

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

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	1.040e+14	4	26	104012	99	1	1973	2	1642	1040	1040	1	6	1	5	5
2	1.040e+14	1	19	104008	99	2	1943	1	99	99	99	1	1	6	4	5
3	1.040e+14	4	98	104003	99	1	1990	1	99	99	99	1	6	3	99	11
4	1.040e+14	2	98	104010	99	1	1983	2	1756	1040	1040	1	6	3	99	11
5	1.040e+14	1	18	104007	99	2	1927	1	99	99	99	1	1	6	4	1
6	1.040e+14	2	19	104007	99	1	1983	1	99	99	99	1	6	2	5	3
7	1.040e+14	4	15	104005	99	2	1970	1	99	99	99	1	1	2	4	6
8	1.040e+14	4	19	104006	99	1	1942	1	99	99	99	2	1	6	5	88
9	1.040e+14	4	15	104005	99	1	1965	2	1040	1276	1040	7	4	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99	99	99	1	1	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99	99	99	1	1	6	77	77
12	1.040e+14	4	18	104005	99	2	1938	1	99	99	99	1	3	6	5	5
13	1.040e+14	7	17	104005	99	1	1945	1	99	99	99	1	1	6	5	4
14	1.040e+14	4	18	104005	99	2	1949	2	1040	1380	1040	1	4	6	5	5
15	1.040e+14	2	15	104003	99	2	7777	1	99	99	99	1	1	2	4	5
16	1.040e+14	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	104006	99	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4	25	104010	99	2	1968	1	99	99	99	1	1	2	5	1
19	1.040e+14	4	40	104007	99	1	1967	1	99	99	99	1	1	2	6	2
20	1.040e+14	1	23	104004	99	1	1932	1	99	99	99	1	1	6	4	3
21	1.040e+14	2	18	104003	99	2	1965	1	99	99	99	1	6	2	5	5
22	1.040e+14	1	27	104011	99	1	1956	1	99	99	99	2	1	1	5	2
23	1.040e+14	4	18	104004	99	2	1923	1	99	99	99	1	3	6	3	3
24	1.040e+14	1	19	104006	99	2	1952	1	99	99	99	1	4	2	5	3
25	1.040e+14	3	16	104004	99	2	1947	2	1276	1040	1040	1	1	6	3	3

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	1.040e+14		26	104012	99	1	1973	2	1642	1040	1040	1	6	1	5	
2				104008	99	2	1943	1	99	99	99	1	1			
				104003	99							1				
				104010	99							1				
				104007	99							1				
				104007	99							1	6			
				104005	99							1	1	2	4	6
				104006	99							2	1	6	5	88
9			15	104005	99					1040		7	4	1		
10	1.040e+14	7	15	104004	99	2				99	99					
11	1.040e+14	7			99	2	7777	1	99	99	99					
12	1.040e+14				99	2	1938	1	99	99	99					
13	1.040e+14				99	1	1945	1	99	99	99					
14	1.040e+14				99	2	1949	2	1040	1380	1040					
15	1.040e+14				99	2	7777	1	99	99	99					
16	1.040e+14				99	1	1974				99					5
17	1.040e+14				99	1						1	6		99	11
18	1.040e+14				99	2						1	1	2	5	1
19	1.040e+14	4		104007	99							1	1	2	6	2
20	1.040e+14	1	23	104004	99							1	1	6		
21	1.040e+14	2	18	104003	99							1				
22				104011	99							99				
					99	2						99				
					99	2	1952	1		99	99					
					99	2	1947	2	1276	1040	1040	1				

What is the study about?
Who created it?

Who was asked?
What is the population?

Mode of collection?

What is the meaning of the codes?

What was the data capture?
(the question)

What was the previous question?

What was the intent/concept of the question?

What do the variables mean?

Who funded it?



