Introduction to DDI: Basic Concepts and How to Develop Skills for Training Researchers

Anja Perry
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany

Jane Fry
Carleton University
Ottawa, ON, Canada

IASSIST
May 29, 2019
Anja Perry

● Economist

● Since 2016 at the GESIS Data Archive in Cologne, Germany

● Tasks
  ○ Consultation and workshops on Research Data Management
  ○ Lead researchers through the ingest process
Jane Fry

- Data Librarian
- MacOdrum Library, Carleton University
  - for almost 20 years!
- Ottawa, Ontario, Canada
- Tasks
  - helping users to discover and use data
  - Research Data Management
Introductions

- Name
- Where you work (name, city, country)
- In 20 words or less,
  - What you do
Outline

● What is metadata? (exercise 1)
● What is DDI and how can it help?
● Challenges
● Coffee break
● History and milestones of DDI
● How to use DDI (exercise 2)
● In the future: Training Library
<table>
<thead>
<tr>
<th></th>
<th>v1</th>
<th>v2</th>
<th>v3</th>
<th>v4</th>
<th>v5</th>
<th>v6</th>
<th>v7</th>
<th>v8</th>
<th>v9</th>
<th>v10</th>
<th>v11</th>
<th>v12</th>
<th>v13</th>
<th>v14</th>
<th>v15</th>
<th>v16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.04e+14</td>
<td>4</td>
<td>26</td>
<td>104012</td>
<td>99</td>
<td>1</td>
<td>1973</td>
<td>2</td>
<td>1642</td>
<td>1040</td>
<td>1040</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1.04e+14</td>
<td>1</td>
<td>19</td>
<td>104008</td>
<td>99</td>
<td>2</td>
<td>1943</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1.04e+14</td>
<td>4</td>
<td>98</td>
<td>104003</td>
<td>99</td>
<td>1</td>
<td>1990</td>
<td>2</td>
<td>1756</td>
<td>1040</td>
<td>1040</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>99</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>1.04e+14</td>
<td>2</td>
<td>98</td>
<td>104010</td>
<td>99</td>
<td>1</td>
<td>1983</td>
<td>2</td>
<td>1756</td>
<td>1040</td>
<td>1040</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>99</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>1.04e+14</td>
<td>1</td>
<td>18</td>
<td>104007</td>
<td>99</td>
<td>2</td>
<td>1927</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1.04e+14</td>
<td>2</td>
<td>19</td>
<td>104007</td>
<td>99</td>
<td>1</td>
<td>1983</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>1.04e+14</td>
<td>4</td>
<td>15</td>
<td>104005</td>
<td>99</td>
<td>2</td>
<td>1970</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>1.04e+14</td>
<td>4</td>
<td>19</td>
<td>104006</td>
<td>99</td>
<td>1</td>
<td>1942</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>1.04e+14</td>
<td>4</td>
<td>15</td>
<td>104005</td>
<td>99</td>
<td>1</td>
<td>1965</td>
<td>2</td>
<td>1040</td>
<td>1276</td>
<td>1040</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>1.04e+14</td>
<td>7</td>
<td>15</td>
<td>104004</td>
<td>99</td>
<td>2</td>
<td>1955</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>1.04e+14</td>
<td>7</td>
<td>15</td>
<td>104003</td>
<td>99</td>
<td>2</td>
<td>1977</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>77</td>
</tr>
<tr>
<td>12</td>
<td>1.04e+14</td>
<td>4</td>
<td>18</td>
<td>104005</td>
<td>99</td>
<td>2</td>
<td>1938</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>1.04e+14</td>
<td>7</td>
<td>17</td>
<td>104005</td>
<td>99</td>
<td>1</td>
<td>1945</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>1.04e+14</td>
<td>4</td>
<td>18</td>
<td>104005</td>
<td>99</td>
<td>2</td>
<td>1949</td>
<td>2</td>
<td>1040</td>
<td>1380</td>
<td>1040</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>1.04e+14</td>
<td>2</td>
<td>15</td>
<td>104003</td>
<td>99</td>
<td>2</td>
<td>7777</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>1.04e+14</td>
<td>4</td>
<td>32</td>
<td>104007</td>
<td>99</td>
<td>1</td>
<td>1974</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>1.04e+14</td>
<td>4</td>
<td>98</td>
<td>104006</td>
<td>99</td>
<td>1</td>
<td>7777</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>99</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>1.04e+14</td>
<td>4</td>
<td>25</td>
<td>104010</td>
<td>99</td>
<td>2</td>
<td>1968</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1.04e+14</td>
<td>4</td>
<td>40</td>
<td>104007</td>
<td>99</td>
<td>1</td>
<td>1967</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>1.04e+14</td>
<td>1</td>
<td>23</td>
<td>104004</td>
<td>99</td>
<td>1</td>
<td>1932</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>1.04e+14</td>
<td>2</td>
<td>18</td>
<td>104003</td>
<td>99</td>
<td>2</td>
<td>1965</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>22</td>
<td>1.04e+14</td>
<td>1</td>
<td>27</td>
<td>104011</td>
<td>99</td>
<td>1</td>
<td>1956</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>1.04e+14</td>
<td>4</td>
<td>18</td>
<td>104004</td>
<td>99</td>
<td>2</td>
<td>1923</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>1.04e+14</td>
<td>1</td>
<td>19</td>
<td>104006</td>
<td>99</td>
<td>2</td>
<td>1952</td>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>1.04e+14</td>
<td>3</td>
<td>16</td>
<td>104004</td>
<td>99</td>
<td>2</td>
<td>1947</td>
<td>2</td>
<td>1276</td>
<td>1040</td>
<td>1040</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Exercise 1

