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
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

- **Conceptual variable**: Common variable specification without a representation
- **Represented variable**: Common variable specification with a code representation
- **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

<table>
<thead>
<tr>
<th>entry</th>
<th>datetime</th>
<th>systolic</th>
<th>diastolic</th>
<th>position</th>
<th>weight</th>
<th>temp</th>
<th>pctO2</th>
<th>pulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>114</td>
<td>70</td>
<td>2</td>
<td>839</td>
<td>36.44</td>
<td>98</td>
<td>70</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>125</td>
<td>86</td>
<td>3</td>
<td>680</td>
<td>37.5</td>
<td>85</td>
<td>92</td>
</tr>
</tbody>
</table>
Example 1: data in long form

<table>
<thead>
<tr>
<th>Identifier举行了</th>
<th>DateTime</th>
<th>Position</th>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>systolic</td>
<td>114</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>diastolic</td>
<td>70</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>weight</td>
<td>83914.60</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>temp</td>
<td>36.44</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>pctO2</td>
<td>98</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>pulse</td>
<td>70</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>away</td>
<td>n</td>
</tr>
<tr>
<td>101</td>
<td>2020-07-14T13:54</td>
<td>2</td>
<td>exposed</td>
<td>n</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>systolic</td>
<td>125</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>diastolic</td>
<td>86</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>weight</td>
<td>68038.90</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>temp</td>
<td>37.5</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>pctO2</td>
<td>85</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>pulse</td>
<td>92</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>away</td>
<td>y</td>
</tr>
<tr>
<td>132</td>
<td>2020-07-14T14:03</td>
<td>3</td>
<td>exposed</td>
<td>n</td>
</tr>
</tbody>
</table>

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


• 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

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