Webinar on Introduction to Metadata for Research Data Management: A Data Documentation Initiative (DDI) Perspective

Alina Danciu, Arofan Gregory

Introduction to Metadata for Research Data management: The DDI Perspective

The DDI Training Working Group

29 April 2021

DATA DOCUMENTATION INITIATIVE

Some Obvious Questions…

- Why should we care so much about metadata?
- Isn’t it just “data about data”?
- Shouldn’t we be focused on the data?

DATA DOCUMENTATION INITIATIVE

If You Care about Data, You Care about Metadata!

- Metadata is what allows people to discover, assess, and navigate through collections of data
- Metadata is what allows people to understand data
- Metadata is what allows machines to process and manage data
- Metadata is what allows people to harmonize, integrate, and reuse data

If you want FAIR data, you need *metadata*!

DATA DOCUMENTATION INITIATIVE

Metadata is the Limiting Factor

- Your data is only as good as the metadata which describes it
- Traditionally, research has focused on data supporting specific research
- Now, the demand for larger, cross-cutting research projects places an emphasis on better data management
- The answer to providing this is to have good, well-managed metadata!

DATA DOCUMENTATION INITIATIVE

Different Understandings and Definitions

- Across research domains, the term “metadata” means many different things
- We need to respect these differences…
- We also need a shared understanding
- The DDI community has spent decades understanding metadata, using it to manage data, and modelling it for implementation in systems
- We do not have the *only* perspective – but we do have a *proven, useful one!*

DATA DOCUMENTATION INITIATIVE

Understanding Metadata

DATA DOCUMENTATION INITIATIVE

Contents

- Defining Metadata
- Data or Metadata?
- Examples
- Roles
- Usage
- Management
- Summary

DATA DOCUMENTATION INITIATIVE

Defining Metadata

DATA DOCUMENTATION INITIATIVE

What is in this can?

DATA DOCUMENTATION INITIATIVE

Now you know!

DATA DOCUMENTATION INITIATIVE

What do the numbers in this table mean?

001	12	01	98
002	36	02	175
003	72	01	94
004	42	01	130
005	18	02	125

DATA DOCUMENTATION INITIATIVE

Metadata tells us!

Respondent ID	Age	Sex	Weight
001	12	01	98
002	36	02	175
003	72	01	94
004	42	01	130
005	18	02	125

Age is in years as of last birthday

Sex: 01 = Male, 02 = Female

Weight is in pounds

Metadata

DATA DOCUMENTATION INITIATIVE

What Does “Meta” Mean, Anyway?

- Metadata are descriptions or information about some thing(s)
- Greek*:* μετά- is a prefix meaning
  - Beyond
  - After
  - Transcending
- Literally – metadata are <u>beyond</u> the thing(s) being described

DATA DOCUMENTATION INITIATIVE

Metadata Clarification

- The numeric table –
  - Metadata are
    - Headings for the columns
    - Descriptions of the values in each column
  - They describe the table and the data in the cells

- People define it as “data about data”, but…

- The food can –
  - Metadata are what is written in the label on the can
  - They describe the contents of the can

- Let’s look at this in more detail …

DDI®
DATA DOCUMENTATION INITIATIVE

Metadata Defined – Not Just *Data about Data*

- Metadata may be about data
  - Numeric table example
- Metadata may be about some other “things”
  - Food can example
- Requires broad view of metadata
  - “Data about some resource”
- Enhanced definition
  - Metadata <-> data intended to be used for describing some object(s)

DATA DOCUMENTATION INITIATIVE

Data or Metadata?

DATA DOCUMENTATION INITIATIVE

Metadata or Data? It Depends!

- Question – How do we intend to use the data?
  - Microdata – data used to represent a population
    - Not each individual respondent, but for producing statistics
  - Administrative data – data used to manage a social program
    - Data used to monitor eligibility and results
  - Primary purpose of data is not to describe
- Metadata are intended to describe
- Data are intended to measure

DATA DOCUMENTATION INITIATIVE

Metadata versus Data

- Consider microdata or administrative data
  - Microdata –
    - data collected from respondents in a statistical survey
    - collected data are a description of each respondent
  - Administrative data –
    - data acquired as a result of managing some social program
    - managed data are a description of each program participant
- Why aren't these metadata?
  - In each case, data describe some things, but…
    - Survey responses are used to *measure* characteristics of a population
    - Administrative data are used to *measure* characteristics of the program participants