● As a researcher, what metadata do you *absolutely* need?

● How do you want it to *streamline* your research?

● What metadata would you like to have, if it is available, but it is *not* *integral* to your research?
“Upstream” vs. “Downstream” Metadata Capture

“Downstream” Metadata Capture

Data → Metadata → Publish/Archive/Deliver

“Upstream” Metadata Capture

Data → Metadata → Publish/Archive/Deliver
Management Patterns for Data and Metadata

**Decentralized:**
Difficult to manage

**Centralized:**
Easier to manage
How can DDI help?
What is DDI?

- Data Documentation Initiative

- “An effort to develop a specification for documenting data files in XML. The DDI Alliance is the organization that created the specification, …”

- More information can be found on the DDI website.
Structure and standardisation

● “Skeletons give us vertebrates shape and structure, markup does the same for text” (Ray, 2003)

● Data documentation contained in Word or PDFs are for human consumption only

● Computers need structure to process documentation

● DDI provides that structure
What is DDI?

- Data Documentation Initiative
  http://www.ddi-alliance.org/

- A **structure** to **consistently** define data and it’s related metadata, for the purpose of supporting the **intelligent use of the data** over time.
A computer does it better, cheaper and faster.

- Improves quality (less error prone)
- Saves person time (resources)
- Optimises use of human skills by automating repetitive, mindless tasks
DDI as a standard

• DDI is a *standard* for metadata

• The standard structure means that all computers, even if they are using different applications, can work on the same data and related information (metadata/documentation)
  - The formats are not proprietary to any specific system
  - Uses generic XML technology as the basis for cross-platform use
Study

Concepts

measures

Universes

Survey Instruments

Questions

using

about

made up of
Questions

collect

Responses

with values of

Variables

made up of

Categories/Codes, Numbers

resulting in

Data Files
Challenges

● Involves initial investment but saves costs in the long-term
  ○ May involve changes to processes and systems
  ○ Investment in technology tools may be considerable

● Legacy metadata may require updating
  ○ Consistency issues
  ○ Format transformations

● Training is required
Milestones
Data Documentation Initiative

1995
1st SGML Codebook Committee meeting

1996
1st DTD prepared at University of Michigan Library

1997
NSF funding received; SGML spec. translated to XML

1998
Beta test of DDI DTD

2000
DDI v1 (DTD based) published

2001
Formal DDI evaluation, funded by NSF

2002
DDI Alliance Charter drafted

2003
DDI Alliance established; DDI 2 published

2003
1st Steering Committee meeting

2005
Final meeting of original Committee

2008
- Public review of DDI 3
- 1st DDI Training Workshop, Schloss Dagstuhl

2009
- 1st NADDI Conf., Lawrence, KS
- DDI RDF Discovery and XKOS vocabularies published
- Sprints held to create model-driven DDI

