



# DDI Common Core

*DDI Alliance*

*Editor: Dan Gillman*

*Technical Direction: ISO Certification Exploration Temporary Working Group (Dan Gillman and Arofan Gregory)*

*Version 1 : 2025*

License: [Creative Commons Attribution 4.0 International \(CC BY 4.0\) License](https://creativecommons.org/licenses/by/4.0/)

## Introduction

The Data Documentation Initiative (DDI) is an ongoing program within the social, behavioural, and economic (SBE) data community for documenting their data. The program is managed by a consortium, the DDI Alliance, that comprises university data libraries, university and national data archives, research centres, national statistical offices, consultancies, and software development organizations. The suite of products under DDI include several standards and other products designed to describe data and the processes used to produce them.

DDI first began in 1995 to build a framework for describing a codebook, the document describing the variables, questions, code lists, classifications, methodologies, usage guidance used to collect, organize, and delimit the data in research studies. This standard was built using XML and first released in 2001 as DDI-Codebook (DDI-C). It has been updated periodically since and the current version is numbered 2.5.

In 2003, the DDI community realized an expanded version of DDI-C that addressed the descriptive needs of data producers such as national statistical offices was needed. These requirements included descriptions of the methodologies used in production, which include questionnaires, sampling, and weighting. These are used in probability based statistical sample surveys, the standard design for surveys conducted by national statistical offices. Further, the ability to reuse descriptions of variables, questions, and other artefacts was seen as necessary to efficiently describe ongoing statistical surveys over time and program in support of comparability, interoperability, and integration. This effort, also developed using XML, was called DDI-Lifecycle (DDI-L). Its current version is numbered 3.3.

Reacting to a dramatic decline in response rates to their surveys, survey organisations began to look to enhance their data by integrating them with data from outside sources. These data come in a variety of formats, with inconsistent quality and coverage, and often aren't based on a probability

sample. A framework of descriptions to support data integration independent of SBE was intended to ease this problem. DDI developed the Cross-Domain Integration (DDI-CDI) standard. This standard was released in early 2025.

Along with the standards, the DDI Alliance advocated the development of several supporting products, which are

- XKOS – eXtended Knowledge Organisation System is an extension of the W3C SKOS (Simple Knowledge Organisation System) used to render concept systems in RDF (Resource Description Framework). XKOS extends SKOS in 2 major ways. It
  - supports levels in hierarchies, which are used to organize statistical classifications
  - includes the semantics for some temporal and sequential relations.
- SDDL – Structured Data Transformation Language is a mid-level language for documenting the processing steps in a data production environment.
- Controlled Vocabularies – A series of category sets and code lists for use as common ways to populate some descriptors in the DDI standards. An example is the names for the kinds of sampling schemes available.

Beginning around 1970, several researchers independently coined the term metadata. This loosely meant “data about data”, and it refers to descriptions of data. Since then, the concept of metadata has expanded, and the term is now applied to descriptions of any object or resource, not just to data. This expanded notion of metadata is how we understand the term. Each of the DDI standards describes more than data.

Given this understanding of metadata, all the DDI standards are metadata standards. They address the organisation and formats of the metadata needed for describing data. This mostly applies to SBE data, but the DDI-CDI standard addresses data independent of the source.

An International Standard is based upon this document that describes several common aspects of all the DDI standards and other products. The DDI standards all use the notion of a variable, and these roughly correspond to a column of data organised in the rectangular format. The common notion of a variable and how its description is organized is the focus.

Each DDI standard and the other products are used to describe parts of the data lifecycle. The phases of this lifecycle are defined, and which phases are addressed by the DDI standards and products are identified.

Of the DDI standards now managed under a UML (Unified Modelling Language) model, those models are independent of each syntax representation (for example, XML, JSON, RDF, etc.) and are called Platform Independent Models (PIM). Each syntax representation uses its own model, and each is an approximation (as close as possible) to the PIM. They are called Platform Specific Models (PSM).

Using the terminology and ideas in ISO/IEC Guide 2, we say each PSM conforms to its PIM.

