MRT: DDI 4 Core Definition of Scope

Ottawa Sprint, 22-24 April 2019

# Overview

This document provides a description of the scope of the work in the MRT Working Group on the DDI 4 Core model. This will be the content of a first production release, following the prototype review of 2018.

The intention is to take a subset of the overall model, and incorporate feedback and comments from the prototype review, so that a useful set of functionality can be brought to market in a short timeframe.

The release date for the DDI Core model and bindings (XML and RDF) is the end of December, 2019. At this time, it is intended that a final version of the DDI 4 Core specification be available for review, and that a production release would follow after issues raised during a subsequent review have been addressed.

The portions of the model which have not been included in the core release will be addressed in subsequent work. The core is meant to provide support for some of the most common functions for which DDI is intended to be used. Significant areas of functionality are not addressed (Data Capture is a good example of this – there are others). While also extremely important, a more incremental approach to providing support for these functions has been chosen.

The working style of the MRT will be to create an initial core release, and to work iteratively to support additional functionality in subsequent releases. Testing and implementation will form part of this cycle. Thus, each delivered part of the model will need to support a meaningful set of functionality from the user’s perspective.

For the DDI 4 Core, the chosen functionality includes data description. It is the nature of the DDI model that this central functionality requires a substantial amount of foundational material, addressing data structures at the logical and physical level, conceptual material, representations of variables, and the relationships of variables at several levels.

The DDI 4 Core will also include support for some applications of the process model. Specific cases regarding data lineage and production processes will be specified. Because neither Data Capture nor Methodology are to be undertaken at this time, however, the application of the process model will be somewhat limited. The fuller use of this model will be included in future releases, but initial implementation of the Data Management View from the prototype release have shown that there is immediate utility in including at least some applications of the process model in the initial DDI 4 Core release.

# The DDI 4 Lion Repository and DDI 4 Core

The DDI 4 work has been using the Lion Repository as a collaborative working platform for the work leading up to the prototype release. The MRT work will focus on a narrower scope and will be using a different modelling platform.

The modelling environment and production platform used for the MRT work has been integrated with that used by the Technical Committee, so that a single production flow will exist at the time that the MRT work goes into production. In most cases, the same tools are being used for producing bindings and documentation, and relevant artefacts can be passed back and forth between these teams without any difficulty.

As a first step, the complete DDI 4 model as contained in the Lion Repository will be extracted and then further refined to contain only those portions of the model identified in this scope document.

# Supported Functionality

Pending a more complete specification of user requirements, the functionality to be supported in the DDI 4 Core model has been decided through discussion among those concerned in the work. This is seen as acceptable because up to this point formal requirements have not been a major feature of DDI development. It is seen as a potentially desirable activity in future, as it would help to inform the further development of DDI work products generally, the DDI 4 Core included, but this effort is not part of the scope of the current MRT work at this time.

There are some features which are seen as easy to include and of sufficient utility to users to be selected, even though they may not represent the more complete coverage of these topics in the full DDI 4 model. In this sense, the scope of the work has been determined opportunistically. The goal is to maximize the utility of the DDI 4 Core release to the user community.

An example of this is seen with methodology. The full DDI 4 model contains both a Methodology Pattern and an implementation of it for Sampling. These have not been included in the DDI 4 Core. Because of the strong relationship between process description and methodology, however, some aspects of methodology are required for a useful implementation of the process model. These have been restricted to summary descriptions and external references for the immediate term.

It is anticipated that those interested in cross-domain data integration will be potential users of the new DDI 4 Core, as has been suggested by discussions at the Dagstuhl workshops on this subject. For this audience, the description of data, foundational metadata, and process description are all of interest. Alignment with some other popular standards is also important in engaging these users.

