# DDI Cross-Domain Integration (DDI-CDI): Complementing the DDI Product Suite

Arofan Gregory

Chair, DDI-CDI Working Group

3 October 2024

# Outline

1. What DDI-CDI is intended to do
2. History and emerging requirements for data sharing/integration
3. What DDI-CDI is and is not
4. Relationship to DDI Codebook, DDI Lifecycle, DDI CVs, SDTL, and XKOS
5. Relationship to other standards
6. Early implementation, tools, and coming projects and documentation
7. Implementation example: UKDA Product Builder
8. Implementation examples: ESS Labs and WorldFAIR SPSS/Stata Converter
9. Summary

# What DDI-CDI Is Intended to Do

- DDI Cross-Domain Integration extends the metadata coverage of DDI specifications to describe data coming from outside the social, behavioral, and economic (SBE) domain
- It allows for the description of data sets to make them “integration ready” in a multi-disciplinary scenario
  - Focuses on variable and datum-level descriptions
  - Provides for lossless, automated re-structuring of data
- It supports the description of the integration (and any non-SBE lifecycle) process flows
- It is not a standard for data collection or management: the focus is on sharing and integration of data
- DDI-CDI is designed to work with non-SBE domain standards, and with generic Web standards and technologies, notably RDF

# History

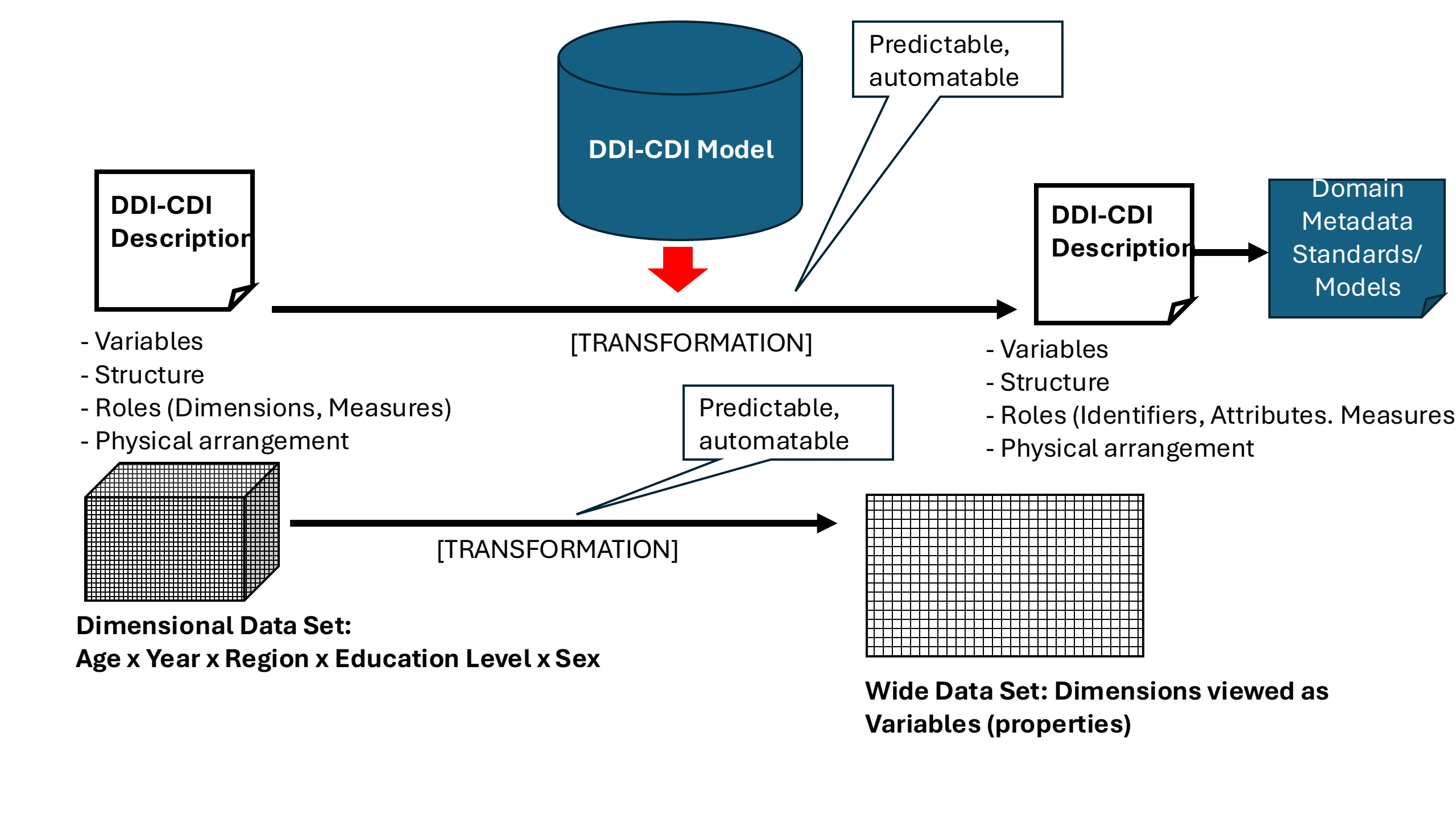
- The “DDI Moving Forward” project was originally envisioned as producing a model-driven replacement for DDI Lifecycle
- Many years were spent developing a very, very large model
  - Some parts were integrated into DDI Lifecycle (variable cascade, etc.)
  - DDI-CDI is the first new product to be based on this model (covers the variable cascade, but also the datum-oriented focus)
- During its development, it became apparent that we did not need a replacement for DDI Lifecycle!
- What we needed was something new: a way of describing non-rectangular data for use in cross-domain scenarios and with new forms of data
  - Aligned with DDI Codebook and DDI Lifecycle
  - Designed to support different requirements (FAIR, etc.)

