Data Integration: Using DDI-CDI with Other Standards

DDI Training Group/DDI-CDI Working Group

DDI Alliance/CODATA

2 September 2021

DATA DOCUMENTATION INITIATIVE

Outline

- Introduction
- History and Background
- Scope and Design
- Foundational Metadata
- Data Structures
- Process
- Helmholtz Prototype Example
- Summary

Introduction

DATA DOCUMENTATION INITIATIVE

DDI-CDI and FAIR

- Many people talk about ***F***indability and ***A***ccess
  - Not so much about ***I***nteroperability and ***R***euse
- DDI-CDI focuses on these aspects of FAIR data
  - It is also quite useful for data discovery
- Interoperability and reuse of data are metadata-intensive
- Historically, these aspects of data management are expensive and have not been fully incentivized by research funders
- Today's focus on FAIR data *demands* that we do more!

Why DDI-CDI?

- DDI-CDI is designed to meet these challenges
- Standard “lingua franca” for describing data of many types
- DDI-CDI is designed to fill the gaps among existing standards and models
  - Alignment and integration
  - Complements existing metadata specifications
- Provides detailed metadata about the data and processes by which it is reused
- Supports an exact understanding of what is required for data reuse
  - Automate the structural transformations
  - Connects structural elements with concepts/vocabularies
  - Support semantic crosswalks between domain ontologies

DATA DOCUMENTATION INITIATIVE

DDI – CDI as a New Type of Work Product

- Earlier DDI standards - DDI Codebook and DDI Lifecycle - are metadata specifications for the Social, Behavioral, and Economic (SBE) Sciences
  - They are generic enough to be used in similar domains (official statistics, public health)
  - They still use terms and models familiar to SBE sciences
- DDI – CDI is different: it is intended to be used across a wider range of domains
  - Different types of data and models
  - More abstract/general terminology
- DDI – CDI is a new type of specification, meant to be used *with* many other standards, in SBE and outside it

DATA DOCUMENTATION INITIATIVE

History and Background

DATA DOCUMENTATION INITIATIVE

History: MRT and DDI Developments

- DDI was working on a “model-driven” version of the standard for many years – “DDI 4” or “DDI Moving Forward”
  - This work is the basis of DDI-CDI
  - This work has also informed developments in DDI Lifecycle
- In the margins of the 2018 European DDI User Conference (Berlin) it was agreed that a “core” of the next-generation/model based DDI work should be brought to market
- A 1-year timeframe was proposed but as usual it took basically forever – thanks COVID! (Qualify w/ first public review version)
- The Modelling, Representation and Testing (MRT) group was formed in early 2019
- The working process was to base models on implementations, tested against real-world use cases
  - ALPHA Network
  - DDI R Libraries (references: 1, 2)
  - Others (U.S. Bureau of Labor Statistics for time series, etc.)

DATA DOCUMENTATION INITIATIVE

Group and Events

- Confluence Sprints Page lists the events in the development of DDI 4 (and CDI)
- MRT: small group (9 members) meeting weekly (and more) for over a year
- No turn-over – members have been extremely focused and disciplined
- Ottawa Sprint in margins of NADDI 2019
- Dagstuhl Sprint in October 2019
- Public Review Release April 2020
- MRT has done 7 webinars to reach out to users in different areas (250+ people)
  - In collaboration with CODATA
  - More are planned
- Communications with management, technical committee work, marketing, and training groups within the DDI Alliance have been emphasized

DATA DOCUMENTATION INITIATIVE

MRT Members

Back row, from left:

Joachim Wackerow Dan

Gillman

Larry Hoyle

Arofan Gregory

Jay Greenfield

Front row, from left:

Hilde Orten

Flavio Rizzolo

Not in picture:

Oliver Hopt

Wendy Thomas

https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/707624961/Dagstuhl+Sprint+October+2019

DATA DOCUMENTATION INITIATIVE

Evolution in Purpose

- DDI-CDI was expected to be the “core” of a model-driven DDI
  - A “next generation” after DDI-Lifecycle
- Implementation cases showed that something else was needed: a focus on data provenance and data integration
- DDI-CDI has emerged as a *companion* to DDI-Codebook and DDI-Lifecycle, not a replacement for them
- The Social, Behavioral, and Economic (SBE) community needs better data integration tools
  - So do other domains!

Scope and Design

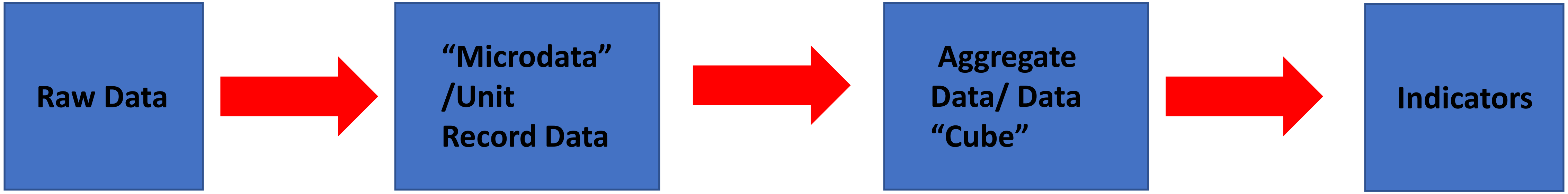
DDI-CDI within the Realm of Metadata Standards

- DDI-CDI does not replace existing domain metadata standards
  - DDI-Codebook or DDI-Lifecycle for SBE sciences
  - Other fine-grained standards according to domain use (EML for ecology, OMOP CDM for clinical records, etc.)
  - It functions as a *complement*
- It adds support for understanding diverse types of data across domain boundaries
- It expands the ability to describe process (provenance)
- It provides a detailed description of integration between disparate types of data and the concepts behind them

The Model is the Main Thing!

- A formal UML class model
- Based on a subset of UML features
- Expressed as Canonical XMI
  - XMI is an XML language for describing/exchanging UML models
  - Canonical XMI is well-supported by a broad range of UML tools
- Representations in specific tools/syntaxes can be generated automatically from the UML (even by users)
- An XML representation is provided, others may be in future
- UML "future-proofs" the standards
  - Against changes in technology tools
  - By being extensible

DDI – CDI Scope (Example)



- DDI – CDI describes the data at each stage, indicating the roles played by each atomic bit of data (“datum”)
- This includes describing classifications, variables, concepts, etc. (“Foundational Metadata”)
- New types of data can be described
- DDI – CDI tracks the processing between each stage (aligns with PROV), reflecting the relationships between atomic datums (uses other standards for describing specific processes - SDTL)

