The Data Documentation Initiative (DDI): A bridge between data producers and data users

Mary Vardigan
DDI Alliance Director and Assistant Director, ICPSR

FedCASIC
March 7, 2007

Value of Metadata

Whereas the creators and primary users of statistics might possess “undocumented” and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal metadata that travels along with the data in order to exploit their full potential. … The metadata provide the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

-– Jostein Ryssevik, "The Data Documentation Initiative (DDI) Metadata Specification"

<ddi>

DDI is…

- An emerging standard for social science metadata – now in Version 3.0
- A life-cycle model
- A project of the international social science research community
- A comprehensive solution for documentation and other products

Features of the Standard

- Written in XML Schemas
  - <Title>Consumer Expenditure Survey</Title>
- Machine-processable with strict data typing
- Modular
- Extensible

More Features

- Aligned with other standards, including ISO 11179, MARC, Dublin Core
- Built to support multiple languages
- Descriptive for comparative data

DDI 3.0 Modules

- Study unit
- Data collection
- Archive
- Logical data product
- Physical data product
- Organizations
- Grouping
- Comparative

Life Cycle Coverage

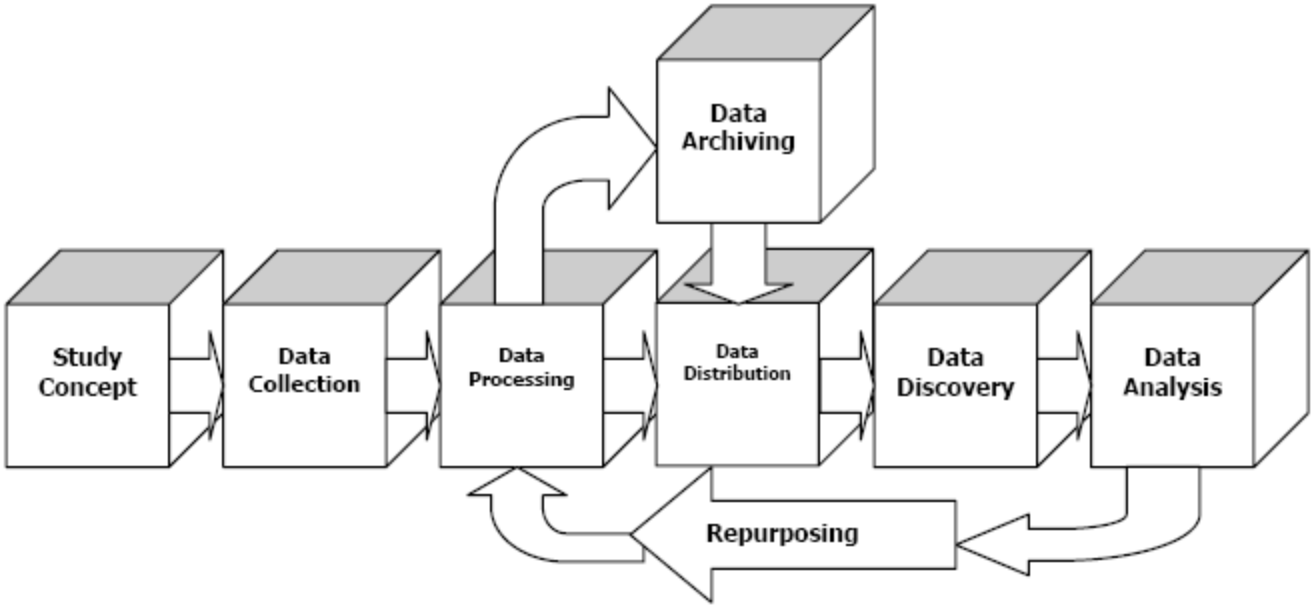

Figure: Combined Life Cycle Model

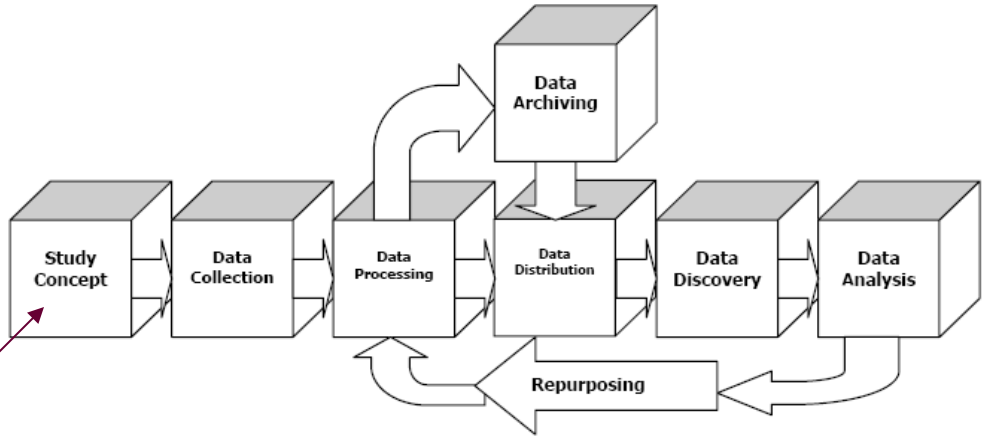

Figure: Combined Life Cycle Model

- In the beginning -- the Study Unit module
  - Research question or purpose of the study
  - Who is proposing the study
  - Research population
  - Background research
  - Formal study proposal reference
  - Funding sources
  - Concepts and definitions

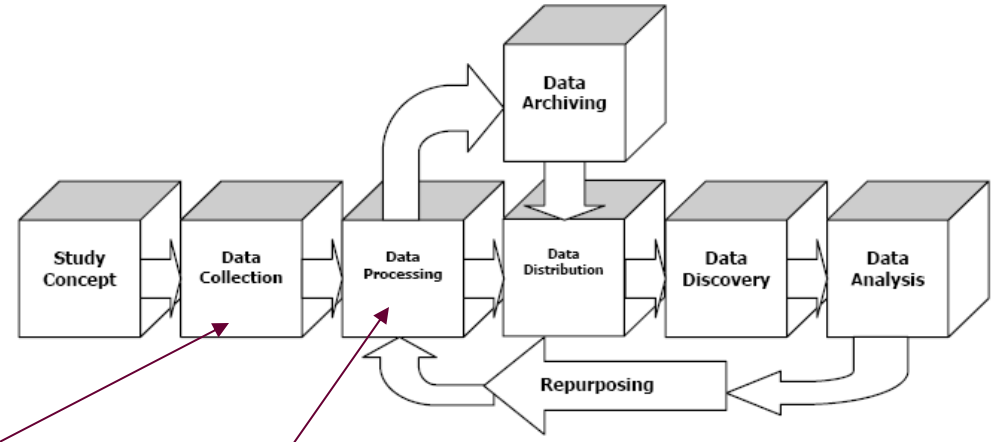

Figure: Combined Life Cycle Model

- Proposal becomes reality – the Data Collection Module
  - Questionnaire development
  - Instrumentation
  - Who collects the data
  - How the data are collected
  - Data cleaning
  - Data processing
    - Recoding and weighting
    - Imputation

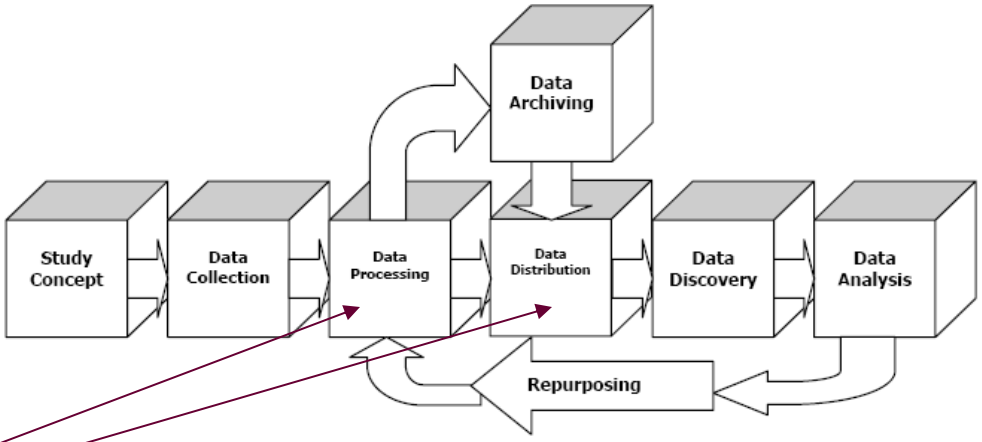

Figure: Combined Life Cycle Model

- Publishing the data -- the logical product module
  - Variables
  - Categories
  - Including links to previously defined
    - Concepts
    - Questions
    - Processing activities

