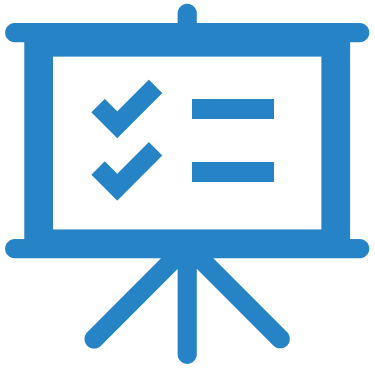

# Introduction to data interoperability across the data value chain

OPEN DATA AND INTEROPERABILITY WORKSHOP
DHAKA, BANGLADESH
28 APRIL – 2 MAY 2019

# Main objectives of this session

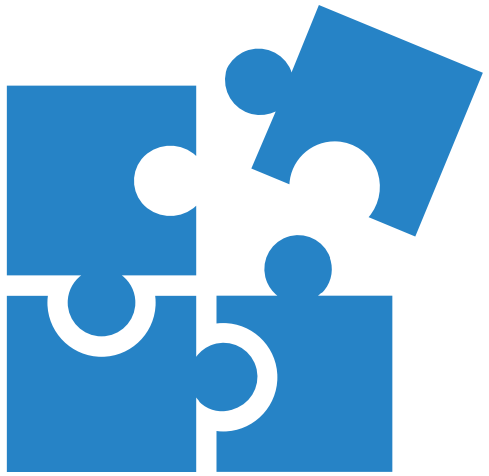


- Have a common understanding of key aspects of **data interoperability**
- Discuss how data interoperability is fundamental to **data quality** and a precondition for **open data**
- Discuss the foundations of data interoperability for **national SDG data** across the complete **data value chain** (from collection to use)

# The SDG data ecosystem is characterized by multiple tensions

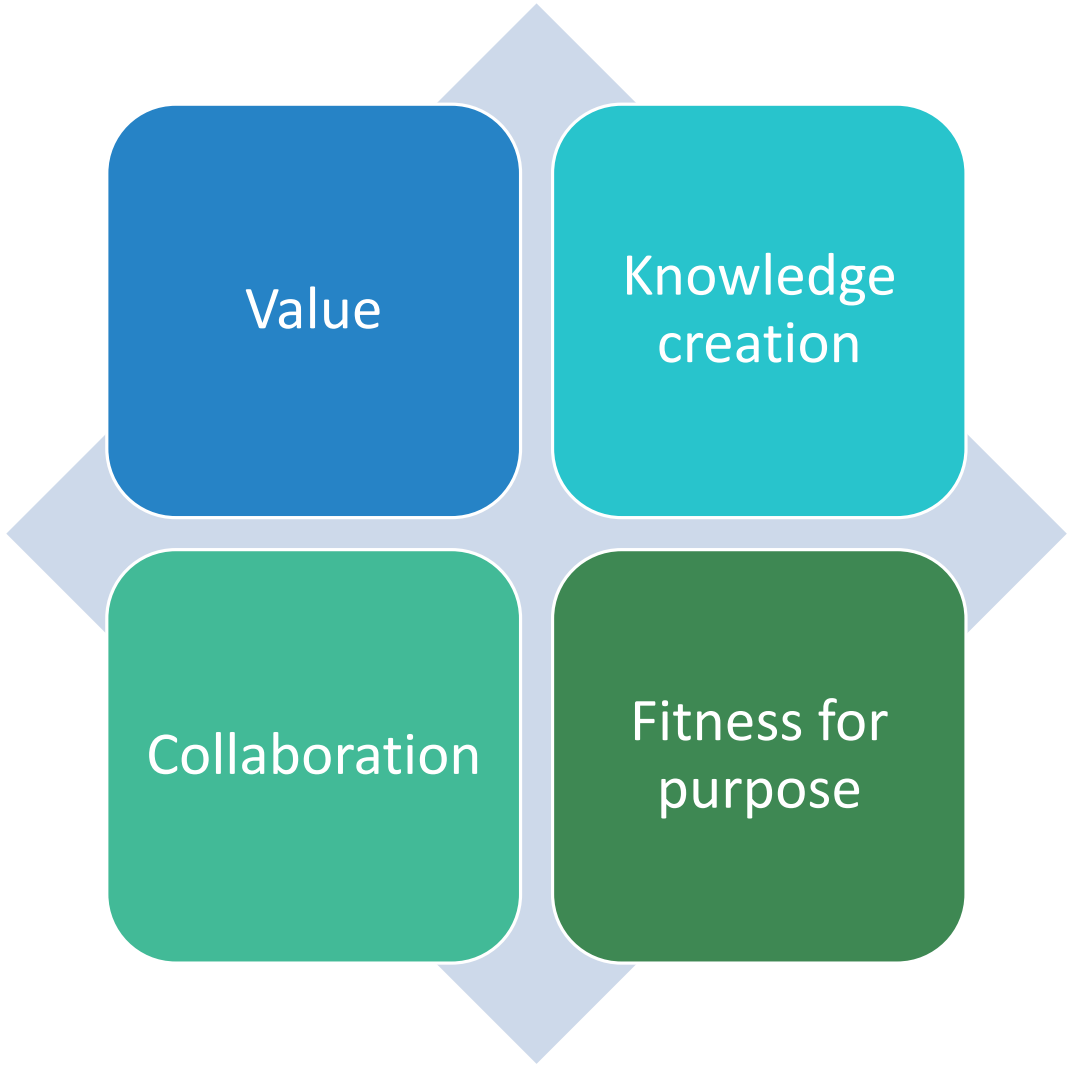
- Global vs local data needs
- Top-down vs bottom-up data producers
- Structured data exchange vs organic data sharing processes
- Sectoral vs. Cross-cutting data applications

