Content revision of “Choosing a DDI Product” page

The Data Documentation Initiative (DDI) is a suite of products that describes metadata about both quantitative and qualitative research data in the social, behavioral, economic, and health sciences. The DDI suite is a set of free standards that document and manage different stages of the research data lifecycle, including conceptualization, collection, process, distribution, discovery, and archiving.

The content areas of DDI cover the following areas:

- Concept: the capture and management of the elements in the Variable Cascade, Unit Cascade, and Value Domains
- Design: approaches to sample selection, data capture, weighting, quality control, quality standards, and process management involved in the design of the study and/or data capture
- Acquire: data capture through the use of questionnaires, measurement, content analysis, and reuse of existing data
- Processing: data capture, data processing, analysis, and data management
- Analysis: analysis resulting in subsidiary data sets, recording analysis, and guidance for informed analysis
- Share / Disseminate: data storage, access management, rights, and restrictions
- Archive: ownership, organization, agent management, relationship between products, versioning, provenance, and preservation

Products within the DDI suite differ in terms of their area of coverage within DDI, supported activities, and required level of infrastructure. From simple descriptive content for human understanding to structures that support metadata-driven statistics production and analysis, DDI addresses a broad area of data management needs. As a suite of standards, DDI

provides a common means of identification for information objects, support for common cross-product content, and an informed means of transforming content between products.

The following table is from the DDI Common Core (link) and indicates which products address specific areas of DDI coverage. The tabs below the table provide additional details on the level of coverage and specific application areas. The product name links to the primary product pages for details on the specification including documentation and usage information.

Phase DDI Product [expression languages]	Concept	Design	Acquire	Process	Analysis	Share / Disseminate	Archive
DDI-Codebook [XML Schema]	x	x	x			x	x
DDI-Lifecycle [XML Schema]	x	x	x	x	x	x	x
DDI-CDI [XML Schema, RDF/OWL, JSON, UML/XMI]	x	x		x		x	
XKOS [RDF/OWL]	x	x				x	

SDTL [RDF/OWL, JSON]			x	x	x	x	

[The following sections will be entered in TABS: note that HubSpot is limited to 6 tabs at each level so I collapsed processing and analysis into a single tab]

Concept

Concept covers the DDI Common Core elements of the Variable Cascade, Unit Cascade, and Value Domain. This is the content published as ISO/PAS 25955:2026. Each of the primary products, DDI-Codebook, DDI-Lifecycle, and DDI-CDI provide varying levels of detail for the elements in these core structures. See the DDI Common Core for details on the common elements.

DDI-Codebook captures descriptive information on each element as expressed in the Instance Variable. The intervening levels of Conceptual Variable, Represented Variable, Unit Type, Universe (as expressed at the study level), and Population (as expressed at the variable level) can be reconstructed from the contents provided in DDI-Codebook. This is important if the content is later translated into DDI-Lifecycle or DDI-CDI for use in additional applications.

DDI-Lifecycle expands on the descriptive capabilities of DDI-Codebook by expressly stating each level of the Variable Cascade, Unit Cascade, and Value Domain. DDI-Lifecycle also provides the structures to support the management and reuse of each of the parts. This has been designed to support Variable Banks, cross-study comparability at various levels of detail, and facilitate searching across multiple studies.

DDI-CDI extends content down through the level of the individual datum. The primary focus of DDI-CDI is the discovery, selection, acquisition, and integration of data from multiple cross-disciplinary data sets. By clearly associating conceptual level information to an individual datum, DDI-CDI allows search systems to accurately identify individual data and to link them to data from other sources in a machine-actionable manner.

All the primary DDI products support the use of external controlled vocabularies to provide cross-data set comparability and searching. DDI provides several **Controlled Vocabularies** to encourage their use.

XKOS is a specialized product focused on the presentation and management of statistical classifications over time to support their accurate use as reusable content for levels in the Unit Cascade or for Value Domains.

Design

DDI-Codebook provides descriptive information on the purpose, design, and methodology used in developing a study or data acquisition process at the study level. It allows for assigning controlled vocabularies at various levels to support identification of broader design principles or specifying levels at which specific methodologies, such as sampling approaches, are applied. External documentation can be identified and associated with design and methodology approaches.

DDI-Lifecycle continues its approach of content management and reuse through identification and versioning of content to support comparison between multiple studies or to provide information on changes within a specific repeated study over time. DDI-Lifecycle allows for the provision of clear, comparable information for exploring why and how a study was designed to accomplish specific goals.

DDI-CDI reflects the approach of DDI-Lifecycle, providing structured information on design and implementation of a study to support comparison and accurate information on effect of design on the comparability of the resulting data.

XKOS focuses on the development and maintenance of statistical classification systems. It provides a structured means of tracking the development of the classification over time.

Acquire

DDI-Codebook provides general information on the source of the data found in the related data set. Descriptive information on the source (population studied, source of secondary data, and process of acquisition) is captured at the

study level. Specific information on the source question for survey data is captured at the variable level. Starting in version 2.6 specific information on how individual variables were selected or generated as the result of extraction or analysis is captured at the data file level by supporting the use of SDTL, a structured language for capturing data processing.