Foundational Metadata

DATA DOCUMENTATION INITIATIVE

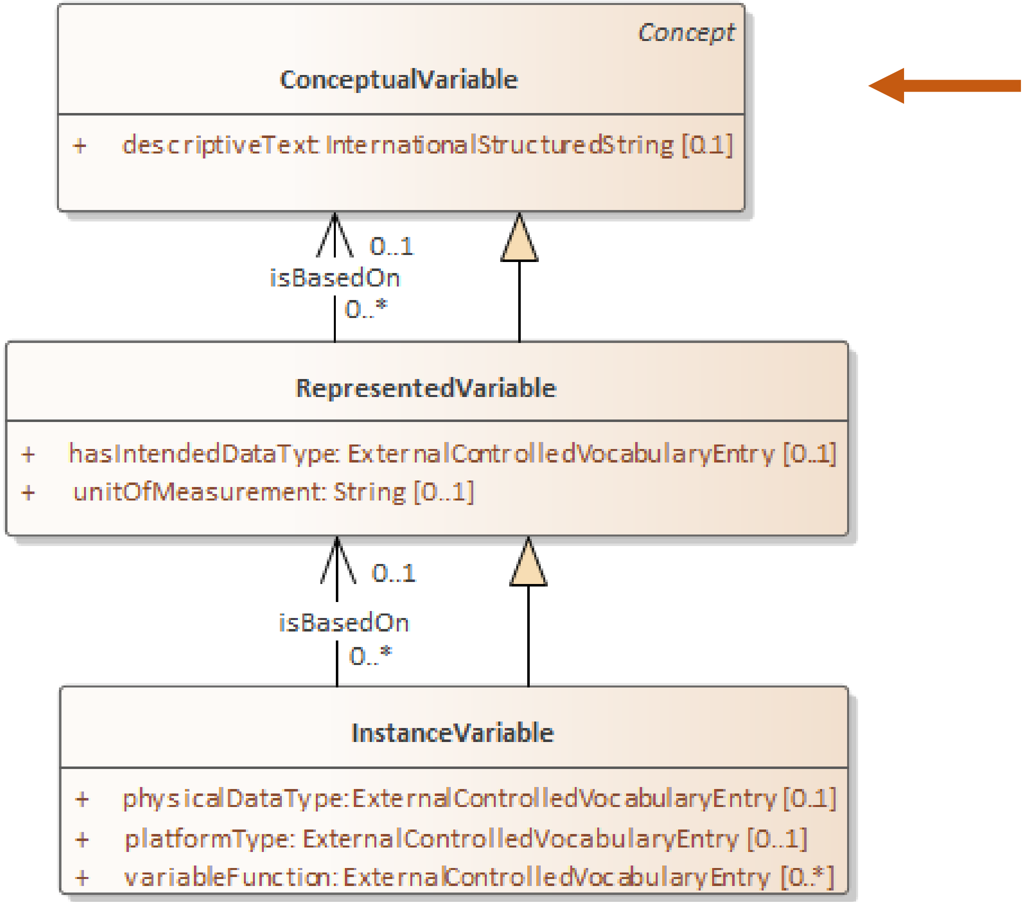
Foundational Metadata (from DDI 4)

- Concepts and Concept Systems
- Variables (of many types!)
- Classifications, Codelists, Categories (etc.)
  - Includes classification management
- Populations, Units, and Universes

Foundational Metadata: The Variable Cascade

- Understanding the roles played by variables is critical in integration of data
- Variables do ***many, many*** different things
- Not all variables are the same!
- We have three levels of variables in our model:
  - Conceptual Variables
  - Represented Variables
  - Instance Variables

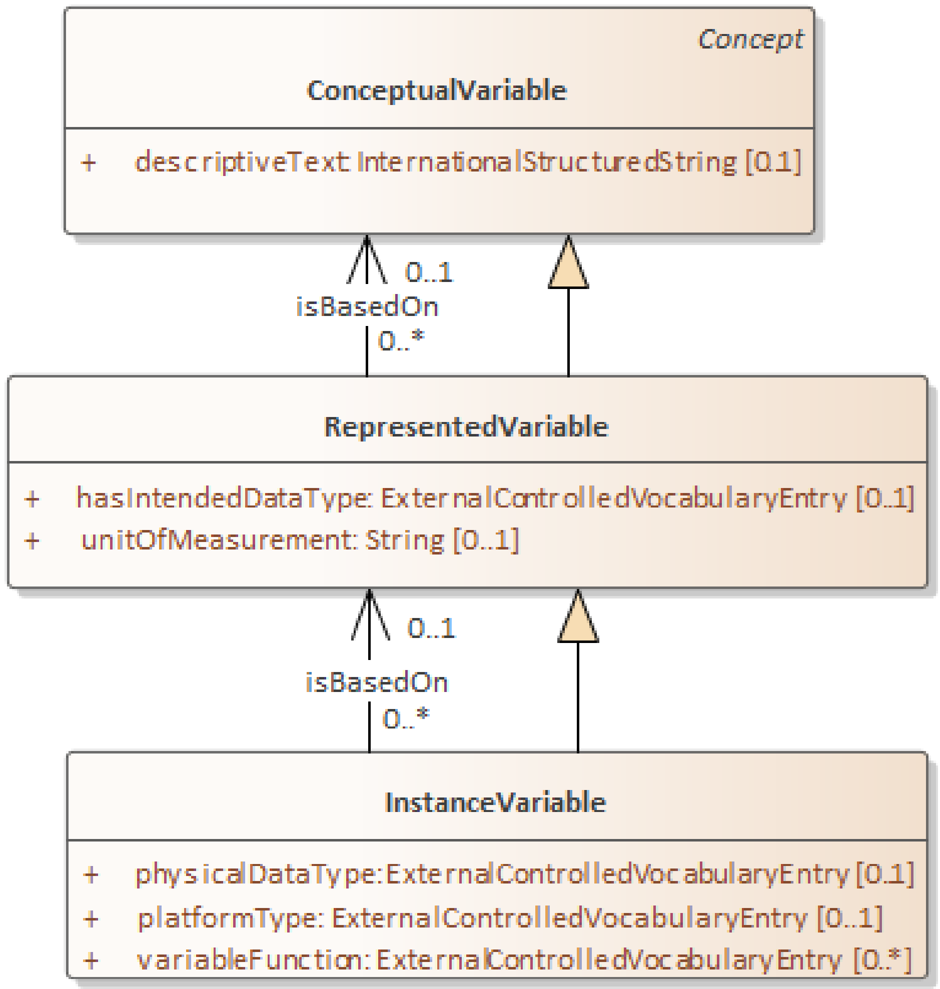
Variable Cascade – Conceptual Variable



Variable descriptions at a high level. Early in designing data collection, broad searches. Broadly reusable.

