CaribData Glossary

Data Sharing for Data Producers in Small Islands

Selvi Jeyaseelan, Ian R Hambleton

2026-01-26

Table of contents

1 Core ideas about data sharing

This theme introduces the foundational concepts that shape how data sharing is understood and practiced. Together, these ideas define what data sharing is (and is not), how value is created from data over time, and how responsibilities are organised.

1.1 Data sharing

1.1.1 Explanation

Data sharing is the practice of making data available for use beyond the purpose for which they were originally collected. This may involve sharing data within an organisation, across institutions, or—under defined conditions—more widely. At its core, data sharing enables reuse while maintaining appropriate oversight.

1.1.2 Why it matters

Responsible data sharing allows data to generate value over time. It supports better decision-making, reduces duplication, improves transparency, and enables collaboration. Sharing ensures that the cost and effort of data collection continue to deliver public benefit.

1.1.3 Common misunderstandings

Data sharing is often assumed to mean making all data open or public. In practice, most data sharing occurs under controlled conditions. Sharing also does not require giving up ownership or control; many models explicitly preserve both.

1.1.4 SIDS note

In small island developing states, data sharing is most effective when it is selective and purposeful. Small populations and limited capacity mean that sharing some data well—clearly documented and appropriately governed—is often more valuable than broad openness.

1.2 Data reuse

1.2.1 Explanation

Data reuse refers to the use of existing data for new purposes beyond their original intent. This may include new analyses, integration with other datasets, or application to different policy or research questions.

1.2.2 Why it matters

Reuse increases the return on investment in data collection. It reduces the need for repeated data gathering, supports innovation, and allows insights to accumulate over time.

1.2.3 Common misunderstandings

Data reuse is sometimes treated as automatic once data are shared. In reality, reuse depends on data quality, documentation, and clarity about conditions of use.

1.2.4 SIDS note

For SIDS, data reuse is particularly valuable because data collection is often costly and infrequent. Well-documented datasets that can be reused safely can support multiple national and regional priorities.

1.3 Data access vs data release

1.3.1 Explanation

Data access refers to the ability to use data under defined conditions, while data release refers to making data publicly available. Access may be time-limited, purpose-specific, or restricted to approved users; release is typically open-ended.

1.3.2 Why it matters

Distinguishing access from release helps clarify options for sharing. Many benefits of data sharing can be achieved through managed access without full public release.

1.3.3 Common misunderstandings

These terms are often used interchangeably, leading to unnecessary resistance. Treating sharing as synonymous with public release can obscure more appropriate models.

1.3.4 SIDS note

In small island developing states, managed access is often preferable to release, particularly for detailed datasets. This distinction allows data to be used while managing disclosure risk.

1.4 Open data

1.4.1 Explanation

Open data are data that are freely available for anyone to access, use, reuse, and redistribute, usually under an open licence. They are typically published online in accessible formats.

1.4.2 Why it matters

Open data can increase transparency, support accountability, and enable innovation by a wide range of users, including researchers, journalists, and communities.

1.4.3 Common misunderstandings

Open data are sometimes treated as the default or ideal form of sharing. In practice, not all data are suitable for open release, and many benefits can be achieved through aggregated or summary data.

1.4.4 SIDS note

For SIDS, open data should be approached selectively. High-value, low-risk datasets and summary products often deliver the greatest benefit while minimising re-identification risk.

1.5 Controlled access data

1.5.1 Explanation

Controlled access data are shared under defined conditions rather than being openly available. Access is limited to approved users, for specific purposes, and may involve agreements or secure environments.

1.5.2 Why it matters

Controlled access provides a practical balance between reuse and protection, allowing sensitive or detailed data to be shared responsibly.

1.5.3 Common misunderstandings

Controlled access is sometimes seen as a barrier to sharing. In reality, it is one of the most common and effective data-sharing models.

1.5.4 SIDS note

In SIDS, controlled access is often the most appropriate default for sharing detailed data, given small populations and heightened disclosure risk.

1.6 Data stewardship

1.6.1 Explanation

Data stewardship refers to the responsible management of data throughout their lifecycle. Stewards ensure data are handled ethically, securely, and in line with legal and organisational obligations.

1.6.2 Why it matters

Clear stewardship supports data quality, continuity, and trust. It underpins effective sharing and reuse by clarifying responsibility and accountability.

1.6.3 Common misunderstandings

Stewardship is often confused with ownership or technical data management. In practice, it includes governance, documentation, access decisions, and communication.