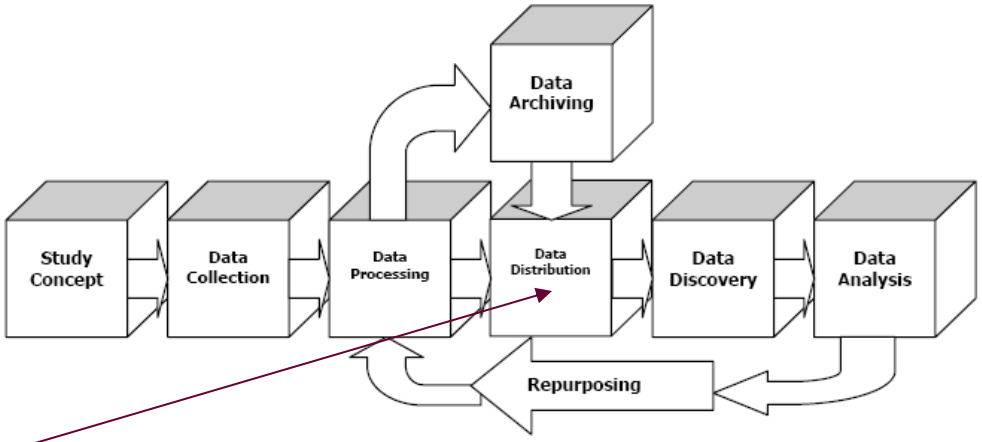

Figure: Combined Life Cycle Model

- The Physical Data Product module
  - Link to a logical product description
  - Physical storage structure and variable locations
  - Information on relational links between record types stored in one or many data files
  - Summary statistics

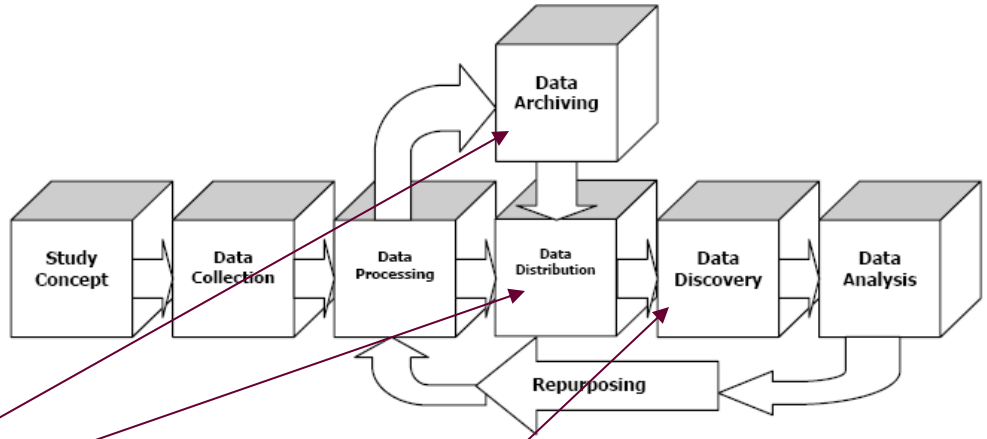

Figure: Combined Life Cycle Model

- Making it visible to the outside world – Data distribution and the Archive module
  - Depositing with a distributor/archive
  - Packaging and publishing as a DDI Instance
  - DDI Instance has identity as an object
  - Metadata can now be “discovered”

DDI Development

- 29 institutions participate in the self-sustaining Alliance – working groups structure
- Effort began in 1995 to create new standard to replace OSIRIS tagged codebooks
- Began as SGML, then converted to Web-friendly XML
- Began as document- and codebook-centric but became broader
- DDI 1 and 2 --- to DDI 3

DDI Alliance

- www.ddialliance.org
- Working Groups structure
  - Technical group
  - Aggregate data, geography, time
  - Instrument documentation
  - Comparative data
  - Usability and outreach
  - New group forming on Survey Design and Implementation

