DDI Working Paper Series -- No. 35

Title:

DDI-CDI and Other Standards

Author:

DDI-CDI Working Group

Principal Editor:

Dan Gillman

DOI:

10.5281/zenodo.11223221

Issue Date:

March 2025

Abstract:

With the growing demand for cross-domain interoperability, it is necessary to discuss how DDI Cross-Domain Integration (DDI-CDI) interacts with a wider ecosystem of standards, each of which performs important functions that complement and are complemented by those that DDI-CDI brings to the table. In an increasingly interdisciplinary world, a single standard cannot cover all aspects of data interoperability, making it necessary to combine multiple standards for cross-domain data integration. We need to understand how standards interact, and this document is an attempt to provide a working description of this, from a DDI-CDI perspective.

DDI-CDI provides the basic descriptive capability to enable interoperability of data and semantics across domains and the integration of heterogeneous datasets. In order to do this, it needs to interact with existing domain standards and those describing semantics. Therefore, we present below the case for why additional standards can and should be used with DDI-CDI.

Keywords:

Interoperability, data integration, variables

License:

Creative Commons Attribution 4.0 International (CC BY 4.0) License

DDI-CDI and Other Standards

DDI-CDI Working Group[1,2]

Abstract

With the growing demand for cross-domain interoperability, it is necessary to discuss how DDI-CDI interacts with a wider ecosystem of standards, each of which performs important functions that complement and are complemented by those that DDI-CDI brings to the table. In an increasingly interdisciplinary world, a single standard cannot cover all aspects of data interoperability, making it necessary to combine multiple standards for cross-domain data integration. We need to understand how standards interact, and this document is an attempt to provide a working description of this, from a DDI-CDI perspective.

DDI-CDI provides the basic descriptive capability to enable interoperability of data and semantics across domains and the integration of heterogeneous datasets. In order to do this, it needs to interact with existing domain standards and those describing semantics. Therefore, we present below the case for why additional standards can and should be used with DDI-CDI.

Keywords: Interoperability, data integration, variables

DDI-CDI - How it emerged

DDI-CDI emerged from the DDI-4 Moving Forward effort. The goal of the Moving Forward project was to build a UML model on which to base further developments of the DDI-Codebook and DDI-Lifecycle standards. This task proved unwieldy, and efforts to integrate data from multiple domains – beyond statistics and social sciences – drove new

[1] Originally conceived and developed at the Dagstuhl Workshop <DDI-CDI: Realising interoperable data services in the metadata ecosystem (https://www.dagstuhl.de/23393)>, held in September 2023.
[2] Dan Gillman was the principal editor for the document.

developments. A pared down model, one with generic data description capabilities, was proposed instead. Further development work was still necessary, but DDI-CDI emerged from this effort.

DDI-CDI is a standard designed to address the needs of data integration, independent of the specific subject-matter domain of the data under consideration. For example, the surveys national statistical offices, polling organizations, and researchers conduct have endured falling response rates for the last 20 years. Augmenting these surveys' data by integrating appropriate data from other sources will improve their estimates. Thus, DDI-CDI is positioned to help significantly enhance the quality of the data from these surveys. There are many other applications as well.

It is somewhat understandable that some people might confuse the original intent of the Moving Forward project with the outcome that is DDI-CDI[3]. It is designed for integrating data. As such it must remain subject-matter independent. Thus the scope of DDI-CDI is much different than that for DDI-Codebook and DDI-Lifecycle.

What does DDI-CDI do?

DDI-CDI is designed to describe, in a subject-matter independent way, the following:

- Structure
- Variables (variable cascade)
- Code lists and classifications
- Processing
- Datums (datum-centred approach)

No other descriptive considerations are part of DDI-CDI.

The standard is designed under the assumption that there are other specifications that can be attached to provide descriptions outside the scope of DDI-CDI.