1.6.4 SIDS note

In SIDS, stewardship roles are often informal or combined with other duties. Clear, simple stewardship arrangements can greatly improve continuity where institutional memory is limited.

1.7 Data governance

1.7.1 Explanation

Data governance refers to the policies, processes, and roles that guide how data are collected, managed, shared, and used. It sets the rules for decision-making about data.

1.7.2 Why it matters

Good governance provides clarity and consistency. It helps balance opportunity and risk, supports trust, and enables sustainable data sharing.

1.7.3 Common misunderstandings

Data governance is sometimes seen as bureaucratic or restrictive. When done well, it enables sharing by providing clear boundaries and expectations.

1.7.4 SIDS note

For SIDS, governance frameworks need to be proportionate and practical. Overly complex models can overwhelm limited capacity and discourage participation.

1.8 Data lifecycle

1.8.1 Explanation

The data lifecycle describes the stages through which data pass, from planning and collection, through use and sharing, to long-term preservation or disposal.

1.8.2 Why it matters

Thinking in lifecycle terms helps ensure that sharing, reuse, and preservation are considered early, rather than as afterthoughts.

1.8.3 Common misunderstandings

The lifecycle is sometimes treated as linear. In practice, data may move back and forth between stages as they are updated or reused.

1.8.4 SIDS note

In small island states, lifecycle thinking is especially important to prevent data loss. Early planning for preservation and reuse can protect valuable national assets.

2 People, ethics, and public interest

This theme focuses on the human dimensions of data sharing. These concepts address rights, responsibilities, trust, and fairness, and help ensure that data sharing serves people and communities rather than undermining them.

2.1 Privacy

2.1.1 Explanation

Privacy refers to the right of individuals to control information about themselves and to be protected from unwanted intrusion or exposure. In data contexts, privacy is concerned with how personal information is collected, used, shared, and safeguarded.

2.1.2 Why it matters

Respecting privacy protects individuals from harm and builds trust in data systems. Without confidence that privacy will be respected, people may refuse to participate in data collection or provide incomplete information.

2.1.3 Common misunderstandings

Privacy is often treated as absolute or as incompatible with data sharing. In practice, privacy can be protected through proportionate safeguards while still

allowing data to be used for public benefit.

2.1.4 SIDS note

In small island developing states, privacy concerns are heightened because people are more easily identifiable. Careful judgement is required, even for data that appear non-sensitive in larger populations.

2.2 Confidentiality

2.2.1 Explanation

Confidentiality refers to obligations placed on those who handle data to protect information from unauthorised disclosure. It is about how data are treated once they have been collected.

2.2.2 Why it matters

Confidentiality underpins trust between data providers and data holders. Breaches can cause direct harm and damage confidence in institutions.

2.2.3 Common misunderstandings

Confidentiality is sometimes confused with secrecy. Protecting confidentiality does not mean that data can never be shared; it means sharing must occur under appropriate conditions.

2.2.4 SIDS note

In SIDS, breaches of confidentiality can have outsized social consequences. Clear rules and consistent practice are especially important where professional and social networks overlap.

2.3 Consent

2.3.1 Explanation

Consent refers to permission given by individuals for their data to be collected and used. Consent may be specific to a single purpose, or broader, allowing future uses under defined conditions.

2.3.2 Why it matters

Consent respects individual autonomy and clarifies expectations about data use. It is a cornerstone of ethical data collection and sharing.

2.3.3 Common misunderstandings

Consent is sometimes treated as a one-time, unlimited permission. In reality, consent has limits and must be interpreted in light of context and purpose.

2.3.4 SIDS note

In small states, consent processes must be clear and realistic. Overly complex consent models may be impractical, while vague consent can undermine trust.

2.4 Public interest

2.4.1 Explanation

Public interest refers to the broader benefit to society that may justify the use or sharing of data. It balances individual interests with collective wellbeing.

2.4.2 Why it matters

Many data-sharing decisions rely on judgements about public interest, particularly when consent is broad or indirect. Clear reasoning helps maintain legitimacy.

2.4.3 Common misunderstandings

Public interest is sometimes invoked without explanation. In practice, it should be defined, justified, and open to scrutiny.

2.4.4 SIDS note

In SIDS, public interest arguments are closely tied to trust in institutions. Transparency about how public benefit is defined and assessed is critical.

2.5 Social licence

2.5.1 Explanation

Social licence refers to the informal approval granted by communities for data practices, beyond what is legally required. It reflects trust and shared expectations.

2.5.2 Why it matters

Even lawful data sharing can fail if it lacks social licence. Public support influences participation, acceptance, and sustainability.

