Data and Metadata Management: A Business Perspective

Wendy Thomas
Minnesota Population Center
IASSIST 2012 Workshop - 5 June 2012
Credits

• Slides 1-9, 11-13, 30-32, 65 Wendy Thomas
• Slide 10 Herve L’Hours (UKDA)
• Remainder of Slides:
  – The slides were developed for several DDI workshops at IASSIST conferences and at GESIS training in Dagstuhl/Germany
  – Major contributors
    • Wendy Thomas, Minnesota Population Center
    • Arofan Gregory, Open Data Foundation
  – Further contributors
    • Joachim Wackerow, GESIS – Leibniz Institute for the Social Sciences
    • Pascal Heus, Open Data Foundation
  – Attribute: http://creativecommons.org/licenses/by-sa/3.0/legalcode
Outline

• Introductions

• Business situation:
  – *What are the primary functions and responsibilities of your organization? What is the size and scope of your organization?*

• Business and environment – internal/external processes models:
  – *What processes take place within the organization? What objects pass through these processes? Who controls the input? Who uses the output and what do they need to do with it?*

• How to integrate DDI into an existing system:
  – Archival model
  – Production model
  – Mixed model
Data/Metadata Mgmt Activities

• Data Capture
  – determining what is to be collected from whom and how

• Data Processing
  – cleaning, normalizing, aggregating, harmonizing, creation of data products

• Data Discovery
  – Finding data, accessing data

• Preservation
  – short term and archival

• Process evaluation and revision
  – quality control, process improvement, evaluation and analysis

• Administrative tracking
  – who has control, where in the process
Introductions

• Who are you?

• What does your organization do?
  – Data collection
  – Data production
  – User access
  – Preservation

• What is the scale of your operations?
Changes in the environment:

• Expanded access to data
• Data available through multiple portals
• Cross portal access
• Linking and layering of data from different sources and different disciplines
• Cost of developing system specific software
• Cost of non-interoperability over the life of the data
Looking at your organization

• What activities take place and what materials do they involve?
• What specific processes take place and in what order?
• Which processes produce metadata of what type?
• What are the critical activities or processes?
• What do you control?
Traditional Process Model

Environment

INPUT

Process

OUTPUT
Archive/Data Discovery and Delivery

- Data and Metadata are generally received from external organizations.
- Focus is on moving data and metadata to a preservation format and supporting discovery and delivery tools.
- Management of ingest process (process management).
- “Value Added” material.
Archive/Data Discovery and Delivery

• Capturing full content
  – Machine actionable, information for discovery, retaining links to other materials, collections and grouping

• Added value metadata from archive
  – Variable, question, and data element groups related to subject and keyword access, linking to a common geography description, linking to an overall organization description, tracking archival management activities and processes
Archive/Data Discovery and Delivery

• How much can you influence depositors?
  – Ingest tools that result in DDI metadata
  – Provision of reusable materials (schemes) or controlled vocabularies
  – metadata management tools
  – Training

• What can be pushed back to long term depositors?
  – Resource package material?
  – Metadata of deposited data so that only differences are reported?
  – Tools to manage change over time?
DDI structures

• Study Unit
• Data Collection
  – Methodology
  – Questions
  – Question flow [optional]
• Variables
• Physical structures
• Group [optional]
• Local Holding Package
Study Unit

Citation / Series Statement
Abstract / Purpose
Coverage / Universe / Analysis Unit / Kind of Data
Other Material / Notes
Funding Information / Embargo

Conceptual Components
Data Collection
Logical Product
Physical Data Product
Physical Instance
Archive
DDI Profile
3.1 Local Holding Package

Citation / Series Statement
Abstract / Purpose
Coverage / Universe
Other Material / Notes
Funding Information / Embargo

**Depository Study Unit OR Group Reference:**
[A reference to the stored version of the deposited study unit.]

**Local Added Content:**
[This contains all content available in a Study Unit whose source is the local archive.]
Study Unit

• Study Unit
  – Identification
  – Coverage
    • Topical
    • Temporal
    • Spatial
  – Conceptual Components
    • Universe
    • Concept
    • Representation (optional replication)
  – Purpose, Abstract, Proposal, Funding

• Identification is mapped to Dublin Core and basic Dublin Core is included as an option
• Geographic coverage mapped to FGDC / ISO 19115
  – bounding box
  – spatial object
  – polygon description of levels and identifiers
• Universe Scheme, Concept Scheme
  – link of concept, universe, representation through Variable
  – also allows storage as a ISO/IEC 11179 compliant registry
Data Collection

