

Data Capture: Tracking Data From Source to Results

Wendy Thomas
NADDI 2015
Madison, WI
April 9, 2015

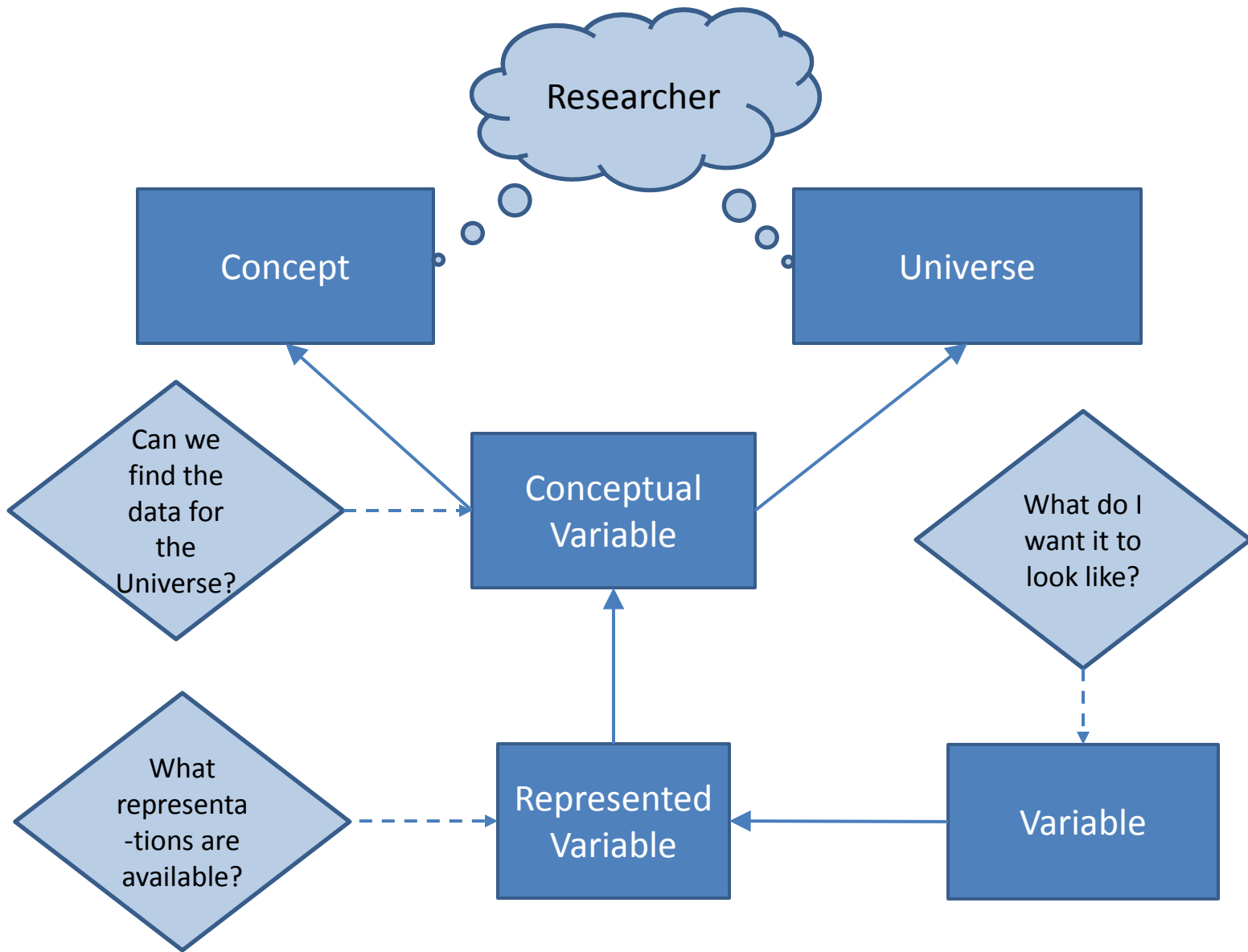
Over the years the focus of DDI has expanded from rendering a completed codebook into a machine processable format to recording data capture and production processes. However, much research involves the identification, restructuring, and reuse of data from existing sources.

