

Data and Metadata Management: A Business Perspective

Wendy Thomas – Minnesota Population Center

Marcel Hebing – German Socio-Economic Panel Study (SOEP), DIW Berlin

EDDI 2012

EDDI-2012_Business by Wendy Thomas and Marcel Hebing is licensed under a
[Creative Commons Attribution-NonCommercial\\-ShareAlike 3.0 Unported License](#).

Credits

UPDATE IN THE END!

- Slides 1-7, 13-14, 64-65 Wendy Thomas
- Slide 53 Herve L'Hours (UKDA)
- Remainder of Slides:
 - The slides were developed for several EDC 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>

PART I

Introduction to DDI

Introduction to DDI

WELCOME AND OUTLINE

Outline

Introduction to DDI

- Welcome
- DDI in 60 Seconds
- Processes
- DDI-Standard
- Tools
- Codebook and Lifecycle

Advanced Topics

- DDI-L in Detail
- Questionnaire example
- Schemes and Reuse
- Comparison
- Question & Answer
- Conclusion

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

The Challenge

- In a large organization, there are many different streams of data production
- “Silos” tend to emerge, each with a different set of systems and processes
- This makes the management of data, metadata, and the production process very difficult – the different systems/silos don’t interoperate easily
- It is difficult to realize the vision of “industrialized” data production, with its attendant efficiencies

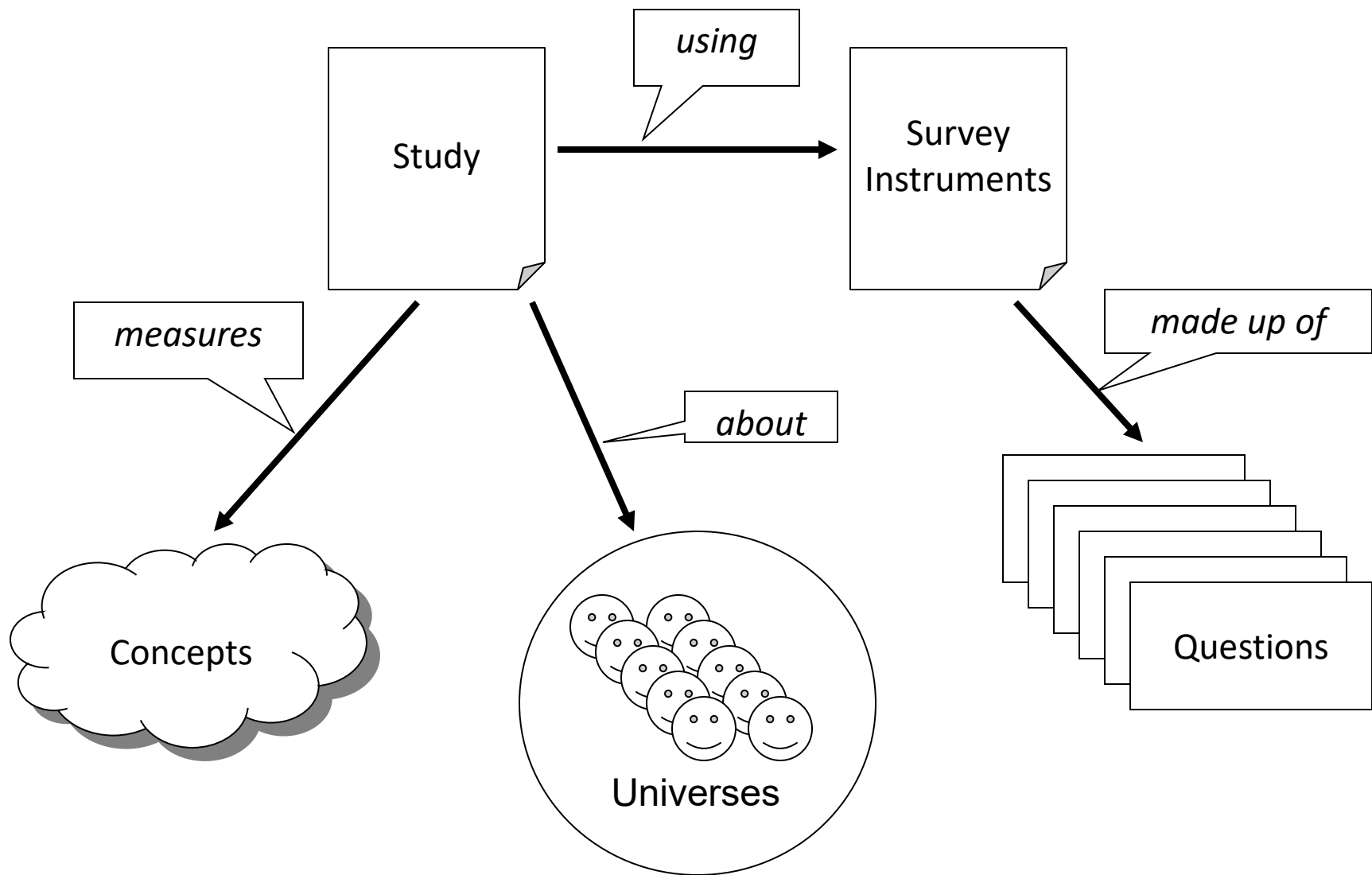
Why Such a Mess?

