Smart Metadata for Data Exchange: DDI-CDI and FAIR Implementation

Arofan Gregory (CODATA)
Chair, DDI-CDI Working Group

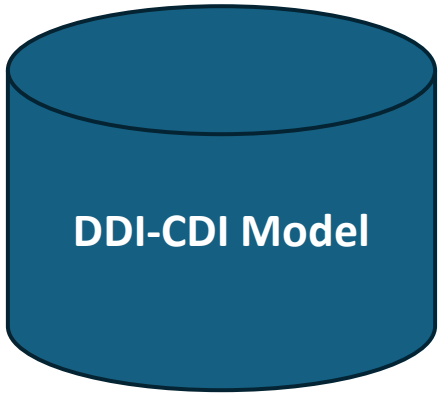
COSMOS Conference

Paris, 11 April 2024

The FAIR Challenge

- FAIR is all about making data more discoverable and reusable, both within and across scientific and policy domains
- There is a strong emphasis on detailed, standard metadata sufficient to drive automated processes for integrating and analyzing data
- The disseminators of data are often not “industrial strength” data producers (unlike the statistical agencies) but research organizations
- The challenge is to make data as useable as possible while requiring the lowest possible effort on the part of data disseminators
  - For stats agencies, this really is more about enhancing data quality/usability, since they are relatively good with standard metadata

Programmatic Restructuring of Data with DDI-CDI



Predictable, automatable

DDI-CDI Description

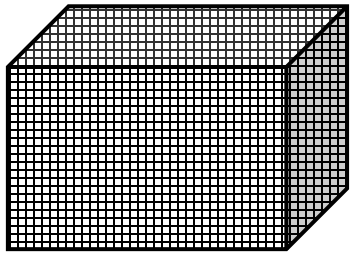
- Variables
- Structure
- Roles (Dimensions, Measures)
- Physical arrangement

[TRANSFORMATION]

DDI-CDI Description

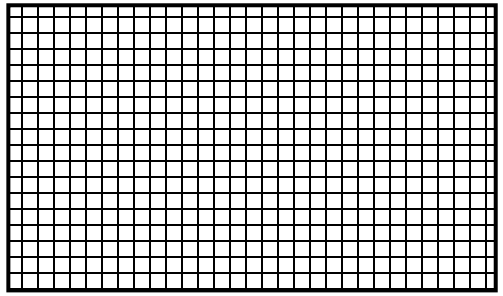
- Variables
- Structure
- Roles (Identifiers, Attributes. Measures)
- Physical arrangement

Domain Metadata Standards/ Models



Predictable, automatable

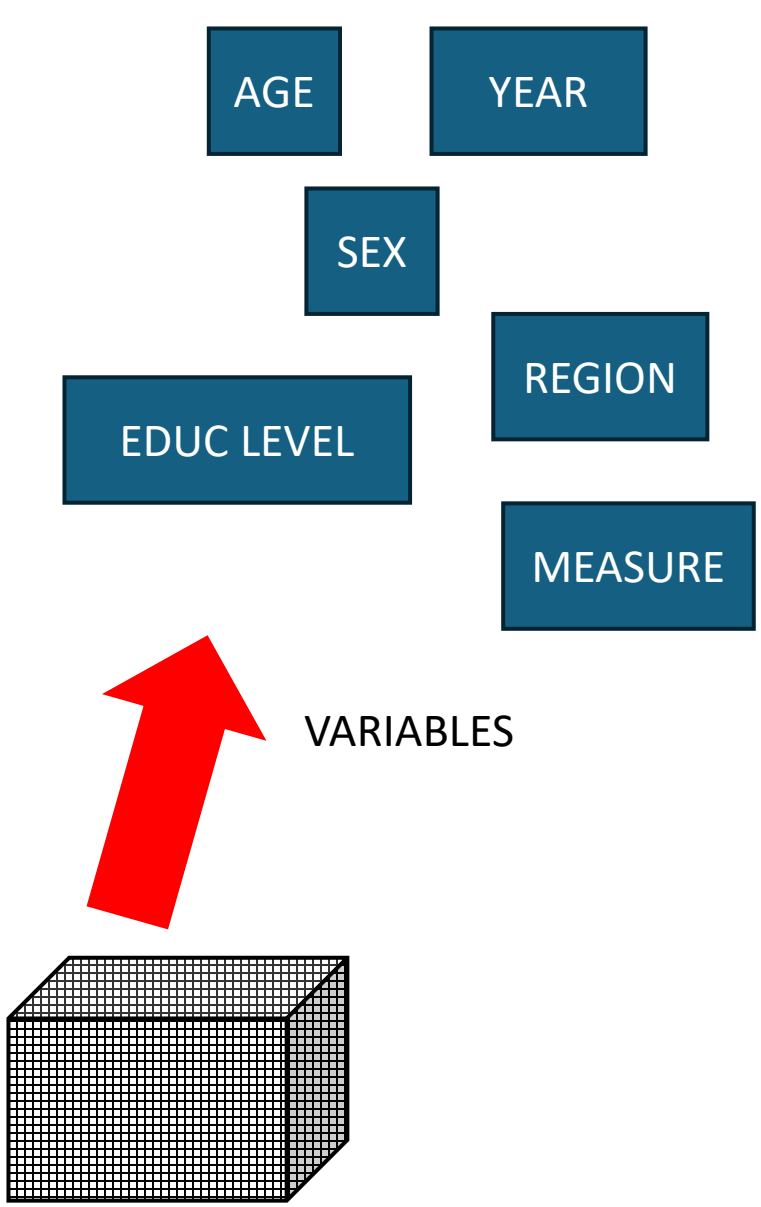
[TRANSFORMATION]