A Comprehensive Solution

- Designed to provide a core document from which other documents can be generated
- Provides for repurposing
- Enhances data discovery through precision in searches
- Forms basis for online analysis
- Carries “intelligence” in the document
- Is optimal for preservation

<ddi>

Typical Example

<var name="**V4439**" ID="**V4439**">
<location StartPos="**544**" EndPos="**544**" width="**1**" />
<labl level="**variable**">**HOW SOON DID POLICE RESPOND**</labl>
<qstn ID="**Q.122**">
Q.122:
<qstnLit>**How soon after the police found out did they respond? Was it within 5 minutes, within 10 minutes, an hour, a day or longer?**</qstnLit>
</qstn>
<invalrng>
<item VALUE="**9**" />
</invalrng>
<txt>**Source code 815**</txt>
<catgry>
<catValu>**1**</catValu>
<labl level="**category**">**Within 5 minutes**</labl>
</catgry>
<catgry>
<catValu>**2**</catValu>
<labl level="**category**">**Within 10 minutes**</labl>
</catgry>
<catgry>
<catValu>**3**</catValu>
<labl level="**category**">**Within an hour**</labl>
</catgry>
<catgry>
<catValu>**4**</catValu>
<labl level="**category**">**Within a day**</labl>
</catgry>
<catgry>
<catValu>**5**</catValu>
<labl level="**category**">**Longer than a day**</labl>
</catgry>
<catgry>
<catValu>**6**</catValu>
<labl level="**category**">**Don't know how soon**</labl>
</catgry>
<catgry>
<catValu>**8**</catValu>
<labl level="**category**">**Residue**</labl>
</catgry>
<catgry missing="**Y**">
<catValu>**9**</catValu>
<labl level="**category**">**Out of universe**</labl>
</catgry>
<codInstr>**MARK (X) FIRST CATEGORY RESPONDENT IS SURE OF**</codInstr>
</var>

Example

- Rendered in PDF through a stylesheet

V4439	HOW SOON DID POLICE RESPOND
Location:	544-544 (width: 1; decimal: 0)
Interval:	discrete
Question:	Q.122:
Literal Question:	How soon after the police found out did they respond? Was it within 5 minutes, within 10 minutes, an hour, a day or longer?
Text:	Source code 815

Value	*Label*
1	Within 5 minutes
2	Within 10 minutes
3	Within an hour
4	Within a day
5	Longer than a day
6	Don't know how soon
8	Residue
9 (M)	Out of universe

Coder Instructions: MARK (X) FIRST CATEGORY RESPONDENT IS SURE OF

Same variable in Online Analysis

V4439 **HOW SOON DID POLICE RESPOND**

Description of the Variable

```
Source code 815

Q.122 How soon after the police found out did they
respond?  Was it within 5 minutes, within 10 minutes,
an hour, a day or longer?

MARK (X) FIRST CATEGORY RESPONDENT IS SURE OF
-----------------------------------------------------
```

Percent	N	Value	Label
17.0	6,623	**1**	Within 5 min
22.1	8,592	**2**	Within 10 min
42.0	16,359	**3**	Within an hour
10.4	4,036	**4**	Within a day
1.6	628	**5**	More than a day
5.8	2,249	**6**	Dont know
1.2	467	**8**	Residue
	123,782	**9**	Out of universe
100.0	**162,736**		**Total**

Properties

Data type:	numeric
Missing-data codes:	9
Mean:	2.81
Std Dev:	1.36
Record/column:	1/1021

Selected Study: NCVS Concatenated Incident-Level File (most recent)

Benefits of Markup at the Source --CAI Systems

- Documentation of instrument
- Codebook
- Data cleaning
- Nonproprietary format
- Generation of setup files
- Promising development – shared SRO-ICPSR database design

Final Word About Metadata

The current metadata movement in the social sciences is predicated on Internet-operable standards for describing and processing information, which will transform the way we think about and use data documentation. The most advanced metadata standard in this area is the Data Documentation Initiative (DDI).

–Raymond F. Currie and Chuck Humphrey, Report of Metadata Group sponsored by Canada's Research Data Centre Network

<ddi>