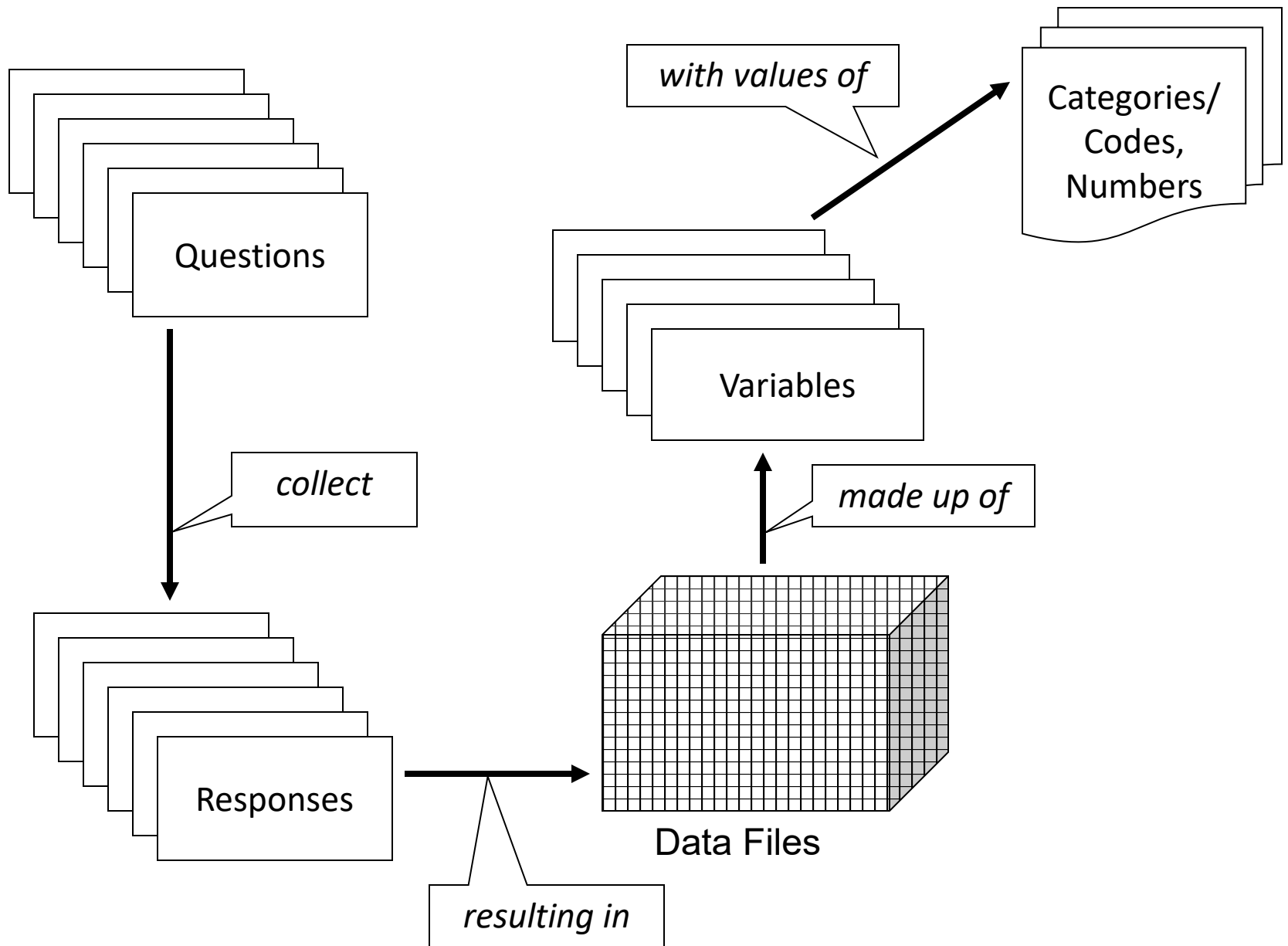
- Everyone thinks their data are special
- And they are right – they are special...
- They're just not as special as they think!
- The point is that good IT solutions for data management can be built on the basis of the *similarities* across the silos
 - Computers can't think!
 - They manipulate the structural aspects of data
 - The structural aspects of data are the same



