

DDI Cross-Domain Integration (DDI-CDI): An Introduction

Arofan Gregory, CODATA
Chair, DDI-CDI Working Group

Outline

- Why DDI-CDI?
 - Relevance for official statistics
- Key features
 - General considerations
 - Foundational metadata
 - Data description
 - Process description
- Examples of use
- Current status/looking forward

Why DDI-CDI?

- Increased focus in the scientific community on cross-domain research
 - CODATA's "Decadal Programme: Making Data Work for Cross-Domain Grand Challenges" (climate change, infectious disease, sustainable cities, disaster risk, etc.)
 - The FAIR principles (Findability, Accessibility, Interoperability, Reusability)
 - Organizations/initiatives: Research Data Alliance (RDA), GO FAIR, European Open Science Cloud (EOSC), etc.
- Relevance for official statistics:
 - Research to support timely policy
 - Integration of research with official data, both as source and as consumer
- DDI perspective: new and "unfamiliar" data being integrated within social, behavioural, and economic sciences
 - Unfamiliar structures and semantics
 - New technology platforms used in other domains
 - Need to describe and integrate based on standard metadata

Requirements for DDI-CDI

- A model which builds on and aligns with existing standards
- Domain-agnostic
- Enables integration of data across domain and institutional boundaries
 - Not focused on data management/archiving
 - Not focused on data collection/production
- Technology-agnostic/model-driven
 - UML model in “canonical” XMI
 - Implementable in different syntaxes and technology stacks
 - RDF, XML, JSON, Python, SQL, etc.
- Emphasis on data structure and data provenance/processing
- Emphasis on scalability through automation

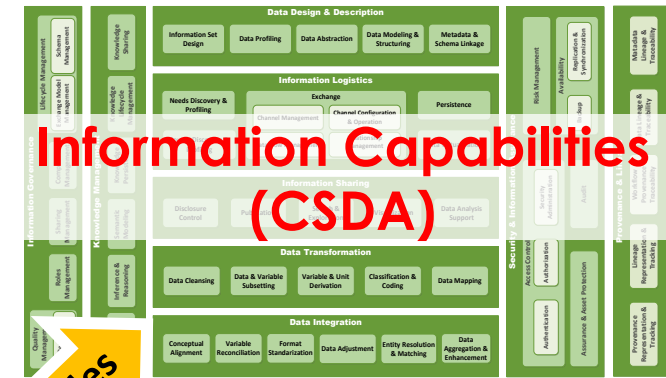
DDI-CDI and other standards

Quality Management / Metadata Management							
Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection exchange	4.1 Create frame and metadata capture	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Create and confirm needs	2.2 Design variable descriptions	3.2 Build process components	4.2 Review and validate	5.2 Review and validate	6.2 Update outputs	7.2 Produce dissemination products	8.2 Contact evaluations
1.3 Establish output objective	2.3 Design collection	3.3 Build dissemination components	4.3 Finalise collection	5.3 Derive new variables and units	6.3 Interpret and plan outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and template	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and input	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Calculate weights	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate aggregates			
		3.7 Finalise production system		5.7 Finalise data files			
				5.8 Finalise data files			

GSBPM



implements



Information Capabilities (CSDA)

other implementation Standards...

integrates

DDI-CDI

integrates

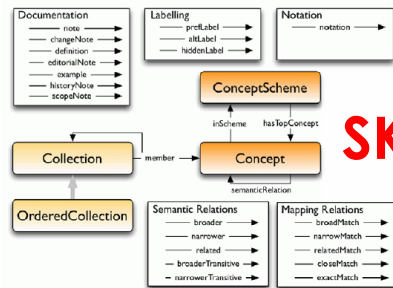


integrates

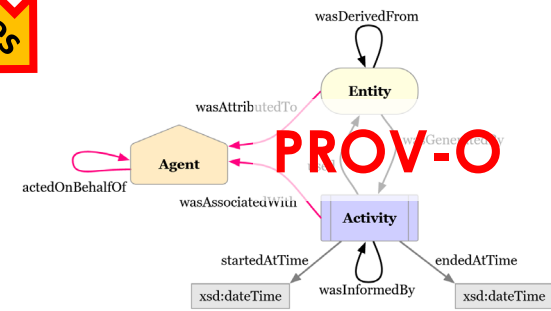
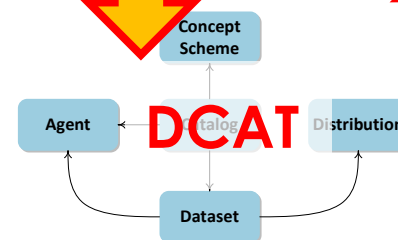
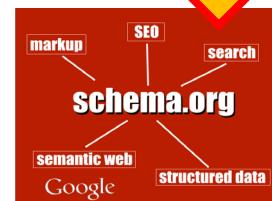
integrates