2.5.3 Common misunderstandings

Social licence is sometimes mistaken for public relations. It is earned through consistent, respectful practice, not messaging alone.

2.5.4 SIDS note

In small island settings, social licence is especially fragile and powerful. Loss of trust can spread quickly, but so can confidence built through good practice.

2.6 Benefit sharing

2.6.1 Explanation

Benefit sharing involves ensuring that the advantages gained from data use are returned to the individuals or communities from which the data originate.

2.6.2 Why it matters

Benefit sharing promotes fairness and counters extractive data practices. It helps ensure that data use contributes to local priorities.

2.6.3 Common misunderstandings

Benefits are sometimes assumed to be indirect or long-term. In practice, benefits should be identifiable and, where possible, visible.

2.6.4 SIDS note

For SIDS, benefit sharing is central to regional and international data partnerships. Visible local benefit strengthens trust and participation.

2.7 Community engagement

2.7.1 Explanation

Community engagement refers to involving communities in decisions about data collection, use, and sharing. It goes beyond consultation to meaningful participation.

2.7.2 Why it matters

Engagement improves relevance, legitimacy, and understanding. It helps align data use with community values and needs.

2.7.3 Common misunderstandings

Engagement is sometimes treated as a one-off activity. Effective engagement is ongoing and responsive.

2.7.4 SIDS note

In small states, engagement is often more direct and personal. This can be a strength, but it also raises expectations for responsiveness and accountability.

2.8 Data ownership vs custodianship

2.8.1 Explanation

Data ownership refers to legal rights over data, while custodianship refers to responsibility for managing data on behalf of others. Many public datasets are better understood as being stewarded rather than owned.

2.8.2 Why it matters

Clarifying custodianship helps resolve confusion about who can make decisions about data sharing and use.

2.8.3 Common misunderstandings

Data are often assumed to be owned outright by the collecting organisation. In reality, obligations to data subjects and the public may limit this.

2.8.4 SIDS note

In SIDS, clear custodianship arrangements help manage shared responsibilities across small institutions and reduce conflict over decision-making.

2.9 Group privacy

2.9.1 Explanation

Group privacy recognises that data can cause harm not only to individuals, but also to communities or groups, even when individuals are not identifiable.

2.9.2 Why it matters

Some data uses can stigmatise or disadvantage groups. Considering group-level impacts is essential for ethical data sharing.

2.9.3 Common misunderstandings

Privacy is often framed only at the individual level. Group harms may still occur even when individual privacy is protected.

2.9.4 SIDS note

In small island societies, where group identities are strong, group-level impacts are especially important to consider when sharing or publishing data.

3 Risk, harm, and protection

This theme covers concepts that help identify, assess, and manage the potential harms that can arise from data use and sharing. These ideas are central to making proportionate, defensible decisions—especially in small populations.

3.1 Disclosure risk

3.1.1 Explanation

Disclosure risk refers to the possibility that information in a dataset could be revealed to unauthorised parties or used in ways that expose sensitive details about individuals or groups.

3.1.2 Why it matters

Understanding disclosure risk helps determine whether data can be shared safely, and under what conditions. It guides decisions about aggregation, access controls, and safeguards.

3.1.3 Common misunderstandings

Disclosure risk is often treated as either present or absent. In practice, it exists on a spectrum and can be reduced, but rarely eliminated entirely.

3.1.4 SIDS note

In small island developing states, disclosure risk is typically higher because fewer data points are needed to identify people or communities.

3.2 Re-identification risk

3.2.1 Explanation

Re-identification risk is the chance that individuals can be identified in a dataset after direct identifiers have been removed, often by combining variables or linking with other data sources.

3.2.2 Why it matters

Re-identification can undermine privacy and expose individuals to harm. Managing this risk is central to responsible data sharing.

3.2.3 Common misunderstandings

Removing names is often assumed to make data anonymous. In reality, indirect identifiers can still enable identification.

3.2.4 SIDS note

Small populations and tight social networks mean re-identification risk is especially pronounced in SIDS, even in de-identified datasets.

3.3 Identifiability

3.3.1 Explanation

Identifiability describes the extent to which data can be linked to a specific individual, directly or indirectly. Data may be fully identifiable, partially identifiable, or effectively non-identifiable depending on context.

3.3.2 Why it matters

Assessing identifiability helps determine appropriate sharing models and safeguards.

3.3.3 Common misunderstandings

Identifiability is often treated as a fixed property of data. In practice, it depends on context, available external data, and user capability.

3.3.4 SIDS note

In SIDS, contextual knowledge can increase identifiability, even when datasets appear anonymous.

