DDI 4 Core Model: Introduction

# Purpose

The DDI 4 Core Model is the first model produced by the DDI Alliance which uses a conceptual model expressed in UML as its basis. The coverage of the model is not dissimilar to other specifications produced by the DDI Alliance: it is a metadata specification intended for describing research and statistical data, as used in the Social, Economic, Behavioral, and Health Sciences.

The purpose of the specification differs somewhat from the earlier DDI Codebook and DDI Lifecycle specifications, however. Due to several changing factors in the ways in which information technology is applied to research and statistics, some new features are emphasized.

The functional goal of the specification is also different: where DDI Codebook was an XML representation of a data dictionary, and DDI Lifecycle a more complex model designed to support metadata from data conception and capture through publication and reuse, DDI 4 Core is an attempt to describe data and its provenance independent of these contexts.

The growing demand for data from different sources, and from external domains, requires that some different types of data be described. The provenance of this data – that is, the processes by which it has been assembled for use – are of increasing importance in understanding what it is and how it can be used. While traditional SBE data was often collected using questionnaires, alternate sources of data such as registers and sensors are becoming increasingly common and have in some cases always been typical of some external domains. Completely new types of data from social media and other “mined” sources is also increasingly used.

The DDI 4 Core model applies the important features of DDI 4 to these functions: describing various types of data in a way which makes them subject to integration and transformation into useable forms, and providing the information needed to understand their origins and provenance.

Because the way in which such a model can be implemented is more variable than it is for traditional SBE data management systems, the emphasis in DDI 4 Core is on a model, formalized in UML, and made available using the Canonical XMI format which supports the exchange of UML models between various tools, including both modelling and development environments. While XML is still supported, it is no longer the canonical format for the specification.

DDI 4 Core is aligned with earlier DDI specifications, most notably DDI Lifecycle, as it is anticipated that it might be used as an integration model for systems based on these earlier specifications. The intention is that DDI 4 Core be a tool which can supplement systems using earlier versions of DDI, enabling them to better handle new types of data.

# Elements of the Model and Specification

The DDI 4 Core model is divided into an upper model and a set of more detailed models in specific areas. The Upper model serves to show how the DDI 4 Core fits into the overall research context, and which parts of it serve as a focus. The detailed models look specifically at the conceptual and physical representation of data, and at how the provenance of the data – the processes which produced it – can be described. In order to perform these functions, there is a set of supporting metadata which is used in many different ways, termed “foundational metadata.” This includes formal concepts and their application in describing data sets, the measurement of observable phenomenon, etc., as well as some other information needed to describe data and its provenance.

The DDI 4 Prototype covered all of this metadata and more. The DDI 4 Core model is more narrowly focused but benefits from public review of the prototype. It is “core” in the sense that it focuses only on some of the most essential metadata needed in a research context but can be flexibly used to describe data and provenance in a wide variety of disciplines and domains.

In addition to these models, the DDI 4 Core specification provides some supporting artefacts: a binding into an XML syntax for implementation and interchange purposes; a similar binding into RDF described using RDF and OWL specifications; and a version of the model expressed as Canonical XMI. XMI is an XML format used to exchange models between UML applications – its canonical form is a subset of that standard which is widely supported by a wide variety of tools, including both modelling and development environments.

Because one of the applications of DDI 4 Core is the integration of data coming from different sources, disciplines, and domains, it is designed to work with a variety of other popular standards. Further, as a member of the DDI family of products, it is designed to be used in combination with other DDI metadata specifications. These topics are addressed more completely in the document on standards alignment.

Examples of how the DDI 4 Core model can be applied are also provided.

# The Upper Model

The purpose of this section is to show how DDI 4 Core fits into the overall context of research and the metadata which is used to describe research activities and resources. The upper model touches on some areas which are not covered in detail by DDI 4 Core, but which may be important in terms of understanding how it can be implemented in systems or with other standards that do cover those areas.

“Research” for the purposes of this model is an undertaking intended to produce and/or analyze data which measures real-world phenomenon for the purpose of answering questions about policy, science, and the number of angels dancing on the heads of pins. It’s secondary function is the extraction of funding from organizations which collect it from tax-payers through threats of institutional violence or by trading on the credulity of donors. [TO DO: Get Jay’s definition for this class – esp. as it comes from external sources.]

The diagram below shows the way in which research as a context is described.