- Methodology
- Question Scheme
  - Question
  - Response domain
- Instrument
  - using Control Construct Scheme
- Coding Instructions
  - question to raw data
  - raw data to public file
- Interviewer Instructions
- Question and Response Domain designed to support question banks
  - Question Scheme is a maintainable object
- Organization and flow of questions into Instrument
  - Used to drive systems like CASES and Blaise
- Coding Instructions
  - Reuse by Questions, Variables, and comparison
Logical Product

- Category Schemes
- Coding Schemes
- Variables
- NCubes
- Variable and NCube Groups
- Data Relationships

- Categories are used as both question response domains and by code schemes
- Codes are used as both question response domains and variable representations
- Link representations to concepts and universes through references
- Built from variables (dimensions and attributes)
  - Map directly to SDMX structures
  - More generalized to accommodate legacy data
Physical storage

• Physical Data Structure
  – Links to Data Relationships
  – Links to Variable or NCube Coordinate
  – Description of physical storage structure
    • in-line, fixed, delimited or proprietary

• Physical Instance
  – One-to-one relationship with a data file
  – Coverage constraints
  – Variable and category statistics
Archive Module

• The Archive module is used to track lifecycle events and provide information about who was responsible for each event
  – The use of this module is optional
  – It provides support throughout the lifecycle, or for just some specific portion of the lifecycle within a single organization

• Lifecycle events are any process step which is significant to the creator of the metadata
  – Can reflect OAIS archiving model, etc.
  – Completely configurable
Archiving and Organizations/Individuals

• Archive contains:
  – Archive-specific information about the holdings in the archive (access, funding information, embargoes, etc.)
  – A list of organizations and individuals, with contact details, etc. (the Organization Scheme)
  – A list of lifecycle events, which reference the acting organization, the date, the type of event, a description of it, and a link to the affected metadata
  – Contains Other Materials and Notes
Lifecycle Events

• Basic information: type of event, date, responsible organization/individual, and description of the event
• Use to list major development activities in the study
• Use to record archival activities such as acquisition, validation, value added, archive management activities, etc.
• May link to specific metadata affected by the event
Mining the Archive

• With metadata about relationships and structural similarities
  – You can automatically identify potentially comparable data sets
  – You can navigate the archive’s contents at a high level
  – You have much better detail at a low level across divergent data sets
Long term data collection process

• Goal may be from cradle to grave or as much as has value to the process
• Data Element management, concepts (and variations on a scheme), questions, question flows, data processing steps and instructions,
• Quality control aspects
• A collection process undergoing change (paper to online collection) - providing a base and then moving it back into development process, providing tools and support for backward integration of processes. Finding the payoff for the business process
Note the similarity to the DDI Combined Lifecycle Model and the top level of the GSBPM
DDI structures

- Schemes
  - Data Element
  - Concepts
  - Geography
  - Questions
  - Variables
- Group
- Comparison
- Control Construct
- Processing Events
- Processing Instructions
Why can DDI 3 do more?

• It is machine-actionable – not just documentary
• It’s more complex with a tighter structure
• It manages metadata objects through a structured identification and reference system that allows sharing between organizations
• It has greater support for related standards
• Reuse of metadata within the lifecycle of a study and between studies
DDI Schemes

• Brief overview of what DDI schemes are and what they are designed to do including:
  – Purpose of DDI Schemes
  – How a DDI Study is built using information held in schemes
DDI Schemes: Purpose

- A maintainable structure that contains a list of versionable things
- Supports registries of information such as concept, question and variable banks that are reused by multiple studies or are used by search systems to location information across a collection of studies
- Supports a structured means of versioning the list
- May be published within Resource Packages or within DDI modules
- Serve as component parts in capturing reusable metadata within the life-cycle of the data
XML Schemas, DDI Modules, and DDI Schemes

XML Schemas

May Correspond

<file>.xsd
<file>.xsd
<file>.xsd
<file>.xsd

DDI Modules

May Contain

Correspond to a stage in the lifecycle

DDI Schemes
XML Schemas, DDI Modules, and DDI Schemes

Instance
- Data Collection
- Logical Product
- Physical Data Structure
- Archive
- Conceptual Component

Reusable
- Ncube
- Inline ncube
- Tabular ncube
- Proprietary
- Dataset

Study Unit

Physical Instance

DDI Profile

Comparative

Physical Instance

S09
XML Schemas, DDI Modules, and DDI Schemes
XML Schemas, DDI Modules, and DDI Schemes

Data Collection
- Question Scheme
- Control Construct Scheme
- Interviewer Instruction Scheme

Logical Product
- Category Scheme
- Code Scheme
- Variable Scheme
- NCube Scheme

Physical Data Structure
- Physical Structure Scheme
- Record Layout Scheme

Archive
- Organization Scheme

Conceptual Component
- Concept Scheme
- Universe Scheme
- Geographic Structure Scheme
- Geographic Location Scheme

Reusable
- Ncube
- Inline ncube
- Tabular ncube
- Proprietary
- Dataset
Why Schemes?