# Data interoperability challenge

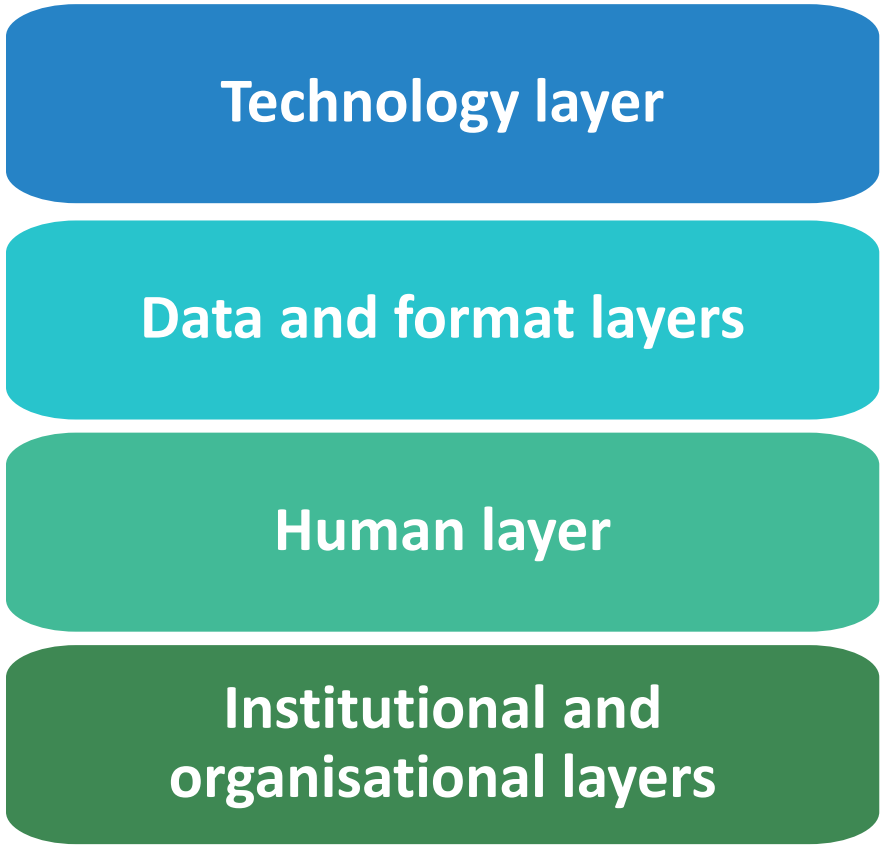


It’s difficult to share, integrate and work with the wealth of data that is available in today’s digital era:

- Divergent needs and capabilities of multiple internal and external constituencies
- Disparate protocols, technologies and standards
- Fragmented data production and dissemination systems