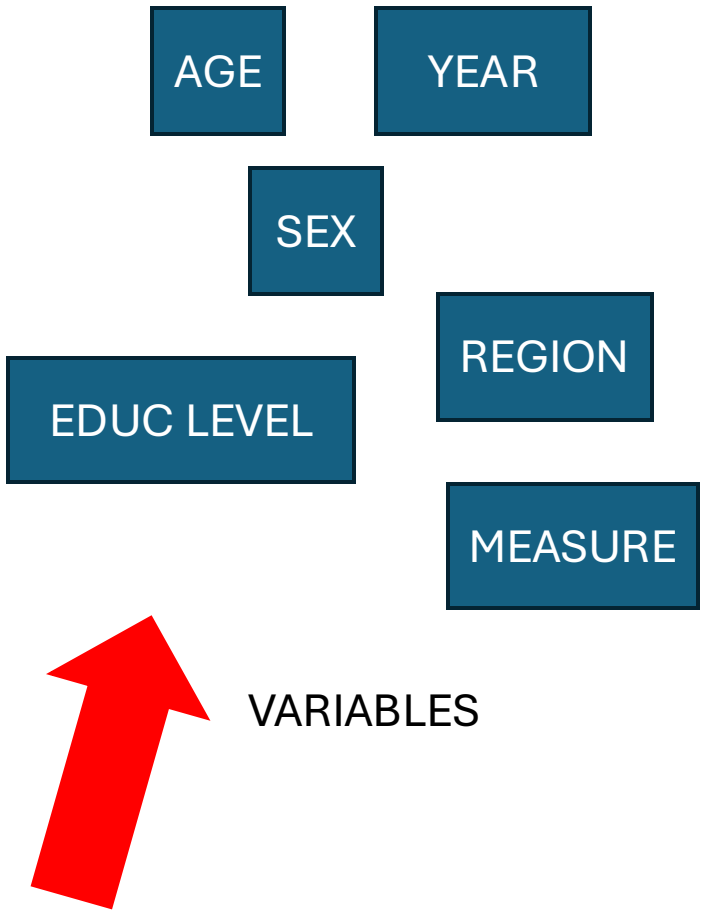
# Emerging Requirements

- Describe data from other domains
  - Might be structured differently (sensor/event data, multi-dimensional data, big data, etc.)
  - Metadata might be encoded using non-XML technology in a different domain standard/model (esp. RDF and Web standards)
  - Might be the output of a different data lifecycle
  - Terminology/semantics would be different
- DDI-CDI had to be *domain-agnostic*
  - Focus on the data, not on the study
  - Direct support for non-DDI standards/models (i.e., SKOS, DCAT, etc.)
- Needed a generic process flow description
  - Data had to be placed in its own context
  - Other domains produce data in unfamiliar ways – this information is needed for accurate integration
- Requirements aligned exactly with those of “cross-domain FAIR” data sharing

# What DDI-CDI *Is*

- A model for describing variables, including concepts and representations
  - Provides a full model of the variable cascade
  - Provides variable descriptions independent of their use in a specific structure
  - Links to externally defined semantics (CV, ontology, classification, etc.)
- A model describing how variables are used in structures
  - Wide/unit-record/rectangular
  - Long/tall (sensor or event data, registers)
  - Multi-dimensional data (superset of SDMX, covers indicators)
  - Key-value (big data, No-SQL, etc.)
- A generic model for describing process flows
- A model for describing datums and their relationships to each other, to variables, etc.
- DDI-CDI provides a superset model intended to be implemented in “profiles” (subsets)


DDI-CDI Model
Predictable, automatable
DDI-CDI Description
- Variables
- Structure
- Roles (Dimensions, Measures)
- Physical arrangement
[TRANSFORMATION]
DDI-CDI Description
Domain Metadata Standards/ Models
- Variables
- Structure
- Roles (Identifiers, Attributes. Measures
- Physical arrangement
Predictable, automatable
[TRANSFORMATION]
Dimensional Data Set:
Age x Year x Region x Education Level x Sex
Wide Data Set: Dimensions viewed as Variables (properties)




| 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.