- You could ask “Why do we have all these annoying schemes in DDI?”
- There is a simple answer: reuse!
- DDI 3 supports the concept of metadata registries (e.g., question banks, variable banks)
- DDI 3 also needs to show specifically where something is reused
  - Including metadata by reference helps avoid error and confusion
  - Reuse is explicit
Designed to Support Registries

- A “Registry” is a catalog of metadata resources
- Resource package
  - Structure to publish non-study-specific materials for reuse
- Extracting specified types of information into schemes
  - Universe, Concept, Category, Code, Question, Instrument, Variable, etc.
- Allowing for either internal or external references
  - Can include other schemes by reference and select only desired items
- Providing Comparison Mapping
  - Target can be external harmonized structure
Management of Information, Data, and Metadata

• An organization can manage its organizational information, metadata, and data within repositories using DDI 3 to transfer information into and out of the system to support:
  – Controlled development and use of concepts, questions, variables, and other core metadata
  – Development of data collection and capture processes
  – Support quality control operations
  – Develop data access and analysis systems
Upstream Metadata Capture

- Because there is support throughout the lifecycle, you can capture the metadata as it occurs
- It is re-useable throughout the lifecycle
  - It is versionable as it is modified across the lifecycle
- It supports production at each stage of the lifecycle
  - It moves into and out of the software tools used at each stage
Metadata Driven Data Capture

• Questions can be organized into survey instruments documenting flow logic and dynamic wording
  – This metadata can be used to create control programs for Blaise, CASES, CSPro and other CAI systems

• Generation Instructions can drive data capture from registry sources and/or inform data processing post capture
Reuse of Metadata

• You can reuse many types of metadata, benefitting from the work of others
  – Concepts
  – Variables
  – Categories and codes
  – Geography
  – Questions
• Promotes interoperability and standardization across organizations
• Can capture (and re-use) common cross-walks
Simple Questionnaire

Please answer the following:

1. Sex
   (1) Male
   (2) Female

2. Are you 18 years or older?
   (0) Yes
   (1) No (Go to Question 4)

3. How old are you? ______

4. Who do you live with?
   ______________________

5. What type of school do you attend?
   (1) Public school
   (2) Private school
   (3) Do not attend school
Simple Questionnaire

Please answer the following:

1. **Sex**
   (1) Male
   (2) Female

2. **Are you 18 years or older?**
   (0) Yes
   (1) No (Go to Question 4)

3. **How old are you?** ______

4. **Who do you live with?**
   __________________________

5. **What type of school do you attend?**
   (1) Public school
   (2) Private school
   (3) Do not attend school
Simple Questionnaire

Please answer the following:

1. Sex
   (1) Male
   (2) Female

2. Are you 18 years or older?
   (0) Yes
   (1) No (Go to Question 4)

3. How old are you? ______

4. Who do you live with?
   ______________________

5. What type of school do you attend?
   (1) Public school
   (2) Private school
   (3) Do not attend school
Representing Response Domains

• There are many types of response domains
  – Many questions have categories/codes as answers
  – Textual responses are common
  – Numeric responses are common
  – Other response domains are also available in DDI 3 (time, mixed responses)
Simple Questionnaire

Please answer the following:

1. Sex
   (1) Male
   (2) Female

2. Are you 18 years or older?
   (0) Yes
   (1) No (Go to Question 4)

3. How old are you? ______

4. Who do you live with?
   ______________________

5. What type of school do you attend?
   (1) Public school
   (2) Private school
   (3) Do not attend school

• Questions

• Response Domains
  – Code
  – Numeric
  – Text

• Statements
Simple Questionnaire

Please answer the following:

1. Sex
   (1) Male
   (2) Female

2. Are you 18 years or older?
   (0) Yes
   (1) No (Go to Question 4)

3. How old are you? ______

4. Who do you live with? __________________

5. What type of school do you attend?
   (1) Public school
   (2) Private school
   (3) Do not attend school
Simple Questionnaire

Please answer the following:

1. Sex
   (1) Male
   (2) Female

2. Are you 18 years or older?
   (0) Yes
   (1) No (Go to Question 4)

3. How old are you? ______

4. Who do you live with?
   ______________________

5. What type of school do you attend?
   (1) Public school
   (2) Private school
   (3) Do not attend school

• Questions
• Response Domains
  – Code
  – Numeric
  – Text

• Statements
• Instructions
• Flow
Statement 1 → Question 1 → Question 2 → Is Q2 = 0 (yes) → No

Yes → Question 3 → Question 4 → Question 5
Process Items

• General Coding Instruction
  – Missing Data (left as blanks)
  – Suppression of confidential information such as name or address