# Interoperability is a characteristic of good quality data




# The interoperability challenge is multi-faceted

# Data interoperability for the SDGs

There are many unrealized opportunities to extract value from data that already exists to meet information needs of the 2030 Agenda

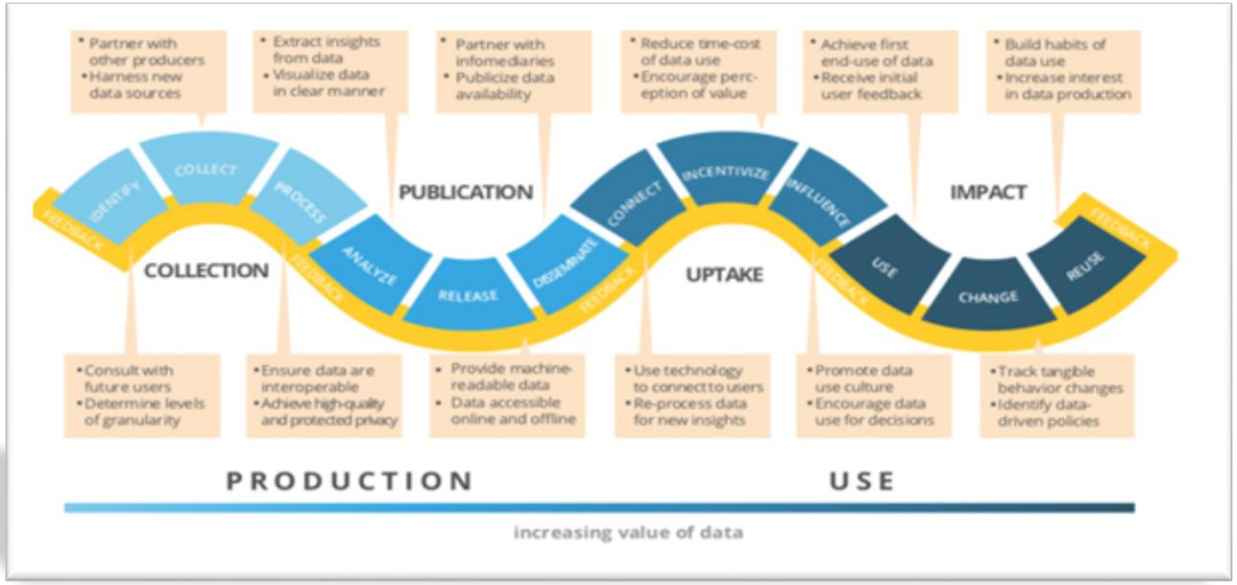
Investing time and resources in the development and deployment of data interoperability solutions will help us make better use of the data that currently sits in sectoral and institutional silos to implement and monitor the SDGs

# Interoperability and Open Data

Many organizations are now adopting open data policies that authorize and facilitate the reuse of their data assets

Open data requires data to be interoperable not only from a technical perspective, but also from a legal and institutional perspective

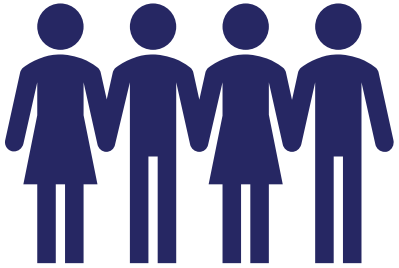
# Interoperability in the data value chain




Source: Open Data Watch

- Interoperability is a key element in data collection
- Planning for interoperability should take place early in the data life cycle
- But interoperability considerations also should inform every step in the data value chain

Global Partnership for Sustainable Development Data



# Collaborative on data interoperability

- **First UN World Data Forum, January 2017:**
  Jointly convened by UN Statistics Division and Global Partnership for Sustainable Development Data

- **Second UN World Data Forum, October 2018:**
  Launch of Data Interoperability Guide

# Vision

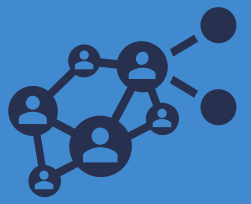
**Integrating and joining up data from multiple sources, and across systems**, to realize the data revolution for sustainable development, where more and better data is **open**, **accessible**, and **used** to the fullest extent possible in improving **people**'s lives