3.4 Small numbers problem

3.4.1 Explanation

The small numbers problem arises when datasets involve very small counts, making individuals or groups easier to identify and statistical estimates less stable.

3.4.2 Why it matters

Small numbers increase both disclosure risk and the chance of misinterpretation.

3.4.3 Common misunderstandings

Small numbers are sometimes treated as a technical nuisance rather than a substantive ethical and analytical issue.

3.4.4 SIDS note

This problem is common in SIDS due to small populations and geographic concentration, and it shapes many data sharing decisions.

3.5 Anonymisation

3.5.1 Explanation

Anonymisation is the process of irreversibly removing identifying information so that individuals cannot reasonably be identified.

3.5.2 Why it matters

True anonymisation allows data to be shared more freely, as privacy risks are greatly reduced.

3.5.3 Common misunderstandings

Anonymisation is often assumed to be easy or guaranteed. In practice, achieving true anonymity is difficult, especially for rich datasets.

3.5.4 SIDS note

In small states, anonymisation is harder to achieve because contextual information can reintroduce identifiability.

3.6 Pseudonymisation

3.6.1 Explanation

Pseudonymisation replaces identifying information with artificial identifiers, allowing data to be linked over time while reducing direct identification.

3.6.2 Why it matters

It supports analysis and linkage while providing some privacy protection.

3.6.3 Common misunderstandings

Pseudonymised data are sometimes treated as anonymous. They are not; re-identification remains possible.

3.6.4 SIDS note

In SIDS, pseudonymised data usually require controlled access and strong governance.

3.7 De-identification

3.7.1 Explanation

De-identification refers to removing or modifying identifiers to reduce identifiability, without guaranteeing anonymity.

3.7.2 Why it matters

It is a common and practical risk-reduction approach used before sharing data.

3.7.3 Common misunderstandings

De-identification is sometimes conflated with anonymisation. The distinction matters for governance decisions.

3.7.4 SIDS note

De-identified data in SIDS often still carry meaningful re-identification risk and should be treated cautiously.

3.8 Statistical disclosure control

3.8.1 Explanation

Statistical disclosure control (SDC) refers to techniques used to reduce the risk of identifying individuals in published data, such as suppression or aggregation.

3.8.2 Why it matters

SDC enables useful data products to be shared while managing risk.

3.8.3 Common misunderstandings

SDC is sometimes seen as purely technical. Judgement about context and harm is equally important.

3.8.4 SIDS note

SDC techniques are especially important in SIDS, where small counts are common.

3.9 Data minimisation

3.9.1 Explanation

Data minimisation means collecting, using, and sharing only the data necessary for a given purpose.

3.9.2 Why it matters

Limiting data reduces risk and simplifies governance.

3.9.3 Common misunderstandings

Minimisation is sometimes interpreted as reducing data quality. In fact, it focuses on relevance rather than volume.

3.9.4 SIDS note

For SIDS, minimisation helps balance limited capacity with responsible sharing.

3.10 Privacy by design

3.10.1 Explanation

Privacy by design involves embedding privacy considerations into data systems and processes from the outset, rather than adding them later.

3.10.2 Why it matters

Early design choices can greatly reduce risk and cost over time.

3.10.3 Common misunderstandings

Privacy by design is sometimes treated as a technical feature rather than an organisational approach.

3.10.4 SIDS note

In small states, early attention to privacy can prevent costly remediation later, when resources are limited.

4 Legal and policy foundations

This theme covers the legal and policy concepts that shape how data can be collected, used, and shared. These ideas provide boundaries and protections, but they also enable data sharing by clarifying responsibilities and permissions.

4.1 Data protection law

4.1.1 Explanation

Data protection law sets out rules for how personal data must be handled. It defines what counts as personal data, establishes rights for individuals, and places obligations on organisations that collect or use data.

4.1.2 Why it matters

Data protection law provides a legal basis for trust. It protects individuals from harm and gives organisations a framework for making defensible decisions about data sharing.

4.1.3 Common misunderstandings

Data protection law is often seen as preventing data sharing. In reality, most laws are designed to allow data use and sharing under defined conditions.

4.1.4 SIDS note

In many SIDS, data protection laws are relatively new or still evolving. Clear guidance and proportionate interpretation are essential to avoid unnecessary caution that limits legitimate data use.

4.2 Lawful basis for data use

4.2.1 Explanation

A lawful basis is a legally recognised reason for collecting or using personal data, such as consent, public interest, or legal obligation.

4.2.2 Why it matters

Identifying a lawful basis helps ensure that data use is legitimate and defensible, particularly when consent is broad or indirect.