Exercise 1

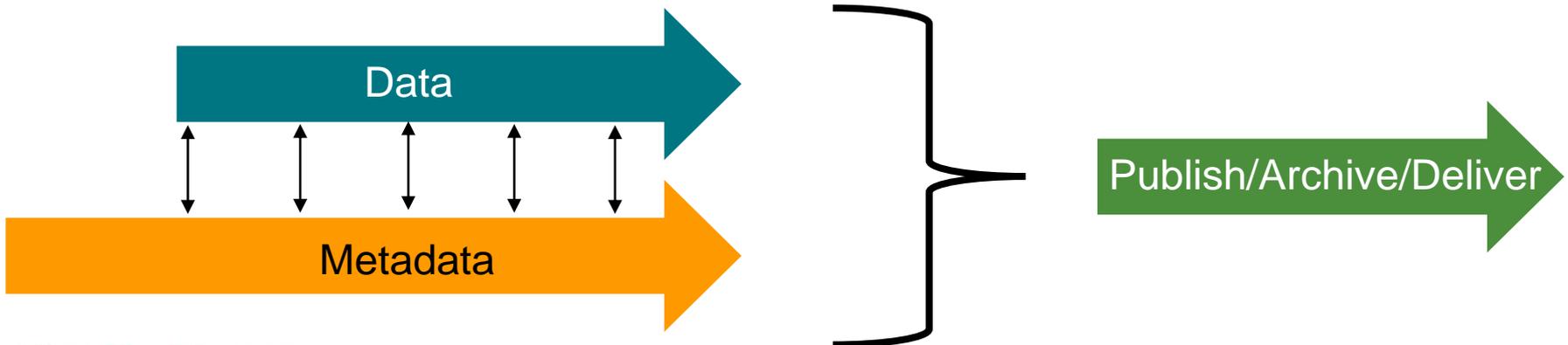
- As a researcher, what metadata do you *absolutely* need?
- How do you want it to *streamline* your research?
- What metadata would you like to have, if it is available, but it is *not integral* to your research?

“Upstream” vs. “Downstream” Metadata Capture

“Downstream” Metadata Capture

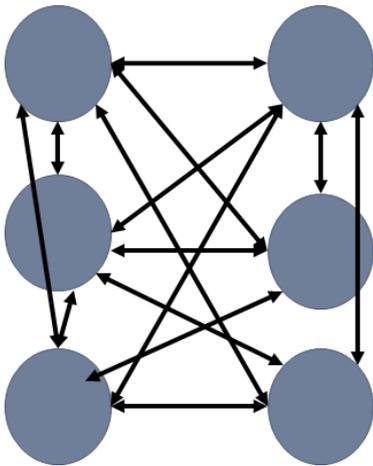


“Upstream” Metadata Capture

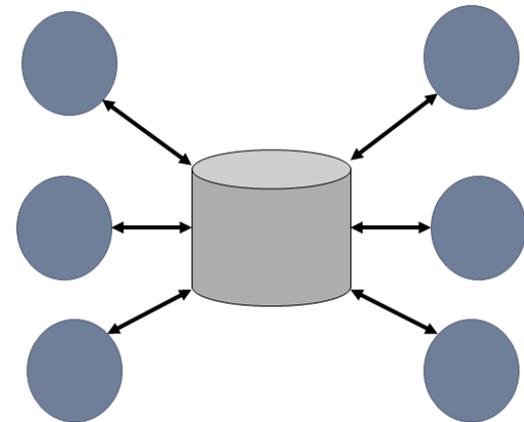


Management Patterns for Data and Metadata

Decentralized:
Difficult to manage



Centralized:
Easier to manage



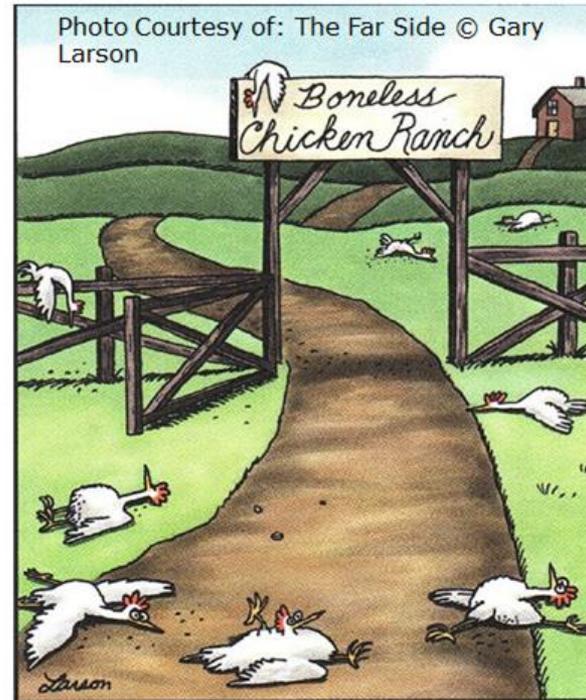
How can DDI help?

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.