Within DDI-CDI, these short statements describe each of the areas listed above:

- Structure
    - The way data, variables, and records are organised logically. The logical view is the way users interact with the data. It may not be the layout in a file.
- Variable cascade
    - Hierarchical description of variables. Each stage in the cascade is designed to maximise shareability. This provides the means to find commonality among data sets across time, programs, subject matter, and data producers.
- Code lists and classifications
    - Managing these provides a way to make lists sharable. Some are well-known, such as ISIC - the International Standard Industrial Classification.

[3] Due to the change in scope of DDI-CDI from the original Moving Forward project to integrating data from multiple sources, we must emphatically state that DDI-CDI is not designed, not intended, and not envisioned to replace other entries in the DDI suite of products, especially DDI-Codebook and DDI-Lifecycle.

- Processing
    - Provides a common model for expressing processing steps, and through a collection of steps, sharing processes in the same way variables can be.
- Datum-centred approach
    - Ability to track datums through processes, data sets, and structures. For complex data management systems, for advanced processing systems, and assessments of data quality, this mechanism is powerful.

Support for Interoperability

Interoperability is the capability of different systems to exchange and process data without requiring direct communication with the data provider. A core design purpose of DDI-CDI is to support this. If using or understanding data requires contacting the producer, a good argument can be made that the schema in the standard is very much incomplete.

The variable cascade provides a full description of a variable. This description allows a user of some data set to interpret the representations of data in that data set. The explanation of allowable values (value domain and sentinel domain) and how to use those data (datatype) are part of the basic information.

Retrieving data from a database requires knowing how the data are structured. Data are structured based on 4 basic structure types described in DDI-CDI: wide, long, key-value, and multi-dimensional. Each of the structure types affords different interpretations of data found in a data set. Importantly, these structure types can be used together in complex cases.

Code lists and classifications are a means for commonly used categories and their codes to be shared and conveyed. Consider a code list for the 50 states in the US. The US Post Office maintains such a list. This list and similar ones are used to standardise the way data are represented. If all variables for US states use the US Post Office list of state codes as their value domain, there is predictability across those variables.

Describing processes supports the ability to share them across applications and promotes reproducibility and replicability. Sharing them across production systems increases consistency. Sharing them for users increases understanding across sources. The datum-centred approach provides the means to track data through processing. Tracking data is important for data quality. Data quality implies data are interpretable, which in turn supports interoperability.

Concepts (meanings) are the basic way understanding is recorded and shared. Each concept is used in possibly several ways, called roles. DDI-CDI supports the use of roles for concepts, and one concept used in different roles means the same thing. For example, the concept adult is the same whether it is the universe for some variable or whether it is a category in a code list. There are several roles for concepts used in DDI-CDI.

An important role for concepts in DDI-CDI is that of category. Essentially, a category is a means of subdividing a population. Consider people in Sweden. Then take the category of

married. People in Sweden can be divided into groups based on whether they are married or not. Married also takes the role of an allowed value, say for a marital status variable.

The use of concepts and their roles is the means to support semantic interoperability. Concepts, as with all managed objects in DDI-CDI, are assigned identifiers. If the identifiers are persistent and globally unique they satisfy a basic criterion for being FAIR.

Case for using other standards with DDI-CDI

Integration with other standards is a core feature of DDI-CDI. By design, DDI-CDI is intended to provide the cross-domain integration machinery where other domain specific standards might be plugged in. Rather than duplicating functionality of other domain-specific standards, DDI-CDI provides a framework for integrating metadata elements from these standards (e.g., data description, provenance) within a cross-domain metadata structure. A large corpus of metadata content is used and maintained in multiple standards across the globe and it's impractical, and in most cases impossible, to transform that content into CDI before use. For instance, existing codelists and classifications in SKOS or XKOS can be used as value domains for DDI-CDI variables by reference without the need to transform the content from XKOS to CDI. Similarly, data catalogued with widely used standards like DCAT and schema.org can be enriched by adding detailed conceptual and variable level descriptions to the metadata provided by those standards.

It's important to note that even though DDI-CDI describes data organisation, structure, and to some extent concepts, there are other aspects of the broader data semantics that are beyond its scope. Things like the methodology used to produce and/or capture the data, e.g. experimental design, sampling, etc. together with its quality can be better described by existing standards and vocabularies like DDI-Lifecycle and the Data Quality Vocabulary.