How well does DDI capture this chain of events and the provenance of individual data objects? Can we create tools or processes to assist researchers in documenting this activity accurately and easily? A recent research project involving 45 indicators and 100-plus data sources covering a 23 year period serves as a case study for examining the following issues: Citing source data; Capturing selection criteria; Citing on-line extraction tools; Recoding, recalculating, transposing, and reformatting process capture; Linking the final research data set to its sources at the cell level; and Original metadata problems.

The presentation will focus on coverage gaps, practices for using existing DDI structures, and recommendations for tools or procedures to assist researchers in capturing this information and incorporating it into their final metadata documents

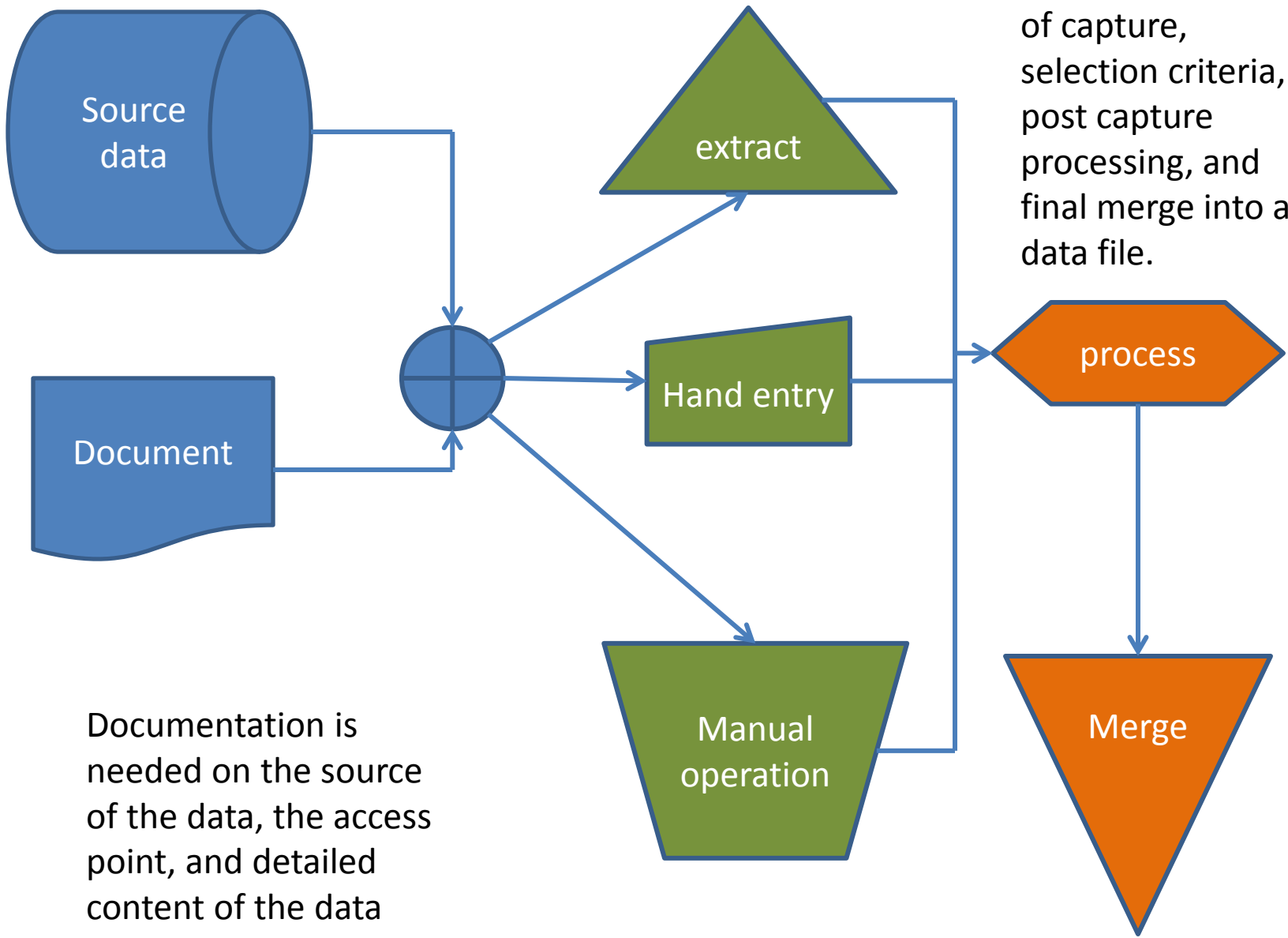
Problem statement

- Most students and many academic researchers base their research on previously collected data from one or more sources
- Data “capture” from secondary sources is not directly addressed in DDI
- Can the process of data capture from extracting data from a typical file to harvesting selected data from an on-line source be captured in DDI 3.2?
- Do the data sources provide sufficient metadata to support detailed provenance trails?



DDI 3.2 objects PRIOR to Data Capture

- Concept
 - Conceptual Variable
 - Represented Variable
- Universe
 - SubUniverse
- Geographic Structure
- Geographic Location

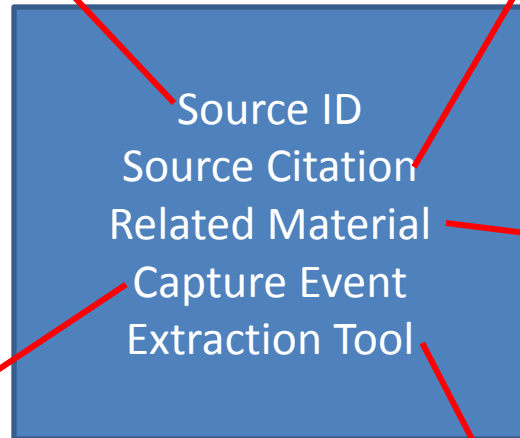


Document means of capture, selection criteria, post capture processing, and final merge into a data file.

Documentation is needed on the source of the data, the access point, and detailed content of the data source.

SOURCE_ID
Source_Selection_Criteria
Source_Quality
Source_Update_Procedure
Purpose
Spatial_Coverage
Temporal_Coverage
Access_Restrictions
Rights

CITATION:
Data_Title
Data_SubTitle
Creator
Producer
Owner
Distributor
Product_Site
Download_Site



RELATED MATERIALS:
Data_Dictionary
Additional_Metadata

CAPTURE EVENT:
Version
Data_Capture_Date
Capture_Agent
Data_File_Name

EXTRACTION TOOL:
Name_of_Tool
Table_Selection
Selected_Objects
Selection_Filter