Dimensional Data Set:
Age x Year x Region x Education Level x Sex

Wide Data Set: Dimensions viewed as Variables (properties)

Programmatic Restructuring of Data with DDI-CDI (Cont.)



AGE	YEAR	SEX	REGION	EDUC LEVEL	MEASURE

The target format is wide, with the dimensions treated as variables (properties): at least one must be a place where integration can be performed with another data set (typically time and geography). Other variables may be combined into compound identifiers, or may be treated as additional descriptors or measures, depending on what they are. The roles of variables Change according to the structures – the meanings/values do not.

Aggregate values would be repeated to align with micro-data records as needed to provide the “complete” records for analysis.

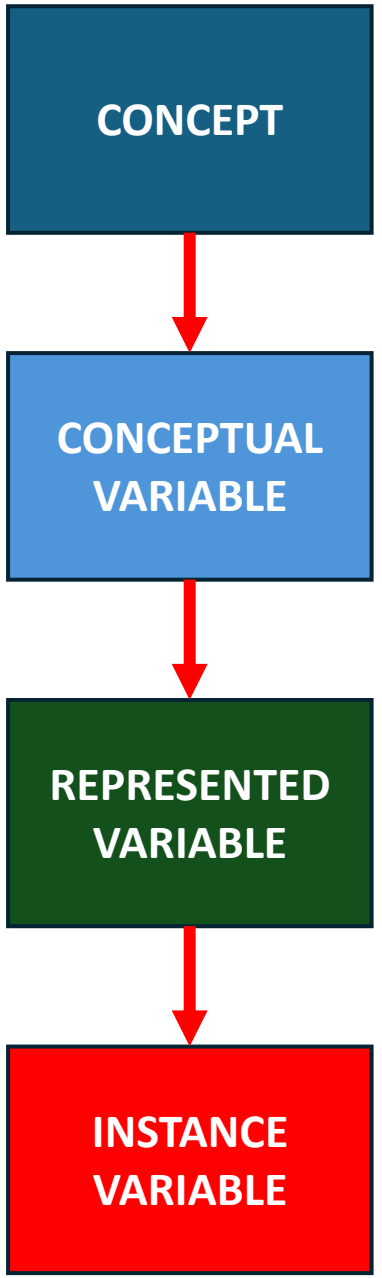
Recoding/semantic transformations are handled separately as appropriate to the structural re-arrangement.

The Evolution of Variable Description

- 20 years ago: a “variable” is a field in a data set (DDI Codebook)
  - ISO-11179 gave us reusable “data elements”
- 10 years ago: GSIM and DDI Lifecycle give us a three-level model:
  - Instance variables (in the data set)
  - Represented variables (reusable, comparable variables/”data elements”)
  - Conceptual variables: reusable with a transformation on the representation
- This was huge step forward, allowing for better data management and production (especially in longitudinal scenarios)
  - Helped users identify comparable variables across waves of data

DDI-CDI and the Variable "Cascade"

- This is the same three-level model, but optimized to meet the FAIR challenge
- Each level has a formal inheritance relationship with those above:
  - It has all of the information about its parent
  - It adds more properties to support its specialization (represented variable has a representation; instance variable has summary statistics, etc.)
- We can treat variables *polymorphically:*
  - An instance variable *is* a represented variable
  - A represented variable *is* a conceptual variable
- These relationships are guaranteed by the formalization in the DDI-CDI model

