2021 Further Development of the DDI Cross Domain Integration Model for FAIR Data Sharing across Discipline and Domain Boundaries

Introduction and Motivation

The Data Documentation Initiative (DDI) Alliance has been a leader in setting metadata standards for the social, behavioural, and economic sciences (SBE) for many years. They have provided specifications which support data collection, management, and dissemination with detailed descriptions of the data typical of those domains. As with many other branches of statistics and research, however, the type, volume, and sources of data have multiplied in the recent past. Many projects are now cross-disciplinary, involving data from different domains. At the same time, computational approaches to analysis of data and the reproduction and origination of research has evolved. These factors combine to highlight the need for an enhanced ability to integrate and understand data across domain boundaries, and to understand the provenance and processing of data, even as more and more of the work is performed programmatically by systems which leverage machine learning and other advanced technology approaches.

The DDI Alliance has recently published a new specification intended to fill this need for integrating data from disparate sources: DDI - Cross Domain Integration (DDI-CDI). Unlike earlier DDI work products, DDI-CDI is not domain-specific, but is designed to be used with

Workshop Summary


DDI-CDI FAIR 0.docx
18 Oct 2021, 09:33 AM

Date and Location

The workshop takes place at Schloss Dagstuhl – Leibniz Center for Informatics on September 20 to 24, 2021. It has the Dagstuhl event number 21383 and a related web page.

See the separate pages with **practical information** and information about **COVID-19**.

Topics

- Modular approach
- Data structure components

research data from any domain. The specification provides a model for understanding and integrating data across a wide range of sources, including big data/no SQL, event history and register data, traditional columnar data, and multi-dimensional data. Further, it provides a way of describing data provenance, with a focus not only on traditional linear processes, but also on declarative "black box" processes employed by many modern systems. DDI-CDI is intended not to replace traditional domain models for data description, but to supplement them when data from different sources and of different types is being integrated. It is designed to work easily with many other popular standards and models, including semantic vocabularies and generic technology specifications for data processing, dissemination, and cataloguing.

With an expected production release at the end of 2021, the current draft of the specification is undergoing finalisation. This workshop will focus on issues of immediate importance leading up to implementation and subsequent revision of DDI-CDI. Experts in the standard and prospective implementers will be in attendance to help refine the development roadmap.

Interoperability, Sustainability, and Alignment with Other Standards

DDI - CDI is fundamentally a model which is intended to be implemented across a wide variety of technology platforms, and in combination with many other standards models, and specifications. To support this use, it is formalized using a limited subset of the Unified Modelling Language (UML). The model is provided in the form of Canonical XMI – an interchange format for UML models supported by many different modelling and development tools. Further, a syntax representation is provided in XML, so that direct implementation of the model is possible if needed.

- UML class model interoperable subset (UCMIS)
- Syntax representations of the model
- Implementation guides
- I-ADOPT and DDI-CDI

Related Material

- DDI Cross Domain Integration (DDI-CDI) Review
- Pre-release (2021-08-25) DDI-CDI model version
  - Enterprise Architect, Canonical XMI, Canonical XMI with unique names
  - Field-level documentation
- Page with links to the EOSC report and DDI-CDI documentation/specification

DDI-CDI Webinars (including slides and recordings)

- Introduction
- Data Description Part 1

The platform-independence of the model makes it more easily applicable across a broad range of applications and helps ensure that it will be sustainable even as the technology landscape evolves. DDI - CDI builds on many other standard models and is aligned with them where appropriate. This is shown in the model itself, where formalizations from other models and specifications are refined, extended, or directly used. The specification includes a description of what these other standards and models are, and how they are used in DDI - CDI.

Topics, Deliverables and Activities

(1) Modular approach: The goal is that specific modules can be used in a flexible way standalone, together with other DDI-CDI modules, or together with other specifications. The work will focus on identification of functional packages, defined function of packages, clear one-way dependencies between packages, separation between functional (core) packages/classes and supporting packages/classes.

