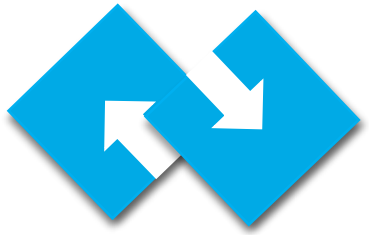
An NSO Perspective on Microdata Modelling Challenges in SDMX

Microdata use-case scenarios, common in statistical agencies, modelling challenges (and possible solutions)

Matjaž Jug, Statistics Netherlands (with the Input from the Cross-Working Groups Task Force on Microdata)



SDMX Global Conference 2025,
29th September to 3rd October 2025
Rome, Italy

Overview

- Intro: Statistics Netherlands' Data Strategy
- Microdata: What is new or different?
- Microdata use-case scenarios, common in statistical agencies
- Modelling challenges (possible solutions in next presentation:-)

Data Strategy: Statistics Netherlands as 'data-centric' organization

- Vision: 'data-centric' data management
  - Data is decoupled from processes, applications, infrastructure and departments silos so that data becomes more shareable;
  - There is optimal coordination of data supply with data demand by sharing data as much as possible and complying with FAIR;
  - We are in full control of our data;
  - Ability to connect to external data ecosystems.

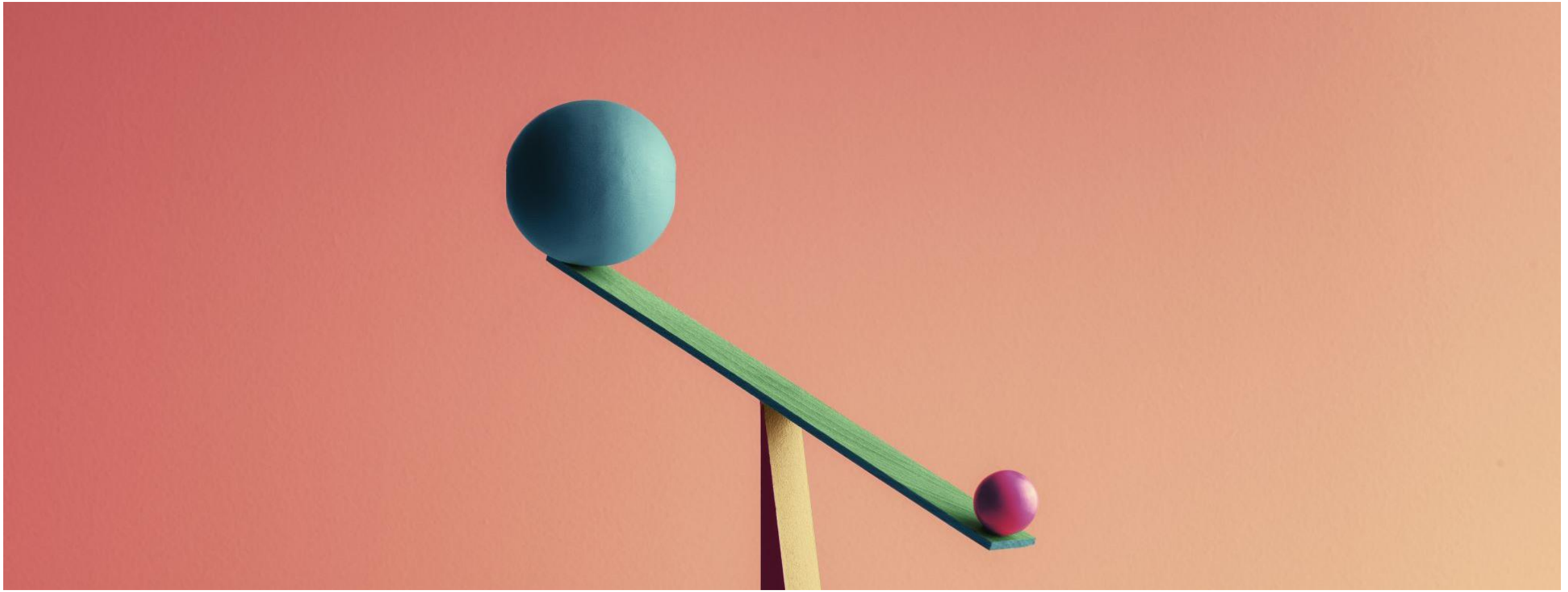
Microdata use-case scenarios in NSOs

- Data-centric approach requires **standardized metadata.**
- End-to-end data lifecycle requires possibility to reuse metadata (**interoperability** across multiple systems).
- SDMX has a potential to be used as **information model**, but it is not (yet) common in NSOs for microdata use cases.
- Potential use-case scenarios:
  - Use of administrative data in statistical production
  - Multi-source / multi-domain data integration
  - Microdata research
  - Data infrastructure: Data Catalogues, Data Lakehouses, Data Fabric..