## Terms

We take a somewhat formal approach here and define the terms used in this document as found in ISO 1087 – *Terminology work and terminology science – Vocabulary*, ISO/IEC Guide 2 –

*Standardization and related activities – General vocabulary*, and the DDI Glossary (<https://ddialliance.org/glossary>). Three additional terms and their definitions are included here:

- concept
  - unit of thought differentiated by characteristics
- platform independent model (PIM)
  - document describing the abstract model of the standardized data exchange process in a platform-independent way
- platform specific model (PSM)
  - model of a software system or business system linked to a specific technological platform

## Metadata

For the DDI standards and related products, metadata are data describing some object(s), and all the standards under DDI are frameworks for organizing metadata. Some metadata are machine-actionable, and others are just readable as text. Machine-actionable metadata are those in a format that can be read and processed using predefined code and logic.

Machine-actionable metadata are desirable, but because of the broad scope of the DDI standards, not all the metadata they prescribe are machine-actionable.

## Variables

A variable is a mapping between a set of units called a sample and a set of permissible (or allowed) values called a value domain. A unit (see DDI Glossary) is an object (as defined in ISO 1087:2019), and each unit in the sample is assigned one element from the value domain. Seen as a function as defined in mathematics, the sample is the domain of the variable, and the value domain is its range.

Variables (see DDI Glossary) are used in efforts to measure a population (see DDI Glossary), such as through experiments, research studies, or surveys. The set of eligible units is the population. The sample, the actual units under study, are selected from the population, and it is a non-empty and not necessarily proper subset of the population. Variables are defined by a characteristic of the population, such as income, gender, number of employees, blood pressure, weight of coal, etc. Populations are not limited to people. In fact, they can be any set of objects of a fixed kind.

Value domains contain the properties corresponding to the characteristic defining the variable. For a variable on marital status, a relevant value domain might include the categories single, married, divorced, and widowed, which serve as properties. One of these is assigned to each unit in the sample. Typically, the name of each value is not what is recorded. Rather, a shortened string called a code is used to represent them. In this case, the codes might be *s*, *m*, *d*, and *w* respectively.

Variables measuring a quantity are represented by numbers and a unit of measure. A quantity such as speed might use miles per hour, metres per second, or furlongs per fortnight as its unit of measure. Each quantity usually has several units of measure from which the designer of the data can choose.

Variables are additionally used to represent address, date and time, name, latitude / longitude / altitude, other positional information. Other possibilities exist as well.

## Variable Cascade

Variables are described through a metadata structure in the DDI standards called the Variable Cascade (VC) (see DDI Glossary). The VC has four levels, each adding to the description accrued in the level above it. Definitions and the features found at each level follow below:

- Concept
  - The concept defining the variable being described.
  - Definition: unit of thought differentiated by characteristics

Note – Variables are described as characteristics of a population above, but characteristics are concepts, too. In this case, the concept defining a variable is a characteristic of another concept, the population.

- Conceptual Variable (CV)
  - Includes the additional concepts used in the allowed values for the variable, such as single, married, divorced, and widowed for marital status or a range of numbers for a measure, and it includes a Unit Type (see Unit Cascade).
  - Definition: description of the semantics of a variable independent of any particular representation or implementation
- Represented Variable (RV)
  - Includes the codes, numerals, formats, or other representations applied to the values, their datatype, and the unit of measure (if applicable).
  - Definition: specification for the encoding of substantive values of a variable, based on a conceptual variable
- Instance Variable (IV)
  - Includes processing details, spatial and temporal aspects of the sample, and links to how the data under the described variable are structured and formatted.
  - Definition: description of a variable in the context of a particular dataset

It bears emphasizing that the levels of the VC are not variables themselves. Each level is part of the description of a variable. In simple terms, an instance of the VC is a description; it is not a variable, which is a mapping between a sample and a value domain.

The reason for the levels in the VC is the need to share common aspects of descriptions of variables across time, data acquisition and processing operations, and databases. In the SBE domain, typical data acquisition and processing operations include statistical surveys and academic research studies. There may be others as well.

In organizations that conduct many surveys or studies, similar variables are often found across the surveys or studies. For example, there are only so many ways a marital status variable can be designed, and marital status is a very common characteristic included in demographic surveys and studies.

Rather than recording substantially the same description each time one is needed, reuse of descriptions

- Is more efficient, as it takes a shorter time to record a shared description
- Reduces errors and inconsistencies

- Promotes finding relevant data across disparate sources
- Enables integration from several sources

Part of the description of a variable includes that of the population from which the sample is drawn. This description shall be included at the CV level of the VC using the Unit Type and more specifically in one or more of the other levels of the VC, except the Concept level. See the Unit Cascade for a fuller description.