2010
- 1st Dagstuhl Workshop on DDI Model
- DDI Lifecycle 3.1 published

2011
- DDI Lifecycle 3.1 published
- DDI Alliance Charter drafted
- Agency registry established
- Tools Catalog released
- External review of DDI

2012
- 1st DDI Codebook 2.5 published as XML Schemas
- 1st set of controlled vocabularies published

2013
- 1st NADDI Conf., Lawrence, KS
- DDI RDF Discovery and XKOS vocabularies published
- Sprints held to create model-driven DDI

2014
- DDI Lifecycle (Model-Driven) products released
- Redesigned DDI Alliance website launched

http://www.ddialliance.org/what/history.html
Who uses DDI

- Norwegian Social Science Data Services
- Harvard University
- American University
- DLI (Statistics Canada)
- Health Canada
- Bureau of the Census
- University of Michigan
- ICPSR
- Bureau of Labor Statistics
- …
What projects use DDI

- CESSDA Data Portal (European quantitative social science datasets)
- Australian Social Science Data Archive
- DAMES Project (UK)
- DataFirst (at University of Cape Town)
- Israel Social Science Data Center
- Philippines National Statistics Office
- Statistics New Zealand
- ICPSR Data Catalog
- Vision of Britain (historical view between 1801 and 2001)
- World Bank (International Household Survey Network)
- ODESI (Ontario Data Portal)

...
DDI Development

● DDI now branched into 2 separate development lines or metadata standards
● DDI Codebook (2003)
  ○ aka DDI C
  ○ Formerly DDI 2
  ○ Built to emulate a physical codebook
  ○ Latest version is 2.5
  ○ Sections
    ■ Document Description • Study Description • Data Files Description • Variable Description • Other Study Related Materials
DDI Development

● DDI Lifecycle (2008)
  ○ aka DDI L
  ○ Formerly DDI 3
  ○ Supports the research data lifecycle
  ○ The one most new users are learning
  ○ Latest version is 3.2
  ○ Sections
    ■ Study Concept • Data Collection • Data Processing • Data Distribution • Data Archiving • Data Discovery • Data Analysis • Repurposing
Comparison

**DDI 1 and 2 (DDI C)**
- Document Description
- Study Description
- Data Files Description
- Variable Description
- Other Study Related Materials

**DDI 3 (DDI L)**
- Study Concept
- Data Collection
- Data Processing
- Data Distribution
- Data Archiving
- Data Discovery
- Data Analysis
- Repurposing

Reference: Jim Jacobs, 2006
Which DDI do I use?

- **DDI C**
  - Relatively straight forward
  - If you want to catalog a dataset
  - If you are describing a single study

- **DDI L**
  - If you are focusing on a lifecycle model
  - Broken down into different functions
  - Are you documenting questionnaires?
  - Are you documenting data?
  - Are you doing both?
Getting started with DDI

- **Daunting at first**
  - Process is broken down into steps
- **Lots of help available**
  - DDI Alliance
  - [http://www.ddialliance.org/training/getting-started](http://www.ddialliance.org/training/getting-started)
  - Colleagues
  - Other researchers
- **DDI List-serv**
- **DDI Best Practices**
  - Work in progress
  - *Feedback always welcome*
Tools to help you get started

http://www.ddialliance.org/resources/tools
Getting Started with DDI

One tool: Nesstar Publisher

- Norwegian Social Science Data Services
- Data management program
- Freeware
- Data and metadata conversion and editing tools
- Enhance datasets
  - Combine catalogue and contextual information
- Merge DDI documents with markup for different sections of the DDI for the same study
  - Merge variable descriptions from SPSS/SAS with DDI
Getting Started with DDI

- **Nesstar Webview**
- Metadata
- Any associated documentation
- Variable groups
- Conduct basic analysis
  - Subsetting
  - Crosstabs
- Bonus
Getting Started with DDI

- Nesstar Webview
- Downloading
  - Documentation
    - PDF format
  - Export files with study descriptions and question text
  - Data exported in format of choice
  - SPSS, SAS, Stata, ASCII, …
Getting Started with DDI

● Check out how Colectica works
  ○ A number of videos to watch
    ■ http://www.youtube.com/user/Colectica/video
  ○ Colectica Questionnaires
    ■ Demo
    ■ Create a Survey and Add
    ■ View a Survey’s Structure
    ■ Add metadata to a Survey
    ■ ...