• Generation Instructions
  – Recodes
    • Review of text answers where items listed as free text result in more than one nominal level variable
      – Create variable for each with 0=no 1=yes
    • Or a count of the number of different items provided by a respondent
  – Aggregation etc.
    • The creation of new variables whose values are programmatically populated (mostly from existing variables)
IPUMS International

- Archival for INPUT, INPUT is data capture for harmonization and creation of data product, PRODUCT is input to archive and data discovery, OUTPUT is delivered to client who treats it as INPUT to their own process
- Focus of IPUMS and the underlying database is the PRODUCT
- We can create a basic DDI instance for the full product and subsets delivered to the client
- Instance does NOT capture the original INPUT structure or changes due to processing
- [codebook content, question bank, comparison, code lists, variable banks, data items]
- [capture metadata from INPUT in a structured way - what can be harvested from current practice, capture change process, focus on actionable metadata and input to system - retain links to source materials, move all used support information to the database to enforce structural consistency, capture change over time for series (censuses within a single country) capturing difs]
DDI Structures

- Question
- Variables
- Interviewer Instructions
- Versioning
- Comparison
- Other Materials
- Organization scheme
Questionnaires

• Questions
  – Question Text
  – Response Domains

• Statements
  – Pre-Post-question text

• Instructions
  – Routing information
  – Explanatory materials

• Question Flow
General Variable Components

• VariableName, Label and Description
• Links to Concept, Universe, Question, and Embargo information
• Provides Analysis and Response Unit
• Provides basic information on its role:
  – isTemporal
  – isGeographic
  – isWeight
• Describes Representation
Representation

• Detailed description of the role of the variable
• References related weights (standard and variable)
• References all instructions regarding coding and imputation
• Describes concatenated values
• Additivity and aggregation method
• Value representation
• Specific Missing Value description (proposed DDI 3.2)
  – Can be used in combination with any representation type
Value Representation

• Provides the following elements/attributes to all representation types:
  – classification level ("nominal", "ordinal", "interval", "ratio", "continuous")
  – blankIsMissingValue ("true" "false")
  – missingValue (expressed as an array of values)
  – These last 2 may be replaced in 3.2 by a missing values representation section

• Is represented by one of four representation types (numeric, text, code, date time)

• Additional types are under development (i.e., scales)
Comparison

• There are two types of comparison in DDI 3:
  – Comparison by design
  – Ad-hoc (after-the-fact) comparison

• Comparison by design can be expressed using the grouping and inheritance mechanism

• Ad-hoc comparison can be described using the comparison module

• The comparison module is also useful for describing harmonization when performing case selection activities
Data Comparison

• To compare data from different studies (or even waves of the same study) we use the metadata
  – The metadata explains which things are comparable in data sets

• When we compare two variables, they are comparable if they have the same set of properties
  – They measure the same concept for the same high-level universe, and have the same representation (categories/codes, etc.)
  – For example, two variables measuring “Age” are comparable if they have the same concept (e.g., age at last birthday) for the same top-level universe (i.e., people, as opposed to houses), and express their value using the same representation (i.e., an integer from 0-99)
  – They may be comparable if the only difference is their representation (i.e., one uses 5-year age cohorts and the other uses integers) but this requires a mapping
DDI Support for Comparison

• For data which is completely the same, DDI provides a way of showing comparability: Grouping
  – These things are comparable “by design”
  – This typically includes longitudinal/repeat cross-sectional studies

• For data which *may* be comparable, DDI allows for a statement of what the comparable metadata items are: the Comparison module
  – The Comparison module provides the mappings between similar items (“ad-hoc” comparison)
  – Mappings are always context-dependent (e.g., they are sufficient for the purposes of particular research, and are only *assertions* about the equivalence of the metadata items)
Comparability

• The comparability of a question or variable can be complex. You must look at all components. For example, with a question you need to look at:
  – Question text
  – Response domain structure
    • Type of response domain
    • Valid content, category, and coding schemes

• The following table looks at levels of comparability for a question with a coded response domain

• More than one comparability “map” may be needed to accurately describe comparability of a complex component
## Detail of question comparability

<table>
<thead>
<tr>
<th>Comparison Map</th>
<th>Textual Content of Main Body</th>
<th>Category</th>
<th>Code Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same</td>
<td>Similar</td>
<td>Same</td>
</tr>
<tr>
<td>Question</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Points to remember

• Few are starting from scratch
• Ongoing processes cannot stop
• You can only act in areas you control
• DDI is not an all or nothing structure
Check list

• Who do you need to interact with in your environment INPUTS and OUTPUTS?
• Where is your focus (may be different for different parts of the organization)?
• What do you control?
• What is your process flow - how far upstream is it practical to insert DDI like structures?
ISO/IEC 11179-1