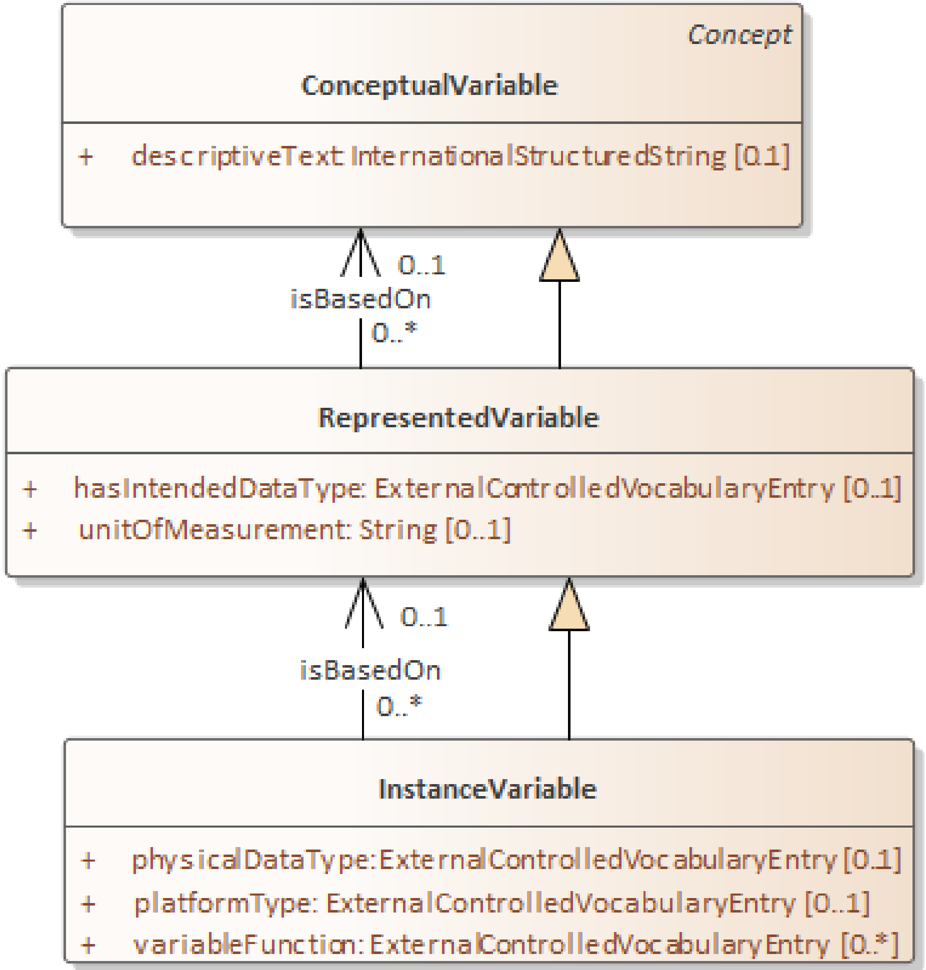
DATA DOCUMENTATION INITIATIVE

Variable Cascade - RepresentedVariable



More specificity about value domain, units of measurement. Still reusable.

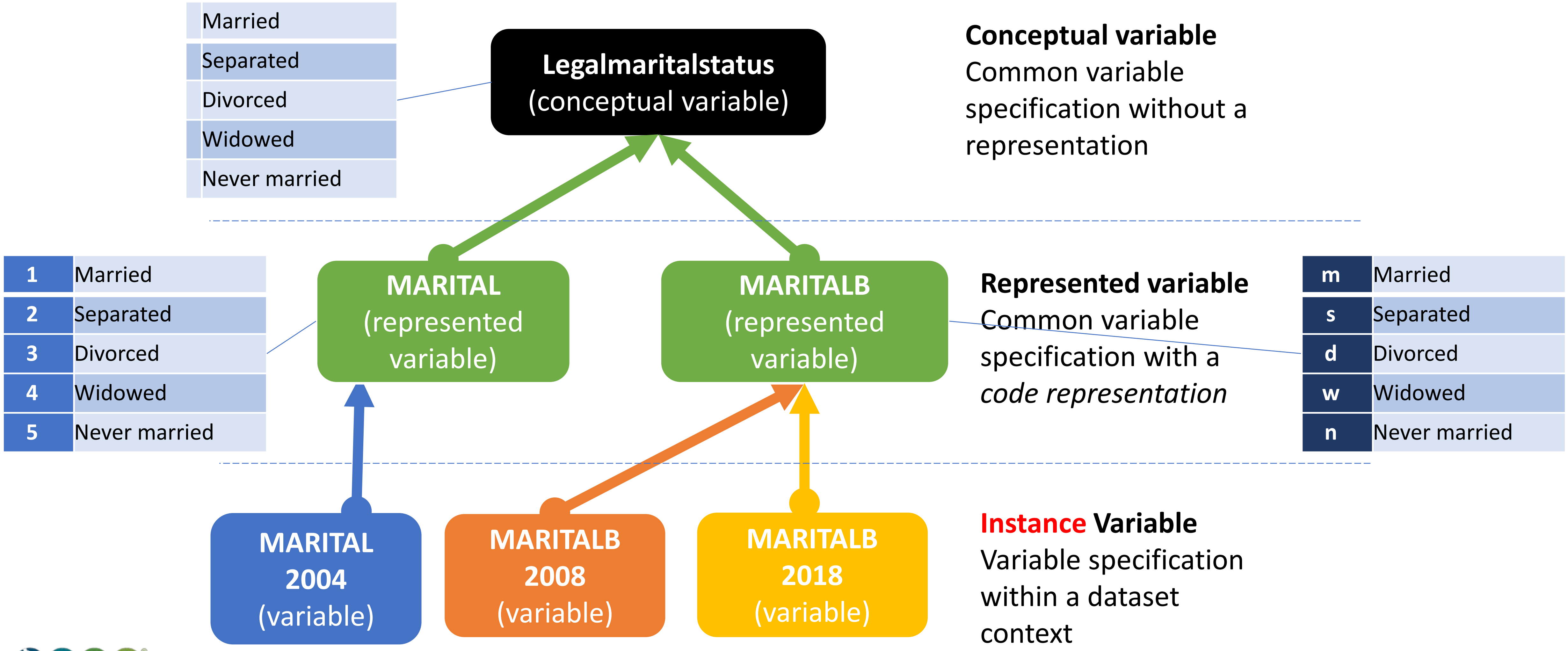
Variable Cascade - InstanceVariable



Describing **collected** data. Physical datatype and platform. Invariant role of the variable (e.g. a weight)

DATA DOCUMENTATION INITIATIVE

Example: Comparability among Variables



DATA DOCUMENTATION INITIATIVE

Application: Recognizing Similar variables in Difficult Cases

- Two variables in different data sets might:
  - Measure the same concept differently
  - Measure the same concept in the same way with different physical representations
  - Exist identically in two data sets, but with no formal link
- In all of these cases, understanding the variables at each level (conceptual, representational, and actual) provides a strong basis for programmatically identifying them as potential points for joining data sets

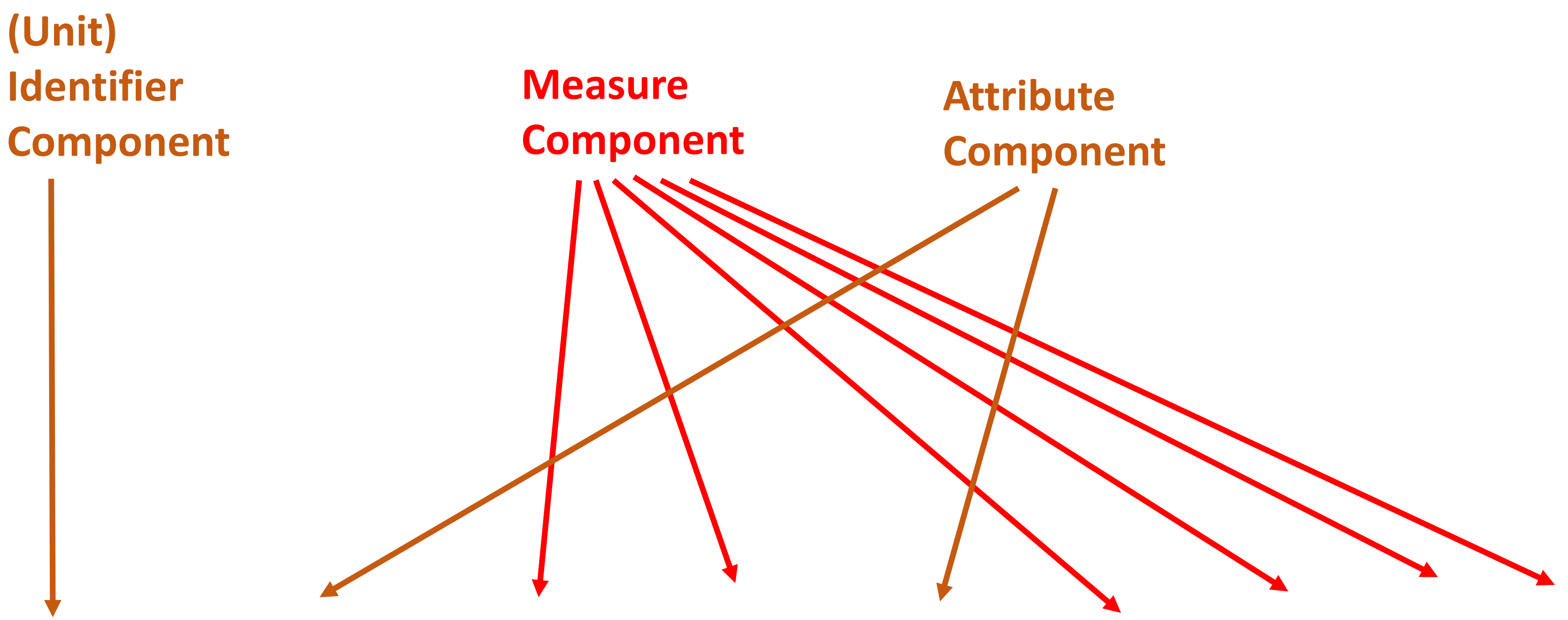
Data Structures

Data Structures

- DDI-CDI currently can describe four different data structures
    - Wide – as with unit records
    - Long – as with event or stream data
    - Key value – as in a key-value store
    - Dimensional – as with aggregate data

