United Nations | Department of Economic and Social Affairs
Statistics

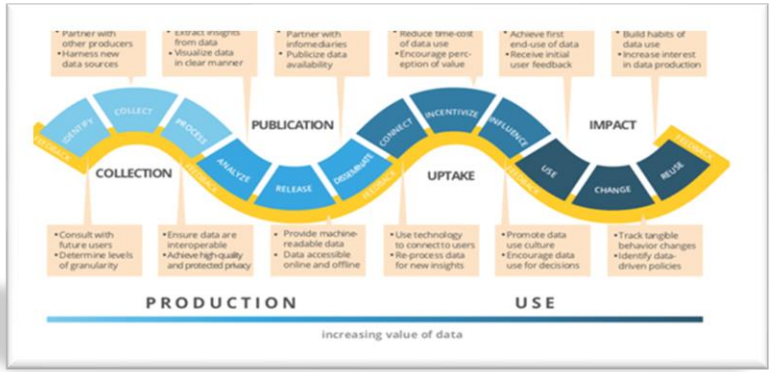
Introduction to SDMX data modeling

Photo by Markus Spiske, unsplash

Interoperability

- Ability to seamlessly share, join, cross-analyse, exchange and re-use data produced from different sources, and at different times, to provide richer information for improved decision making
- It is a crucial characteristic of good quality data and of effective data management systems

Interoperability should be understood along the whole “data value chain”, from collection to use



Interoperability and data modelling

- Interoperability is highly dependent on data and metadata modelling decisions and practices
- The same information content is often represented in variety of ways across different systems and organizations.
- There is usually no single “right” way of representing information
  - Some data structures are better suited for managing transactional processes (e.g., capturing data from a survey or maintaining a civil registration database)
  - Others are better suited for analyzing and communicating data to users (e.g., for the creation of data visualizations in a monitoring dashboard).

What is data modeling?

- A process focused on:
    1. Clearly and unambiguously **identifying things** that a dataset aims to capture
    2. Selecting the key properties that should be captured to **describe those things** in a meaningful way
    3. Deciding **how things relate** to each other
    4. Deciding how this information should be **formally codified**

→ *This is the essence of the Entity-Relationship model, which underlies most modern database management systems and applications*

Examples

- The content of a dataset may refer to **entities** such as “city”, “person”, or “activity”,
- These entities may be usefully described with **attributes** like “name”, “age”, or “industry”.
- In a specific application, it could be useful to capture **relationships** among entities and attributes, e.g., the fact that
  - one or more persons may live in a city,
  - a person may be employed in one or more types of activity …

Canonical data and metadata models

- Models that follow specific standardized patterns, making them highly **reusable** and conducive to data sharing.
- Provide a **common template** to which different datasets can be mapped
- Help develop a common understanding of how the various components of a dataset relate to each other and to the components of other datasets
- Reduce the number of transformations that user applications need to perform on their own to integrate the data from those sources

Standardization is not for free

- The underlying principle is **to hide from user the internal complexity of the operational data models** (e.g., which are optimized to avoid data redundancy and ensure data consistency validations), so they can concentrate on using data rather than spending time trying to understand the intricacies of internal data structures
- Data **providers need to take responsibility for mapping the data** from its original, operational structures, into commonly agreed presentations for dissemination and distribution purposes
- This may entail the need to undertake so-called “Extract-Transform-Load”, or ETL, procedures, **hidden from the view of users**

The multi-dimensional ‘data cube’ model

- Presents all relevant data about a population of interest in a **simple**, **self-contained** tabular view
- Each data point is characterized by
  - **Measures**: Observed values on one or more variables interest
  - **Dimensions**: A set of uniquely identifying characteristics
  - **Attributes**: A set of additional characteristics that further describe it

Time period	Location	Sex	Unemployment rate	Unit of measurement
2016	Urban	Male	3.4	percent
2016	Urban	Female	3.2	percent
2016	Rural	Male	2.5	percent
2016	Rural	Female	2.3	percent
2017	Urban	Male	3.7	percent
2017	Urban	Female	3.6	percent
2017	Rural	Male	2.7	percent
2017	Rural	Female	2.4	percent
2018	Urban	Male	3.8	percent
2018	Urban	Female	3.6	percent
2018	Rural	Male	2.7	percent
2018	Rural	Female	2.6	percent

Domains of dimensions, attributes and measures

- Each dimension, measure and attribute encapsulates **a concept**
- Concepts can be:
  - drawn from a code list (for e.g., “country ISO code”)
  - required to adhere to a specific data format (e.g., “YYYY” for years)
  - required to be contained within a specific range of values (e.g., “numerical values between 0 and 1”).
  - drawn from a type of values (e.g., “text”)