The following list provides a summary of the functionality which will be supported as forming the most useful core functionality for DDI users:

* **Describing data** – users will be able to describe data sets (variables, representations, associated definitions/concepts, identification). This covers the data dictionary portion of a codebook.
* **Describing concepts and their relationships (in non-data-description) roles** – concepts and concept systems are used in many specific ways related to data, but not specifically in the roles
* of categories or variables. Examples are their use in specialized geography descriptions, describing coverage, and supporting data discovery.
* **Describe logical organization of data** – the logical organization/structuring of data in several common forms (rectangular/unit-record data, cube/aggregate/time series data, event data) with reference to the logical content, but independent of how it might be physically formatted/stored.
* **Describe physical organization** – the physical organization/structure of data.
* **Description of transformation/relationships between data organization styles, to accommodate different uses of the data** – the use of the datum as a pivot point between different styles of data organization/structure, informing how the same datums can perform different roles in different uses of the data (flexible viewpoints used to assign roles to different variables for different purposes).
* **Describing data lineage processes** – the actions performed on data set to produce related data, with summary information regarding methodology and data capture, but without a full description of these in a detailed way (this is an extension point for further work).
* **Describing collections of items and their internal structures/relations** – A general pattern of collections will be presented, along with some specific implementations of the pattern as relevant to support other stated functionality.
* **Alignment with selected common standards** – [Add ISO 19115 family] Some specific standards have been identified which will be used in combination with DDI. The alignment with these standards needs to be supported, as appropriate to their intended use. These include the DCAT vocabulary, the PROV-O vocabulary, and the GSBPM. In some cases (PROV-O, DCAT) there are hand-off point between the standards which should be identified to enable their combined use. In the case of GSBPM, the description of data lineage might consist of a process which navigates the GSBPM framework, in which case they are complementary descriptions of the same basic information, used for different purposes (DDI for processing/documentation, GSBPM for communication).

# DDI 4 Core Packages and Classes

This section details the contents of the initial package pulled from the Lion Repository, to include *at least* those needed to support the functionality described above. This section is intended to rercord what was done, and to facilitate further work on the DDI 4 Core model, which will start from the set of classes listed, but will very likely exclude many of them if they are unused. Open issues where new classes will likely be needed are indicated in the first section – these are the subject of analysis proposals found in other working MRT documents.

## Open Issues for December Release

From Datum-based Examples work:

* For Tall format Value columns, we need a generic datatype to describe datums from different instance variables
* For Tall format VariableRef columns, we need to be able to reference Value Mappings (and by association to Instance Variables)
* For Viewpoints applied to Data Cubes, we need to add a **dimension** role.
* Should a variable be allowed to have more than one role in the same Viewpoint?

From other places:

* How to handle Annotations and Citation Information if Identifiers change as a result of modelling work?

As an initial activity, the contents of the Prototype were reviewed and an initial pull was made from the Lion Repository to include *at least* the full set of classes to be included in the DDI 4 Core. (Because changes are being made to resolve modeling issues this includes some classes on which there are currently dependencies which may or may not be included in the published DDI 4 Core.)

At the same time, a more complete extraction was made from the Lion Repository of all of the Prototype Review classes as well as other parts of the model which were not published for review.

## Initial DDI 4 Core Extraction

The following table is a list of all classes extracted for inclusion in the initial pull, organized as an alphabetical inventory. Note that many of these classes are included as a result of dependencies on other objects in their packages, and that the contents of the production DDI 4 Core model are likely to be different.