DATA DOCUMENTATION INITIATIVE
Getting Started with DDI

- Check out how Nesstar Webview works
  - Using the ODESII data repository
    - [http://www.library.carleton.ca/help/odesi-how-to-use-odesi](http://www.library.carleton.ca/help/odesi-how-to-use-odesi)
  - Navigating the ODESII repository
  - Searching for variables
  - Finding, subsetting and downloading
  - Creating a cross tabulation
  - Downloading a full dataset
Getting Started with DDI

- **Colectica Reader**
- Free tool
  - To view the metadata
  - No specialized software is needed
- Generates documentation for variables and code lists
  - PDF, Word, HTML
Getting Started with DDI

- **Another tool: Colectica for Excel**
  - Tools to help with your metadata
    - [https://www.colectica.com/software/](https://www.colectica.com/software/)
    - [http://www.ddialliance.org/node/893](http://www.ddialliance.org/node/893)
  - Documents variables and datasets directly from within Excel
  - Can be used to produce detailed (item-level) metadata for studies already completed
  - Creates metadata and documentation for surveys
  - DDI version 3.1, 3.2
  - Saves metadata directly in the Excel file
    - *When the file is shared, so is the metadata*
Getting Started with DDI

● **Nesstar drawbacks**

● For advanced statistical analysis -
  ○ *it is best to download the data and use a statistical analysis package*

● Must have access to a server to publish the dataset

● Not intuitive when starting to markup datasets

● Not intuitive for first-time user in Webview

● Downloading into SAS not user friendly

● **Not a drawback, just a consideration**

● Uses DDI Codebook standard
Why Use DDI?

DDI encourages comprehensive description of data for discovery and analysis and supports effective data sharing. Because DDI is a structured standard, it facilitates machine-actionability and interoperability and it can actually be used to drive systems. Another feature of DDI is its focus on metadata reuse; “enter once, use often” means you can reuse metadata over the course of the data life cycle to avoid costly duplication of effort.

DDI has advantages for several different audiences:

+ Librarians
+ Managers
+ Repositories
+ Researchers
+ Developers

Question 1: How do these audiences use DDI differently?

Question 2: Are there any audiences missing?
“Standards are important to the effective functioning of libraries. Using a standard vocabulary to document research data leads to consistency and improved interoperability.

DDI is designed to make research data independently understandable. DDI provides a standard structure for all of the metadata that accompanies a dataset and helps users of that dataset to interpret its contents. This is useful when assisting patrons and data analysts.

DDI is an open, non-proprietary standard and anyone can use it.”
Managers

- “Metadata are expensive to produce, so reusing structured, standardized metadata makes good business sense.

- DDI promotes interoperability and thus supports partnerships with others that involve data and metadata exchange.

- DDI’s structure can enable effective search and discovery, subsetting, generation of syntax files, and flexibility in display, resulting in many efficiencies.”
Repositories

- “Codebooks have long been used to interpret data files, but PDF and Word codebooks are not “intelligent.” In contrast, DDI codebooks are structured and can be interactive, enabling users to navigate through a collection.

- DDI can serve as a foundation for data catalogs as it provides a standard structure for searching at both the study and variable levels to enable users to discover data of interest.

- Using DDI throughout the archiving life cycle can streamline the repository’s workflow, leading to efficient ingest, management, and preservation of data.”

http://www.ddialliance.org/training/why-use-ddi
Researchers

- “Recent open access mandates from funders require that data be shared in order to validate results and to encourage new discoveries. This means that data must be well-documented, which is DDI’s strength.
- Complex, longitudinal data projects require additional levels of data management. DDI can support this and can enable creation of reports, displays, and tools that leverage the richness of the data. Some examples are question banks, concordances, and interactive codebooks.
- The structure of DDI can support data comparison and harmonization.”
Developers

- “Using a structured standard optimizes machine-actionability and makes programming against the structure possible.

- DDI can actually drive process, leading to greater efficiencies.

- DDI can be used with relational databases to increase flexibility.”
In the future: a Training Library

- For anyone to use
  - we want you to use them
  - so you don’t have to develop your own slides

- Content
  - different topics
  - for different audiences
  - let us know if there are other topics you would like added

- Release Date: 2019 / 2020
Thank you!

Contact information:
Anja Perry
Anja.Perry@gesis.org

Jane Fry
jane.fry@carleton.ca