DDI’s Current Product Line: DDI Codebook, DDI Lifecycle and related products

Wendy Thomas
IASSIST 2018
This is about DDI...

• ...and how it has responded to:
  • Community changes
  • Expanding needs
  • Technical changes
  • Data changes

• DDI started in the late 1990’s in response to the need for structured metadata by archives and producers in order to
  • Reuse a single source of metadata for different products
  • Provide a machine readable base of data to support search and retrieval of data
  • Encourage the creation of structured metadata by the data producer
What is DDI?

• A collection of products that supports the structured capture and use of metadata surrounding the creation, preservation, and dissemination of data in the social, behavioral, economic, and health sciences
  • The DDI Standard in various versions
  • Controlled Vocabularies
  • XKOS – an independent publication targeted to a specific community to support the management and publication of Statistical Classifications
  • DDI Agent Registry – to support the DDI Identification structure

• These products reflect the needs of the community as well as the technical environment in which they developed
### Type of use

<table>
<thead>
<tr>
<th>User Community</th>
<th>Archives/Preservation</th>
<th>Data Discovery Systems</th>
<th>Data Producers</th>
<th>Specialized Metadata Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Sciences</td>
<td></td>
<td></td>
<td>Economics</td>
<td>Educational Testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Comparative Surveys</td>
<td>Statistical Systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Health</td>
</tr>
</tbody>
</table>

### DDI Products

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DDI Codebook</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Beta</td>
<td>[XSD]</td>
<td>V 3.0</td>
<td>V 3.0</td>
<td>V 3.1</td>
<td>V 3.2</td>
<td>V 3.2</td>
<td>V 3.2</td>
<td>V 3.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDI Lifecycle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controlled Vocabularies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISCO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XNOS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agent Registry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDI 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Funding:

- US National Science Foundation
- Health Canada
- DDI Alliance

### Formats:

- DTD
- XML Schema
- UML
- RDF

### Comparatives Surveys

- Health Statistical Systems
DDI Codebook

• **Target:**
  • Human Reader

• **Perspective:**
  • Retrospective
  • Document
  • Descriptive
  • Discovery

• **Assumptions:**
  • Documents managed in XML
  • Codebook still the format
  • Content could be identified and used in different ways
  • XML was a version of another source
  • The data already existed
DDI Codebook coverage: 1996-2000

- Focus on unit data collected by surveys
- Core of information needed to inform the end user to support intelligent use of the related data
- Structure is focused on information that is presented in different formats for different uses (published codebook, set-up files for statistical software, programmer access)
- Background information on the development and implementation of the study is primarily found in external “Other” material
DDI Codebook development: 2000 - 2012

• Additional data types:
  • Expanded to cover statistical (tabular, aggregate, structured, dimensional) data

• Interaction with related communities:
  • Support for spatial search systems
  • Geographic information needed to integrate data into a GIS system
  • Addition of content to support additional GSBPM content

• Technical adaptation:
  • Ability to apply variable descriptions to more than data store by supporting recording of physical data location information outside of the variable
  • Support for broader use of Controlled Vocabularies
  • Content to support transition to DDI Lifecycle
  • Moved schema to GITHUB for development version control
First transformation period: 2003-2010

• Still things on the list to do:
  • Complex data files – formats other than archival formats
  • Repeated surveys
  • Questionnaire content and flow

• Added data producer needs:
  • Capture from the idea all the way through the data and metadata lifecycle
  • Content management
  • Quality control
  • Data capture and processing
DDI Lifecycle

• **Target:**
  • Human Reader, Computer Ingest

• **Perspective:**
  • Retrospective – but progressive
  • Document
  • Metadata driven statistics

• **Assumptions:**
  • Documents managed in XML
  • Reusability
  • Metadata as a product
  • Management of metadata
  • Grouping of studies
  • Focus on a slice – i.e. Questions
  • Same data stored in multiple structures
DDI Lifecycle coverage: 2003-2009

• Focus on capturing metadata at “point of origin” and building on it through the process resulting in a data set – content versioning

• To support the reuse of metadata particularly in repeated surveys

• Better support for structured conceptual data that could be reused and provide implicate comparability

• Support for the management of data files in archives supporting additional storage formats and archival information

• Clearer record descriptions and record linkages

• More background and development metadata brought into the structured content for potential reuse (data cleaning, recoding, derivation instructions, etc.)
DDI Lifecycle development: 2010 - 2018

• Additional depth in class types:
  • Expanded representations to cover scales, image based domains, direct use of geographic codes, etc.
  • Added question grids and blocks
  • Added non-question measures

• Interaction with related communities:
  • Addition of input / output parameters and binding (OWL-s) to track datum flow
  • Support for full ISO-11179 structure
  • Statistical Classification – XKOS content, GSIM structure

• Technical adaptation:
  • Consistency within types of groupings (Schemes, groups, etc.)
  • Unique element names within the set of schemes
  • Consistency for reference naming (adding abstract classes as needed)
  • Generation of documentation from schema and structured documents
  • Moved schemas to GITHUB for development version control
DDI 3.3 schema available at: https://bitbucket.org/DDITC/ddi-l_3