|  |  |  |
| --- | --- | --- |
| **Class Name** | **Definition (Partial)** | **Package** |
| Access | Describes access to the annotated object. This... | Discovery |
| Act | An Act is an indivisible, atomic step, i.e.... | Workflows |
| Agent | An actor that performs a role in relation to a... | Agents |
| AgentListing | A listing of Agents of any type. The AgentList... | Agents |
| AgentRelationStructure | Defines the relationships between Agents in a... | Agents |
| Algorithm | An algorithm is an effective method that can... | MethodologyPattern |
| AlgorithmOverview | High level, descriptive, human informative,... | SimpleMethodologyOverview |
| AnnotatedIdentifiable | Used to identify objects for purposes of... | Identification |
| AppliedUse | Links the guidance instructions to specific... | Methodologies |
| AttributeRole | An AttributeRole identifies one or more... | LogicalDataDescription |
| AuthorizationSource | Identifies the authorizing agency and allows... | Representations |
| BoundingBox | A type of Spatial coverage describing a... | Discovery |
| BusinessAlgorithm | A Business Algorithm is used to express the... | BusinessWorkflow |
| BusinessFunction | Something an enterprise does, or needs to do,... | Methodologies |
| BusinessProcess | BusinessProcesses could be Generic... | BusinessWorkflow |
| Category | A Concept whose role is to define and measure... | Conceptual |
| CategoryRelationStructure | Relation structure of categories within a... | Conceptual |
| CategorySet | Specialization of a Concept System focusing on... | Conceptual |
| ClassificationFamily | A Classification Family is a collection of... | Representations |
| ClassificationIndex | A Classification Index is an ordered list (... | Representations |
| ClassificationIndexEntry | A Classification Index Entry is a word or a... | Representations |
| ClassificationItem | A Classification Item represents a Category at... | Representations |
| ClassificationItemRelationStructure | A complex RelationStructure for use with... | Representations |
| ClassificationSeries | A Classification Series is an ensemble of one... | Representations |
| ClassificationSeriesRelationStructure | Describes the complex relation structure of a... | Representations |
| Code | A type of Designation that relates a... | Representations |
| CodeList | A list of Codes and associated Categories. May... | Representations |
| CodeRelationStructure | Relation structure of codes within a codelist... | Representations |
| CollectionMember | Generic class representing members of a... | CollectionsPattern |
| Comparison | The minimal pattern for a comparison including... | CollectionsPattern |
| ComparisonMap | Provides a basic pattern for a comparison map... | CollectionsPattern |
| ComputationAction | Provides an extensible framework for specific... | Workflows |
| Concept | Unit of thought differentiated by... | Conceptual |
| ConceptRelationStructure | Relation structure of concepts within a... | Conceptual |
| ConceptSystem | A set of Concepts structured by the relations... | Conceptual |
| ConceptSystemCorrespondence | Relationship between 2 or more ConceptSystems... | Conceptual |
| ConceptualDomain | Set of Concepts, both sentinel and substantive... | Conceptual |
| ConceptualVariable | The use of a Concept as a characteristic of a... | Conceptual |
| ConditionalControlStep | Type of WorkflowControlStep in which the... | Workflows |
| ControlledVocabulary | The specification of a controlled vocabulary... | CustomMetadata |
| CorrespondenceTable | A Correspondence Table expresses relationships... | Representations |
| Coverage | Coverage information for an annotated object.... | Discovery |
| CustomInstance | A set of CustomValues to be attached to some... | CustomMetadata |
| CustomItem | A custom item description. This allows the... | CustomMetadata |
| CustomItemRelationStructure | Contains a set of CustomItemRelations which... | CustomMetadata |
| CustomStructure | A Collection containing a set of item... | CustomMetadata |
| CustomValue | An instance of a key, value pair for a... | CustomMetadata |
| DataPipeline | A DataPipeline is a single traversal of the... | BusinessWorkflow |
| DataPoint | A DataPoint is a container for a Datum. | LogicalDataDescription |
| DataPointRelationStructure | A means for describing the complex relational... | FormatDescription |
| DataStore | A DataStore is either a SimpleCollection or a... | LogicalDataDescription |
| DataStoreLibrary | A DataStoreLibrary is a collection or, again,... | LogicalDataDescription |
| DataStoreRelationStructure | A structure for describing a complex relation... | LogicalDataDescription |
| Datum | A Datum is a designation (a representation of... | LogicalDataDescription |
| Design | The design pattern class may be used to... | MethodologyPattern |
| Designation | A sign denoting a concept. | Representations |
| DesignOverview | High level, descriptive, human informative,... | SimpleMethodologyOverview |
| EnumerationDomain | An abstract base to allow all codifications (... | Representations |
| ExternalMaterial | ExternalMaterial describes the location,... | Utility |
| GeographicExtent | Defines the extent of a geographic unit for a... | GeographicClassification |
| GeographicUnit | A specific geographic unit of a defined Unit... | GeographicClassification |
| GeographicUnitClassification | Describes the classification of specific... | GeographicClassification |
| GeographicUnitRelationStructure | Defines the relationships between Geographic... | GeographicClassification |
| GeographicUnitTypeClassification | A structured collection of Unit Types defining... | GeographicClassification |
| GeographicUnitTypeRelationStructure | Defines the relationships between Geographic... | GeographicClassification |
| Goal | Goals are the "things" a method aims to... | Methodologies |
| Guide | Provides a guide for the usage of a result... | Methodologies |
| Identifiable | Used to identify objects for purposes of... | Identification |
| IdentifierRole | An IdentifierRole identifies one or more... | LogicalDataDescription |
| IfThenElse | IfThenElse describes an if-then-else decision... | Workflows |
| IndexEntryRelationStructure | Structures relationship of Classification... | Representations |
| Individual | A person who may have a relationship to... | Agents |
