the *TDASP* may have in such *Algorithm*, whether in whole or in part.
8. The *Data Processor* shall be transparent about the collection, use and disclosure of an *Algorithm*.
9. The *Data Processor* is committed to respecting the privacy rights of all applicable individuals granted by privacy laws and to keeping abreast of all applicable data protection/privacy laws of the jurisdiction(s) in which it operates.
10. The *Data Processor* shall keep each *Algorithm* secure.
11. The *Data Processor* is aware of the obligations set out in this MOU and shall follow necessary data protection procedures when handling an *Algorithm.*
12. Relevant security procedures are in place and guidance is issued to staff of the *Data Processor* explaining its data protection and security obligations.

Appendix C

Access, Security and Disclosure of an Algorithm relating to the Data Processor and a TDASP

1. A *Data Controller* nor 3rd party will not be given or able to access or review an *Algorithm* provided by a *TDASP*.

2. Access to an *Algorithm* will be strictly limited by the *Data Processor* and the *Data Processor* will implement appropriate physical, technical and organisational measures ("**Protection Measures**") to ensure security.

3. At the end of the research period the *Algorithm* will be destroyed (with evidence provided for its destruction).

4. The *Data Processor* will not transfer the *Algorithm* outside the [local jurisdiction].