Repeatable?	FIELD	Flag	URL/Filename	Text	Instructions
N	SOURCE_ID				TPOPyymmddnnn - a unique number comprised of a set preface, date, sequential 3 digit number. This becomes the base number for captured data or files that require the creation of a file name. Each source is a collection event from a specific source site.
N	Source_Selection_Criteria				Text description and/or URL/filename
N	Source_Quality				Text description and/or URL/filename
N	Source_Update_Procedure				Text description and/or URL/filename
N	Purpose				Text description and/or URL/filename
Y	Spatial_Coverage				Text description and/or URL/filename
Y	Temporal_Coverage				Text description and/or URL/filename
N	Access_Restrictions				Text description and/or URL/filename
N	Rights				Text description and/or URL/filename. Property, reuse rights.
	CITATION:				
N	Data_Title				If title is descriptive (not obtained from the source) Flag=A (assigned)
N	Data_SubTitle				For example a Table Identification or name
N	Creator				
N	Producer				
N	Owner				
N	Distributor				
N	Product_Site				
N	Download_Site				Specify Flag F=Full file S=Selected Variables/Records A=Analysis Tool (recoded/created variables)
	RELATED MATERIALS:				
N	Data_Dictionary				Enter file name in URN/Filename field and the title of the document in the Text Field. If you assign the file name use the Source_ID plus "_DD". If the data dictionary is structured in a standard format such as DDI place an "X" in the Flag field.
Y	Additional_Metadata				Flag defines Type: M=Methodology; C=Codelists; S=Setup files; D=General Documentation
	CAPTURE EVENT:				
N	Version				
N	Data_Capture_Date				ISO format yyyy-mm-dd or if time is important (dynamic data sets) yyyy-mm-ddThh:mm:ss-06:00
N	Capture_Agent				Name of person or machine used to capture data
Y	Data_File_Name				Note file name in URL/Filename and file format in Flag. NOTE to repeat this object ALL files must be captured in the same process at the same time (for example ACS summary files for all states, US - 52 files)
	EXTRACTION TOOL:				
N	Name_of_Tool				
Y	Table_Selection				Use if full tables are selected.
N	Selected_Objects				Separate object names with " "
Y	Selection_Filter				Object name and selection criteria using AND OR NOT etc. for example: SEX=M AND AGE >= 25 AND <=64