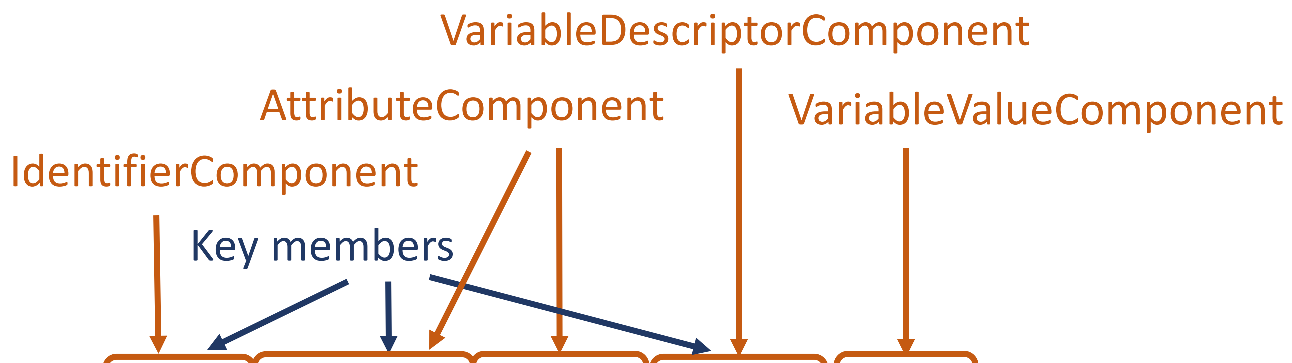
We'll now show some examples of data and their representations in the different structure types.

Roles: The COVID example- Wide structure



entry	datetime	systolic	diastolic	position	weight	temp	pctO2	pulse
101	2020-07-14T13:54	114	70	2	83914.6	36.44	98	70
132	2020-07-14T14:03	125	86	3	68038.9	37.5	85	92

Roles: The COVID example- Long structure

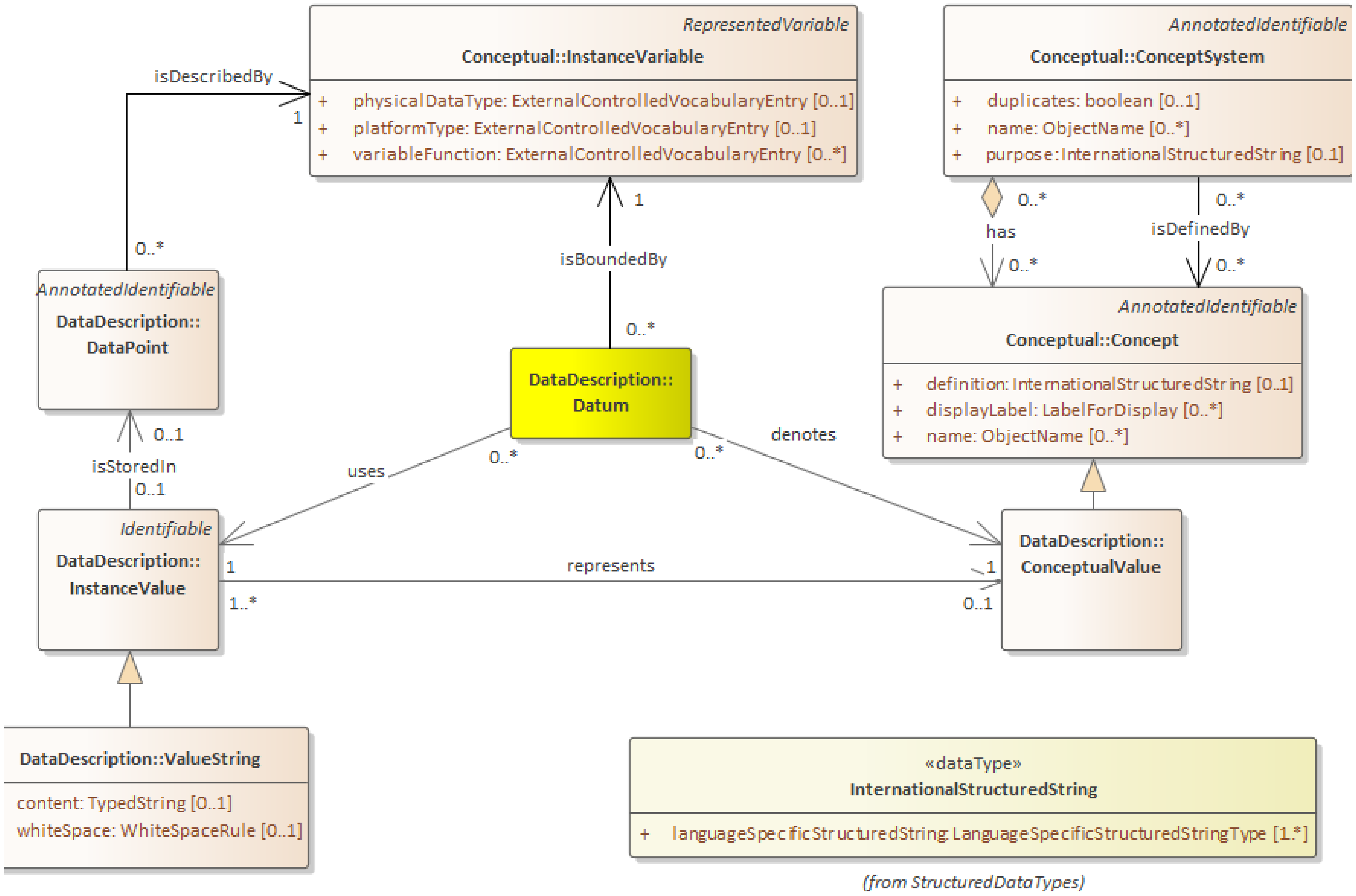


Entry	DateTime	Position	Measure	Value
101	2020-07-14T13:54	2	systolic	114
101	2020-07-14T13:54	2	diastolic	70
101	2020-07-14T13:54	2	weight	83914.60
101	2020-07-14T13:54	2	temp	36.44
101	2020-07-14T13:54	2	pctO2	98
101	2020-07-14T13:54	2	pulse	70
101	2020-07-14T13:54	2	away	n
101	2020-07-14T13:54	2	exposed	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.

The Datum Approach



Entry	DateTime	Position	Measure	Value
101	2020-07-14T13:54	2	systolic	114

Cross Domain – Content and Structure

- Integrating data across domains involves both dealing with different kinds of discipline’s structures and vocabularies
  - Sensor data streams in tall structures
  - Survey data in wide structures
  - Administrative summary data in cubes
- A standard also needs to be discipline agnostic.
  - Vocabularies need to be referenced, not built in
    - (e.g. “question”)
- A standard needs to be able to at least reference metadata in other disciplines standards.
  - This, of course, presents challenges for machine actionability.

