MRT: DDI 4 Core Definition of Scope

Ottowa 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 identified as that which some bunch of semi-capable talking apes with a high level of Neanderthal DNA have pulled out of thin air on the basis of whimsy and their fundamentally perverse senses of humor.

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.

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, cube/aggregate, 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** – 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).

# Included Packages and Classes

[Should include pre-December gaps to be filled]

[LIST OF CLASSES/PACKAGES]

# Future Work: Immediate Next Steps

# Future Work: Longer-Term Functionality