DATA DOCUMENTATION INITIATIVE

Metadata – A Role

- Metadata is a role for data
- No data are always metadata
  - Data are metadata when they are used that way
- Most (if not all) data can be used as metadata
- Sometimes, data could be both metadata and data
  - Depends on usage
  - Same data, different perspective

DATA DOCUMENTATION INITIATIVE

Examples

DATA DOCUMENTATION INITIATIVE

Dublin Core Metadata Example

- Dublin Core
  - Commonly used metadata standard
  - Dublin Core Metadata Element Set
    - Originally 15 core elements
    - Over 40 additional elements
  - Developed by
    - Online Computer Library Center – Dublin, OH
  - Used in many digital libraries
    - Museums, Book libraries
  - Objects of interest:
    - Museums – paintings, bones, pictures, documents, tools – not data
    - Book libraries – books and magazines

DATA DOCUMENTATION INITIATIVE

Telephone Call Example (1)

- Another example – telephone company data
  - Telephone company knows
    - Who called whom
    - Whether connection is made or why it failed
    - Length of connection
  - But they don't have the *content* of each conversation
- Telephone data contains:
  - One number was used to call another
    - Including subscriber information of caller
  - Date/time this occurred
  - Whether a connection was made
  - How long connection lasted

DATA DOCUMENTATION INITIATIVE

Telephone Call Example (2)

- Use as data
  - Build network of calls
    - Nodes are phone number / subscribers
- Use as metadata
  - Part of the description of a particular phone call

- The role the data play is crucial for making the distinction

DATA DOCUMENTATION INITIATIVE

Classification Scheme Example

- Classification scheme entry
    - Concept, code, definition, description, relationships

- Data for a classification database
- Data for construction of n-cubes, as element of a dimension
- Data for construction of stratified sample

- Metadata for user of a table as description (meaning) of a cell
- Metadata for user of data based on the stratified sample

DATA DOCUMENTATION INITIATIVE

Roles

DATA DOCUMENTATION INITIATIVE

Roles of Metadata

- Metadata supports 3 main uses:
  - Discovery - find relevant resources (e.g., data)
  - Understandability - convey the semantics of resources
  - Usage - describe limits of reasonable use
- Above uses are roles
  - Variables apply in all three cases
  - So do data sets
  - Therefore, structures of data do as well
- But these are definitely not the only roles

DATA DOCUMENTATION INITIATIVE

Usage

DATA DOCUMENTATION INITIATIVE

What Do We Want to Describe?

Metadata may be used to describe …

- Data
- Variables
- Segments
- Value Domains
- Classification schemes
- Code lists
- Questions
- Questionnaires
- Instruments
- Collection process
- Sample design
- Weighting
- Data transformations
- Editing procedures
- Machine Learning methods
- Allocation
- Estimation
- Disclosure control

DATA DOCUMENTATION INITIATIVE

Uses of Metadata - Simple

- Document descriptions of statistical objects
  - Promote records management
- Provide means of comparison across
  - Objects (e.g., classification schemes)
  - Programs, experiments, studies
  - Time
  - Organizations
  - Subject areas (including within broad domains)
  - Cultures (especially languages)
  - Political entities (e.g., countries)

DATA DOCUMENTATION INITIATIVE

Uses of Metadata - Enhanced

- Create new descriptions of more complex objects
  - Descriptions of variables used to describe each data set
  - Descriptions of questions used to describe question form or questionnaire
- Promote reuse – “describe once, use many”
  - Increase comparability
  - Reduce inconsistency
    - Avoid gratuitous differences
  - Increase quality
  - Reduce future burden
  - Reduce costs

DATA DOCUMENTATION INITIATIVE

Uses of Metadata - Advanced

- Promote machine-readable metadata
  - Avoid using long prose documents
  - Implement machine-readable formats (e.g., XML, JSON)
  - Strive towards metadata-driven processing
  - Machine-actionable metadata
- Promote concept management
  - Note how and where concepts are used
    - E.g., Concept, Universe, Category, Variable
  - Link similar meanings
  - For definitions, use principles defined in
    - ISO 704 (Terminology – Principles and methods)
    - ISO/IEC 11179-4 (Information technology — Metadata registries (MDR) — Part 4: Formulation of data definitions)
  - For terminological approach to data and metadata, based on ISO 704