In the space of provenance and lineage, popular models include the Business Process Modelling and Notation (BPMN) standard and the PROV Ontology (from W3C). Besides that, there are a multitude of implementation languages for driving data transformation, cleaning, and analysis, such as R, Python, SAS, and SPSS to name a few, and a couple of emerging standard languages for describing such processes, namely SDTL and VTL. Provenance and lineage expressed in these languages at varying levels of detail can be linked to the structural data descriptions of DDI-CDI to provide a more in-depth understanding of the meaning, origin, and derivation of the data.

How might DDI-CDI typically be used with other standards?

DDI-CDI is intended to be complementary to and used in combination with other standards and models that focus more on domain-specific aspects (such life-cycle models) or focus on the specifics of cross-domain problems (such as semantics, discovery, or provenance). Such generic elements as classifications and variables are given a detailed formal treatment in DDI-CDI but are agnostic as to the domain. It is left to the user to employ whatever domain semantics are demanded by the data with which they are working.

This feature of the specification makes it well-suited to combining data coming from more than one domain or system, to allow a description that supports systems that perform data integration, harmonisation, and similar functions.

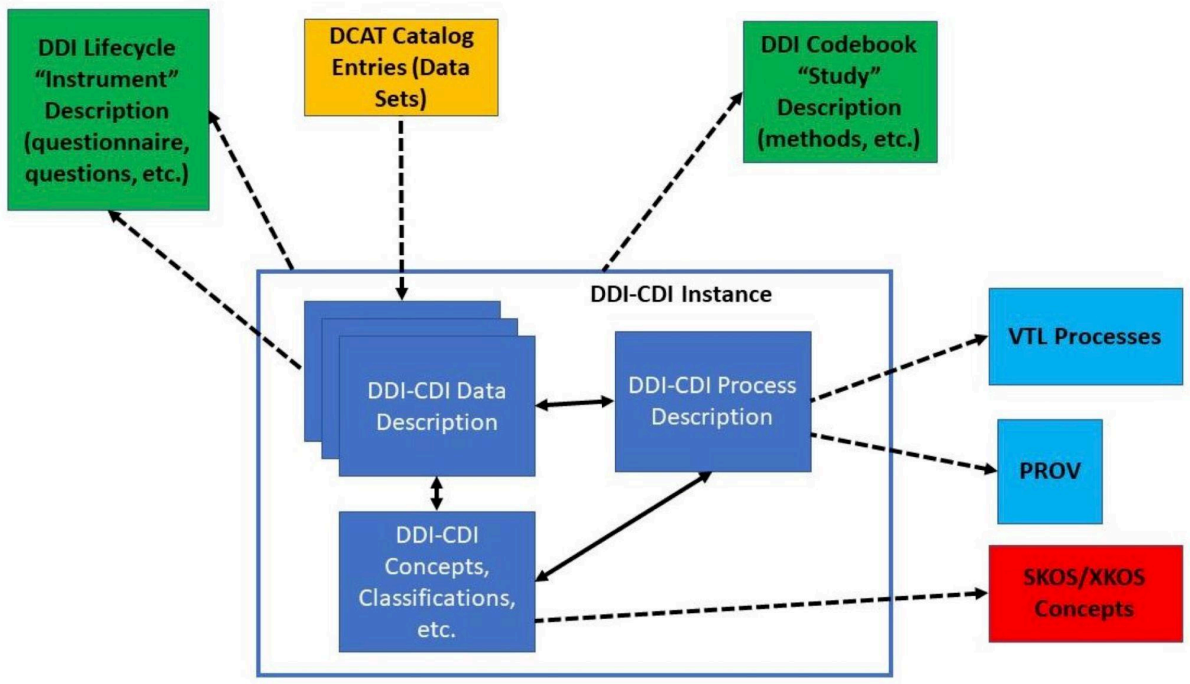
The illustration below shows an example of how DDI-CDI metadata might fit into a larger standards-based cross-domain implementation. The areas labelled “DDI-CDI” are covered by the DDI-CDI model; all other boxes are *examples* of the type of non-DDI-CDI metadata standards with which it is intended to be integrated.