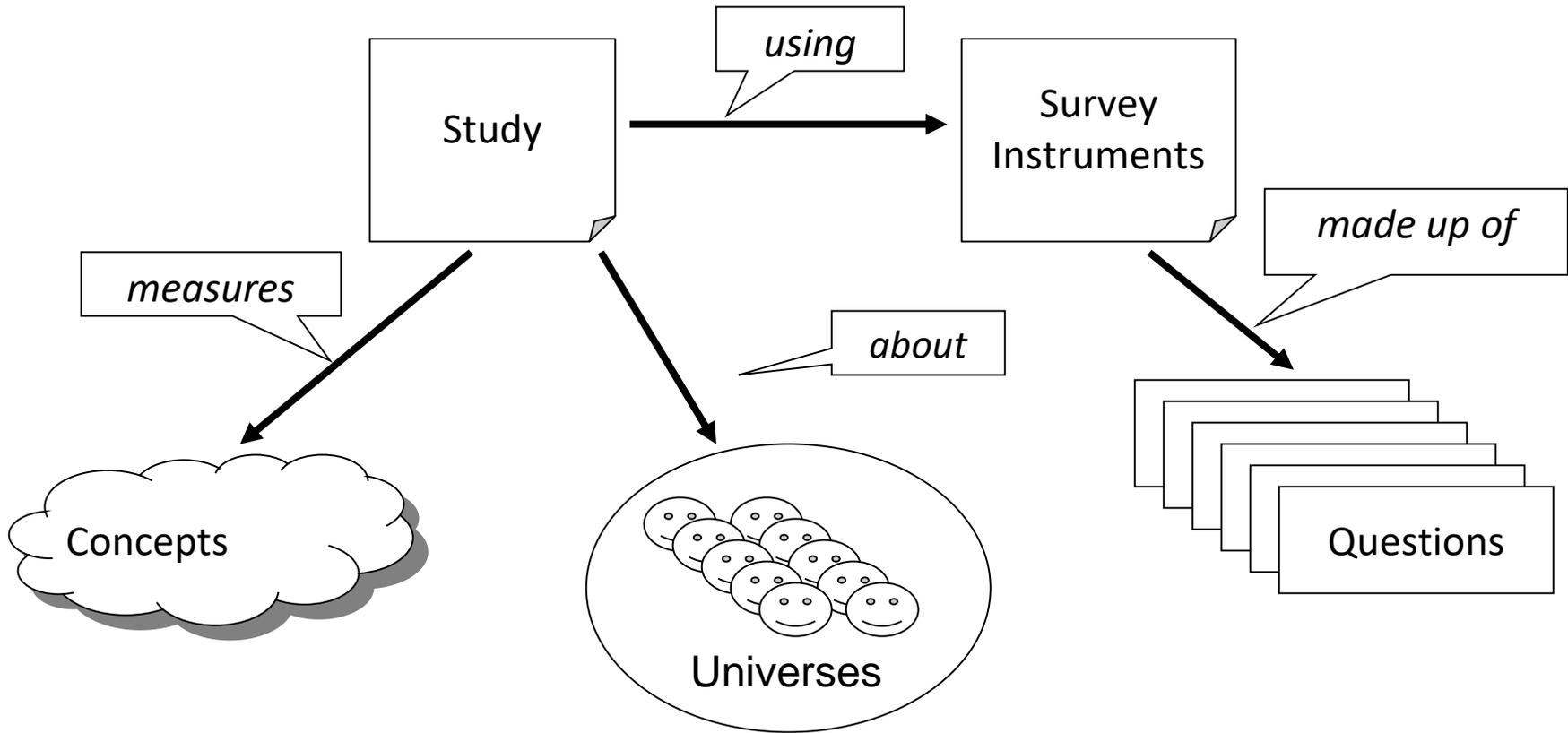
Why computers?