integrates

integrates



SKOS



PROV-O

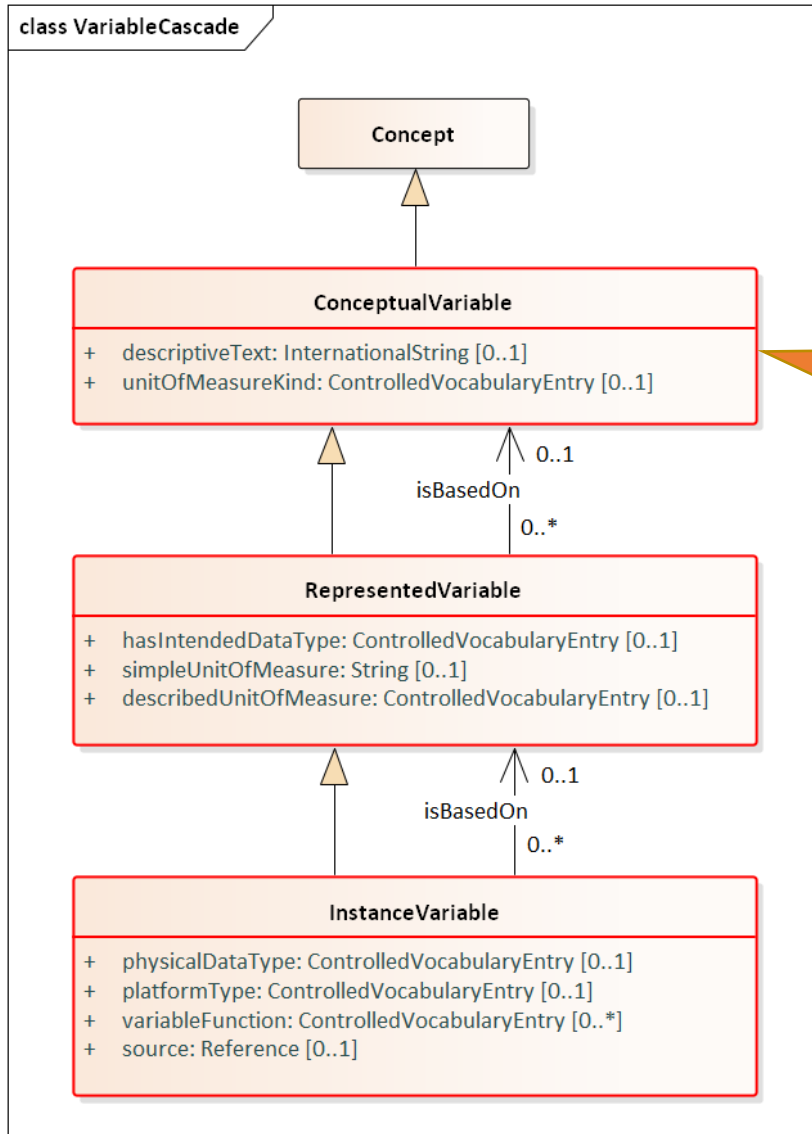
Key Features

- Foundational metadata (concepts, classifications, variables)
- Structural metadata (wide data, long/streaming data, key-value “big” data, multi-dimensional data)
- Process metadata
 - Framework for understanding how data sets are transformed and related
 - Reliance on other standards (PROV, VTL, SDTL, etc.)