4.2.3 Common misunderstandings

Lawful basis is sometimes assumed to mean consent only. Most data protection frameworks recognise multiple lawful bases.

4.2.4 SIDS note

In SIDS, where consent processes may be informal or resource-limited, clarity about alternative lawful bases is especially important.

4.3 Freedom of Information (FOI)

4.3.1 Explanation

Freedom of Information laws give the public a right to access information held by public bodies, subject to defined exemptions.

4.3.2 Why it matters

FOI promotes transparency and accountability, and intersects directly with decisions about data release.

4.3.3 Common misunderstandings

FOI is sometimes interpreted as requiring the release of all data. In practice, exemptions protect privacy, confidentiality, and sensitive information.

4.3.4 SIDS note

In small states, FOI requests may involve highly contextual data. Careful balancing of transparency and privacy is essential.

4.4 Research ethics approval

4.4.1 Explanation

Research ethics approval involves independent review of proposed research to ensure that data collection and use are ethical and minimise harm.

4.4.2 Why it matters

Ethics review protects participants and supports public trust in research and data sharing.

4.4.3 Common misunderstandings

Ethics approval is sometimes seen as a one-time hurdle. Ongoing ethical responsibility continues throughout the data lifecycle.

4.4.4 SIDS note

In SIDS, ethics committees may be small and overburdened. Clear protocols and realistic expectations help ensure timely and consistent review.

4.5 Data sharing agreement (DSA)

4.5.1 Explanation

A data sharing agreement is a formal document that sets out the terms under which data are shared between parties, including purpose, access conditions, and responsibilities.

4.5.2 Why it matters

DSAs provide clarity and accountability, reducing uncertainty and risk for all parties.

4.5.3 Common misunderstandings

DSAs are sometimes seen as overly legalistic. In practice, simple agreements can be highly effective.

4.5.4 SIDS note

For SIDS, lightweight and standardised DSAs can support sharing without overwhelming limited legal capacity.

4.6 Data use agreement (DUA)

4.6.1 Explanation

A data use agreement focuses on how data may be used once accessed, including restrictions, reporting requirements, and obligations to protect confidentiality.

4.6.2 Why it matters

DUAs help ensure that data are used only for agreed purposes and support enforcement if misuse occurs.

4.6.3 Common misunderstandings

DUAs are often confused with DSAs. While related, DUAs typically focus on the user's responsibilities rather than the act of sharing.

4.6.4 SIDS note

In SIDS, DUAs help manage reputational risk by clarifying expectations for external users.

4.7 Licensing

4.7.1 Explanation

Data licences specify how data may be used, reused, and shared. They are especially important for open or publicly released data.

4.7.2 Why it matters

Clear licensing removes ambiguity and enables lawful reuse.

4.7.3 Common misunderstandings

Licences are sometimes omitted or treated as optional. Without a licence, reuse may be legally unclear.

4.7.4 SIDS note

For SIDS, standard licences can simplify sharing and reduce the need for bespoke legal advice.

4.8 Intellectual property (IP)

4.8.1 Explanation

Intellectual property law governs rights over creative and informational outputs, including databases in some jurisdictions.

4.8.2 Why it matters

Understanding IP helps clarify who can authorise sharing and under what conditions.

4.8.3 Common misunderstandings

Data are often assumed to be automatically owned or unowned. IP rights can be complex and context-specific.

4.8.4 SIDS note

In SIDS, IP considerations often intersect with public sector data and donor-funded projects, requiring clear upfront agreements.

5 Access models and governance tools

This theme describes the practical mechanisms used to decide who can access data, under what conditions, and with what safeguards. These models translate principles about risk, trust, and legality into workable arrangements.

5.1 Open access

5.1.1 Explanation

Open access refers to data that are available to anyone, without the need for registration or approval. Users are generally free to access and use the data under an open licence.

5.1.2 Why it matters

Open access maximises reach and reuse. It supports transparency, public engagement, and innovation by lowering barriers to data use.

5.1.3 Common misunderstandings

Open access is sometimes assumed to be appropriate for all data. In practice, many datasets require restrictions to protect privacy or manage sensitivity.

5.1.4 SIDS note

In small island developing states, open access is best suited to high-level, aggregated, or low-risk data products rather than detailed microdata.

5.2 Registered access

5.2.1 Explanation

Registered access requires users to create an account or identify themselves before accessing data. Access is still relatively open, but usage can be monitored.

5.2.2 Why it matters

Registration introduces accountability and allows data holders to understand who is using data and for what general purpose.