DDI 3.2 objects from Data Capture

- Methodology
 - Data collection methodology
 - Quality Statement
- Collection Event
 - Data Collector Organization Reference
 - Data Source
 - Data Collection Date
 - Mode Of Collection
 - Collection Situation
 - Quality Statement
 - Action to minimize loss
- Instrument with Control Constructs (?)
- Variable level information on source
- Physical level information on the source

The only way to link the Methodology information to the Collection Event is to create multiple Data Collection Modules

The quality of the data collected is dependent on the quality of the data source. Perhaps creating multiple data collection modules is a better solution than creating multiple data collection statements that are less than ideal solutions.

files) to create a limited DDI profile of your source data?

Capturing Data Processing

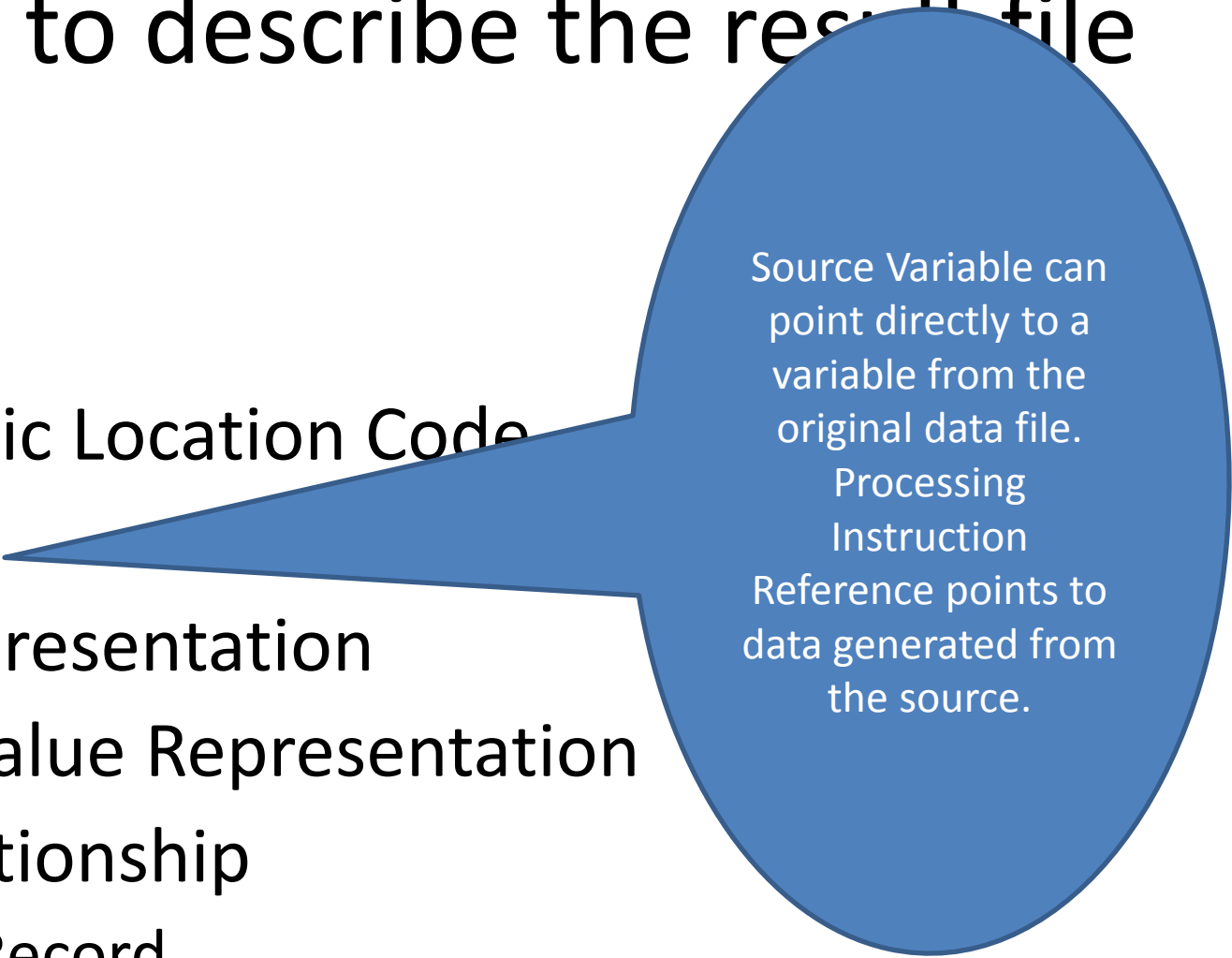
- Microdata to Aggregate data
 - Capturing filters and calculation processes from statistical software
- Selection, transposition, reordering of result data
 - Handling missing cases
 - Handling low n values