## Unit Cascade

Populations, in a similar way as with variables, can be described with reusable parts. The Unit Cascade (UC) contains the three levels of description, and they are described and defined as follows:

- Unit Type
  - One of possibly many generic collections of units within the context of a data acquisition organization, such as people, households, and business establishments for national statistical offices.
  - Definition: class of units defined by essential characteristics
- Universe
  - A specialization of a Unit Type by applying socio-economic categories as delimiting characteristics, such as women are people (a Unit Type) who are female (a characteristic) and at least 18 years old (another characteristic).
  - Definition: class of individuals that share a unit type and typically have other characteristics in common, exclusive of time and geography
- Population
  - A specialization of a Universe by applying geographic and time constraints, such as women (a Universe) in the United States (geographic constraint) in 2025 (time constraint).
  - Definition: universe in which the individuals share time and geography

Each level in the UC is known generically as a Unit Class. A Unit Class is used to understand the kind of units each variable measures, classifies, or identifies. This information is vital for analysing data and drawing conclusions.

Any Unit Class may be associated with each descriptive level in the VC. The needs and details of the situation determine which Unit Class is appropriate to use.

## Value Domains

A value domain is the set of permissible values for a variable. The purpose is to inform a user of some data the values some variable can take. A permissible value contains both a representation and a meaning. The representation is what is recorded for the value in a file. The meaning is the semantics of the value.

Value domains have two main typologies, one based on structure, the other based on usage.

### Value Domain Structures

Value domains are structured in three possible ways: enumeration, range, and rule. Value domains specified by a range or a rule are not enumerated and referred to as described.

## Enumerated Value Domains

An enumerated value domain contains a complete predefined list of all the permissible values. Each permissible value is an ordered pair containing a representation and its meaning. For a marital status variable that uses an enumeration of permissible values, the marital status categories (value domain) might be these:

<s, Single>

<m, Married>

<d, Divorced>

<w, Widowed>

Note, the words “Single”, “Married”, “Divorced”, “Widowed” in this example are substitutes for their definitions, which are the verbal way to record and convey a meaning. The representation for each permissible value in the example is the letter on the left-hand side of the comma. These simple shortened representations are called codes. Codes are easier to record than full names and are easier to process.

## Described value domain – Range

Numeric values are described by a range. Ranges consist of an interval of numbers and how the numbers are represented, e.g., decimal (Arabic) digits, hexadecimal digits, binary digits. Counts, quantities, monetary values, and date/times are the 4 main kinds of ranges. Counts are specified by a range of non-negative integers. Quantities are specified by real or floating-point numbers with a given precision. Monetary values are specified by a scaled value, usually using two digits to the right of the decimal. Date/times are specified based on the format used to specify date/times, which is a kind of precision. The format possibilities are

- *yyyy*
- *yyyy:mm*
- *yyyy:mm:dd*
- *yyyy:mm:dd-hh*
- *yyyy:mm:dd-hh:mm*
- *yyyy:mm:dd-hh:mm:ss*
- *hh*
- *hh:mm*
- *hh:mm:ss*

Here, *yyyy* is the 4-digit year, *mm* is the 2-digit month, *dd* is the 2-digit day, *hh* is the 2-digit hour, *mm* is the 2-digit minutes, and *ss* is the 2-digit seconds. Upper and lower bounds mean the same as start and stop.

Typically, the upper and lower bounds of a range are part of its description. However, it is more efficient to place these constraints on the description of variables. This greatly reduces the number of value domains needed to describe the variables for an organization.

The unit of measure for quantity variables is likewise more efficiently placed with the description of variables.

## Described value domain – Rule

Other kinds of data are specified by rules. Identifiers, locators, names, addresses, etc. are best described this way. When there are no predefined lists or controlled vocabularies from which to enumerate values, a rule for forming those values is the best description possible.

Sometimes the rules are formalizable. Regular expressions and EBNF (extended Backus–Naur form) are some means to do this. The rules for the formation of URIs (Uniform Resource Identifiers) in RFC 3986 is a practical example.

## Value Domain Usages

Value domains specify permissible values in two situations: one describing values pertaining to subject matter (substantive usage), and the other for values pertaining to processing (sentinel usage). All the discussion in Value Domain Structures referred to substantive values.

Sentinel values are for processing distinctions. For SBE data, sentinel values often refer to missing data or data that were refused to be given. Other possibilities include transcription errors and network errors. There are many possibilities.