Formal review period will begin in the next 2 weeks
Controlled Vocabularies coverage: 2010 -

• Controlled Vocabularies (CVs) are created by the DDI to support commonly used vocabularies among DDI users and CESSDA members

• CVs are published at
  • http://www.ddialliance.org/controlled-vocabularies

• Currently published in:
  • Customized Genericode format (XML)
  • XLS spreadsheet
  • HTML version for viewing

• CVs may be used by any version of DDI or other standard
Controlled Vocabularies development: 2016-

• Development of close ties with CESSDA development in this area
• Movement to a new development and management platform created within the CESSDA work plan
• Expansion of bindings to include:
  • SKOS
  • DDI Lifecycle CodeList
  • DDI 4 Custom Vocabulary
XKOS coverage: 2011-2016

• RDF Vocabulary
• XKOS standardizes the representation of statistical classifications as linked metadata
• Builds on the SKOS W3C Recommendation and implements the Neuchâtel and GSIM statistical models
XKOS development: 2017-2018

• Resolution of comments from the 2016 public draft review
• Refinements and documentation
• Publication of XKOS v1.2 in June 2018
Second Transformation

• Expanded User Community – new needs:
  • Metadata driven statistical production
  • Data integration
  • Mixed capture research
  • Automate DDI production process

• Technology was changing...even faster
  • RDF, Linked data
  • Metadata managed in data bases – still need to transfer
  • New data storage and access structures
  • Unstructured/undocumented data sources
DDI 4

• **Target:**
  - Human, Computer Understanding

• **Perspective:**
  - Prospective, Event, Retrospective
  - Descriptive + Computational
  - Data Linking

• **Assumptions:**
  - Metadata managed in databases
  - Multiple bindings
  - Transport and preservation medium
  - Metadata captured at origin
  - Access from the study/research area down and the datum up
DDI 4 development: 2012-2018

• Production process
  • UML to documentation content and multiple bindings

• Patterns
  • Structural consistency for collections
  • Expanded use of a process pattern

• Data description
  • Expanded to cover specific case identification and individual datum

• Portions of DDI 3.2 content

• Commonly used DDI 2.5 content
Development Summary

• Target
  • Expanded include Human to Computer understanding of the metadata

• Perspective
  • Move from retrospective to full range of viewpoints
  • From description to data linking

• Assumptions
  • Increasing demands from access from different directions and perspectives
  • Metadata as a requirement to meet the needs of producers and users of data

• Development changes
  • Automatic generation of documentation and related bindings from UML
  • Use of the binding that works for the job at hand (roundtripping of metadata)
  • Use the level of metadata that meets your needs
  • DDI production needs to be iterative in updating its content and automated in its production process
  • Version control in development (scheduled for 2018)
Infrastructure profiles: DDI Codebook

• Identification
  • Only requires instance identifier
  • Content is nested limiting the need for internal references
  • Uses standard ID and IDRef (supported by standard XML validation tools)
  • Supports capture of standard DDI URN

• XML only binding

• Current tools for creation of an instance, catalog of instances, and transformation to PDF document or web site

• Assumes XML instance is a publication in itself and will be managed as such
Infrastructure profiles: DDI Lifecycle

• Identification
  • Required by most classes
  • Registry of DDI Agent identification
  • Uses DDI structure for identification that resolves to a URN (requires secondary validation tool)
  • Allows and encourages reuse of metadata between instances

• XML only binding

• Current tools for creation of an instance, repository of DDI objects, transformation to PDF document or web site, and creation of questionnaires

• Assumes management of metadata content in XML – although this is loosening up in v3.3
Infrastructure profiles: DDI 4

- Identification
  - Same as Lifestyle (across bindings)

- UML based
  - Currently testing canonical XMI (expression of UML) for portability across UML tools

- XML needs secondary validation for:
  - Identification
  - Cardinality enforcement
  - Support of internal continuity when needed

- RDF
  - Has not had external review
  - Cardinality and type enforcement by validation with ShEx

- Assumes management of metadata is some form of “DDI aware” data management system
What do I use?

• What is your technical infrastructure?
  • Codebook has the lowest infrastructure requirements and is very suitable for individual researchers or anyone with infrastructure constraints

• What does your data look like?
  • Codebook can describe
    • Unit and dimensional data
    • Basic capture information (questions, derivation codes, secondary use of source data)
    • Archival data file structures – limited relational information
    • Focus is on the individual study/data set
  • Lifecycle can describe
    • Questionnaire structure
    • Multi-wave studies and their internal relationships
    • Cross study relationships
    • Common conceptual material
  • XKOS describes formal, managed Statistical Classifications
  • DDI 4 is in the prototype stage and not ready for implementation
Continuing development themes

• Ensure content is not lost as you move from earlier to later versions of the central standard
• Support for user communities in terms of both content and infrastructure needs and restrictions
• Clear lines of transition between versions
  • A repository should be able to take an earlier version and populate a newer version programatically
  • A repository should be able to identify the content it supports (in the same or earlier version structure) and populate a different version to the extent of its capabilities
  • Flexibility of use over time – future use of metadata may require transformation into a different version
DDI Alliance
http://ddialliance.org

Wendy Thomas wlt@umn.edu