Dimensional Example

Aggregate fisheries statistics, Year 2019 – Norwegian boats – Round weight in tons

	Fishing equipment, main group			
Species, main group	Purse seine	Conventional	Trawl	Other
Pelagic fish	587465	1371	712873	359
Cod and cod species	33554	397254	237496	11
etc.				

Multi-dimensional "keys" are identifiers formed from the set of dimension values (variables in long table).

Long table (micro-data set)

Boat name	**Nationality of boat**	**Landing date**	**Spieces**	**Round weight in ton**	**Fishing equipment**
Lønningen	NO	07.04.2019	Blue whiting	2917,75	Trawl
Ligrunn	NO	29.08.2019	Herring	23034,41	Trawl
Ligrunn	NO	29.08.2019	Pollock	307,1	Trawl
Ligrunn	NO	25.09.2019	Mackerel	1621,68	Seine net
Vikingbank	NO	05.11.2019	Norwegian spring breeding herring	2932,16	Net equipment

DATA DOCUMENTATION INITIATIVE

Key-Value example

Semantic Sensor Network Ontology (SSN) example

Wide

SensorIDz	Property	Time	ResultingValue
sensor/35-207306-844818-0/BMP282	atmosphericPressurehPa	2017-06-06T12:36:12Z	1021.45

Key-Value

Key	Value
sensor/35-207306-844818-0/BMP28/atmosphericPressurehPa/2017-06-06T12:36:12Z	1021.45

Application: Automating Data Integration

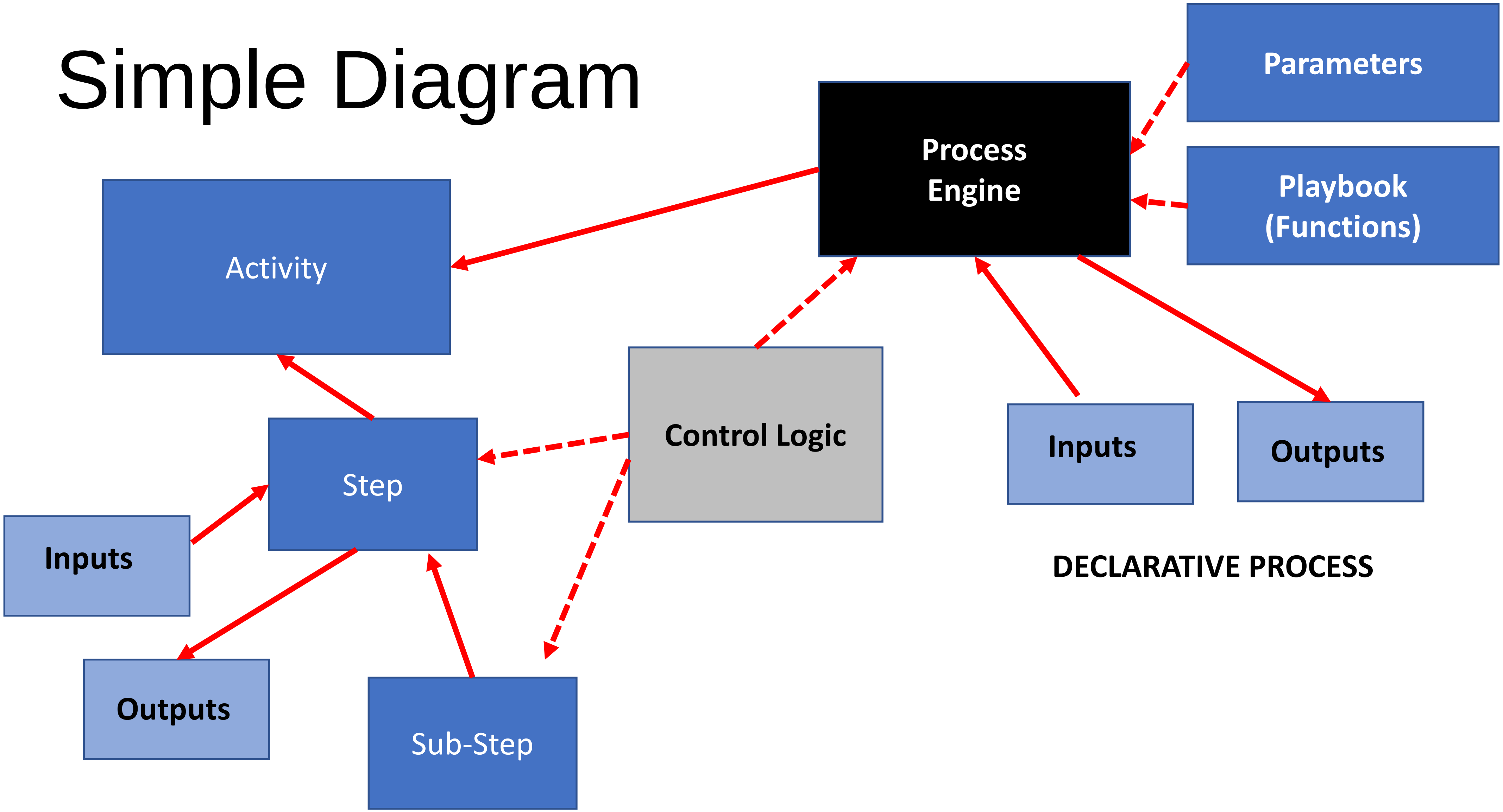
- If I understand the role played by any given data point in its data set of origin, I can predict what role it must play in the data set I need to transform it into for integration purposes
- The DDI-CDI model shows us how these relate, and can avoid manual intervention in performing the needed structural transformations
- Reduces the (up to 80%) resource burden on projects for preparing data for analysis

Process

The DDI-CDI Process Model

- Describes the use of individual processes, and how they fit together
- Supports standard descriptions (SDTL, VTL) and specific languages (SQL, R, STATA, SPSS, Python, SAS, etc.)
- Three “modes”:
    - Procedural: Step-wise, with decision points
    - Declarative: “Black box” multi-threaded, uses a “playbook” and configurations
    - Hybrid approaches of the two