DATA DOCUMENTATION INITIATIVE

Management

DATA DOCUMENTATION INITIATIVE

Managing Metadata

- Metadata are data
  - Metadata can be managed in a database
    - A metadata repository is a database of metadata
    - Metadata is also often stored in data repositories
- Relational, Object-oriented, Graph / Network, Other?
  - Depends on several factors
    - Available software
    - Current expertise
    - Integration with other systems
    - Time frame
    - Costs
    - Risks

DATA DOCUMENTATION INITIATIVE

Benefits of Managed Metadata

- More consistency in the production and dissemination of data
  - More comparable data
  - More understandable data
  - More FAIR data
- Assessing and ensuring data quality
- Organizational efficiency
  - Reuse of metadata assets
  - Better design, implementation, and execution of processes
  - Increased automation

DATA DOCUMENTATION INITIATIVE

Summary

- Definition of metadata
  - data intended to be used for describing some object(s)
- Compare with data
  - When are data metadata? It depends.
- Roles
  - Metadata can be used for multiple purposes (administrative, semantic, etc.)
- Uses
  - Human readable documentation
  - Sharing metadata and reusable descriptions
  - Concept management and machine-readable metadata
- Management
  - Metadata are data, so they may be managed similarly

DATA DOCUMENTATION INITIATIVE

DDI Products and Evolution

DATA DOCUMENTATION INITIATIVE

DDI Metadata – Evolution

- DDI has been producing specifications for metadata in RDM for more than 2 decades
- Our understanding of metadata has evolved
- Our suite of products has grown to support more aspects of RDM
- The evolution is ongoing

DATA DOCUMENTATION INITIATIVE

DDI: Major Specifications

- DDI Codebook (aka “DDI 1.0”, “DDI 1.2”, “DDI 2.5”, etc.)
- DDI Lifecycle (aka “DDI 3.0”, “DDI 3.1”, “DDI 3.3”, etc.)
- DDI Cross Domain Integration (aka “DDI-CDI”)

DATA DOCUMENTATION INITIATIVE

DDI Codebook

- An XML description of a “codebook” (a data dictionary)
  - Rectangular files
  - No concept of metadata reuse
  - Based on models in existing analysis tools (Stata, SPSS, SAS, etc.)
- Included Dublin Core and descriptive “study-level” metadata
- Machine-readable (*slightly* machine-actionable…)
- Described data for a single study (one point in time)
- After-the-fact description to support archiving and reuse

