

DDI-CDI: A New Type of Metadata Specification for Data Sharing and Integration

Arofan Gregory, ilg21@yahoo.com

Abstract

DDI Cross-Domain Integration (DDI-CDI) is a new work product from the DDI Alliance. It is designed to enable the description of disparate data types, and their integration at a structural level. In the modern age, many different types of data are used to complement or replace traditional survey data, including administrative registers, sensors, social media data ("big data") and others. DDI-CDI provides an ability to integrate these different types of data within systems, relying on the description of how variables, concepts, and different types of classifications and controlled vocabularies are associated with data in different structures. At the heart of this model is the variable cascade, based on the GSIM model.

The DDI-CDI model is first and foremost a UML model, with a broad range of possible syntax representations, including a canonical XML, but also RDF, JSON, and others. It is intended to act as a companion-piece to other more comprehensive metadata models, allowing data to be exposed, reused, and losslessly exchanged by systems which do not share the same underlying data structures. There is an increased demand for the sharing of data coming from the scientific community and the data produced by official statistical organizations, and DDI-CDI hopes to provide a basis on which such exchanges can be better supported. It acts as a supplement to, rather than a replacement for, other common standard metadata models.

This presentation will describe the core features of the DDI-CDI model, as well as its relationship to the standards within the Modern Statistics community to which it relates (i.e., GSIM, GSBPM, SDMX). The real-world need for this functionality will also be covered in terms of how data is increasingly being shared between official statistics and scientific research. Some early implementations will also be described.