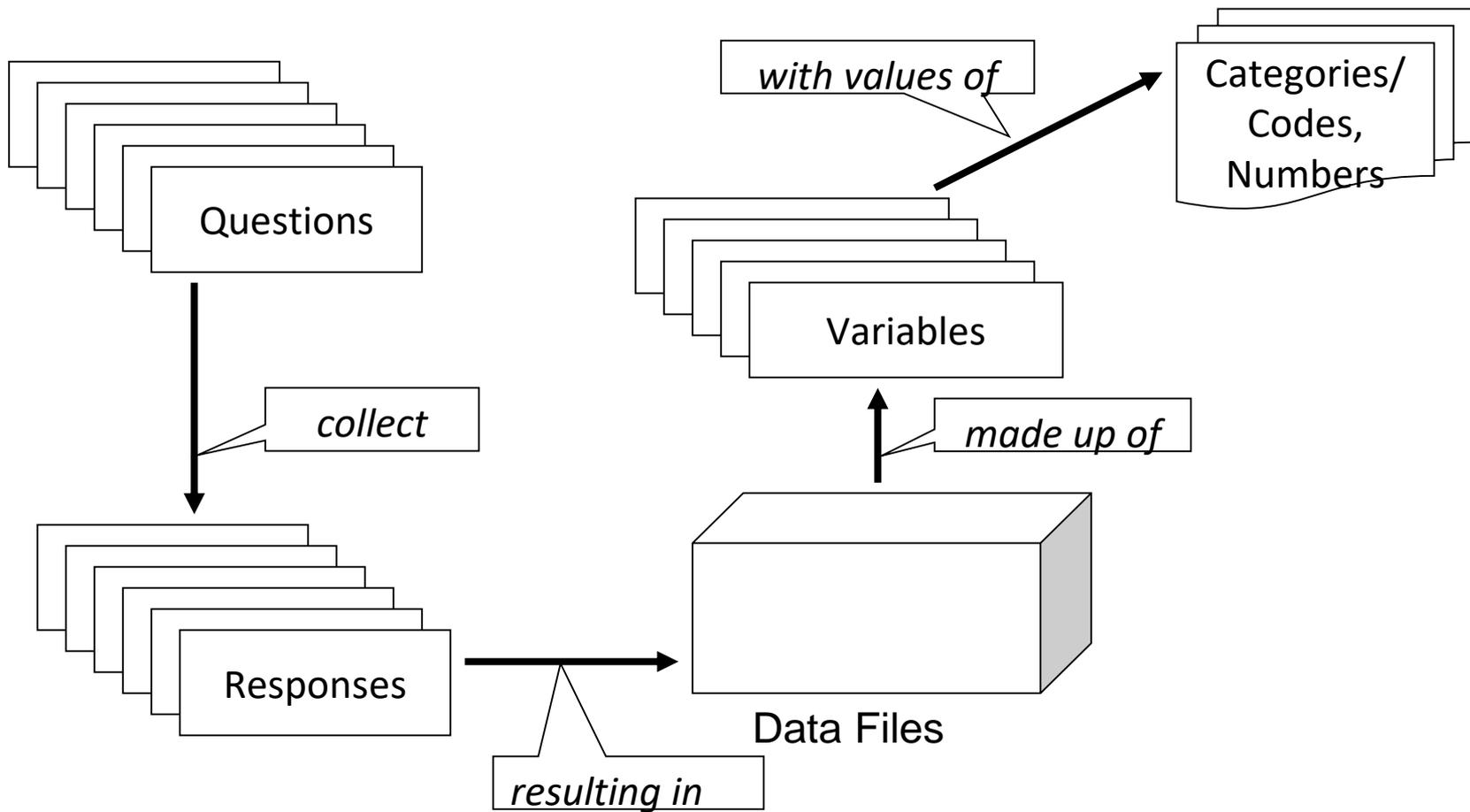
A computer does it better, cheaper and faster.