Foundational Metadata

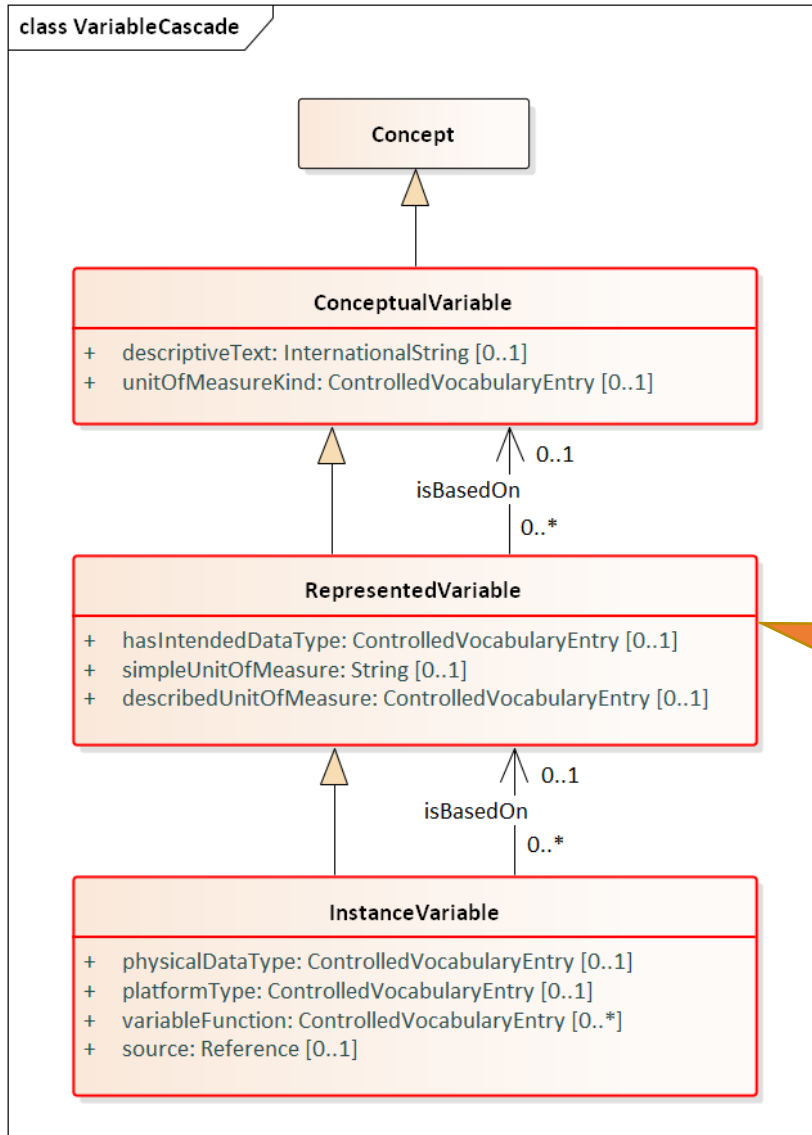
- Draws heavily on GSIM, Neuchatel, etc.
- Integrates the W3C “Simple Knowledge Organization System” (SKOS) and XKOS
- Implements the GSIM variable cascade as a key way of describing data of different types