Getting the **governance and institutional framework** right

Designing **data structures** with users in mind

Standardizing the **data content**

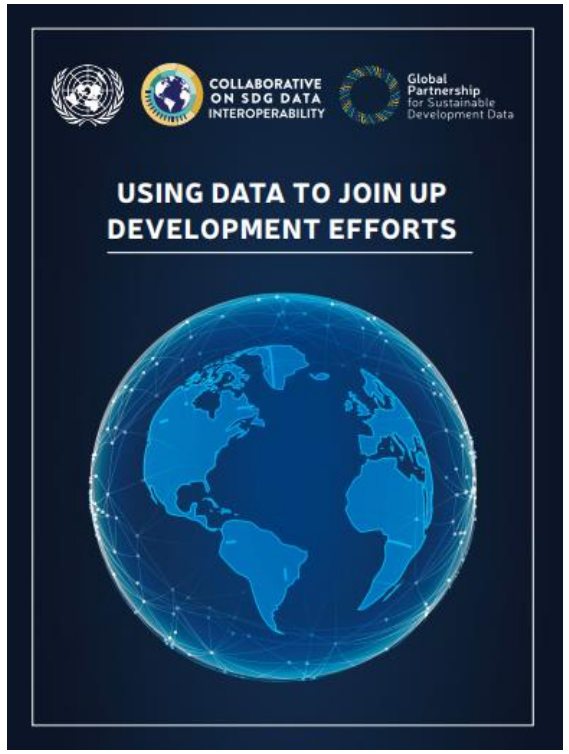
Providing standard **interfaces** to access and work with data



Disseminating **linked open data** for knowledge creation

# Pathway to data interoperability

# Practitioner's guide to joining-up data in the development sector




1. Data management and governance
2. Canonical data and metadata models
3. Classifications and vocabularies
4. Standardised interfaces
5. Linked open data

# Data management and governance

# Data management and governance

**Data management:**

*"The development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles"*

DAMA (2017)

# Data management and governance

**Data governance:**

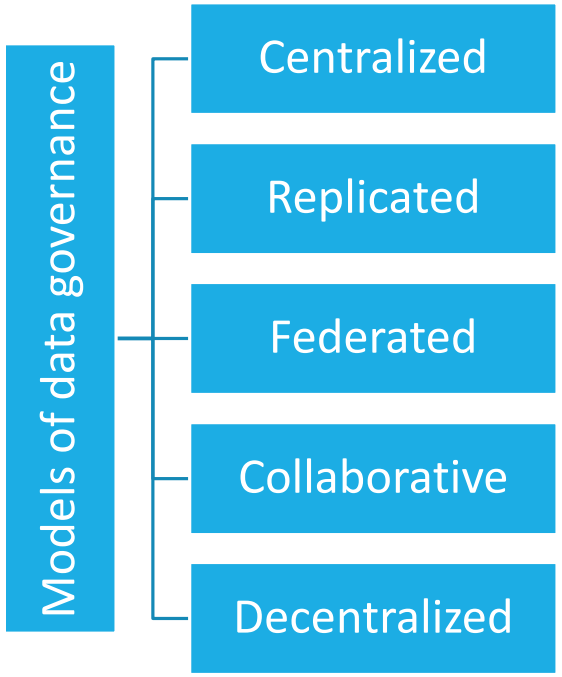
*“How decisions are made about data, and how people and processes are expected to behave in relation to data”*

DAMA (2017)

# Data management and governance

- Very often interoperability is less a technology problem and more a data management and governance issue
- To be effective, data management requires that data be governed – well controlled with oversight and accountability – across its lifecycle.
- Institutional frameworks play a key role in creating the environment where data, technology, and business processes fit with each other

# Data management and governance




- ✓ Too much decentralization does not work well in volatile environments that require data standards and coordination to tackle global information sharing challenges.
- ✓ Conversely, too much centralization can hinder experimentation and the creativity needed to innovate and to respond to emerging needs of data users and the quickly changing technological landscape.

# Data management and governance

Legal and regulatory frameworks are crucial to interoperability

- They set the boundaries of what is acceptable conduct and what is not
- They specify how data can be shared across organizations (e.g., standards for data reporting, security and protection)
- They determine what data can, or cannot, be shared and integrated (for example, data protection and privacy laws).