Microdata: What is new (or different)?

- **Conceptual level**: conceptual metadata model has additional (or different) conceptual elements and there are terminological challenges such as the use of terms “variable” and “dimension”;
- **Logical level**: need to document complex relationships within and between datasets such as primary-foreign key relationships and cardinality;
- **Presentational level**: requirements for data selection and filtering such as classification levels and permanent constraints and sentinel values;
- **Technical level**: support for large volume micro datasets, live access to database / masterdata, direct links to other standards. (not covered in this presentation!).

Concepts



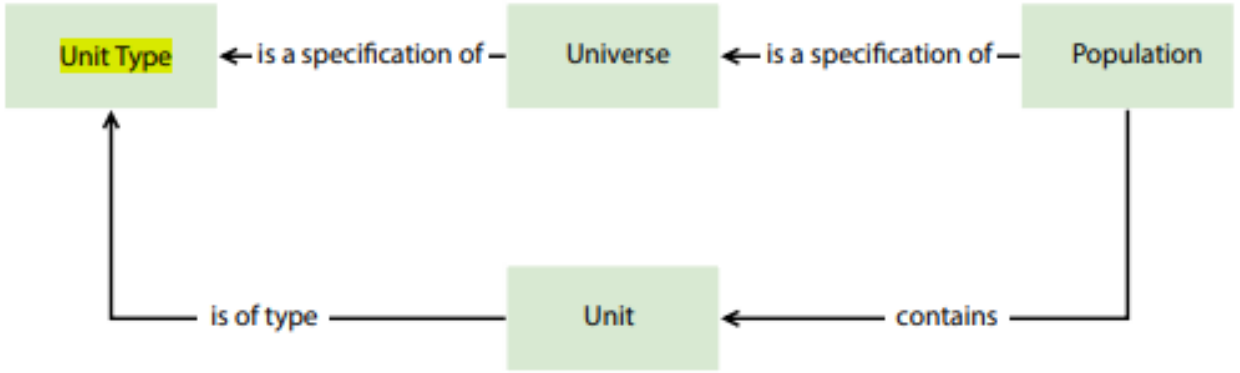
Definition of "Microdata"?

"..sets of observations or records pertaining to **individual statistical units**, such as people, households, or businesses. GSIM categorizes microdata as a **type of information object** containing **individual-level data points**, allowing for detailed analysis of **specific entities** rather than aggregated statistics."

Generic Statistical Information Model (GSIM) Version 2.0

3.2 *Population*

Figure 11 ***Population*** **and** ***Unit***



Source: UNECE.

Terminology!

GSIM User Guide.pdf

SDMX glossary

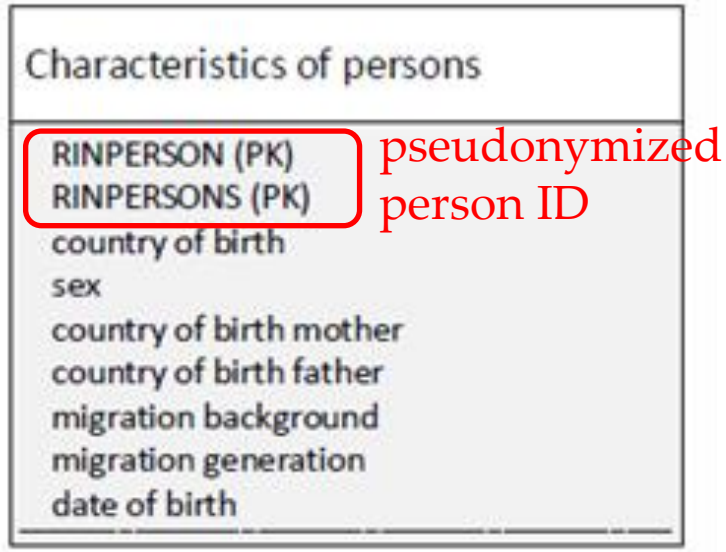
Primary Keys and Dimensions

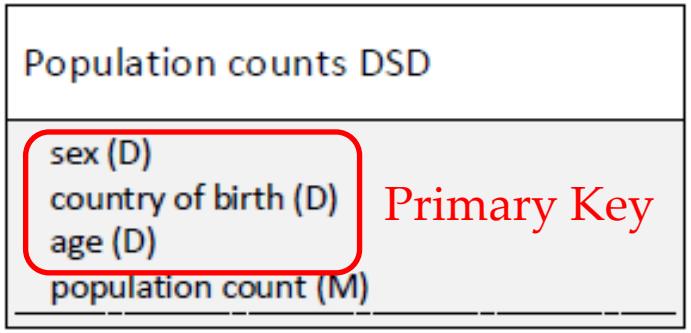
Microdata:

- Data granularity at the level of statistical units
- PK is often defined by identifier(s)
- Difficult to define what is Dimension and what is Measure

Macrodata:

- Granularity varies
- Dimension (usually categorical variable) is part of the primary key
- Dimensions and Measures are well defined

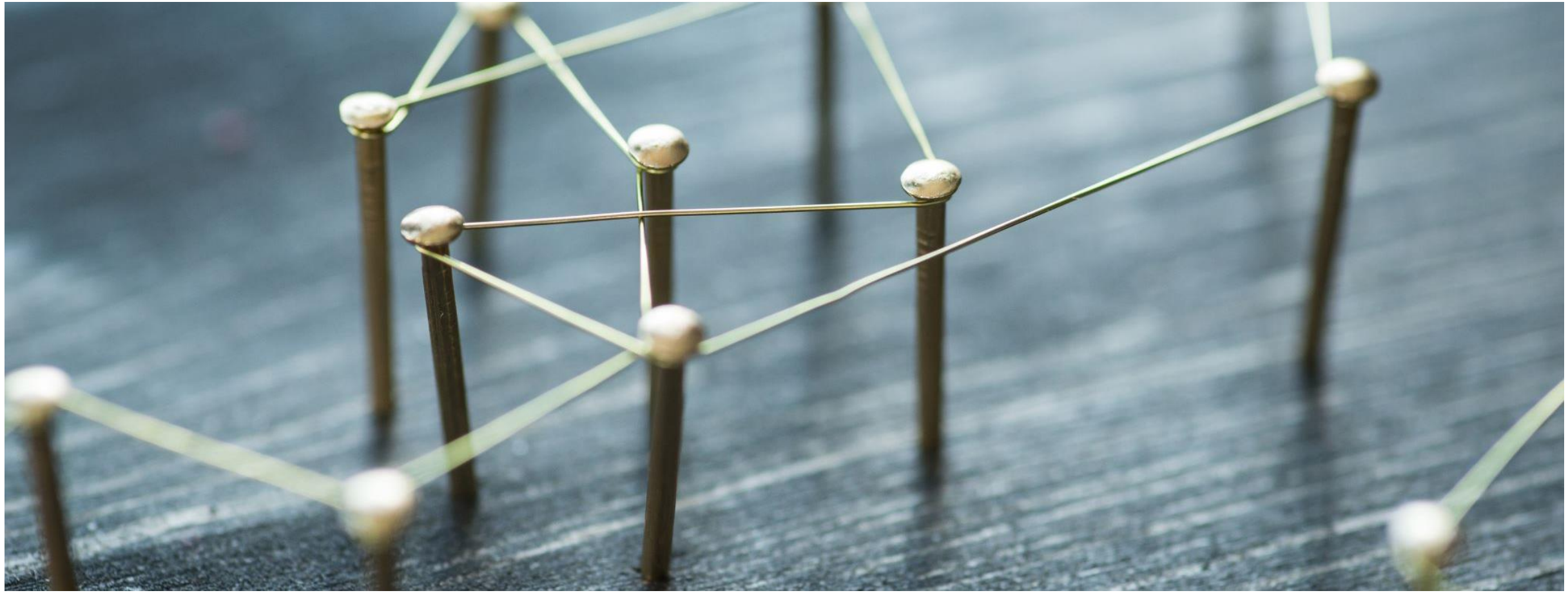




Modelling challenges

- The lack of agreed way (and place in the standard) to record information about statistical unit types (or other entities).
- Definition of “primary key” designation for microdata – in micro datasets usually defined by “identifier(s)” vs. the combination of “dimensions”, commonly used in macro datasets.
- Criteria for "measure" vs. "attribute" in microdata not clear.