- 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





Challenges

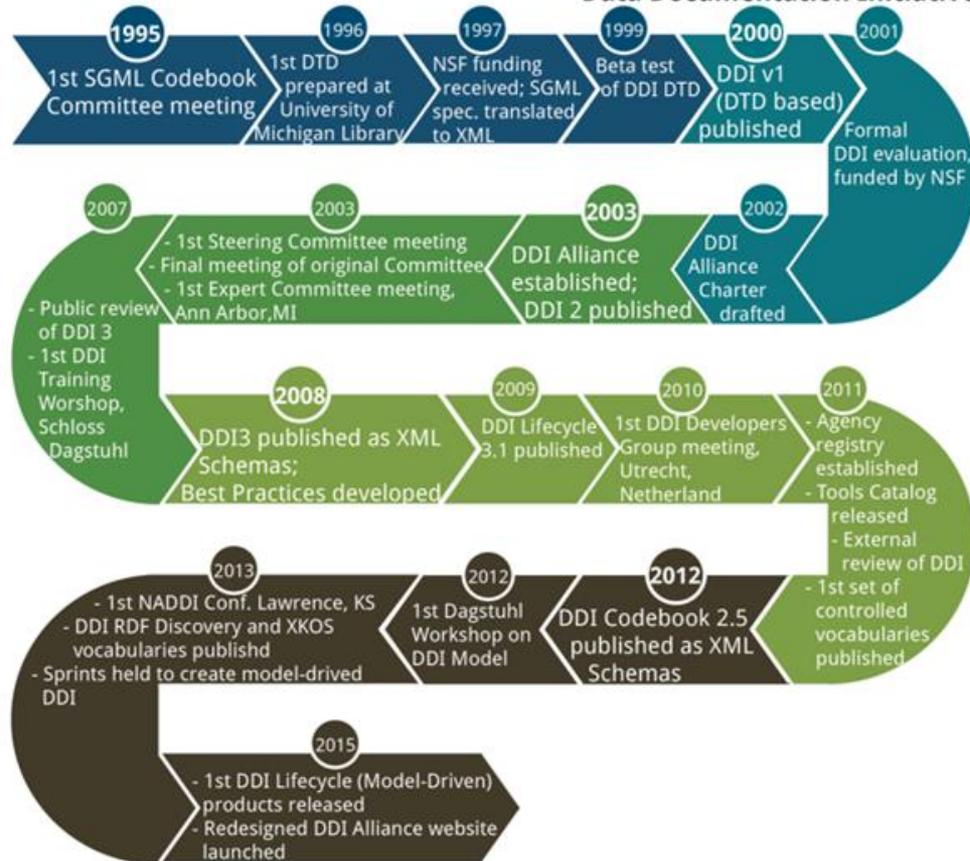
- Involves initial investment but saves costs in the long-term
 - May involve changes to processes and systems
 - Investment in technology tools may be considerable
- Legacy metadata may require updating
 - Consistency issues
 - Format transformations
- Training is required





Milestones

Data Documentation Initiative



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

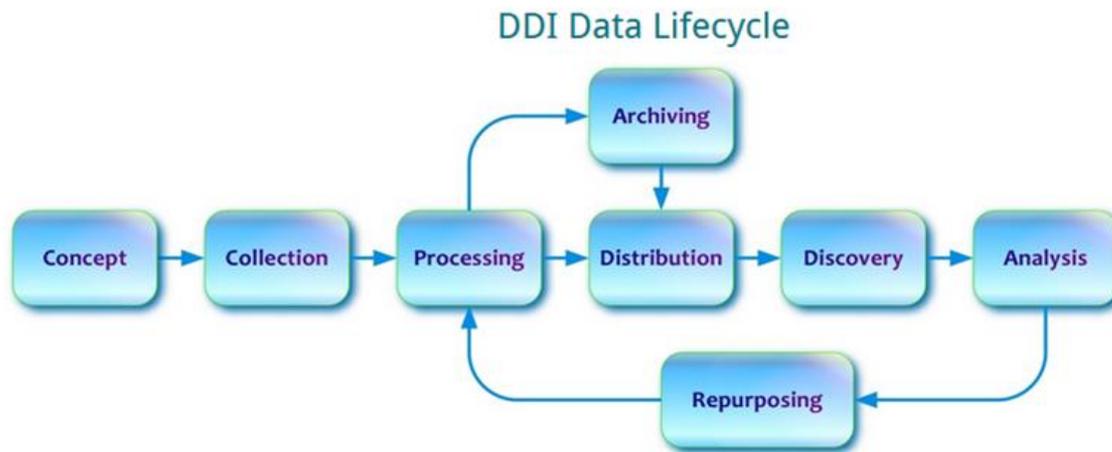
- **DDI now branched into 2 separate development lines or metadata standards**
- **DDI Codebook (2003)**
 - aka DDI C
 - Formerly DDI 2
 - Built to emulate a physical codebook
 - Latest version is 2.5
 - Sections
 - *Document Description • Study Description • Data Files Description • Variable Description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

DDI Research Lifecycle



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

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>
 - Colleagues
 - Other researchers
- **DDI List-serv**
- **DDI Best Practices**
 - Work in progress
 - *Feedback always welcome*

Tools to help you get started

Tool Purpose: DDI version(s) supported: Availability: Supported Operating Systems:

Search Tools by Name:

Name	Version(s) supported	Availability	Description	Purpose
------	----------------------	--------------	-------------	---------

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*
 - ...

Getting Started with DDI

- **Check out how Nesstar Webview works**
 - Using the ODESI data repository
 - <http://www.library.carleton.ca/help/odesi-how-to-use-odesi>
 - *Navigating the ODESI 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/>
 - <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

Exercise 2

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?

Librarians

- *“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.”*

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