Simple Diagram



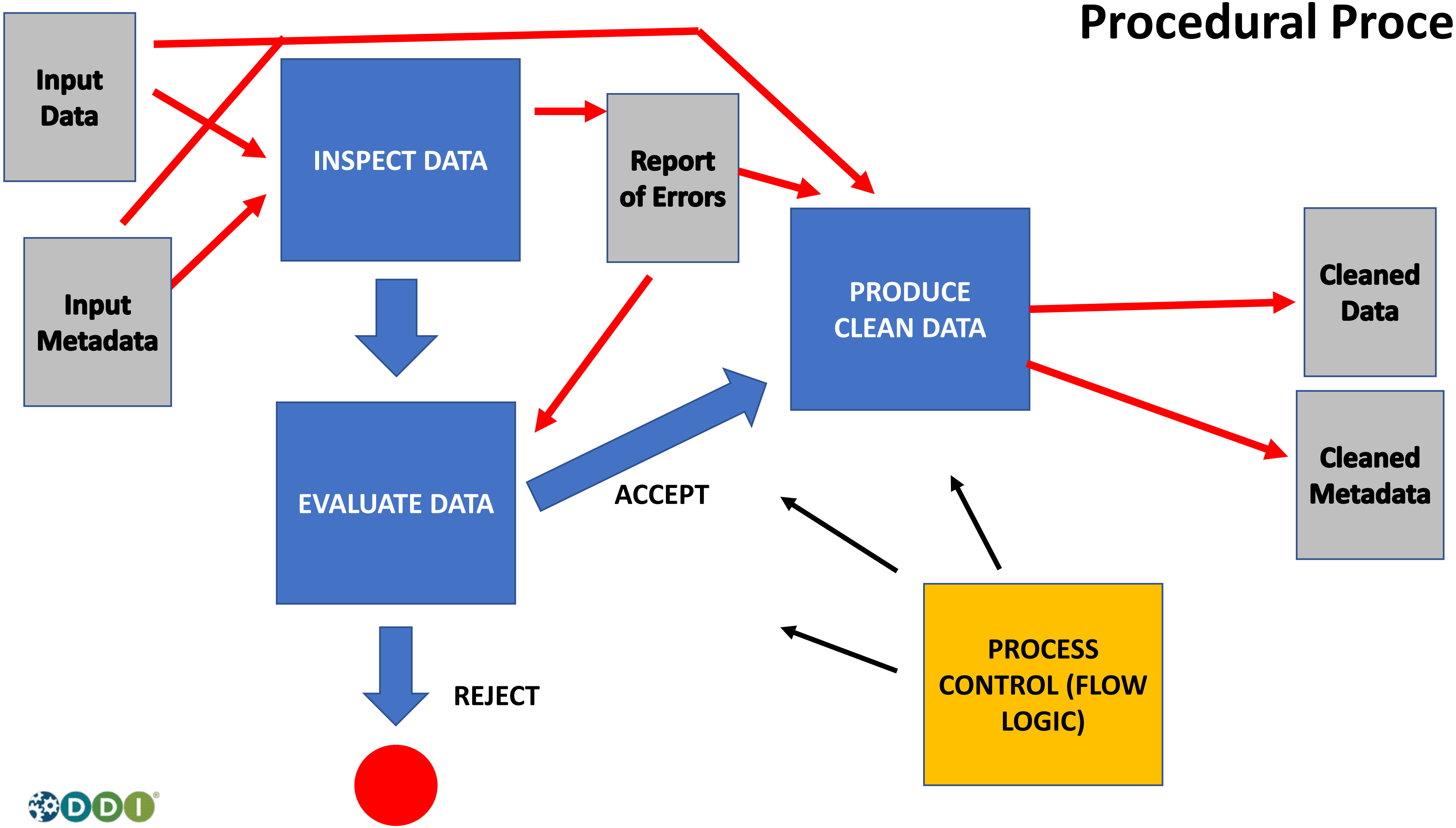
DECLARATIVE PROCESS

PROCEDURAL PROCESS

Process Example: Data Cleaning Activity

- Step 1: Inspect data and generate a report with:
  - Impossible cases (e.g., 3-year-olds giving birth; people marrying themselves)
  - Improbable cases (e.g., 12-year-olds with an earned income of 250,000 EUR per annum)
- Step 2: Evaluate quality
  - Accept: Data has low percentage of impossible/improbable cases - continue
  - Reject: Data has high percentage of impossible/improbable cases – exit process with errors
- Step 3: Produce cleaned data set
  - Trim out impossible/improbable cases

Procedural Process



The Helmholtz Prototype Example

The Helmholtz Prototype Example

- https://www.helmholtz.de/en/
- The Helmholtz Institute is the largest German research institution
  - They have a number of projects looking at “grand challenges”
  - Climate change is one of the areas of focus
- The Helmholtz Metadata Collaboration (https://www.hmc-plattform.org/en) has started a FAIR data-sharing project among the involved institutions, some dealing with oceanographic data
- This example comes from an on-going exploration of how they will be moving into the second phase of the project
  - First phase focused on basic metadata and discovery
  - Second phase will look at detailed descriptions of the data

Phase I: Basic Metadata

- In Phase I, Schema.org is being used to support Discovery and other basic data-sharing functions (e.g., access)
- In FAIR terms, this is the “F” and the “A”
- Schema.org is implemented using JSON-LD
  - Schema.org is a set of metadata elements for data discovery supported by many search engines (Google’s Dataset Search is important here)
  - JSON is a generic way of describing information using nested Javascript arrays
- This approach has been used in other, similar projects (e.g., UN Decade of Ocean Science - https://www.oceandecade.org/)