Relationships



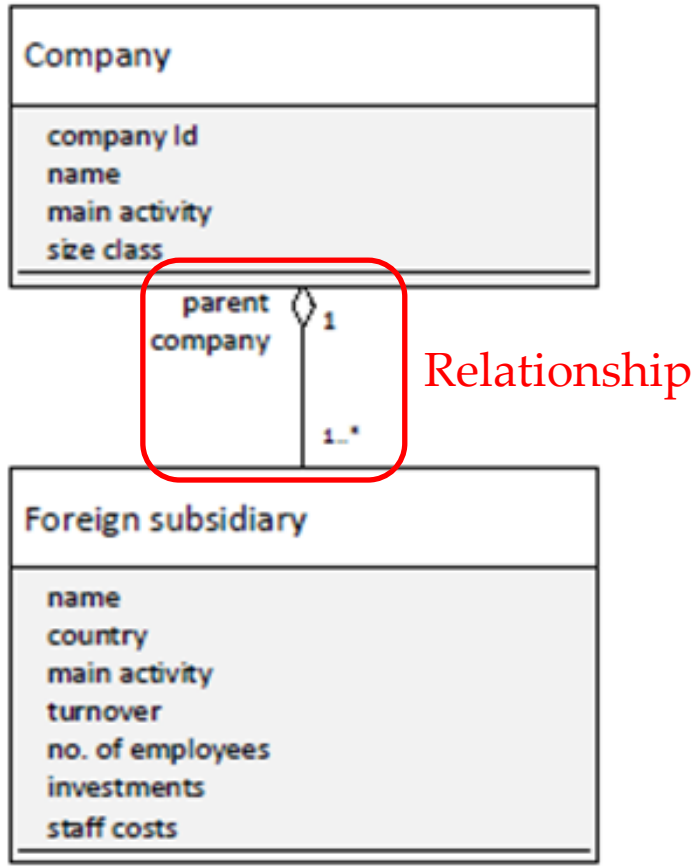
Relationships in microdata

Examples of relationships:

- Entity-Relationships (between statistical unit types)
- Micro-Macro relationships

Examples of use cases:

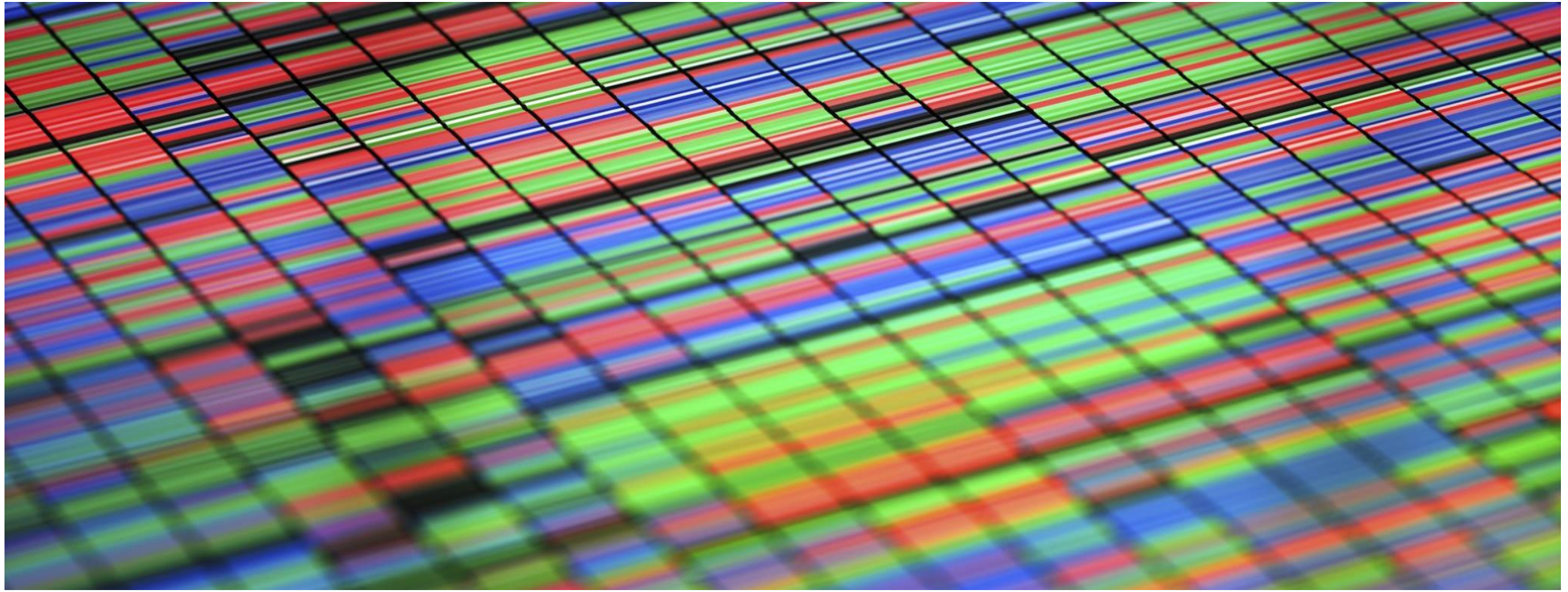
- Querying for concepts in data catalogues
- Querying for data in data stores
- Documenting data semantics
- Documenting data transformations (data lineage)