Deliverable(s): detailed plan on the topic along the listed subitems, description of some related rules (as modeling guidelines)

Activities:

- Identify target functions for modules and areas of functional overlap
- Propose a set of packages to support functions as identified
- Design a mechanism/pattern to be implemented for cross-package dependencies
- Produce a draft model/example which implements the package structure and dependency mechanism (e.g., DCAT/Schema.org, PROV-O)
- Document the approach

- Data Description Part 2
- DDI Training Webinars 2021
- Data Integration: Using DDI-CDI with Other Standards

Organizers and Participants

Organizers

- Simon Cox, CSIRO Australia and W3C Dataset Exchange Working Group
- **Arofan Gregory, Consultant and DDI Alliance**
- **Simon Hodson, CODATA - Committee on Data of the International Science Council (ISC)**
- Steven McEachern, Australian National University and DDI Alliance
- **Hilde Orten, Norwegian Center for Research (NSD) and DDI Alliance**
- **Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences and DDI Alliance**

Participants list (tba)

(2) Data structure components (toolkit): Review an approach for building new data structure types (in addition to the existing traditional wide/rectangular data, long [event] data, multi-dimensional data, and NoSQL/key-value data). Possible additional data structure types include graphs, text, any object in a “cell” (tables, text, binary objects, arrays of arrays, etc.)

Deliverable(s): related description and guide, formal description of additional data structure types

Activities: Describing Graphs

- Identify the functional requirements of a structural DDI-CDI description of data expressed as a graph (what will the structural description be used for?)
- Work with example(s) to propose extensions to the existing DDI-CDI data structure descriptions (NGSI-LD, CSV on the Web WG approaches, possibly DataCube)
- Document proposal and examples

Activities: Describing Nested Arrays

Identify the functional requirements of a structural DDI-CDI description of data expressed as a set of nested arrays (what will the structural description be used for?) These datasets are often very large – what is the intended support? Subsets extracted using queries/services? Entire databases?

- Work with example(s) to propose extensions to the existing DDI-CDI data structure descriptions (NetCDF)
- Document proposal and examples

Activities: Documenting the “Toolkit”

- Identify which features of the DDI-CDI model are used in describing new/hybrid data sets
- Propose a methodology for combining these, based on some examples (taken from above but also including “hybrid” long/wide data, for example)

- Document the methodology in a form which could be included in the specification in future

---

(3) UML class model interoperable subset (UCMIS): The strict use of UCMIS enables a robust model which can be imported in many UML tools and represented in object-oriented syntax representations. The focus here will be the relationship to other specifications (in the light of the modular approach) on the model level and syntax representation level. See documentation and spreadsheet of previously named “Practitioner's Subset for Data Modeling”.

Deliverable(s): detailed description of using the UML approach with Abstraction stereotypes for relating to classes of other specifications, description of how this gets realized on the level of syntax representations

Activities:

- Review the existing draft and evaluate suitability for publication as a stand-alone work product
- Create examples of how Abstraction stereotypes can be used to relate to other classes in external specifications, and document the approach
- Create examples of how UCMIS binding into specific syntax representations can be expressed – document the approach

---

(4) Syntax representations of the model: Exploration and decisions on OWL/RDF-S, JSON-LD, SheX (as constraint language for RDF). The work will build on an existing mapping from UML to OWL/RDF-S.

Deliverable(s): documentation on decisions and mapping

Activities:

- Identify requirements for syntax expression in RDF and related technologies, based on existing examples (Helmholtz, NGSI-LD (INTERSTAT, github), etc.)
- Consider the existing draft and evaluate in terms of the identified requirements
- Modify mapping to reflect the findings

---

(5) Implementation guides: Identify the methodology by which a community of users will specify how they will employ the model in their own implementations, such that they become more easily interoperable. Intersection with other machine-processible descriptions of data-sharing resources and methods within the community will be a focus.

Deliverable(s): design and document the methodology for defining community implementation guides and provide whatever tools/templates/examples might be useful.

Activities:

- Identify the required functionality for Implementation Guides (e.g., specifying a subset of the model, indicating syntax expressions)
- Develop practical approaches to the creation of IGs – how the analysis within a community can be conducted and documented
- Draft documentation and templates for applying and publishing IGs

---

(6) I-ADOPT and DDI-CDI: The I-ADOPT specification provides a model of how data can be described with clusters of variables and captures information about the data which is not expressed in DDI-CDI. This activity is aimed at looking at the intersection of the two specifications and determining how they can best be

used to solve real-world problems in cross-domain data sharing.

Deliverable(s): a mapping of DDI-CDI and I-ADOPT, with a documented example or examples showing how the two specifications can be combined to support cross-domain data-sharing requirements.

Activities:

- Identify the use cases for which the combined use of DDI-CDI and I-ADOPT is appropriate
- Develop required mapping between DDI-CDI and the I-ADOPT metamodel
- Apply the model to example use cases
- Document the mapping and examples