Sentinel values should be managed in their own (sentinel) value domain. Separating substantive and sentinel values into their own value domains reduces the overall number of value domains necessary to manage.

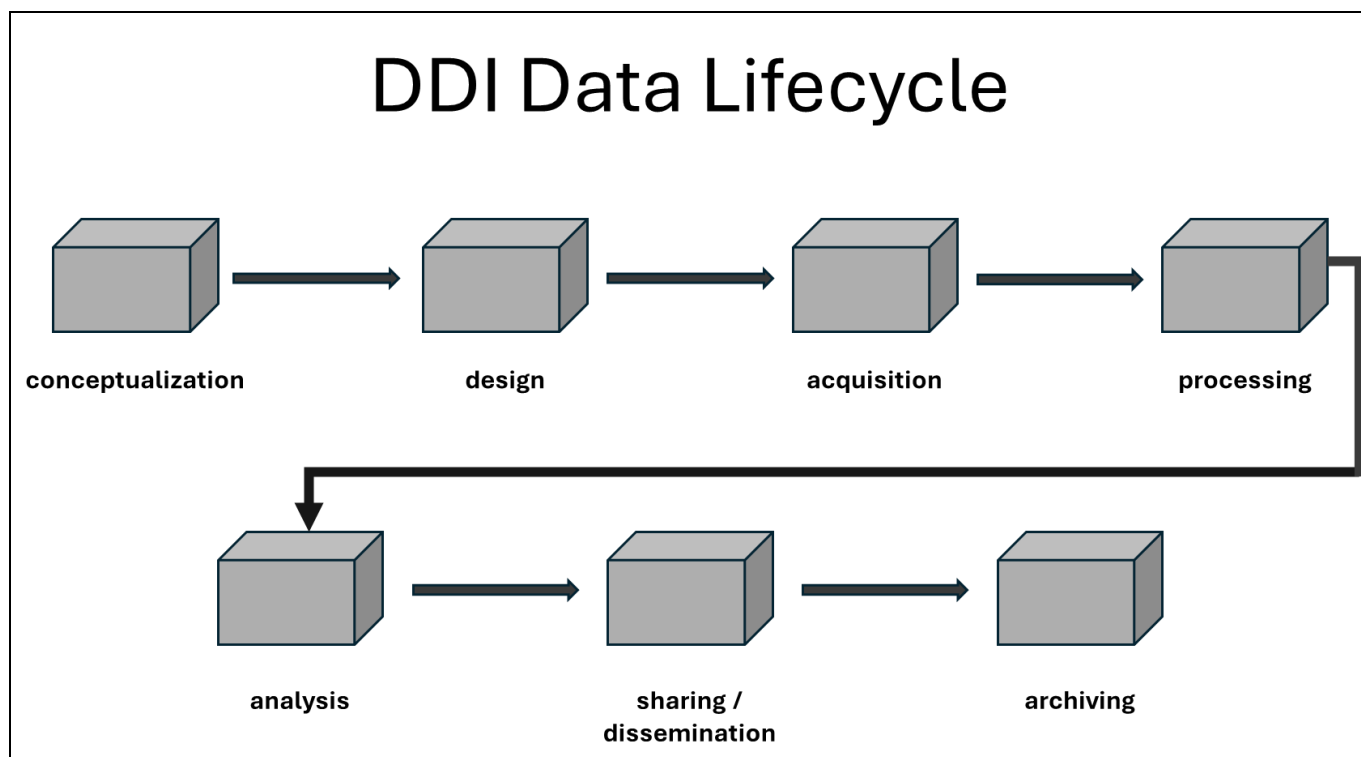
Almost all sentinel value domains are enumerated. It is possible to expect a range of sentinel values, but this situation is rare.

## Data Lifecycle

The DDI Data Lifecycle (see DDI Glossary) (hereafter Data Lifecycle) covers the following phases:

- Conceptualization –
  - Establishment of goals, essential concepts, and strategy
- Design –
  - Specification of the strategy to achieve the goals
- Acquisition –
  - Activities to acquire data, either pre-existing or generated
- Processing –
  - Activities to transform data to be suitable for analyses
- Analysis –
  - Application of (statistical) adjustments and transformations to produce outputs
- Sharing / Dissemination –
  - Exchange and other procedures for users to obtain data
- Archiving –
  - Activity supporting future use of data through preservation and providing access

See Figure 1 below:



**Figure 1: DDI Data Lifecycle**

This outline of the Data Lifecycle presumes many detailed activities under each phase. An example of a standard that provides this enhanced detail is the Generic Statistical Business Process Model under the UN Economic Commission for Europe. It has a similar set of phases, each with several more detailed activities described under it.