5.2.3 Common misunderstandings

Registered access is sometimes mistaken for controlled access. It usually involves minimal review and few restrictions.

5.2.4 SIDS note

For SIDS, registered access can offer a low-burden way to increase oversight without discouraging legitimate use.

5.3 Restricted / controlled access

5.3.1 Explanation

Restricted or controlled access limits data availability to approved users, specific purposes, or defined time periods. Approval processes and agreements are typically required.

5.3.2 Why it matters

Controlled access allows sensitive or detailed data to be reused while managing privacy and reputational risk.

5.3.3 Common misunderstandings

Restricted access is often seen as a failure of openness. In reality, it is a widely used and effective sharing model.

5.3.4 SIDS note

In SIDS, controlled access is frequently the most appropriate model for sharing individual-level or detailed datasets.

5.4 Trusted Research Environment (TRE)

5.4.1 Explanation

A Trusted Research Environment is a secure setting where approved users can analyse data without removing them from the environment. Outputs are checked before release.

5.4.2 Why it matters

TREs reduce disclosure risk while allowing detailed analysis, particularly for sensitive data.

5.4.3 Common misunderstandings

TREs are sometimes assumed to require complex infrastructure. In practice, the concept can be implemented at varying levels of sophistication.

5.4.4 SIDS note

For SIDS, shared or regional TREs may be more feasible than national systems, allowing economies of scale while preserving control.

5.5 Data access committee

5.5.1 Explanation

A data access committee is a group responsible for reviewing and approving requests to use data. It considers purpose, risk, and public benefit.

5.5.2 Why it matters

Committees provide consistency, transparency, and accountability in access decisions.

5.5.3 Common misunderstandings

Access committees are sometimes seen as bureaucratic barriers. Clear criteria and proportionate processes can make them enabling rather than obstructive.

5.5.4 SIDS note

In small states, access committees are often small and multi-disciplinary. Simple procedures help ensure decisions are timely and trusted.

5.6 Tiered access

5.6.1 Explanation

Tiered access provides different levels of access to the same dataset, depending on user role, purpose, or risk level.

5.6.2 Why it matters

Tiered models allow broad access to low-risk data while protecting more sensitive information.

5.6.3 Common misunderstandings

Tiered access is sometimes viewed as complex. In practice, even two-tier systems can be highly effective.

5.6.4 SIDS note

For SIDS, tiered access supports flexibility and proportionality without requiring multiple separate datasets.

5.7 Embargo period

5.7.1 Explanation

An embargo period is a defined time during which data are not shared, often to allow primary analysis or reporting by the data producers.

5.7.2 Why it matters

Embargoes balance incentives for data collection with longer-term sharing goals.

5.7.3 Common misunderstandings

Embargoes are sometimes treated as indefinite restrictions. They should be time-limited and clearly defined.

5.7.4 SIDS note

In SIDS, embargoes can protect small teams' ability to publish or report findings before wider reuse, supporting sustainability.

6 Data quality, documentation, and reuse

This theme focuses on whether data are understandable, reliable, and usable beyond their original purpose. High-quality data sharing depends as much on documentation and clarity as it does on access.

6.1 Data quality

6.1.1 Explanation

Data quality refers to how well data represent what they are intended to measure. Common dimensions include accuracy, completeness, consistency, timeliness, and validity.

6.1.2 Why it matters

Poor-quality data can mislead decision-making and undermine trust. Sharing data without understanding their quality can do more harm than good.

6.1.3 Common misunderstandings

Data quality is sometimes treated as a single score or threshold. In practice, quality is relative to purpose: data may be suitable for some uses but not others.

6.1.4 SIDS note

In SIDS, data collection is often resource-intensive. Clear understanding of data quality helps ensure that limited data are used appropriately and not over-interpreted.

6.2 Fitness for purpose

6.2.1 Explanation

Fitness for purpose describes whether data are suitable for a specific use, given how they were collected, processed, and documented.

6.2.2 Why it matters

Assessing fitness for purpose prevents misuse. It shifts attention from whether data are "good" to whether they are appropriate for a particular question.

6.2.3 Common misunderstandings

Fitness is sometimes assumed to be inherent to the dataset. In reality, it depends on context and intended use.

6.2.4 SIDS note

For SIDS, explicit discussion of fitness for purpose is important when data are reused regionally or internationally, where context may be lost.

6.3 Data provenance

6.3.1 Explanation

Data provenance describes the origin of data and the processes that led to their current form, including collection methods, transformations, and analysis.