The dark blue box illustrates the metadata for three different datasets modelled using DDI-CDI. This includes the data description (comprising the variable cascade and the structural description), the DDI-CDI concepts and classifications that are used in this, and the process description, which in this example might describe the steps needed to integrate these datasets.

In this illustration, the data description references DCAT catalogue entries describing these three data sets (orange box), and detailed metadata contained in other DDI products (green boxes). Here this additional metadata from DDI-Codebook or DDI-Lifecycle comprises descriptions of the overall research effort that produced the data, giving information on the methods used in collecting and producing the data (DDI-Codebook); and granular description of the questionnaire which was used during data collection (DDI-Lifecycle).

The underlying concepts used in DDI-CDI constructions, such as classifications, categories, and code lists, are expressed in SKOS and XKOS (red box). The DDI-CDI Process Description references additional information described according to PROV (for a step-by-step description of the process) and VTL (for descriptions of specific functions which were applied to the data) as it was transformed from one data set to another at specific points in the processing (light blue boxes).

The format of external references to other standards will vary depending on what syntax representation is used, but will typically consist of URIs.



The purpose of this diagram and the accompanying descriptions is to illustrate how the highly-detailed, fine-grained metadata presented in DDI-CDI draws on and references metadata in other standards. It is thought that this is likely to be a relatively typical example of how DDI-CDI will be used in relation to other standards, for the purposes of data integration.

How does one choose another standard to use?

If DDI-CDI does not meet all the descriptive needs of a proposed system, then additions to the descriptive capabilities are needed. Standards are a useful source of such specifications.

The choice of a standard is based on 4 dimensions: domain, community, function, and purpose[4]. Here is what is meant by each:

- Domain refers to the types of materials the standard is intended to be used with or could potentially be useful for.
- Community refers to the groups that currently or potentially use the standard.
- Function refers to the role a standard plays in the creation and storage of metadata.
- Purpose refers to the general type of metadata the standard is designed to record.

The assessment of these criteria will guide the developer to the potential standards for inclusion in a systems architecture. There are many metadata standards in existence. An example of more than one standard addressing the statistical lifecycle is the case of DDI-Lifecycle and the Generic Statistical Information Model (GSIM).

These standards address the same domain, function, and purpose.The community criterion is different, though. DDI-L was developed by the DDI Alliance, a consortium of data archives,

[4] https://jennriley.com/metadatamap/seeingstandards.pdf

libraries, and producers. GSIM was developed by the UNECE, which oversees cooperative work among national and international statistical organisations.

There is overlap between the two communities, but there are big differences as well. It is these differences between the communities that help explain differences between the two standards. The needs of the DDI Alliance led to the development of a specification rendered in XML. The UNECE community was interested in developing a conceptual model. Part of that decision was guided by the speed with which GSIM was developed - in one year. This also means GSIM is not fully specified.

Knowing these community-based differences is important for figuring out which specification to select.

Examples of standards to use with DDI-CDI

I-ADOPT

I-ADOPT is an interoperability framework for representing observable properties developed by the RDA InteroperAble Description of Observable Property Terminologies (I-ADOPT) WG and available as RDA endorsed recommendations[5]. It is a domain-agnostic model that was originally developed for and tested with the environmental research community but is getting adoption beyond that scope and will be helpful in many other domains. It offers a FAIRer representation of variables/observable properties compared to some other standards, which all use one concept for this concept. The I-ADOPT framework provides rich semantic context by decomposing its description into atomic parts in a systematic way reusing FAIR terminologies for each of the components. The framework is based on a ontology[6] consisting of four main classes (Variable, Property, Constraints, and Entity) and six relations (hasProperty, hasObjectOfInterest, hasContextObject, hasMatrix, hasConstraint, constrains) where a variable requires at least an ObjectOfInterest and a Property but can be further qualified by a Matrix or a ContextObject, that can be further constrained by a Constraint. The benefit of this approach is that the plain concepts used in the ontologies mentioned before can be augmented by the richer I-ADOPT variable concepts.