In the table below, we show how each of the DDI standards and related products fits within the Data Lifecycle just described. We show which phases of the Data Lifecycle each addresses. See Table 1 below:

Phase DDI Product	Concept	Design	Acquire	Process	Analysis	Share / Disseminate	Archive
Codebook	x	x	x			x	x
Lifecycle	x	x	x	x	x	x	x
CDI	x	x		x		x	
XKOS	x	x				x	
SDTL			x	x	x	x	

**Table 1: Mapping DDI Products onto Data Lifecycle**



## Model Independence

The DDI standards use the distinction between a platform independent model (PIM) and platform specific model (PSM) – i.e. platform-specific for a specific target language. Some of the DDI standards and products – DDI-L, DDI-CDI, and SDTL – are maintained through a Unified Modelling Language (UML) model. Migration of DDI-C to a UML model basis is underway. These UML models are independent of any syntax representation, i.e., a format suitable for building a database. Each UML model is a PIM.

DDI standards and products use syntax representations of the models, like XML Schema, JSON-Schema, or RDF (Resource Description Framework) syntaxes. Each syntax representation has its own model equivalent (often the schema for the syntax representation), and this is a PSM. For a standard based on a PIM, each PSM conforms to the PIM.

Each of the standards has conforming syntax representations developed, in development, or planned. Each is or will be a representation of the UML model (the PIM).

Originally, DDI-C and DDI-L were designed using XML (eXtensible Markup Language) and the schema is expressed in XML Schema. XML is often used as the language to store metadata using DDI standards. This tradition was extended to DDI-CDI, also. XKOS is based on the abstract syntax of RDF. XKOS is expressed in Turtle syntax.

## Semantic Interoperability

Many of the elements in the DDI standards and products require user defined input, usually in the form of text. These elements include sampling procedure, datatype, language proficiency, data collection mode, etc. Left to their own devices, users will enter whatever value they think is appropriate, so there is no control over variation. This leads to a lack of comparability between applications and a loss of semantic interoperability. This is especially acute when machines are being depended on to make comparisons.

The use of controlled vocabularies, a set of values from which all users choose their input, eliminates the issues described above. Semantic interoperability, maintaining the same meanings across the usage of the same terms and using the same terms for the same situations, is achieved.

The DDI Alliance maintains a collection of Controlled Vocabularies for this purpose.

## Bibliography

- [1] DDI Alliance (2025). <https://ddialliance.org/>
- [2] DDI-C (2025). DDI-Codebook <https://ddialliance.org/ddi-codebook>
- [3] DDI-CV (2025). DDI-Controlled Vocabularies. <https://rdf-vocabulary.ddialliance.org/>
- [4] DDI-CDI (2025). DDI-Cross Domain Integration. <https://ddialliance.org/ddi-cdi>
- [5] DDI-L (2025). DDI-Lifecycle. <https://ddialliance.org/ddi-lifecycle>
- [6] DDI-SDTL (2025). DDI-Structured Data Transformation Language. <https://ddialliance.org/sdtl>
- [7] DDI-XKOS (2025). DDI-eXtended Knowledge Organization System. <https://ddialliance.org/xkos>
- [8] ISO Online Browsing Platform (OBP) - [Online Browsing Platform \(OBP\)](#)
- [9] ISO/IEC 11179-3:2023, Information technology — Metadata registries (MDR) — Part 3: Metamodel for registry common facilities
- [10] ISO/IEC 19501:2005, Information technology — Open Distributed Processing — Unified Modeling Language (UML) Version 1.4.2
- [11] JSON (2025). JavaScript Object Notation. <https://www.json.org/json-en.html>
- [12] RFC 3986:2005. Uniform Resource Identifiers. <https://www.rfc-editor.org/info/rfc3986>
- [13] UNECE GSBPM:2019. Generic Statistical Business Process Model. <https://statswiki.unece.org/spaces/GSBPM/pages/243269812/GSBPM+v5.1>
- [14] W3C XML:2006, Extensible Markup Language (XML). <https://www.w3.org/XML/>
- [15] W3C RDF:2014, Resource Description Framework (RDF). <https://www.w3.org/RDF/>