6.3.2 Why it matters

Provenance supports transparency, reproducibility, and trust. It helps users understand how data were generated and modified.

6.3.3 Common misunderstandings

Provenance is often treated as a technical detail. It is equally important for interpretation and accountability.

6.3.4 SIDS note

In small states, where institutional memory may be informal, documenting provenance helps preserve knowledge when staff change.

6.4 Metadata

6.4.1 Explanation

Metadata are information that describe a dataset, including its content, structure, methods, coverage, and limitations.

6.4.2 Why it matters

Metadata make data findable, interpretable, and reusable. Without metadata, sharing has limited value.

6.4.3 Common misunderstandings

Metadata are sometimes added at the end of a project. Effective metadata are developed alongside data collection and use.

6.4.4 SIDS note

For SIDS, metadata are a low-cost way to improve data sharing and coordination, even when access to the data themselves is restricted.

6.5 Data dictionary / codebook

6.5.1 Explanation

A data dictionary or codebook defines variables in a dataset, including meanings, units, categories, and coding schemes.

6.5.2 Why it matters

Clear definitions prevent misinterpretation and support reuse by people unfamiliar with the original data.

6.5.3 Common misunderstandings

Codebooks are sometimes assumed to be needed only for surveys. They are valuable for administrative and operational data as well.

6.5.4 SIDS note

In SIDS, where datasets are often reused across multiple roles and institutions, codebooks help maintain consistency over time.

6.6 FAIR principles

6.6.1 Explanation

The FAIR principles describe data that are Findable, Accessible, Interoperable, and Reusable. They provide guidance rather than strict rules.

6.6.2 Why it matters

FAIR principles help ensure that data can be discovered and reused responsibly, increasing their long-term value.

6.6.3 Common misunderstandings

FAIR is sometimes equated with open data. In fact, data can be FAIR while access remains controlled.

6.6.4 SIDS note

For SIDS, adapting FAIR principles pragmatically—focusing first on findability and clarity—can deliver benefits without excessive burden.

6.7 Interoperability

6.7.1 Explanation

Interoperability refers to the ability of datasets or systems to work together through shared formats, standards, or definitions.

6.7.2 Why it matters

Interoperability enables data linkage, comparison, and regional analysis.

6.7.3 Common misunderstandings

Interoperability is often seen as purely technical. Shared meaning and agreed definitions are equally important.

6.7.4 SIDS note

In SIDS, interoperability supports regional collaboration and reduces duplication across small national systems.

6.8 Dataset versioning

6.8.1 Explanation

Dataset versioning involves tracking changes to data over time, so users know which version they are using and how it differs from others.

6.8.2 Why it matters

Versioning supports reproducibility and prevents confusion when datasets are updated or corrected.

6.8.3 Common misunderstandings

Versioning is sometimes overlooked for data that are "final." In practice, many datasets evolve.

6.8.4 SIDS note

In small teams, clear versioning helps maintain continuity when responsibilities shift between staff.

6.9 Data citation

6.9.1 Explanation

Data citation provides a standard way to reference datasets, similar to citing publications, acknowledging data creators and sources.

6.9.2 Why it matters

Citation supports attribution, accountability, and recognition for data work.

6.9.3 Common misunderstandings

Data are often used without formal citation. This reduces visibility and discourages sharing.

6.9.4 SIDS note

For SIDS, data citation helps ensure that local data producers receive recognition when data are reused externally.

7 Infrastructure and technical basics

This theme introduces core technical concepts that support data sharing in practice. These ideas are included at a high level, focusing on what decision-makers and practitioners need to understand rather than on technical implementation details.

7.1 Open vs proprietary formats

7.1.1 Explanation

Open formats are file types that can be used and read by a wide range of software without restriction, while proprietary formats are controlled by specific vendors or tools.

7.1.2 Why it matters

Using open formats makes data easier to access, share, and preserve over time. They reduce dependence on specific software and lower barriers for reuse.

7.1.3 Common misunderstandings

Open formats are sometimes assumed to be lower quality or less secure. In practice, they are often more durable and widely supported.

7.1.4 SIDS note

For SIDS, open formats reduce long-term costs and help ensure data remain usable even as software and staff change.

7.2 Data repository

7.2.1 Explanation

A data repository is a system or service designed to store, manage, and provide access to datasets over time. Repositories may be institutional, national, regional, or thematic.

7.2.2 Why it matters

Repositories support preservation, controlled access, and reuse. They provide a stable home for data beyond individual projects.

7.2.3 Common misunderstandings

Repositories are sometimes confused with simple storage. Effective repositories include governance, documentation, and access controls.