| InstanceVariable | The use of a Represented Variable within a... | Conceptual |
| InstanceVariableRelationStructure | A realization of RelationStructure that is... | LogicalDataDescription |
| LevelStructure | The LevelStructure describes the nesting... | Representations |
| LogicalRecord | The LogicalRecord is a record definition. It... | LogicalDataDescription |
| LogicalRecordRelationStructure | Allows for the complex structuring of... | LogicalDataDescription |
| Loop | Iterative control structure to be repeated a... | Workflows |
| Machine | Mechanism or computer program used to... | Agents |
| MeasureRole | A MeasureRole identifies one or more... | LogicalDataDescription |
| MemberIndicator | Provides ability to declare an optional... | CollectionsPattern |
| MemberRelation | Defines one kind of relationship between one... | CollectionsPattern |
| MetadataDrivenAction | MetadataDrivenActions are Acts in which their... | Workflows |
| Methodology | Methodology brings together the design,... | MethodologyPattern |
| MethodologyOverview | High level, descriptive, human informative... | SimpleMethodologyOverview |
| Organization | A framework of authority designated to act... | Agents |
| Parameter | An Input or Output to a Process Step defined... | Workflows |
| PhysicalDataSet | The information needed for understanding the... | FormatDescription |
| PhysicalLayoutRelationStructure | A realization of RelationStructure that is... | FormatDescription |
| PhysicalOrderRelationStructure | PhysicalStructureOrder orders... | FormatDescription |
| PhysicalRecordSegment | A description of each physical storage segment... | FormatDescription |
| PhysicalSegmentLayout | The PhysicalSegmentLayout is an abstract class... | FormatDescription |
| PhysicalSegmentLocation | Among other things defines the location of a... | FormatDescription |
| Population | Set of specific units (people, entities,... | Conceptual |
| Precondition | A precondition is a state. The state includes... | Methodologies |
| Process | Process is an implementation of an algorithm.... | ProcessPattern |
| ProcessControlStep | A Process Step that controls the ordering of... | ProcessPattern |
| ProcessOverview | Process is an implementation of an algorithm.... | SimpleMethodologyOverview |
| ProcessStep | One of the constituents of a Process. It can... | ProcessPattern |
| RecordRelation | The RecordRelation object is used to indicate... | LogicalDataDescription |
| RelationStructure | The set of MemberRelations used to structure a... | CollectionsPattern |
| RepeatUntil | Iterative control structure to be repeated... | Workflows |
| RepeatWhile | Iterative control structure to be repeated... | Workflows |
| RepresentedVariable | A combination of a characteristic of a... | Conceptual |
| Result | Describes the results of a process for the... | Methodologies |
| SegmentByText | Defines the location of a segment of text... | FormatDescription |
| SentinelConceptualDomain | Description or list of possible sentinel... | Conceptual |
| SentinelValueDomain | The Value Domain for a sentinel conceptual... | Representations |
| Service | A means of performing a Process Step as part... | ProcessPattern |
| Sign | Something that suggests the presence or... | SignificationPattern |
| Signified | Concept or object denoted by the signifier... | SignificationPattern |
| Signifier | Concept whose extension includes perceivable... | SignificationPattern |
| SimpleCollection | Simple Collection container (set or bag) that... | CollectionsPattern |
| SpatialCoverage | A description of spatial coverage (geographic... | Discovery |
| Split | The components of a Split consists of a number... | Workflows |
| SplitJoin | SplitJoin consists of process steps that are... | Workflows |
| StandardWeight | Provides an identified value for a standard... | SimpleCodebook |
| StatisticalClassification | A Statistical Classification is a set of... | Representations |
| StatisticalClassificationRelationStructure | A structure for describing the complex... | Representations |
| StructuredCollection | Structured Collection container extends a... | CollectionsPattern |
| SubstantiveConceptualDomain | Set of valid Concepts. The Concepts can be... | Conceptual |
| SubstantiveValueDomain | The Value Domain for a substantive conceptual... | Representations |
| TemporalCoverage | Describes the temporal coverage of the... | Discovery |
| TemporalRelationControlStep | Defines complex synchronous or asynchronous... | Workflows |
| TopicalCoverage | Describes the topical coverage of the module... | Discovery |
| Unit | The object of interest in a process step... | Conceptual |
| UnitDataRecord | Gives a UnitDataRecord structure to a Logical... | LogicalDataDescription |
| UnitDataViewpoint | The assignment of measure, identifier and... | LogicalDataDescription |
| UnitSegmentLayout | UnitSegmentLayout supports the description of... | FormatDescription |
| UnitType | A Unit Type is a type or class of objects of... | Conceptual |
| Universe | A defined set or class of people, entities,... | Conceptual |
| ValueAndConceptDescription | A formal description of a set of values. | Representations |
| ValueDomain | The permitted range of values for a... | Representations |
| ValueMapping | Provides physical characteristics for an... | FormatDescription |
| VariableCollection | A collection (group) of any type of Variable... | Conceptual |
| VariableRelationStructure | RelationStructure for use with any mixture of... | Conceptual |
| VariableStatistics | Contains summary and category level statistics... | SimpleCodebook |
| ViewpointRole | A ViewpointRole designates the function an... | LogicalDataDescription |
| VocabularyEntry | One entry term and its definition in an... | CustomMetadata |
| VocabularyRelationStructure | Contains the Vocabulary Relations used to... | CustomMetadata |
| WorkflowControlStep | A subtype of WorkflowStep which controls the... | Workflows |
| WorkflowProcess | A Workflow Process is a realization of Process... | Workflows |
| WorkflowService | A means of performing a Workflow Step as part... | Workflows |
| WorkflowStep | One of the constituents of a Workflow. It can... | Workflows |
| WorkflowStepSequence | A WorkflowStepSequence controls the order of... | Workflows |