# What DDI-CDI *Is* (continued)

- A UML model expressed in Canonical XMI
  - Syntax representation in XML (XML Schema)
  - Syntax representation in RDF (OWL/RDF-S vocabulary, expressed as Turtle and JSON-LD)
- A specification designed to be directly combined with other specifications
  - DDI Codebook/Lifecycle
  - SKOS/XKOS
  - DCAT
  - Process descriptions (SDTL, VTL, PROV, programming code, etc.)
  - Any other specification (esp. in RDF)

# What DDI-CDI is *Not*

- A replacement for DDI Codebook or DDI Lifecycle for data production, management, or archiving
- An XML representation of human-readable documentation at the study level
- A monolithic model for describing the data lifecycle
  - Does not provide information to manage longitudinal studies
  - Does not present the “data lifecycle” in a specific form

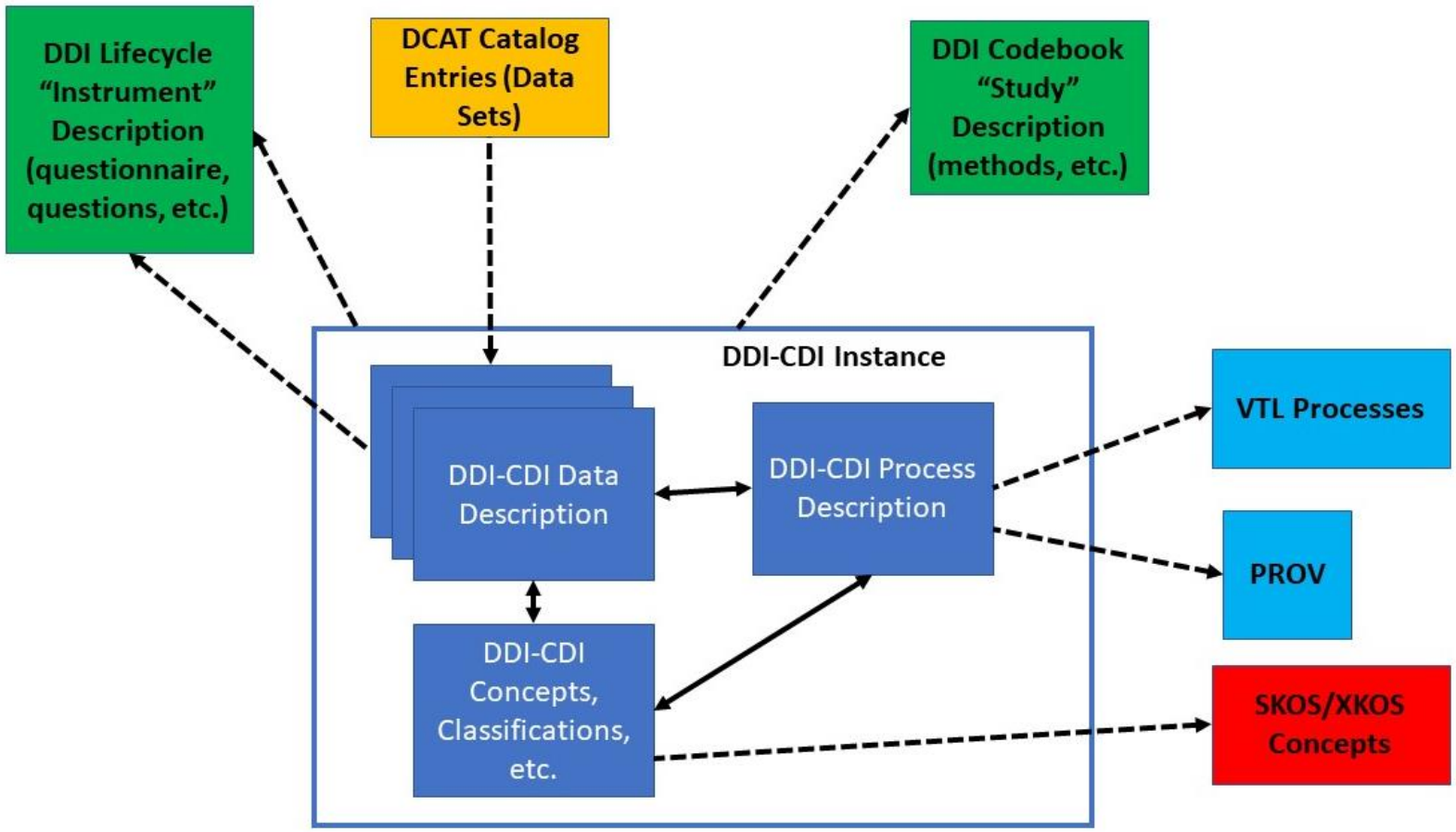
# DDI-CDI and DDI Specifications

- Metadata can be programmatically derived from DDI Codebook and Lifecycle
- Process descriptions can reference DDI Codebook/Lifecycle metadata (datasets, variables)
- Instrument and Study descriptions in DDI Codebook/DDI Lifecycle can be referenced
- Can use DDI CVs (esp. in SKOS) to define semantics
- Can describe processing with SDTL/SDTH
- Can represent codelists/categories/classifications with SKOS/XKOS when represented in RDF
  - *Instead of* DDI-CDI classes

# DDI-CDI with other Domain Standards

- DDI-CDI is designed to be populated not only from DDI Codebook and Lifecycle metadata, but from similar metadata in any domain specification
  - Domain-agnostic terminology
  - Generic description of many different data structures
  - Generic process flow description
  - External semantic definitions for concepts
- Leverages RDF design to allow combination with metadata in other common Web (or domain) vocabularies
- Acts as a “common language” for moving between domain specifications
  - Used in CDIF for this purpose
  - Example: SDG Indicators disaggregations: SDMX-to-DDI-CDI-to-Schema.org in Google Data Commons

# DDI-CDI with Other Standards - Example



# Early Implementations

- UKDA Product Builder (variables and structures)
- EOSC Future “ESS Labs” integration of European Social Survey with Climate and Environmental Data (process)
- US Bureau of Labor/BLS (indicators, variable cascade )
- UN Statistical Division SDG Indicator disaggregations (multi-dimensional data and variables)