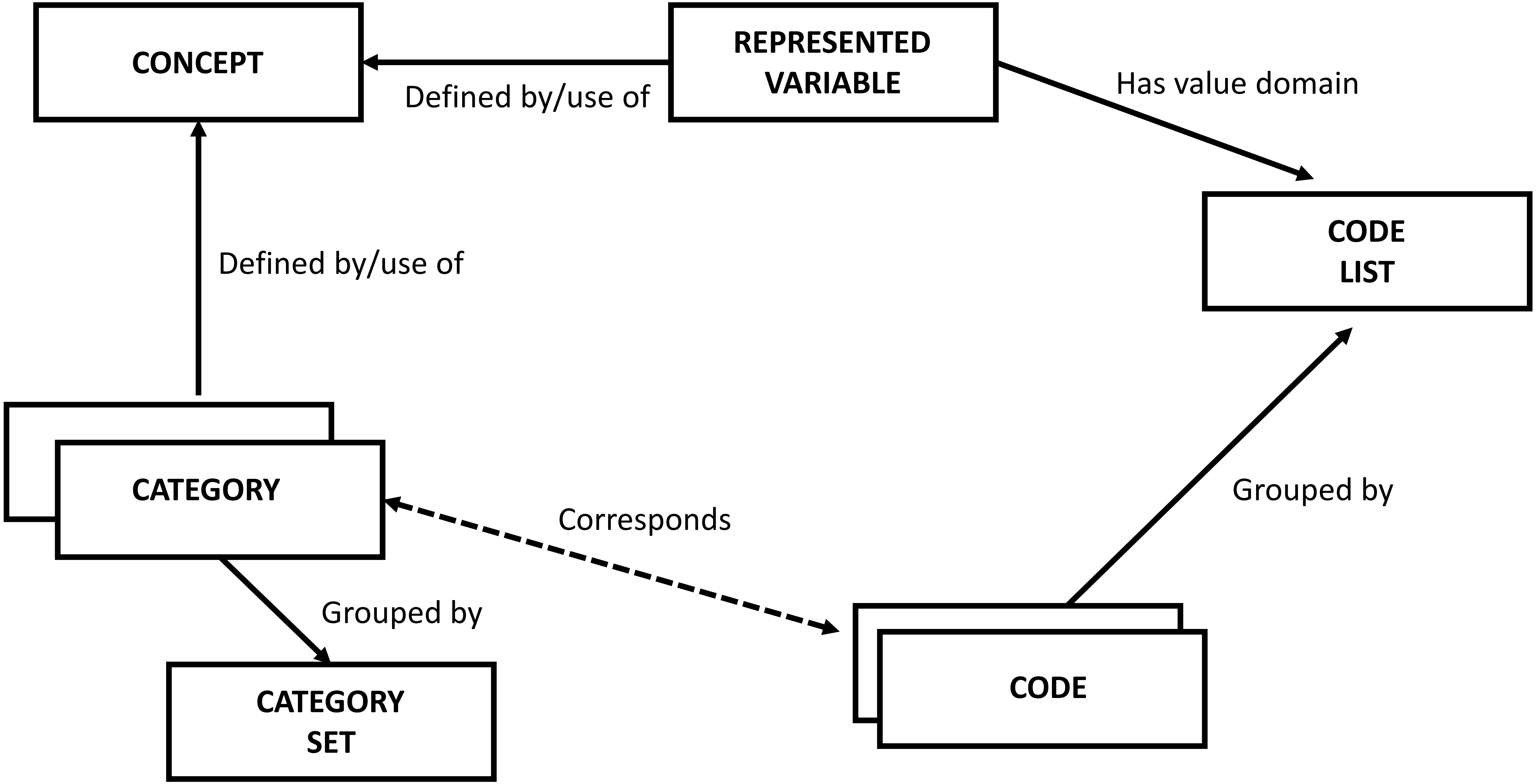
DATA DOCUMENTATION INITIATIVE

Phase 2 (Exploration): Using DDI-CDI

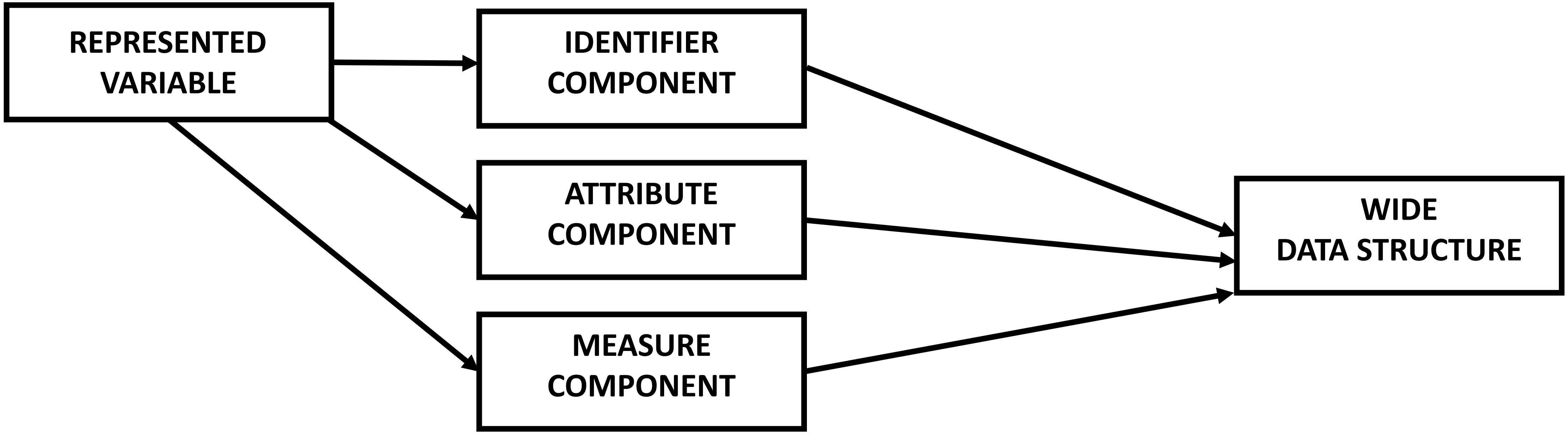
- First step is to analyze the data and select the correct classes from DDI-CDI for describing it
- Second step is to identify how the selected classes and relationships will be expressed in a syntax for the data itself
- This would be documented in an implementation guide so it could be shared, making the data accessible to systems which want to use it

Phase 2 Technologies and Standards

- The model is being subsetted using Protégé, a popular tool among ontologists
- The subset will be expressed as OWL
- Concepts are likely to be in SKOS
  - DDI-CDI supports the use of SKOS concepts
- The RDF classes/relationships described in the OWL will be implemented using JSON-LD
- In this example, we see the DDI-CDI model, OWL, SKOS, and JSON-LD as all playing a role in the overall approach.
  - Other domain-specific standards/ontologies could supplement this basic model



DATA DOCUMENTATION INITIATIVE



Notes:

The identifiers plus the Measure or Attribute identify the location of cells (Data Points) in the data. Data Points hold Datums.

A Helmholtz Data Example

Not logged in

PANGAEA.

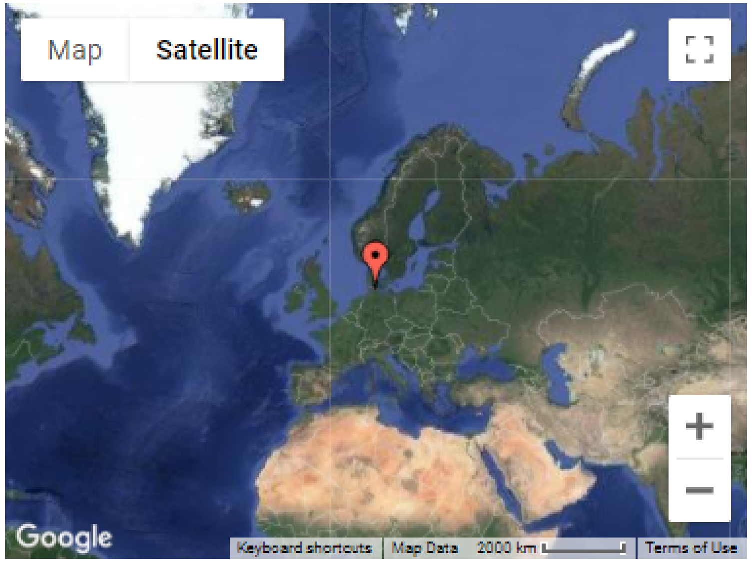
Data Publisher for Earth & Environmental Science

SEARCH SUBMIT HELP ABOUT CONTACT

Citation: **Bange, Hermann Werner; Malien, Frank (2015):** Hydrochemistry from time series station Boknis Eck from 1957 to 2014. *PANGAEA*, https://doi.org/10.1594/PANGAEA.855693

Always quote citation above when using data! You can download the citation in several formats below.

RIS Citation | BibTeX Citation | Copy Citation | Facebook | Twitter | Show Map | Google Earth



Related to: **Lennartz, Sinikka T; Lehmann, Andreas; Herrford, Josefine; Malien, Frank; Hansen, Hans Peter; Biester, Harald; Bange, Hermann Werner (2014):** Long-term trends at the Boknis Eck time series station (Baltic Sea), 1957–2013: does climate change counteract the decline in eutrophication? *Biogeosciences*, **11(22)**, 6323-6339, https://doi.org/10.5194/bg-11-6323-2014

Source data set: **Bange, Hermann Werner; Malien, Frank (2014):** Boknis Eck Timeseries Database. http://www.bokniseck.de/

Coverage: *Latitude:* 54.529500 * *Longitude:* 10.039330

Date/Time Start: 1957-04-30T00:00:00 * *Date/Time End:* 2014-12-16T11:06:33

Minimum DEPTH, water: 1 m * *Maximum DEPTH, water:* 35 m

Schema.org describes this type of metadata (Phase I)

“Parameters” (DDI-CDI “variables”)

Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method/Device	Comment
1	DATE/TIME	Date/Time		Bange, Hermann Werner		Geocode
2	Latitude of event	Latitude				
3	Longitude of event	Longitude				
4	DEPTH, water	Depth water	m	Bange, Hermann Werner		Geocode
5	Cast number	Cast		Bange, Hermann Werner		
6	Sample code/label	Sample label		Bange, Hermann Werner		
7	Chlorophyll a	Chl a	µg/l	Bange, Hermann Werner		
8	Nitrate	[NO3]-	µmol/l	Bange, Hermann Werner		
9	Flag	Flag		Bange, Hermann Werner		NO3
10	Nitrite	[NO2]-	µmol/l	Bange, Hermann Werner		
11	Flag	Flag		Bange, Hermann Werner		NO2
12	Oxygen	O2	µmol/kg	Bange, Hermann Werner		
13	Flag	Flag		Bange, Hermann Werner		Oxygen
14	Phosphate	[PO4]3-	µmol/l	Bange, Hermann Werner		
15	Flag	Flag		Bange, Hermann Werner		PO4
16	Salinity	Sal		Bange, Hermann Werner		
17	Silicon dioxide	SiO2	µmol/l	Bange, Hermann Werner		
18	Flag	Flag		Bange, Hermann Werner		SiO2
19	Temperature, water	Temp	°C	Bange, Hermann Werner		