DDI-Lifecycle provides the ability to fully capture the development and organization of a questionnaire. It captures detailed question information, the use of a question within a questionnaire, and full routing information. DDI-Lifecycle metadata can be used to management and create questionnaires in multiple modes and track data from its point of capture through the storage of data within a defined variable. This includes processing for derived variables, capture of non-question-based measurements. DDI-Lifecycle supports the development and maintenance of Question Banks as well as the generation of multi-model and multi-lingual questionnaires based on the consistent content for the question and questionnaire structure.

SDTL (Structured Data Transformation Language) was designed as an independent intermediate language for representing data transformation commands. Statistical analysis packages (e.g., SPSS, Stata, SAS, and R) provide similar functionality, but each one has its own proprietary language. SDTL consists of JSON schemas for common operations, such as RECODE, MERGE FILES, and VARIABLE LABELS. SDTL provides machine-actionable descriptions of variable-level data transformation histories derived from any data transformation language. Provenance metadata represented in SDTL can be added to documentation in DDI and other metadata standards. SDTL can be used within processing structures in DDI-Codebook, DDI-Lifecycle, and DDI-CDI.

Process & Analysis

DDI-Lifecycle focuses much of its processing content on the structure of a questionnaire and the movement of data from the point of capture to the point of storage. The details in this content support creating, fielding, and processing questionnaires and the resultant data. Version 3.3 added support for describing and implementing sampling activities. It captures simple and complex sampling strategies and the issues of sampling across longitudinal studies. In addition, by capturing input/output flows of data through capture and processing, DDI-Lifecycle provides complex provenance information at the level of the individual datum. SDTL can be used at many points to capture detailed computations in processing and analyzing data.

DDI-CDI extends process information capture into generalized process systems. [provide examples]

SDTL (Structured Data Transformation Language) was designed as an independent intermediate language for representing data transformation commands. Statistical analysis packages (e.g., SPSS, Stata, SAS, and R) provide similar functionality, but each one has its own proprietary language. SDTL consists of JSON schemas for common operations, such as RECODE, MERGE FILES, and VARIABLE LABELS. SDTL provides machine-actionable descriptions of variable-level data transformation histories derived from any data transformation language. Provenance metadata represented in SDTL can be added to documentation in DDI and other metadata standards.

Share / Disseminate:

A key goal of all DDI products is support for accurate discovery data and the ability to share, disseminate and analyze that data.

DDI-Codebook was originally developed to provide a consistent structured codebook for social science data to facilitate the ability of archives and producers to share and disseminate data and metadata for analysis using common statistical software. Key elements were tagged to allow accurate translation into and out of this software. Additional content was provided in human readable structures to provide analysts with the information needed to understand the provenance, strengths, and limitations of the data. DDI-Codebook has been enhanced over time to increase the use of controlled vocabularies for cross-study comparability at increasing levels of detail. Version 2.6 has expanded the description of data access with additional structured content for automated access and added similar description for metadata access. This information can now be referenced by the study, data file, variable, variable groups, questions, and dimensional data structures down to individual categories in value domains.

DDI-Lifecycle provides additional structure to similar information and expands support for the reuse of metadata as well as data providing additional options for discovery, sharing, and dissemination. DDI-Lifecycle took the content of DDI-Codebook and provided the structure needed to accurately program access to the data and metadata represented there.

DDI-CDI is designed for the purpose of cross-domain integration. It is modeled to support the goals of discovering and accessing data from multiple disciplines and providing the pieces of information required to accurately integrate that data in an automated way. The datum is core to the DDI-CDI model with clear links to the information within the Variable Cascade, Unit Cascade, and Value Domains. It includes processing information on the creation of the data as well as processing integration actions. Links are provided to existing DDI-Codebook or DDI-Lifecycle metadata for added information on provenance, methodologies, and data capture processes.

XKOS manages the management, versioning, and publication of Statistical Classifications. The online provision of structured RDF/OWL statistical classifications encourages their use by a wide range of data producers. Statistical classifications published by the managing organization ensure that the content is up-to-date supporting access to and integration of data from multiple sources.

SDTL tracks the relationship between original data sets and their subsidiaries by capturing the process of data selection, integration, and the generation of new data sets.

DISCO, an unpublished product of DDI, is an RDF expression of key elements from DDI-Codebook and DDI-Lifecycle commonly found in search systems used to identify, locate, and access data.

Archive

DDI-Codebook was developed by data archivists and producers initially for use within the data archive community. In addition to the traditional “codebook” content of variable names, labels, universe, unit type, value domain, and storage locations, DDI-Codebook provided information on the storage location of the data set, the purpose of the data, topical, spatial, and temporal coverage of the data, source and producer information, and information on the archive managing the data and making it available to researchers. Throughout DDI-Codebook, the archive was able to tag added content provided by the archive and differentiate it from that obtained from the source.

DDI-Lifecycle added an additional section of metadata that allowed the archive to track and manage their metadata and provide a complete provenance chain as well as access to the individuals and organizations who contributed to or

managed the data and metadata. It allows capture of access restrictions imposed by the producer or implemented by the archive and links content elements to those restrictions or procedures for access. DDI-Lifecycle also added a section that supports clear content management of the deposited metadata as well as the additions or alterations made by the archive over the life of the data/metadata.