Modelling challenges

- Ability to capture and describe relationships between entities (such as statistical unit types) including cardinality.
- Ability to describe structural data transformations (such as micro -> macro).

Presentation



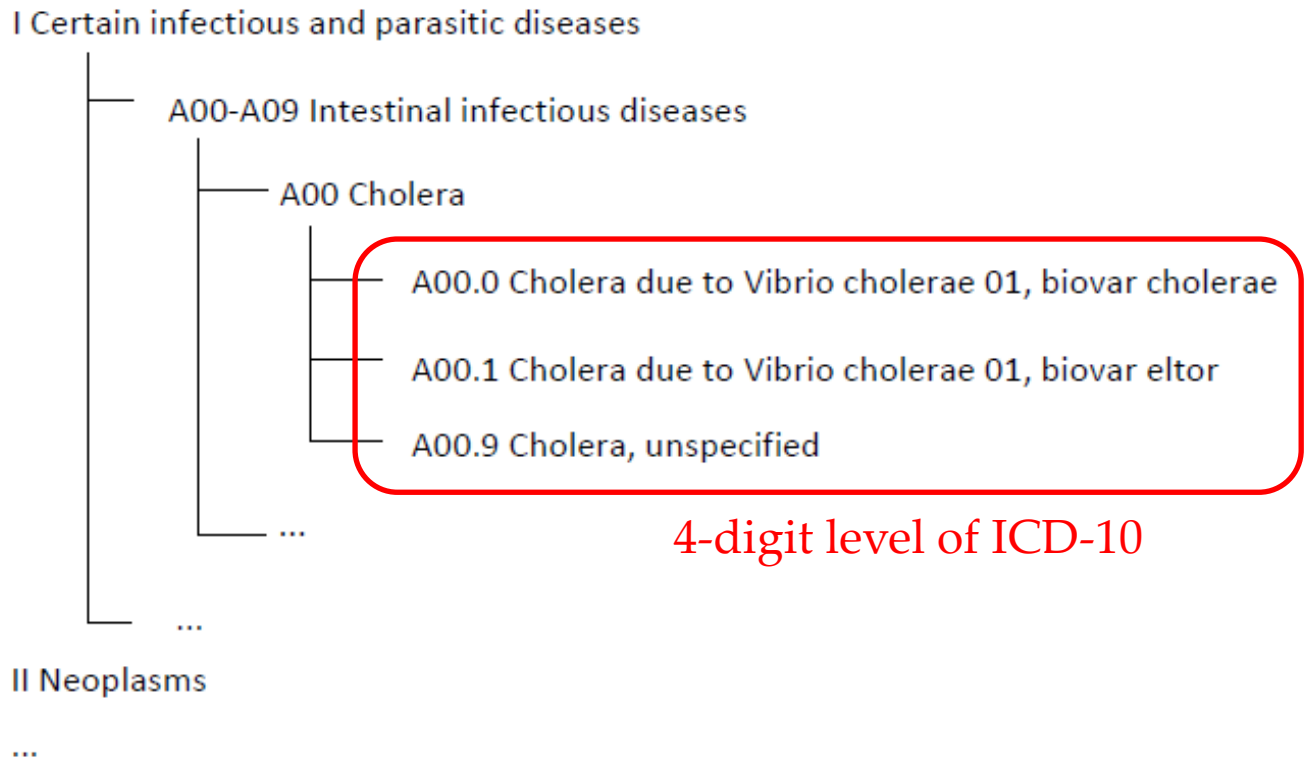
Data & metadata presentation

Examples:

- Selection
- Filtering
- Sentinel values

Examples of use cases:

- Ad-hoc aggregation using hierarchical classifications
- Datasets with subpopulations
- Use of special codes for sentinel values



Causes of death: ICD-10 classification (13.860 codes)

Modelling challenges

- Ability to restrict the value range of a variable (concept) based on a single level of a hierarchical code list.
- Ability to durably record a filtering operation.
- Ability to define sentinel values.

Questions?