Introduction to DDI

DDI 3 IN 60 SECONDS





Introduction to DDI

PROCESSES

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?

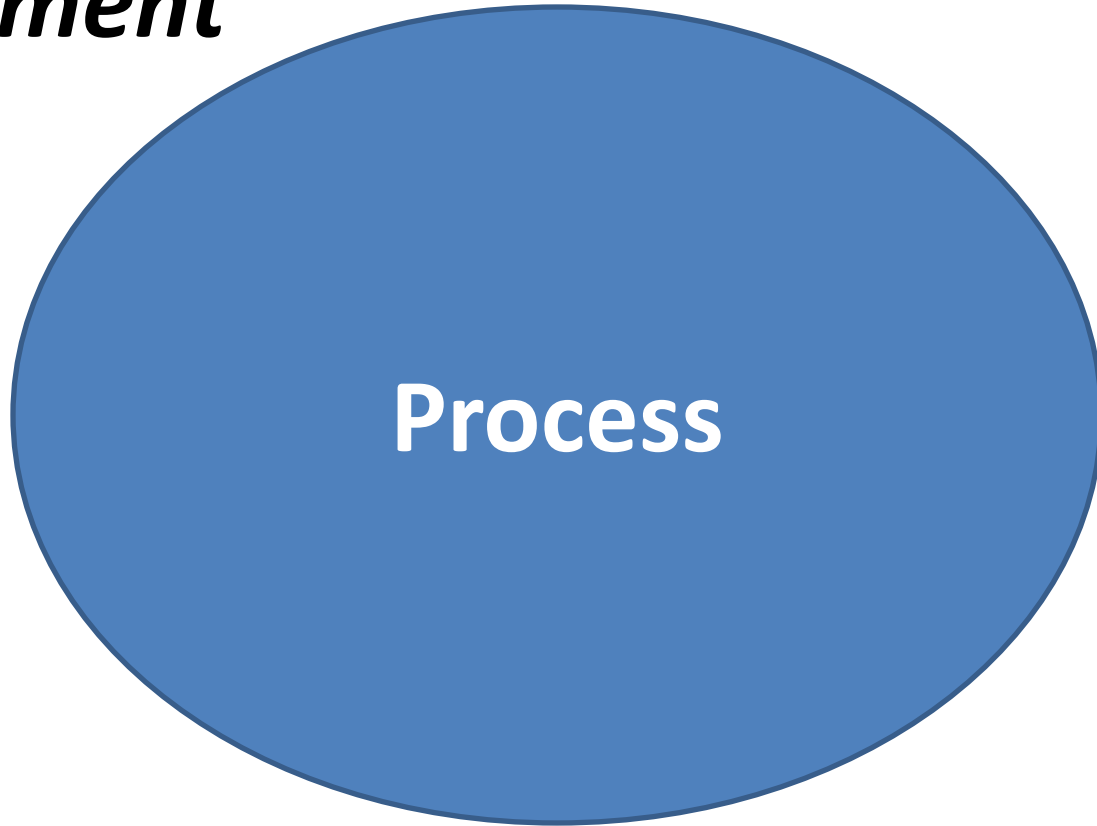
