DDI 4 Core Overview and Scope[[1]](#footnote-1)

# Purpose

## Note on Process

The work on producing the DDI 4 Core was launched so that, following the Prototype review, some of the core features of the DDI 4 work could be made ready for production release, recognizing that with available resources a more narrow scope was desirable. Emphasis was placed on short-term delivery: the Modeling, Representation, Testing (MRT) working group has allowed itself a year to complete the work on the initial core release, with delivery of a final deliverable ready for review and release at the end of December 2019.

## Goals

DDI has always faced the requirement of dealing with a large range of data, both for archival purposes and to provide support for the entire production lifecycle to large studies and statistical agencies. The result of this work is a model which in many important respects is domain-independent. Recent developments in the research world are placing a greater emphasis on cross-domain integration of data, and data coming from new sources, some of it in unfamiliar forms (e.g., “big data,” social media, sensor data, etc.). Social Science is no different from other domains – the DDI community is faced with a requirement for a more flexible ability to describe and manage, now available in a wider variety of forms.

The DDI 4 Core is intended to provide useful functionality in response to this requirement. In presentations, review comments, and discussions it has become apparent that some aspects of the DDI 4 model included in the DDI 4 Prototype are of especially high value. Identified features include the conceptual aspects of variables and classifications, the datum-oriented description of data, and the use of the process model to describe data lineage (the processing involved in the provenance of data). These same features have been identified as of interest by participants at recent Dagstuhl workshops on the subject of cross-domain data integration, with a further emphasis on alignment with external standards and the use of a UML model as a primary deliverable.

The DDI 4 Core will include not only the XML and RDF syntax representations, but will also deliver the UML from which they are derived in the form of Canonical XMI, a portable, tool-independent expression of the model. This not only makes DDI available for representation in other syntaxes and systems, but provides a stable basis for the maintenance of the model into the future.

User-oriented subsets of the DDI 4 model are provided by the inclusion of Functional Views, organized to support the application of DDI to specific tasks. This approach was employed in the Prototype, and will be carried forward into DDI 4 Core in a refined form, having both a technical and documentary expression. In addition, high-level documentation aimed at introducing the model to adopters has been added. Together, these should make the DDI 4 Core more adoptable and easier to approach.

Because of the use of DDI 4 Core for cross-domain integration, and for other purposes, some key external standards have been selected as candidates for specific alignment (PROV-O for provenance, GSIM as a higher-level model of core metadata, GSBPM as a framework for process description, and DCAT for data discovery). Documentation of alignment with this small set of selected standards will be part of the deliverable package. The use of existing RDF vocabularies in the RDF syntax representation of the model is anticipated, as a needed feature of alignment with standards/best practice in the Linked Data domain.

The idea that DDI 4 Core be re-branded to reflect its intended use has been discussed: DDI-Codebook and DDI-Lifecycle have brands which reflect their intended use, while DDI 4 Core does not. A re-branding would communicate to users the purpose of the new release, and minimize confusion as to which version of the standard is best suited for their applications - currently, the use of the version number indicates an erroneous relationship between versions 3 and 4 which is causing some confusion among the potential users of the new standard release, as it did when earlier releases were referred to as “DDI 2” and “DDI 3”. (Suggestions have been along the lines of "DDI - Integration," "DDI - Cross-Domain," etc. Identifying a better name for the DDI 4 Core will need further exploration and conversation with the Marketing group and others.)

*Please see the following page for a list of features.*

# Features

The DDI Core will have the following specific features:

* Conceptual and data description
  + Ability to describe variables as they are used and re-used at different process stages (the "variable cascade") with a strong connection to a conceptual layer
  + Ability to describe traditional and non-traditional data formats (unit-record/rectangular data files, multi-dimensional/aggregate data, event-history ("tall-skinny") data, and other data formats (analysis of no-SQL is on-going)
  + Ability to describe classifications as used in description of data
* Process description
  + The ability to describe in detail the processes to which data have been subject, as an aspect of data provenance ("data lineage"); data capture and methodology are not described in detail
* User-oriented subsets
  + Definition of Functional Views/subsets intended for specific applications, in the form of relevant technical specification and documentation
* Technical Deliverables
  + A UML model expressed in Canonical XMI, a portable flavor of that standard which is supported by a large number of tools which process UML. (This will support various validation and further processing like the use of DDI in program syntaxes as automatically generated by some UML tools.)
  + XML Schema for use in XML implementations for preservation and exchange
  + OWL specification with supporting validation resources (i.e., ShEx) for use in RDF implementations for discovery in the Web of Linked Data
* Documentation Deliverables
  + Complete documentation of the syntax representations for XML Schema and OWL
  + Complete documentation of the classes and relationships in the UML model
  + High-level documentation, including an explanation of design, features, and scope, and introductory material for the major features of the model (the variable cascade, the datum-based data description, and the process model for describing provenance/lineage)
* Documentation for implementation, alignment, and/or integration with external standards (PROV-O, DCAT, GSBPM, GSIM)

1. The term “DDI 4 Core” is used throughout as the current working name. The issue of a re-naming to bring the work in-line with other DDI products is discussed in the text. [↑](#footnote-ref-1)