Laws and regulations exist at many different levels

- International normative frameworks, international laws and domestic laws
- Memoranda of understanding (MOUs), data sharing agreements
- Licenses
- Corporate policies, protocols and procedures for data sharing within the parameters of the law

# Data management and governance

Embedding data interoperability as a guiding principle in an organization also requires appropriate oversight and accountability

- Ensure that data is comprehensive, timely and supported by metadata
- Ensure conformity with standards and compliance with any applicable laws and regulations
- Ensure coordination with other organizations and entities on the best approaches to sharing and exchanging data

# Data management and governance

The appropriate accountability model will vary from one organization to another, depending on:

- Organization size and complexity
- Availability of resources
- Management structure
- Role of the organization in the broader data ecosystem

# Data management and governance

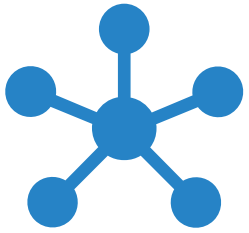
**Questions for discussion:**

- What model of data governance would work best to ensure interoperability of SDG data in Bangladesh?
- What mechanisms of oversight and accountability are in place in your organization with respect to the implementation of interoperability standards?
- Are internal data availability assessments/audits conducted on a regular basis to determine what data is held and how it is handled over its lifecycle?
- Does your organization conduct comprehensive data quality assessments and audits in collaboration with other stakeholders within the national statistical system? Do these assessments include interoperability as a quality dimensions?
- Is there a Monitoring, Evaluation and Learning framework in place that includes indicators on data governance issues?

# Data management and governance

**Common pitfalls in data governance:**

- Failing to take an organisational approach to data management and governance issues and relegating data interoperabliity issues to the IT department
- Not developing/enforcing a clear chain of accountability specifying roles and responsibilities across departments when it comes to the effective governance of data across/between organisations
- Overlooking/not considering interoperability issues as a requirement when updating or procuring new IT systems; resulting in internal data silos, and multiple types of data held in incompatible formats and schemas
- Not making the best use of legal and regulatory tools and frameworks that can create a safe and structured environment in which data can be shared and integrated while respecting privacy, data protection and security considerations



# Data and metadata models

# Data and metadata models

- SDG data and metadata is often represented in variety of (usually incompatible) ways across different organizations within the National Statistical System
  - ✓ Prioritizing internal operational needs
  - ✓ Having a specific applications in mind
- Interoperability is highly dependent on data and metadata modelling decisions and practices
  - ✓ Producers and users of data must have a common understanding of how it is structured in order to effectively exchange it across systems.
  - ✓ They must also share a common understanding of how the various components of a dataset relate to each other and to the components of other datasets.
- Data and metadata modelling decisions can ensuring that systems are designed with interoperability in mind from the outset

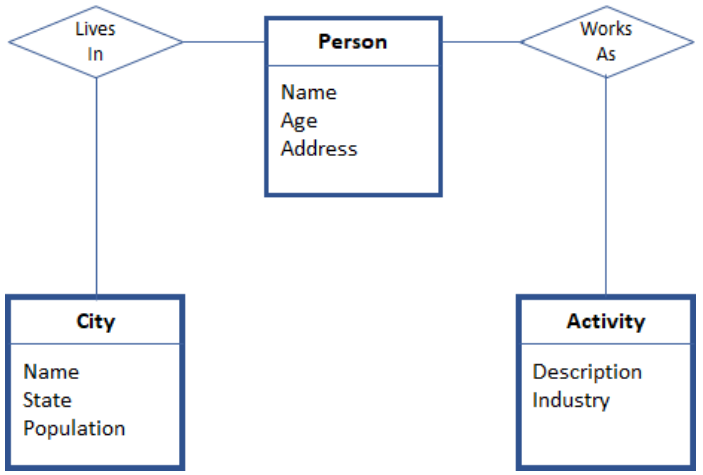
# Data and metadata models

- There is no single “right” way of representing information
  - Some data structures are better suited for operational processes (e.g., capturing data from a survey or maintaining a civil registration database)
  - Others are being better suited for data sharing and dissemination (e.g., for the creation of data visualizations)

*Can we ensure that all SDG data across the National Statistical System are mapped to a common data and metadata structure?*