## Prototype Review Classes Excluded from the DDI 4 Core Initial Extraction

This section describes which packages, classes, and relationships, included in the Prototype Review, were not included in the initial pull from the Lion Repository. These were manually trimmed out of the initial pull as they would result in the inclusion of large numbers of unwanted classes.

**Relationships:**

***From InstanceVariable:***

Removed for Core Build as no capture information in included in core:

Relationship: sourceCapture (target Capture) - The source of capture for the values that populate this InstanceVariable. This may be any class using Capture as an extension base, currently RepresentedQuestion and RepresentedMeasurement. This is the direct source of capture for the content of the data. If a transformation is required the source is not multiple captures but the transformation.

***From DataStore:***

Removed the following content in preparation for Core Build:

Relationship: isInStudy (target Study) - A Study has at most one DataStore. Many studies can have the same set of record types.

***From DataStoreLibrary:***

Removed following from Core Build:

Relationship: isInStudySeries (target StudySeries) - Whereas a DataStore is associated with a Study, a DataStoreLibrary is associated with a StudySeries. Each StudySeries has at most one DataStoreLibrary.

***From DataPipeline:***

Removed the following for Core Build:

Relationship: isInStudy (target Study) - A study has at most one DataPipeline. The same DataPipeline can be used in many studies.