Could also be expressed in Schema.org

The Data

1 Date/Time	2 Latitude	3 Longitude	4 Depth water [m]	5 Cast	6 Sample label	7 Chl a [µg/l]	8 [NO3]- [µmol/l]	9 Flag (NO3)	10 [NO2]- [µmol/l]	11 Flag (NO2)	12 O2 [µmol/kg]	13 Flag (Oxygen)	14 [PO4]3- [µmol/l]	15 Flag (PO4)	16 Sal	17 SiO2 [µmol/l]	18 Flag (SiO2)	19 Temp [°C]
1957-04-30T00:00:00	54.5295	10.0393	1	1	1						321.9		0.000		15.30			7.70
1957-04-30T00:00:00	54.5295	10.0393	5	1	1						325.0		0.010		15.30			5.40
1957-04-30T00:00:00	54.5295	10.0393	10	1	1						325.0		0.020		15.70			6.10
1957-04-30T00:00:00	54.5295	10.0393	15	1	1						318.8		0.030		16.40			4.50
1957-04-30T00:00:00	54.5295	10.0393	20	1	1						300.0		0.060		17.00			4.30
1957-04-30T00:00:00	54.5295	10.0393	26	1	1						281.3		0.240		17.40			4.30
1957-05-14T00:00:00	54.5295	10.0393	1	1	2								0.020		15.40			8.70
1957-05-14T00:00:00	54.5295	10.0393	5	1	2								0.070		15.40			8.70
1957-05-14T00:00:00	54.5295	10.0393	10	1	2								0.010		15.80			7.20
1957-05-14T00:00:00	54.5295	10.0393	15	1	2								0.150		17.00			6.40
1957-05-14T00:00:00	54.5295	10.0393	20	1	2								0.170		17.40			6.30
1957-05-14T00:00:00	54.5295	10.0393	26	1	2								0.410		18.40			6.50
1957-06-04T00:00:00	54.5295	10.0393	1	1	3								0.040		14.70			13.40
1957-06-04T00:00:00	54.5295	10.0393	5	1	3								0.010		14.70			13.20
1957-06-04T00:00:00	54.5295	10.0393	10	1	3								0.120		15.00			10.20
1957-06-04T00:00:00	54.5295	10.0393	15	1	3								0.190		15.60			8.80
1957-06-04T00:00:00	54.5295	10.0393	20	1	3								0.460		17.60			6.70
1957-06-04T00:00:00	54.5295	10.0393	26	1	3								0.750		18.40			6.40
1957-06-14T00:00:00	54.5295	10.0393	1	1	4						437.5		0.010		14.80			13.90

Described in DDI-CDI

DDI-CDI Identifier Components

DDI-CDI Measure Components

1 Date/Time	2 Latitude	3 Longitude	4 Depth water [m]	5 Cast	6 Sample label	7 Chl a [µg/l]	8 [NO3]- [µmol/l]	9 Flag (NO3)	10 [NO2]- [µmol/l]	11 Flag (NO2)	12 O2 [µmol/kg]	13 Flag (Oxygen)	14 [PO4]3- [µmol/l]	15 Flag (PO4)	16 Sal	17 SiO2 [µmol/l]	18 Flag (SiO2)	19 Temp [°C]
1957-04-30T00:00:00	54.5295	10.0393	1	1	1						321.9		0.000		15.30			7.70
1957-04-30T00:00:00	54.5295	10.0393	5	1	1						325.0		0.010		15.30			5.40
1957-04-30T00:00:00	54.5295	10.0393	10	1	1						325.0		0.020		15.70			6.10
1957-04-30T00:00:00	54.5295	10.0393	15	1	1						318.8		0.030		16.40			4.50
1957-04-30T00:00:00	54.5295	10.0393	20	1	1						300.0		0.060		17.00			4.30
1957-04-30T00:00:00	54.5295	10.0393	26	1	1						281.3		0.240		17.40			4.30
1957-05-14T00:00:00	54.5295	10.0393	1	1	2								0.020		15.40			8.70
1957-05-14T00:00:00	54.5295	10.0393	5	1	2								0.070		15.40			8.70
1957-05-14T00:00:00	54.5295	10.0393	10	1	2								0.010		15.80			7.20
1957-05-14T00:00:00	54.5295	10.0393	15	1	2								0.150		17.00			6.40
1957-05-14T00:00:00	54.5295	10.0393	20	1	2								0.170		17.40			6.30
1957-05-14T00:00:00	54.5295	10.0393	26	1	2								0.410		18.40			6.50
1957-06-04T00:00:00	54.5295	10.0393	1	1	3								0.040		14.70			13.40
1957-06-04T00:00:00	54.5295	10.0393	5	1	3								0.010		14.70			13.20
1957-06-04T00:00:00	54.5295	10.0393	10	1	3								0.120		15.00			10.20
1957-06-04T00:00:00	54.5295	10.0393	15	1	3								0.190		15.60			8.80
1957-06-04T00:00:00	54.5295	10.0393	20	1	3								0.460		17.60			6.70
1957-06-04T00:00:00	54.5295	10.0393	26	1	3								0.750		18.40			6.40
1957-06-14T00:00:00	54.5295	10.0393	1	1	4						437.5		0.010		14.80			13.90

DDI-CDI Data Point

DDI DATA DOCUMENTATION INITIATIVE

Summary

DDI-CDI Promotes FAIR Data Sharing

- By supporting the interoperability and reuse of data across domains
  - Differences in structure
  - Differences in process
- By providing a concept-rich, detailed descriptions of data
  - Supports mapping of domain ontologies/semantics
- Designed to provide machine-actionable metadata for leveraging sophisticated data integration processing

DATA DOCUMENTATION INITIATIVE

Looking toward the Future

- DDI-CDI is emerging as an important component in what is termed “AI-Ready Data”
- If automation is the answer to challenges of scale in data reuse within and across domains, we need better metadata for the machines to act on
- A rich model like DDI-CDI helps to meet that need, by adding intelligence which is described in a common way across all forms of data
- Ultimately, this approach seems like a realistic way to make FAIR data-sharing a reality

Where Are We?

- Engagement with external domains and reviewers has been high
  - Webinars on specific topics/domains
  - EOSC Project to Recommend Applications of DDI-CDI
  - International FAIR Convergence Symposium/GO FAIR/CODATA
- Next Steps
  - Focus on examples (UOM, Process)
  - Look at external data cases (e.g, NetCDF, graphs)
  - Examples with DCAT, Schema.org, SKOS, etc.
  - Integration with FAIR FIPs/FDOs
  - Methodology for creating user community implementation guides
- Release of final specification in coming months
  - Core specification in first release
  - Supporting examples and implementation guidance to follow

Credits: DDI Training Working Group

Florio Orocio Arguillas
Alina Danciu
Adrian Dusa
Jane Fry
Martine Gagnon
Dan Gillman
Arofan Gregory
Taras Günther
Lea Sztuk Haahr
Simon Hodson
Chifundo Kanjala
Kaia Kulla

Kathryn Lavender
Amber Leahey
Marta Limmert
Jared Lyle
Alexandre Mairot
Lucie Marie
Hayley Mills
Laura Molloy
Hilde Orten
Anja Perry
Flavio Rizzolo
Knut Wenzig