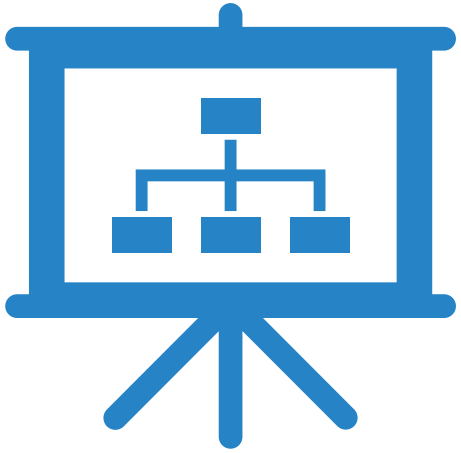
# Data and metadata models

- **Data modelling:**
  Process of clearly and unambiguously identifying things (**entities**) that a dataset aims to capture, and then selecting the key properties (**attributes**) that should be captured to describe those entities in a meaningful way
- It requires deciding:
  - How entities and attributes relate to each other (**relationships**),
  - How their information content should be formally codified within a dataset



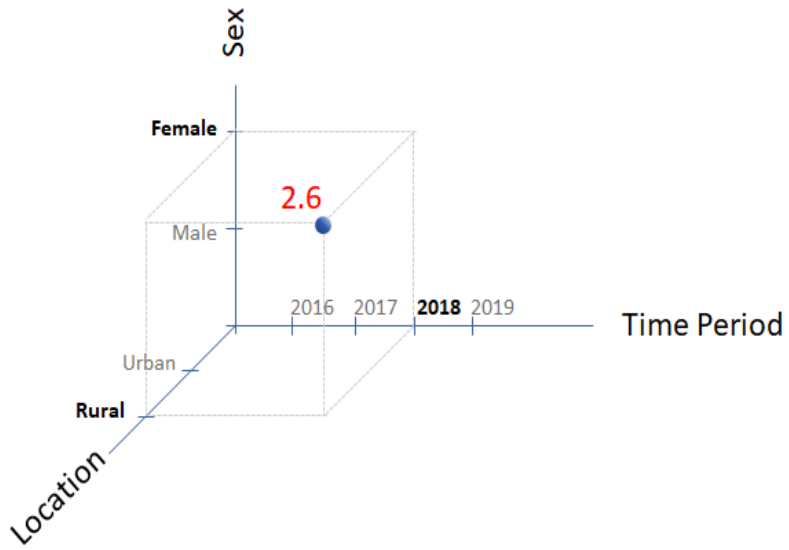
# Data and metadata models

- **Canonical data and metadata models**
  - ✓ Follow specific standardized patterns
  - ✓ Highly reusable and conducive to data sharing
  - ✓ Can be used to represent multiple sources of data and metadata using common patterns, thus making data integration simpler and more efficient



# Data and metadata models

- In SDMX, a “Data Structure Definition” (DSD) describes the elements of a statistical dataset through:
  - **Dimensions:** Concepts that are needed to uniquely identify an individual observations
  - **Attributes:** Concepts that provide more information about some part of the dataset.
  - **Primary measure(s)**: “observation values”




There are various globally agreed DSDs in different domains of application (e.g., National Accounts, Balance of Payments, Price Statistics, International Merchandise Trade, Energy, and SDG indicators*)