DDI-CDI variable cascade – Conceptual



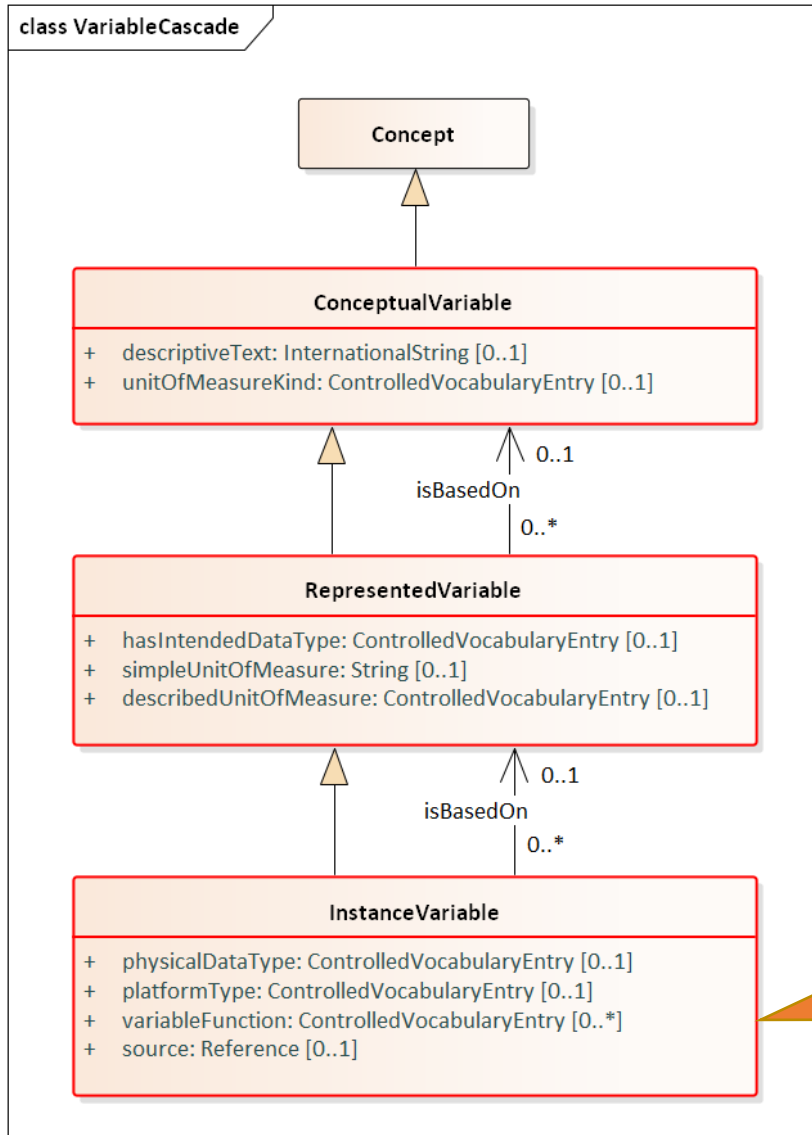
- Variable descriptions at a high level, e.g. conceptual domains
- Early design data capture/ intake
- Broad search and discovery
- Least specific/Most reusable

DDI-CDI variable cascade – Representation



- Variable descriptions at a detailed level, e.g. value domains
- Advanced design for all stages of data lifecycle
- Specific search and discovery
- More specific/Less reusable

DDI-CDI variable cascade – Instance



- Physical data description, e.g. physical data types
- Use of a variable in specific data instances
- Data search and discovery
- No reusable

Example: comparability and traceability

Married
Separated
Divorced
Widowed
Never married

Legalmaritalstatus
(conceptual variable)

Conceptual variable
Common variable
specification without a
representation

1	Married
2	Separated
3	Divorced
4	Widowed
5	Never married

MARITAL
(represented
variable)

MARITALB
(represented
variable)

Represented variable
Common variable
specification with a
code representation

m	Married
s	Separated
d	Divorced
w	Widowed
n	Never married

MARITAL
2004
(variable)

MARITALB
2008
(variable)