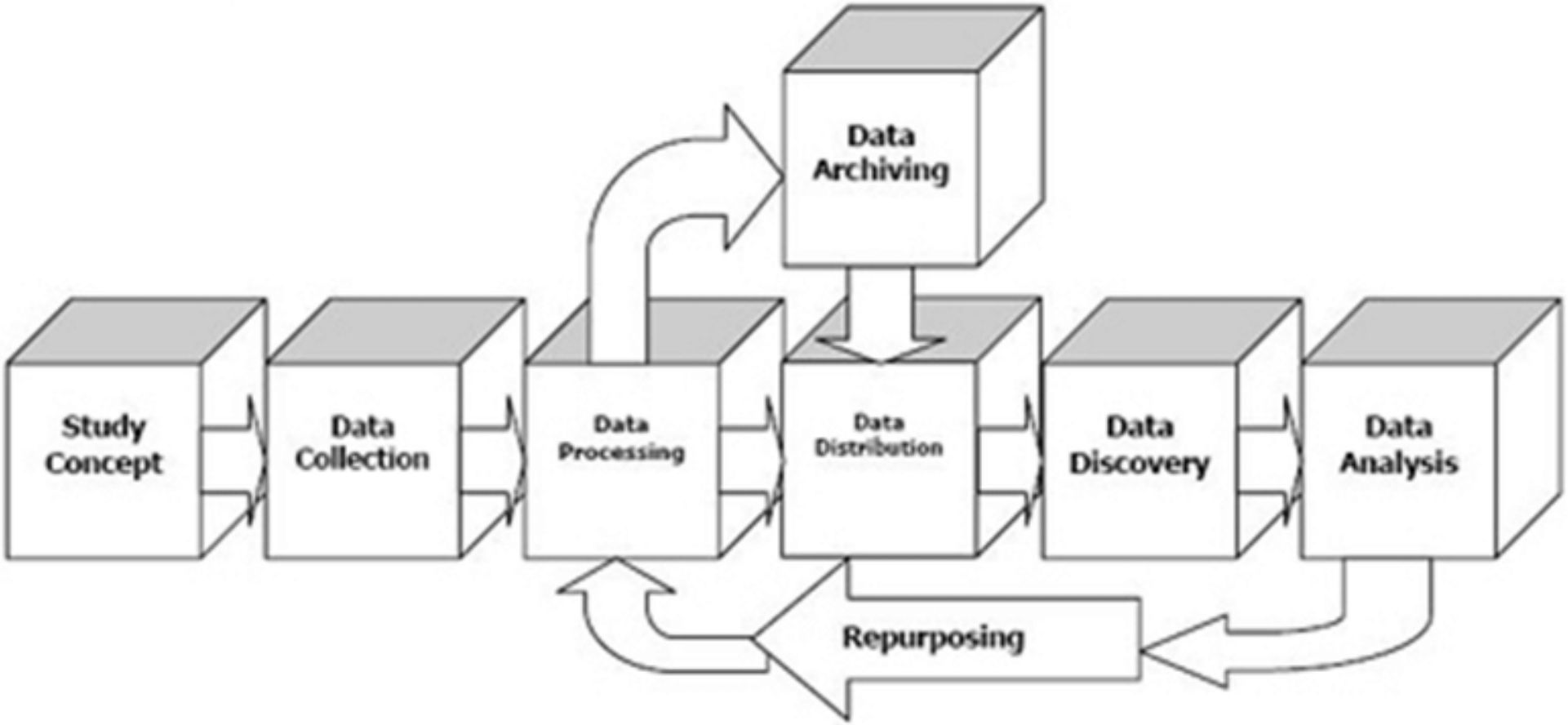
DATA DOCUMENTATION INITIATIVE

DDI Lifecycle

- Major expansion
  - Describe multiple waves for longitudinal/repeat data collection
  - Describe comparison and harmonization
  - Describe data collection and survey instruments
  - Describe the entire data lifecycle
- Reuse of metadata was central to these functions
  - Support for centralized metadata management
- Focus still primarily on rectangular data
- XML encoding
  - Machine-readable
  - Machine-actionable

DATA DOCUMENTATION INITIATIVE

The DDI Lifecycle Diagram (Original Version)



DATA DOCUMENTATION INITIATIVE

An Important Change…

- DDI Codebook allowed you to reference Concepts from variable descriptions
- DDI Lifecycle provided full-blown support for describing Concepts and reusing them
  - Referenced by Variables
  - Referenced by Categories in Classifications/Codelists
  - Referenced by Units/Populations/Universes
- With the popular “semantic” technologies, Concepts become central
  - SKOS is the most-used vocabulary in the RDF world
  - Basis of sematic mapping between organizations/domains

DDI Cross-Domain Integration (DDI-CDI)

- An extension of the metadata set found in DDI-C and DDI-L
  - Not a replacement!
  - Will be released Summer 2021
- Provides support for additional types of data
  - Long data/sensor data/event data
  - Multi-dimensional data/data “cubes”
  - Key-Value data/No SQL data/”big” data
- Provides support for describing process and provenance across data sets as data is reused/harmonized/integrated
- *Very* Concept-rich
- Focus is on individual “Datums”
- Model-based (UML), not just XML
  - Emphasis on machine-actionable metadata!

DATA DOCUMENTATION INITIATIVE

Contact

E-Mail:

ddi-train@googlegroups.com

Contact form:

https://ddialliance.org/learn/request-a-training-session

DATA DOCUMENTATION INITIATIVE

Questions?

DATA DOCUMENTATION INITIATIVE

Credits: DDI Training Working Group

Florio Orocio Arguillas
Alina Danciu
Adrian Dusa
Jane Fry
Martine Gagnon
Dan Gillman
Arofan Gregory
Taras Günther
Lea Sztuk Haahr
Chifundo Kanjala
Kaia Kulla
Kathryn Lavender
Amber Leahey
Jared Lyle
Alexandre Mairot
Laura Molloy
Lucie Marie
Hayley Mills
Hilde Orten
Anja Perry
Knut Wenzig

DDI®
DATA DOCUMENTATION INITIATIVE