# Data and metadata models

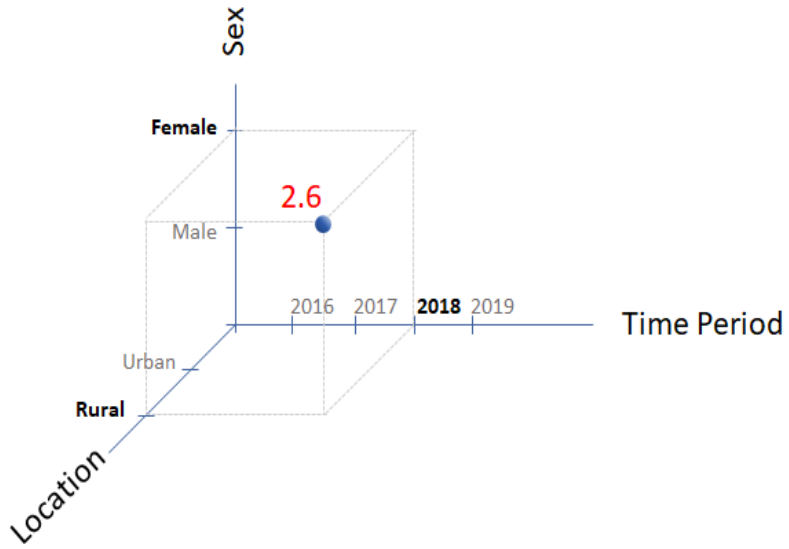
## SDG Data Structure Definition

- **Generic** template for the integration of SDG indicator data
- Highly **reusable** and conducive to **data sharing**
- **Focused on simplicity**, so data is easily understood by a wide range of users and applications
- **Self-contained** and **stable** over time
- Incorporates **standard definitions and classifications**
- **Extensible** to include national disaggregations
- Data platform and **technology-independent**

→ **Already tested** in by various countries

# Data and metadata models

- ✓ SDMX modelling guidelines and DSD guidelines provide a step-by-step introduction to data modelling and the creation of DSDs.
- ✓ The guidelines contain numerous links to further, more detailed guidelines and templates that can be used in the modelling process.



# Data and metadata models

- Standard metadata schemas
  - A metadata schema specifies the metadata elements that should accompany a dataset within a domain of application.
  - For instance, W3C's Data Cube Vocabulary (W3C 2014) recommends that each dataset should be accompanied by **title**, **description**, **date of issue or modification**, **subject**, **publisher**, **license**, **keywords**, **etc**.

# Data and metadata models

**The Data Catalog (DCAT) vocabulary:** a Canonical metadata schema specification designed to organize reference metadata at the level of the dataset and above:

- improve discoverability of datasets
- support dataset management and address operational issues
- promote the re-use of metadata elements from existing “namespaces”.

- Well-documented, flexible and practical metadata standard grounded on the foundations of widely used vocabularies
- Used by major engines for data portals
- Adopted by the European Commission to publish datasets pertaining to the public sector

# Data and metadata models

**Generic data modelling checklist:**

- Starting from a set of source tables, identify elementary datasets to be modelled (variables or indicators).
- Identify key entities that are described in the information contained in the dataset (e.g., places, people, businesses…):
  - Identify dimensions and attributes needed to describe each entity at the target level of granularity (e.g., location, time period, sex…);
  - To the extent possible, re-use standard dimensions and naming conventions from existing data models (e.g., from existing SDMX data structure definitions);
- Consider merging or splitting columns from original tables to define more useful dimensions for data exchange.
- Create a separate table of distinct values for each dimension assigning a unique ID to each row.

# Data and metadata models

**⚠ Caution!**

A common mistake in modelling datasets for sharing and dissemination is to try to replicate internal data structures from operational database systems.

The focus should be on producing simple, self-contained datasets that are easy to understand and manipulate by users and client applications.

# Standard classifications and vocabularies

# Standard classifications and vocabularies

- Classification systems shape the way data is collected, processed, analyzed and shared with users
- The use of standard classifications and vocabularies to identify, label, and catalogue individual data points and datasets has an impact on the ability of people (and machines) to easily find, access and integrate different datasets and information across data platforms
  - ✓ They allow data producers to express the meaning of data without ambiguities and
  - ✓ They enable users to find and link related pieces of information, from the unit record level to the dataset level, across different information systems.
  - ✓ They constitute the basis for data management and data interoperability.

# Standard classifications and vocabularies

- To expose data without ambiguities and ensure semantic interoperability, it is crucial to adopt standard vocabularies and classifications at the design phase of any new data collection, processing or dissemination system.
- However, the use of customized classifications and vocabularies is sometimes unavoidable
  - legacy data management systems,
  - specific needs in their primary domain of application
- Often it is necessary to “standardize after the fact,” mapping “local” terminology to standard vocabularies and taxonomies

# Standard classifications and vocabularies

- To meet the needs of a continuously changing data ecosystem, classifications and vocabularies need to adapt over time and be continuously “mapped” to each other by establishing associations of correspondence between their elements.
- Moreover, they need to be publicly available and accessible in open, machine-readable formats, such as CSV, JSON or RDF

# Standard classifications and vocabularies

**Examples:**

The SDMX content-oriented guidelines

Commonly used geographic referencing and coding standards

- Standard country or area codes for statistical use (M49)
- ISO 3166  standard for country codes and codes for their subdivisions.