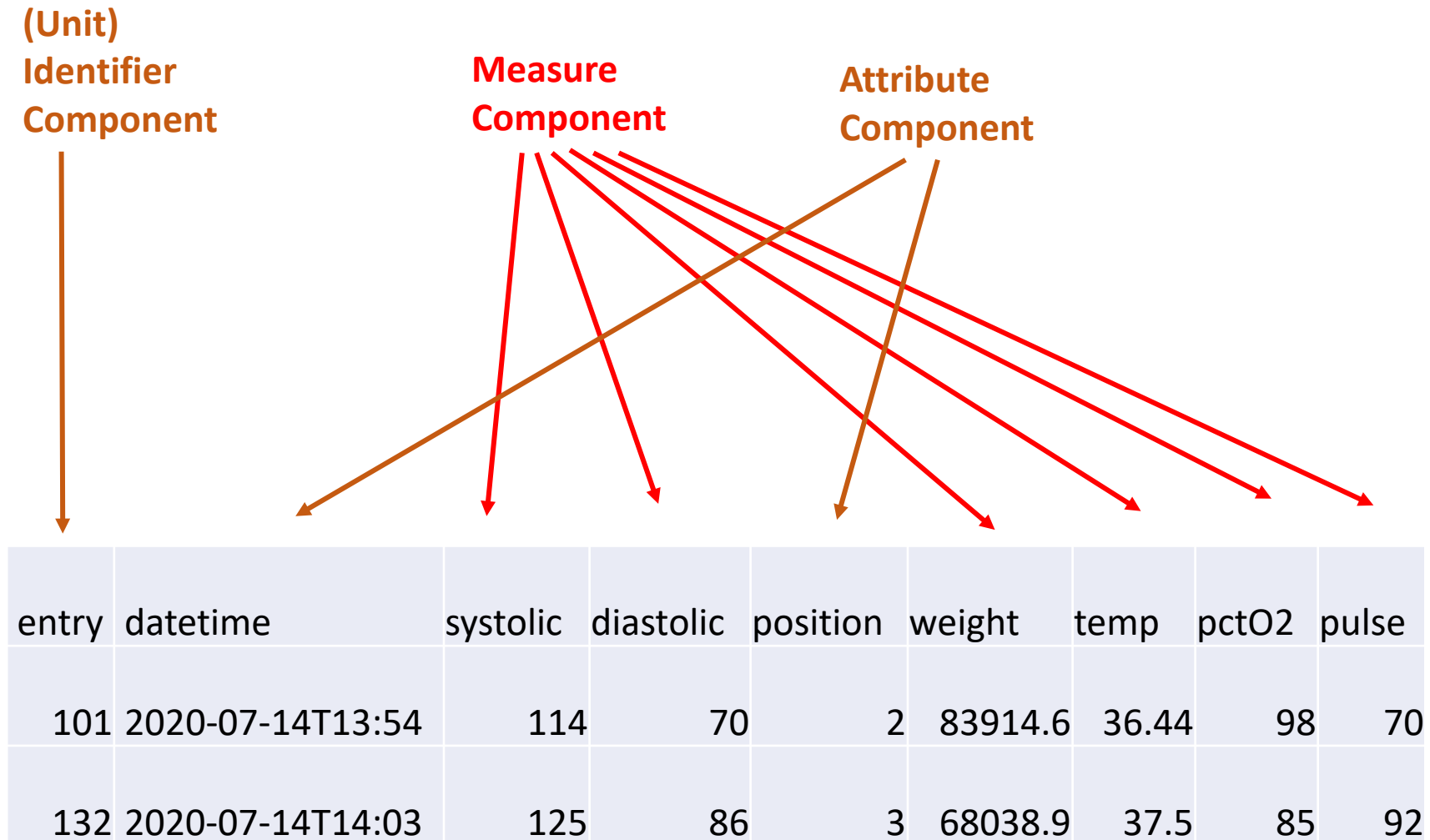
MARITALB
2018
(variable)

Instance Variable
Variable specification
within a dataset
context

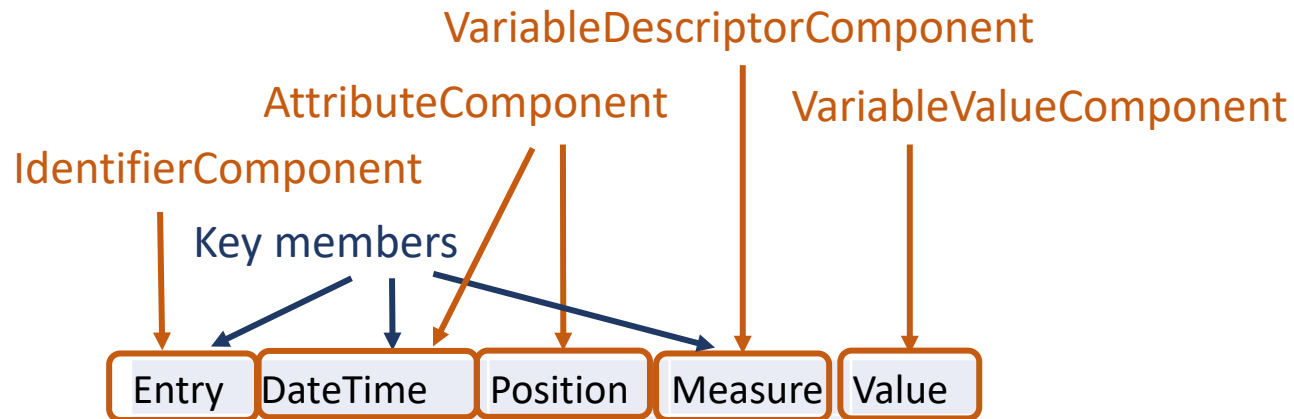
DDI-CDI and Data Structure Description

- Assigns roles to the atomic pieces of information within any given data set (identifier, measure, attribute, etc.)
- Associates the represented variables with specific roles, defines keys
- Allows for automatic transformation of data from one structure to another without loss
- Four basic types of data:
 - Wide – as with unit records
 - Long – as with event or stream data, sensor data
 - Key value – as in a key-value store (“big data”)
 - Dimensional – as with aggregate data, also indicators
 - Supports description of SQL databases using tables connected with keys