DDI 3.2 objects Post Capture

- Processing Event
 - Cleaning Operation
 - Control Operation
 - Data Appraisal Information
- Processing Instruction
 - General Instruction (for general processes)
 - Generation Instruction (for item specific processes)

DDI 3.2 to describe the result file

- Category
- CodeList
- Geographic Location Code
- Variable
- Value Representation
- Missing Value Representation
- Data Relationship
 - Logical Record



Source Variable can point directly to a variable from the original data file. Processing Instruction Reference points to data generated from the source.

DDI 3.2 objects to relate the result files

- Study Unit
- Abstract
- Purpose
- Funding Information
- Analysis Units Covered
- Kind of Data
- Data Relationship
 - Record Relationship

Issues with DDI 3.2

- Ideally want documentation for source data in DDI format in order to leverage reuse and provenance links on data objects
- Describing “data capture” for non-questionnaires requires a modified use of control constructs and instrument
- Difficult to bundle the following phases:
 - Data exploration
 - Data evaluation
 - Data capture
 - Data processing
- Need a real process model

Issues with data sources

- Documentation only available as a web page
- Limited information on variables
 - Source
 - Generation information
- Limited information on on-line calculation (aggregation, imputation, etc.) processes
- Lack of standardized structure for documentation
- Incomplete documentation (data dictionary only, separate study level documentation)
- PDF documentation

Recommendations

- Capture your foundational information in DDI
- Capture your search process and decisions regarding sources
- Use a standard capture form for each external source
 - Go with a tool that is familiar to the user
 - Transform captured source information into DDI structure

Recommendations (cont.)

- Use repetition of modules to bundle sets of information about sources together
- Create at least basic DDI Study Units for each source file
 - Not all information has to be in line, reference external documents by creating OtherMaterial descriptions and attaching to the appropriate element

Recommendations (cont.)

- Use Group to pull all of the source DDI Study Units and resulting files together
- The goal is to be able to track individual cell information back through its processing to the source
 - And to provide a link back to the provenance information provided by your data source
- Provide clear information on the quality of metadata obtainable for your source

Confessions of a busy data archivist...

- All of this is theoretical based on the case study noted in slide one and requests from specific projects to identify what they need to capture in order to provide provenance information
- None of these have been implemented
- There is already push-back in terms of determining the minimal amount of information they can get away with capturing