Standard vocabularies used to represent key metadata annotations

- Dublin Core
- Schema.org

# Standard classifications and vocabularies

Controlled vocabularies and classifications need to be continuously updated and adapted to the diverse and changing requirements of data producers, data managers, and data users.

- Starting point should always be the reuse of existing classifications and vocabularies
- New classifications and vocabularies should only be suggested where a genuine gap exists or where old systems are redundant and no longer fit for purpose.
- Once a standard classification or vocabulary is created and implemented; policies, methods and procedures for its maintenance need to be established

→ Part of data governance

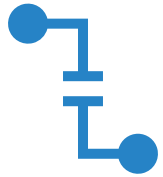
# Standard classifications and vocabularies

**Standardization vs. customization**

As information systems become larger, more complex and more interconnected, there is a growing tension between the need to use standard classifications and vocabularies to enhance interoperability, and the need to devise specialized ones for specific user groups.

Thought needs to be put into how this balance is set and to the extent possible, efforts should be made to connect new classifications to existing ones to ensure continuity and interoperability down the line.

Classifications and controlled vocabularies should not be seen as static; they need to be flexible and be continuously updated and adapted to the diverse and changing interests of data producers, data managers, and data users.



# Open data formats and standard interfaces

# Open data formats and standard interfaces

- Data needs to be easily available and accessible to a variety of user groups.
- Interoperability is not only about standardized data production, but also about standardized “data logistics” (Walsh and Pollock)
- There is need for common “pathways” to get data from providers to users in a fast, convenient, effective, and efficient manner.

# Open data formats and standard interfaces

Open data formats

- A first step is to make the open data available through bulk downloads in open data formats (such as CSV, JSON, XML, and GeoJSON, etc )
- Data interoperability is greatly enhanced when electronic data files are made available using openly documented, non-proprietary formats.
  - ✓ Human-editable and machine-usable
  - ✓ Agnostic to language, technology and infrastructure.

# Open data formats and standard interfaces

- CSV
  - ✓ Easy-to-use data format for developers and non-developers alike
  - ✓ Probably the most widely supported format across different technological platforms
  - ✓ Although it does not incorporate a schema for validation, there are recent alternatives that combine CSV tabular data with additional schema information
- JSON or XML
  - ✓ Useful to structure data in a text format and to exchange data over the internet
  - ✓ Allow producers and users to encode common data elements and sub-elements in such a way that data and metadata are linked together but clearly distinguishable from each other.

# Open data formats and standard interfaces

- Application Programming Interfaces (APIs)
  - ✓ Highly-reusable pieces of software that enable multiple applications to interact with an information system.
  - ✓ Provide machine-to-machine access to data services
  - ✓ Enable users to focus on the data rather than spend their time collecting it. APIs
  - ✓ Allow to automate data flows that involve repetitive and frequent data sharing and exchange operations, avoiding costly and error-prone manual intervention.
  - ✓ Provide the building blocks for users to easily pull the data elements they need to build their applications.

# Open data formats and standard interfaces

- API documentation
  - ✓Technical contract between a data provider and its users.
  - ✓One good practice is the consistent use an API description language (e.g., Swagger)
  - ✓It is also crucial for the documentation of an API to keep track of its different versions

# Open data formats and standard interfaces

- The Open API specification is a standard format to document all the functionality of a web REST API, describing, in a way that is both human and machine readable:
  1. The resources provided by the API
  2. The operations that can be performed on each of these resources
  3. The inputs needed for, and outputs provided by, each of these operations
- It also can be used to document user authentication methods, and to provide additional information like contact information, license, terms of use, etc.
- Example: UNSD'S Global Sustainable Development Goal Indicators API.

# Open data formats and standard interfaces

- It is good practice to manage all of an organization's APIs as a single product.
- To improve interoperability, all APIs in an organization's porfolio should:
  - Be standardized
  - Have mutually consistent documentation and functionality
  - Implement common design patterns
  - Be built from reusable components.
- To facilitate discoverability, the description of the API portfolio should be served by a specific end.

# Open data formats and standard interfaces

- System interfaces should **prioritize interoperability and flexibility over specificity and optimization**.
- A balance must be struck between specific user group needs and broader usability.
- Over-customization of an interface can inhibit its accessibility, usability and interoperability with other systems.

# Thank you