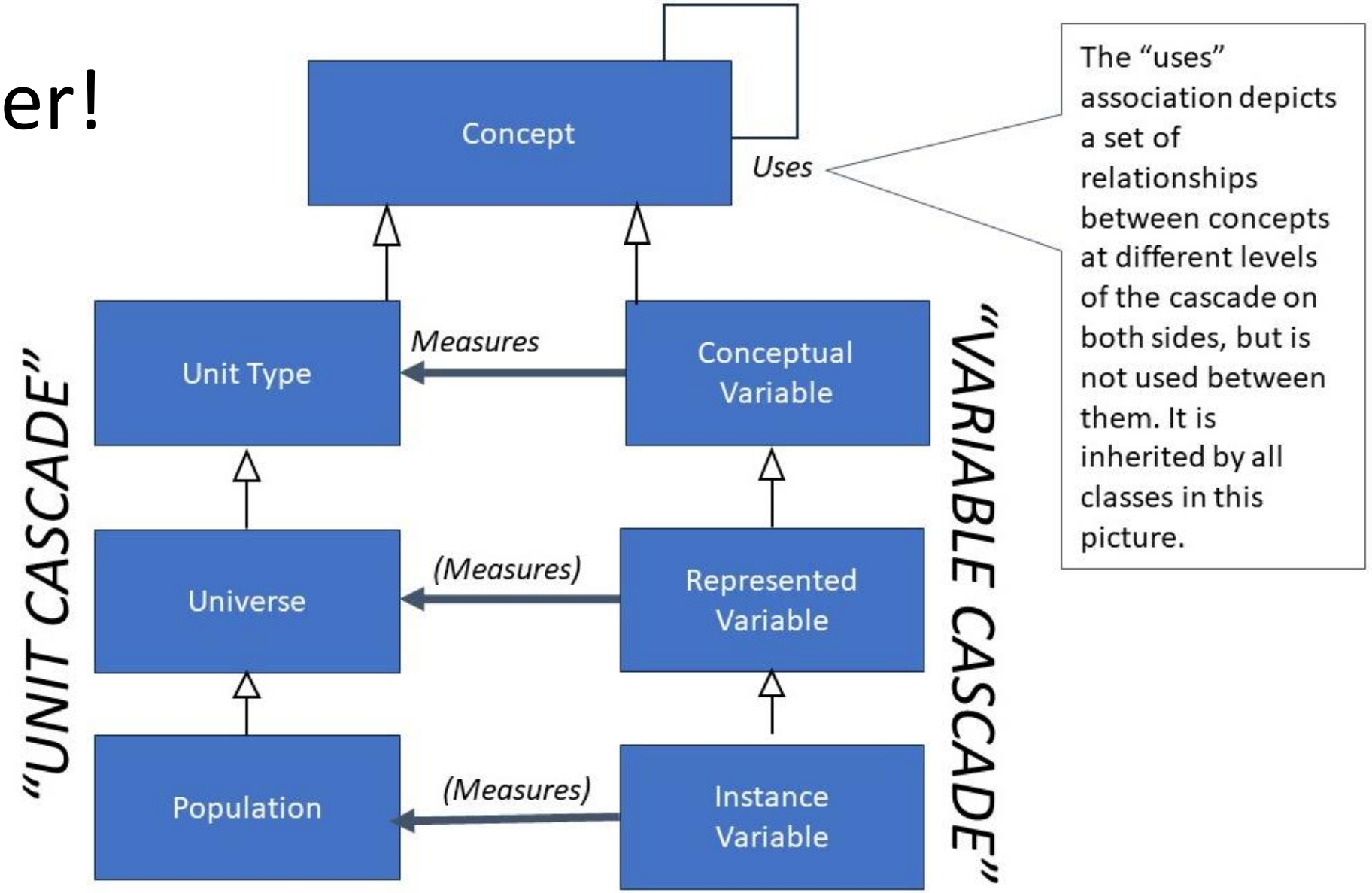


The Benefit

- The data disseminator describes their instance variables
- The data user can “blow up” the metadata for processing as if it was anything in the entire cascade

- Minimum cost for disseminator – maximum utility for the user

Even better!



The "uses" association depicts a set of relationships between concepts at different levels of the cascade on both sides, but is not used between them. It is inherited by all classes in this picture.

Advances in Techology

- The polymorphic relationships in the DDI-CDI model can be made explicit in RDF instances using standard features of RDFS and similar vocabularies
  - They become available for querying and processing using SPARQL, etc.
  - This type of metadata expression does not require the “polymorphic” use of OO technology
- Textual descriptions in the “Unit Cascade” are much more useful as a consequence of LLM technology
  - This is a new area, but it holds great promise
  - Very helpful in a FAIR scenario where unfamiliar data are being discovered and compared/integrated

Summary

- DDI-CDI provides a foundation for meeting the “FAIR challenge”
  - Disseminators describe the data as they understand it within their systems
  - Users can programmatically transform it into something easier to use
- The richness of the variable model optimizes the metadata description
  - Disseminators provide instance variable descriptions
  - Users ban expand the metadata set to all levels of the cascade for processing/integration
  - Relationships to the Unit Cascade provide even more information, especially for LLM systems