- WorldFAIR Cross-Domain Interoperability Framework (CDIF)
  - Minimum data description profile (approx. 30 classes for long, wide, multi-dimensional)

# Tools

- Field-level documentation tool
  - Subsets of selected elements
  - Browseable documentation of selected classes
- Postman-funded “High Value Data Services” (in testing)
  - Wrappers municipal data in Socrata in DDI-CDI and DDI Codebook
  - Python libraries for working with DDI-CDI metadata
- Dataverse implementation (design stages)
- WorldFAIR (Sikt) Stata/SPSS-to-DDI-CDI Converter (to the Datum level)
- “Nectar Publisher” CSV/Excel-to-DDI-CDI converter (under development)
  - Work in the DDI Developer’s Group
- SDMX-to-DDI-CDI Converter
  - Prototyped by UNSD for CDIF
- ESS Labs Process Browser (prototype)

# Upcoming Projects

- EOSC CLIMATE ADAPT
  - Horizon-funded
  - Integration of environmental and social data
- EU OSCAR X-Ray Absorption Spectroscopy (XAS) project
  - Example of hard-science implementation for exchanging measuring device data on chemical spectrum with other domains
- Other WorldFAIR+
  - Second round of World FAIR
  - Many CDIF implementations for climate data, life sciences, hard science physical data, medical and clinical data, others
- Others (?)

# Existing & Upcoming Documentation

- High-level documentation (part of 1.0 spec)
- Field-Level Documentation (part of 1.0 spec)
  - Integrated with syntax representations
  - Subsetting for community profiles
  - https://ddi-cdi.github.io/ddi-cdi_v1.0-rc3/field-level-documentation/
- Business case document (nearly complete)
- DDI-CDI with other Standards (nearly complete)
- DDI-CDI Reference Profile for RDF (50% complete)
- Creating DDI-CDI Implementation Guides (50% complete)
- Standing up a Git repository for DDI-CDI Resources (profiles, application code, etc.)

# Examples

- UKDA Product Builder
- Sikt: ESS Labs and WorldFAIR SPSS/Stata converter

# Summary

- DDI-CDI is designed to accommodate new types of data and data from other domains – domain-agnostic
- Does *not* replace DDI Codebook or DDI Lifecycle!
- Works with many other DDI products (DDI Codebook, DDI Lifecycle, DDI CVs, XKOS, SDTL)
- Works with RDF and generic Web standards, and also domain standards
- Already tested in early implementations
- Very popular in the cross-domain FAIR space (CDIF)
- Many tools under development
- Good documentation and support now in development



# DDI - CDI: Integrating Data for Better Science!