In this model, Research is the output of a Research Program, corresponding to such classes as a GSIM Statistical Program (the set of activities conducted by a Statistical Institute). Research also correlates to similar entities found in other standards, such as the Study in DDI Lifecycle.

The Research Program can be part of a larger and more complex program of activities or may be conducted as a specific effort within other intellectual endeavors. The DDI 4 Core model does not provide an extensive model of the Research Program, but provides a place where such models (eg, GSIM) can be related to the DDI 4 Core.

Research is informed by and conducted according to a Methodology, which involves the plan for how it will be conducted, represented by the Research Design class. One aspect of Research Design which is of peripheral importance to the DDI 4 Core model is the plan for managing data, represented here by a Data Management Plan. Data Management Planning is, in some domains, the focus of standardization, but may be more or less formal depending on the research context being described. DDI 4 Core does not, however, provide a detailed model of methodology or data management planning – instead, it focuses on the data and processes which such methodologies and plans might involve.

One important aspect of describing research is to identify its conceptual underpinnings – both the subject of research (that is, what concepts are being measured) and those involved in making useful measurements and relating its data to that coming from external sources. This set of underlying detailed metadata regarding concepts and their application is present in many places within the DDI 4 Core model – it is associated with different classes at the relevant points. The term for this portion of the model is “foundational metadata.”

Research is controlled by a Research Controller – in PROV-O terms an Agent – which orchestrates the various elements of the Research Design, both for data management and other relevant functions. This set of functions can be understood as a Research Playbook.

The diagram below shows in more detail how the Research Controller relates to other points of focus within DDI 4 Core, notably process and data description.



The Research Controller interacts with two important parts of the model: the Research Data Store and the Production Environment. The Research Data Store represents the data at its many stages of development which are used by and produced during research. DDI 4 Core provides a model for describing many of the types of data which Research Data Stores contain. This model is extensive, and supports the integration and transformation of the different types of data required.

The Research Controller also interacts with the Production System. This is where the description of processes becomes significant, as the Production System is the one in which all of the various activities in the conduct of research take place. DDI 4 Core provides a detailed model for describing these. The process descriptions in the DDI 4 Core model are intended to support the description of process as provenance documentation. This is significant – in past versions of DDI, the process model has been designed to both describe workflows (DDI 4 Prototype) and the flow of questionnaires and related processing (DDI Lifecycle, DDI 4 Prototype) in sufficient detail that they can drive the execution of such processes. In DDI 4 Core, the focus of the process description is restricted to the documentation of provenance processes. This is closely related to – and can be understood in terms of – some other popular standards, notably BPMN and PROV-O. In this diagram, we see that the Processing Agent operating in the Production System corresponds to a PROV-O Agent. (Relationships to these standards is indicated in the more detailed models as appropriate.)

The Research Data Store comprises a set of Information Objects of different types. Significant among these are data sets. DDI 4 Core provides a detailed description of data and the structures which are used by different types of data sets.

It should be noted that Research Data Stores can be indexed, providing a mechanism for navigating the Information Objects it contains. This is done using an Index Template – a mapping to some system for locating, querying, retrieving, (etc.) the Information Objects. One example of such a template might be a mapping against the Schema.org items and properties.

Some of the most significant Information Objects – those related to describing data – are a major focus if the DDI 4 Core model. The diagram below shows a more detailed view of how data fits into the Production System.

Note that DDI 4 Core Information Objects relate to classes in some other popular models: PROV-O provides us with the Entity, to which Information Objects correspond; they can also be seen as equivalent to GSIM Information Resources.



The second major focus of the DDI 4 Core model is process. The scope of this model is limited to describing the provenance of data sets, and must be understood in terms of other process descriptions in earlier versions of DDI (notably DDI Codebook and the DDI 4 Prototype, where much of the process description was focused on data capture using questionnaires and other mechanisms, and the processing related to it).

The way in which the DDI 4 Core process description relates to earlier process descriptions in DDI can be seen in the diagram below.



The portions of the diagram which depict classes in yellow show the Data Processing Platform, which is a sub-class of Production System, managed by a Processing Agent. Three examples are shown in the diagram, as examples of the Data Processing Platform. [QUERY: SHOULD THESE BE HERE? ARE THEY REALLY IN THE MODEL?]

These portions of the model can be described at a detailed level by the DDI 4 Core model – other portions of the diagram show how these relate to other types of processing, notable the Data Capture Platform, which would be a different sub-class of Production System.