7.2.4 SIDS note

For SIDS, shared or regional repositories can offer sustainable solutions where national infrastructure is limited.

7.3 Data catalogue

7.3.1 Explanation

A data catalogue is a structured listing of available datasets, often including descriptions, metadata, and access information, without necessarily hosting the data itself.

7.3.2 Why it matters

Catalogues improve visibility and coordination by helping users discover what data exist and how they might be accessed.

7.3.3 Common misunderstandings

Catalogues are sometimes assumed to require full data sharing. In reality, they can function even when data access is restricted.

7.3.4 SIDS note

In SIDS, data catalogues are a practical first step toward sharing, helping overcome fragmentation with minimal risk.

7.4 Secure storage

7.4.1 Explanation

Secure storage refers to protecting data from unauthorised access, loss, or corruption through physical, technical, and organisational safeguards.

7.4.2 Why it matters

Secure storage protects confidentiality and data integrity and is foundational to trust in data systems.

7.4.3 Common misunderstandings

Security is often seen as purely technical. Human processes and clear responsibilities are equally important.

7.4.4 SIDS note

In small states, simple, well-understood security practices are often more effective than complex systems that cannot be maintained.

7.5 Data archiving and preservation

7.5.1 Explanation

Data archiving and preservation involve keeping data accessible and usable over the long term, beyond the life of a project or system.

7.5.2 Why it matters

Preservation protects data from loss and ensures that valuable information remains available for future use.

7.5.3 Common misunderstandings

Archiving is sometimes treated as passive storage. Effective preservation requires planning and active management.

7.5.4 SIDS note

For SIDS, where data collection is costly and infrequent, long-term preservation is essential to protect national knowledge assets.

8 Culture, incentives, and practice

This theme focuses on the human and organisational factors that shape whether data sharing actually happens. Even when legal, technical, and ethical foundations are in place, culture and incentives often determine success or failure.

8.1 Data literacy

8.1.1 Explanation

Data literacy is the ability to understand, interpret, and use data appropriately. It includes knowing what data can and cannot say, recognising limitations, and asking informed questions of data.

8.1.2 Why it matters

Data sharing is only useful if people can engage with data responsibly. Data literacy supports better decision-making, reduces misinterpretation, and strengthens trust in evidence.

8.1.3 Common misunderstandings

Data literacy is often equated with technical skills or statistics. In practice, it is as much about critical thinking and context as it is about analysis.

8.1.4 SIDS note

In small island developing states, data literacy has multiplier effects. Individuals often work across multiple roles, so strengthening literacy in one place can benefit many systems.

8.2 Incentives for data sharing

8.2.1 Explanation

Incentives for data sharing are the motivations that encourage individuals and organisations to make data available for reuse. These may include recognition, funding requirements, professional norms, or demonstrated impact.

8.2.2 Why it matters

Without incentives, data sharing competes with other priorities and is often deprioritised. Clear incentives help make sharing sustainable rather than exceptional.

8.2.3 Common misunderstandings

Incentives are sometimes assumed to be financial. Recognition, attribution, and reduced duplication can be equally powerful.

8.2.4 SIDS note

For SIDS, visible recognition of data contributions—especially at regional or international level—can be a strong motivator and help retain expertise.

8.3 Barriers to data sharing

8.3.1 Explanation

Barriers to data sharing include fear of misuse, uncertainty about rules, limited capacity, lack of documentation, and competing demands on staff time.

8.3.2 Why it matters

Understanding barriers helps shift the conversation from blame to problem-solving. Many barriers are structural rather than attitudinal.

8.3.3 Common misunderstandings

Resistance to sharing is often framed as unwillingness. In practice, it is frequently driven by risk aversion, workload, or unclear expectations.

8.3.4 SIDS note

In small states, barriers are often amplified by limited staffing and overlapping responsibilities. Addressing even one barrier can unlock wider progress.

8.4 Responsible data use

8.4.1 Explanation

Responsible data use refers to using data in ways that are ethical, lawful, and respectful of context. It includes careful interpretation, transparent methods, and appropriate communication of findings.

8.4.2 Why it matters

Responsible use protects individuals and communities from harm and sustains trust in data systems. Misuse can undermine confidence even when sharing was appropriate.

8.4.3 Common misunderstandings

Responsibility is sometimes assumed to rest only with data holders. In reality, users share responsibility for how data are interpreted and communicated.

8.4.4 SIDS note

In SIDS, responsible use is especially important because misinterpretation can have immediate reputational or policy consequences. Clear expectations for users help protect both data producers and communities.