To demonstrate its usefulness for the DDI-CDI community we will explore use cases from official statistics which often use variables with long labels, for example "number of nights in a 3-star hotel". The I-ADOPT variable decomposes this into:

- nights -> object of interest
- number -> property
- hotel -> context object
- 3-star -> constraint

This kind of decomposition could be useful to derive semantic similarity links between variables and associated contexts. In DDI-CDI, the variable above would be linked to a

[5] https://doi.org/10.15497/RDA00071
[6] https://w3id.org/iadopt/ont/

similar concept ("number of nights in a 3-star hotel"), and this concept would link to a small graph representing the I-ADOPT decomposition. Techniques like graph embedding (http://rdf2vec.org) could then be used to evaluate the semantic distance between the initial variables. Note that the I-ADOPT decomposition of complex labels could be automated with NLP methods.

For the I-ADOPT community in turn, we will work on a use case from experimental observations where measure variables are dependent on environmental conditions that can be used as qualifiers. Users needing to apply detailed property descriptions in more complex use cases require recommendations on how to deal with variable dependencies as I-ADOPT focuses only on single variable descriptions. We see CDI as a complementary standard that can help to guide the I-ADOPT implementation by using variable relations and structure components.

The Cross-Domain Interoperability Framework (CDIF)

In the WorldFAIR Project, a major focus was identifying a set of standards which could be used in cross-domain scenarios to share data and metadata. The primary issue in such cases is that, while the formats and structure of data might be well understood within a domain, and described with the metadata standards familiar to domain practitioners, users from outside the domain might not have such understanding. As multi-disciplinary research grows, integrating data from different sources becomes increasingly challenging.

CDIF gives guidance on a set of recommended fields in different standards for the discovery and cataloguing of FAIR resources intended for sharing in these scenarios, and gives a similar description for how access conditions, controlled vocabularies, and the data itself can be described. Data description uses a “minimum profile” of DDI-CDI expressed in JSON-LD, and specifies how such metadata descriptions should be published on the Web so that it can be located programmatically.

CDIF uses a set of different standards for different purposes. Notably, Schema.org and DCAT are used for cataloguing and making data discoverable, ODRL is used to describe conditions of access, and SKOS - supplemented by XKOS - is employed to add additional metadata for formal statistical classifications where those are used. DDI-CDI is employed for describing variables and the structures of data, covering the wide, long, and multi-dimensional cases.

In CDIF, DDI-CDI does not use the native elements for describing codelists and classifications: it assumes that SKOS will be used for this purpose. Further, domain semantics can be referenced, as these will generally be available in some standard form on the Web (SKOS or possibly expressed as an OWL ontology, etc.) Where such definitions are used for the purposes of defining variables and other relevant concepts, the URI is employed to reference them.

In the case of CDIF, it is recognised that standards must be generic in cases where the resources they describe are being produced by members of one domain or discipline but being consumed by members of a different one. For this reason, the set of standards listed - many coming from W3C or the Web community - are seen as more appropriate for use than others that have been produced within a specific domain for their own use. In order to make

adoption and conformance with the CDIF recommendations easier, W3C standards that are already in widespread use have been preferred. Further, RDF-based syntaxes are preferred to XML within the FAIR community, and the JSON-LD expression is one that is familiar to the largest possible number of developers, based on the popularity of JSON for Web development.

In the case of data description, however, it was found that no standard aside from DDI-CDI provided the structural metadata needed, as well as providing a sufficient platform to connect all of the disparate standards used to support other needed functions. The design of DDI-CDI to integrate other standard metadata is critical in making it suitable for this scenario.

CDIF is only in its first published version, and it will be expanded in the future to cover provenance and data “context”, mapping of semantics and structures in data integration, and other topics. Additional standards such as PROV, I-ADOPT and SSSOM are seen as likely candidates for similar inclusion in the recommendations in due course. Thus, the capabilities for DDI-CDI to combine with other specifications and standards is one which will continue to prove valuable.