**Packages:**

The following packages found in the Prototype have been removed from Core Build

DataCapture

SamplingMethodology

StudyRelated

**Classes:**

The following classes have been removed:

From DataCapture:

 BooleanResponseDomain

 Capture

 CodeListResponseDomain

 ConceptualInstrument

 ExternalAid

 ImplementedInstrument

 InstanceMeasurement

 InstanceQuestion

 Instruction

 InstrumentCode

 InstrumentComponent

 NumericResponseDomain

 RankingResponseDomain

 RepresentedMeasurement

 RepresentedQuestion

 ResponseDomain

 ScaleResponseDomain

 Statement

 TextResponseDomain

From SamplingMethodology:

 SampleFrame

 SamplePopulationResult

 SamplingAlgorithm

 SamplingDesign

 SamplingGoal

 SamplingProcedure

 SamplingProcess

From StudyRelated:

 Budget

 ComplianceStatement

 Embargo

 ExPostEvaluation

 QualityStatement

 Standard

 Study

 StudyControl

 StudyRelationStructure

 StudySeries

From Utility:

 DocumentInformation

 FundingInformation

The following StructuredDataTypes have been moved to the unpublished package ClassesOnHOLD:

 ConditionalText

 ContentDateOffset

 DynamicText

 DynamicTextContent

 FixedText

 LiteralText

 StudyIndicator

 TargetSample

# Post-December Work: Next Steps

Some important parts of the DDI model have been excluded from the DDI 4 Core for the reasons given above. It is recognized that integrating these parts of the model is important as a next step, after the December 2019 delivery, so that the corresponding user functionality is supported. In discussion, these functional areas included the following:

* **Data Capture –** The collection of data from questionnaires, registers, and other sources is a major use-case for DDI. As modelled in the prototype, this is an important implementation of the process-related portions of the model. Although process is included, as implemented in some other applications, data capture is an important and high-priority addition. Some work has already been done on the integration of the data capture model with the data description model – this needs to be finalized and agreed and is estimated to be a significant amount of work.
* **Methodology –** The prototype contained both a pattern for describing methodology, and a set of concrete classes which are related the pattern. From these, a model exists specifically for sampling. More complete support for methodology is contemplated, with specifics and priority yet to be determined.
* **Study Description (Purpose, Data Dictionary, Related Information) -** The prototype contained a model of the Study which is the basis for many of the real-world “codebook” implementations of DDI seen in previous versions of the standard. While the DDI 4 Core contains the heart of the information needed (the data dictionary portion) it does not cover some other types of information needed to represent codebooks. For migration purposes, this is an important model to finalize. Note, however, that it is a non-trivial effort, as the range of codebook metadata expressed in earlier versions of DDI is extensive.

Other areas could also be usefully included in a next release – identifying these is not within the scope of current MRT efforts.

# Post-December Work: Longer-Term Functionality

There are many areas which have been considered for inclusion in DDI 4 over the course of its development. Of these, some have been the subject of considerable work, and others have not gone much beyond the planning stages.

Identifying these is not within the scope for current MRT efforts but should be considered at this point in general terms, to make sure that future directions are not in some way unintentionally limited. One activity has been to make sure that any of the content within the Lion Repository which may in future become part of a production DDI 4 model has been preserved, in the case that the Lion Repository itself is taken offline.

Functionality mentioned here may also become the focus of more immediate work – it was not included in the Prototype Review package, and so is assumed to be of lower priority for release. It has been the focus of much past work, however, and represents a significant resource should it become a priority for inclusion.

**Qualitative Data Description** – a great deal of work has gone into describing the different types of metadata associated with non-quantitative data, including the addressing of segments in various types of files such as images, interview transcripts, videos, etc. Linking various types of information to these, and further documenting the process by which they are quantified and further analyzed is also supported. The interest in standardization in this area has recently become a topic of discussion within the Social Sciences and related domains, reflecting renewed interest. The REFI standard (an effort driven by QDA) is the most prominent example of this: <https://www.qdasoftware.org/products-project-exchange/>.