Traditional Process Model

Environment

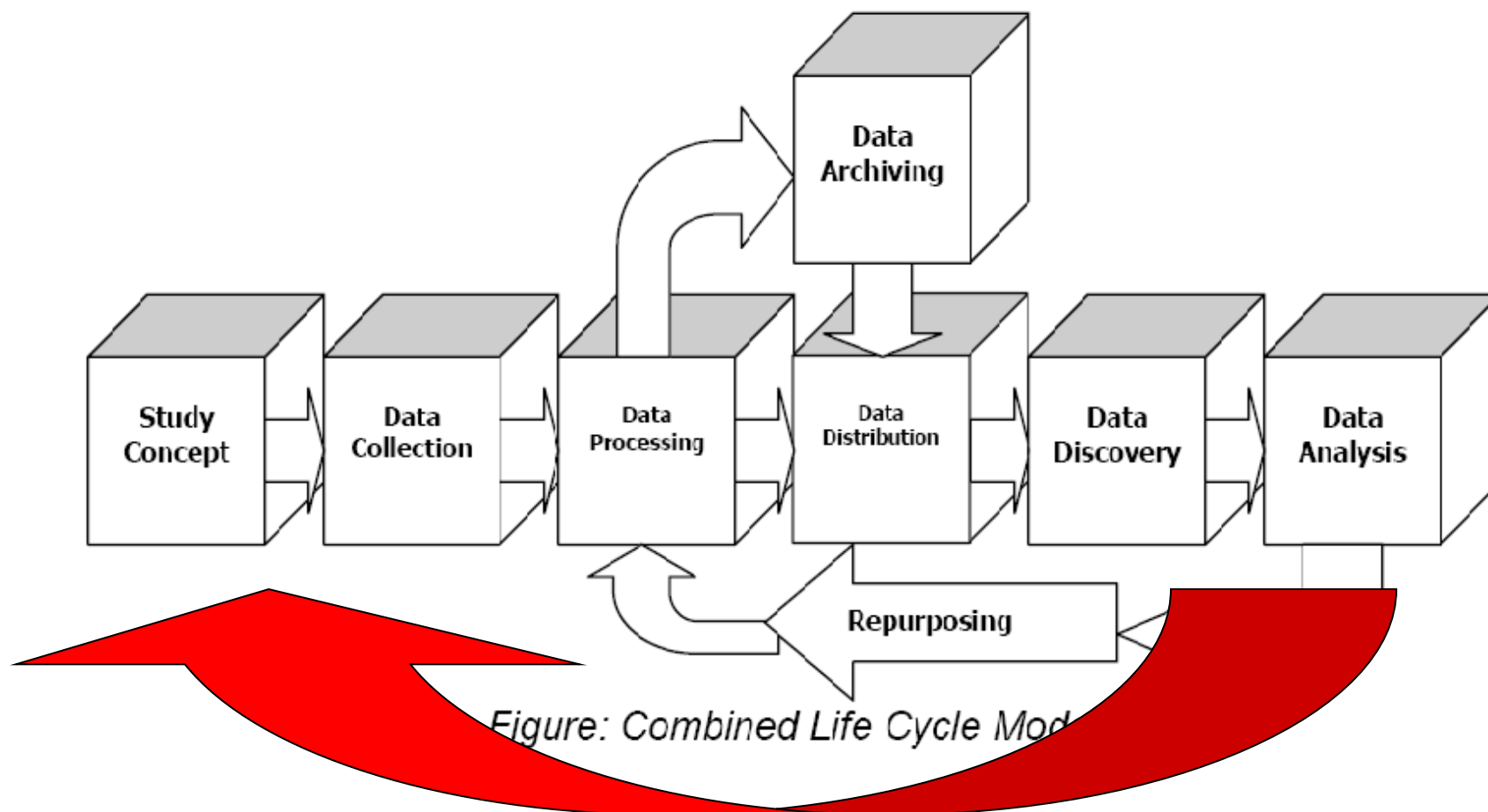
INPUT

Process

OUTPUT



DDI Lifecycle Model



Metadata Reuse

Data/Metadata Life Cycle Orientation