Example 1: data in wide form



Example 1: data in long form

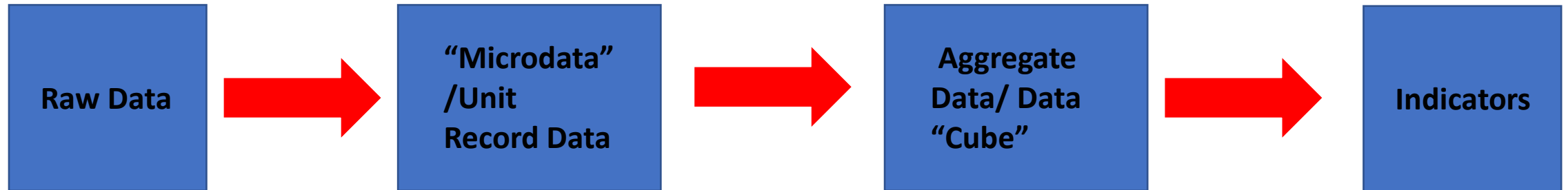


Entry	DateTime	Position	Measure	Value
101	2020-07-14T13:54	2	systolic	114
101	2020-07-14T13:54	2	diastolic	70
101	2020-07-14T13:54	2	weight	83914.60
101	2020-07-14T13:54	2	temp	36.44
101	2020-07-14T13:54	2	pctO2	98
101	2020-07-14T13:54	2	pulse	70
101	2020-07-14T13:54	2	away	n
101	2020-07-14T13:54	2	exposed	n
132	2020-07-14T14:03	3	systolic	125
132	2020-07-14T14:03	3	diastolic	86
132	2020-07-14T14:03	3	weight	68038.90
132	2020-07-14T14:03	3	temp	37.5
132	2020-07-14T14:03	3	pctO2	85
132	2020-07-14T14:03	3	pulse	92
132	2020-07-14T14:03	3	away	y
132	2020-07-14T14:03	3	exposed	n

The Variable Descriptor Component has values taken from the list of non-Unit Identifiers in the wide data set.

The “key” for each value is composed from the Identifier and the Variable Descriptor, and may include non-transposed components, e.g. DateTime.

Typical Data Transformations



- DDI – CDI describes the data at each stage, indicating the roles played by each atomic bit of data (“datum”)
- DDI – CDI tracks the processing between each stage (implements PROV), reflecting the relationships between atomic datums (uses other standards for describing specific processes – VTL, SDTL, proprietary)
- Supports both declarative and procedural process description

Some Implementation Examples

- European Social Survey (ESS) Multilevel Application:
 - Pan-European social survey
 - Integrates 30+ “context” variables (Eurostat, OECD, IMF, etc.)
 - Now integrating environment data using DDI-CDI
- US Bureau of Labor Statistics: Indicators
- UK Smart Energy Research Laboratory (SERL)
 - Combines surveys, energy meter data, climate data
- Interstat project: DDI-CDI as intake model for open government data to support NGSI-LD domain models
- Others

<https://codata.org/the-role-of-ddi-cdi-in-eosc-possible-uses-and-applications/>

Current Status/Looking Forward

- Release anticipated by end of summer 2022
 - Will include XML reference syntax
 - RDF syntax soon to follow
 - Enhanced documentation/implementation guidance to follow
- Will form key part of the “Cross-Domain Interoperability Framework” (CDIF) being developed through the EU-funded WorldFAIR project (<https://codata.org/initiatives/decadal-programme2/worldfair/>)

Contact

Arofan Gregory, Chair, DDI-CDI WG

ilg21@yahoo.com

CODATA (Simon Hodson, Executive Director)

simon@codata.org