pre-production

production

post-production

secondary use

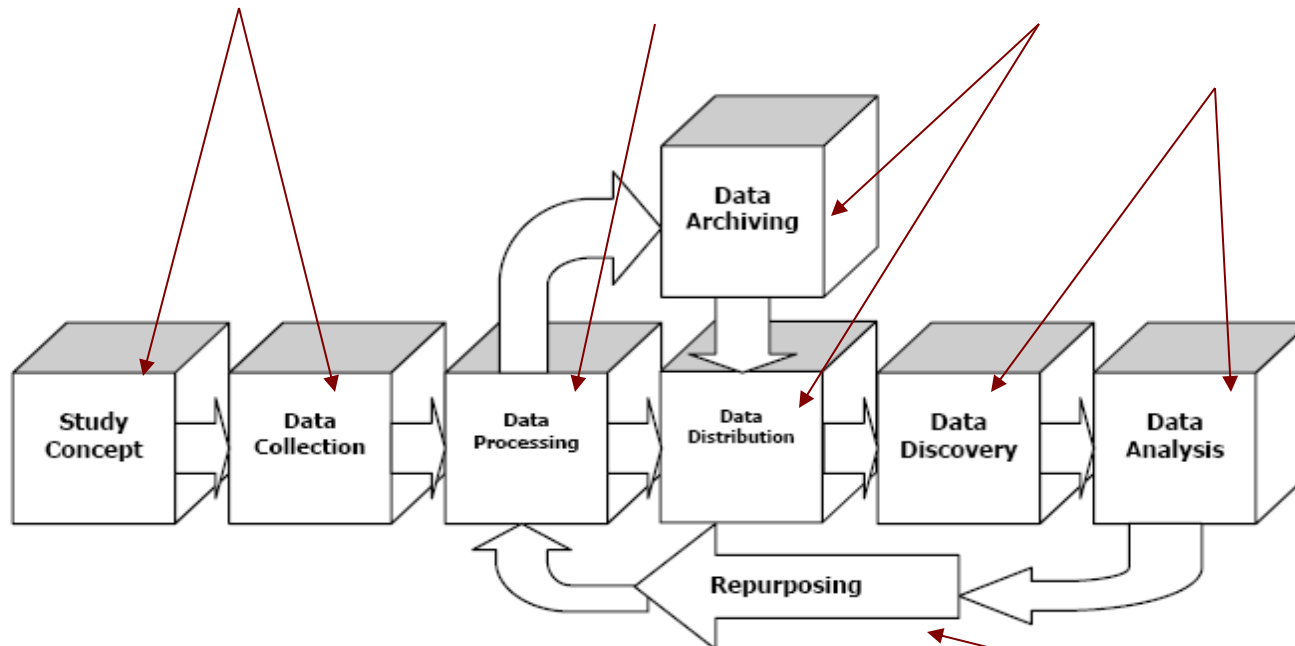
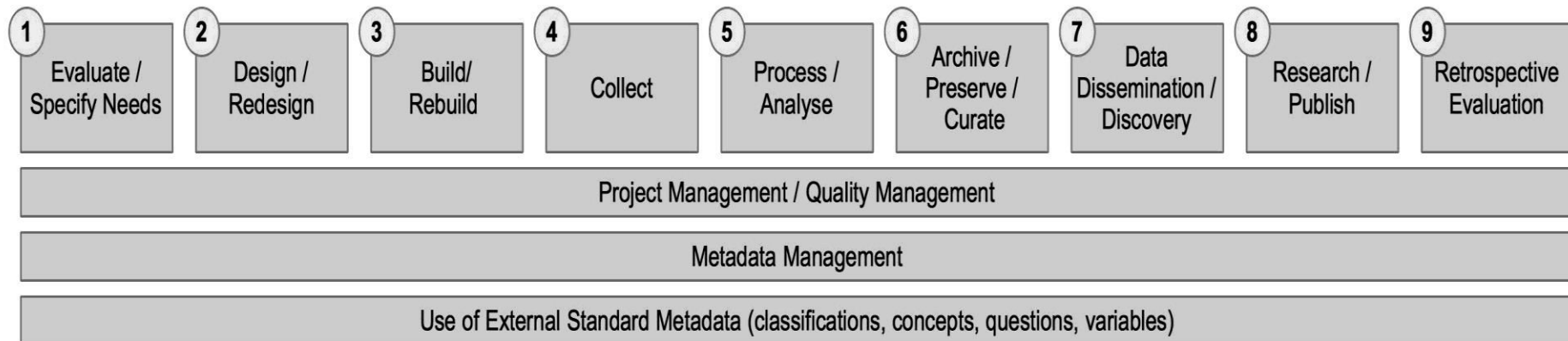


Figure: Combined Life Cycle Model

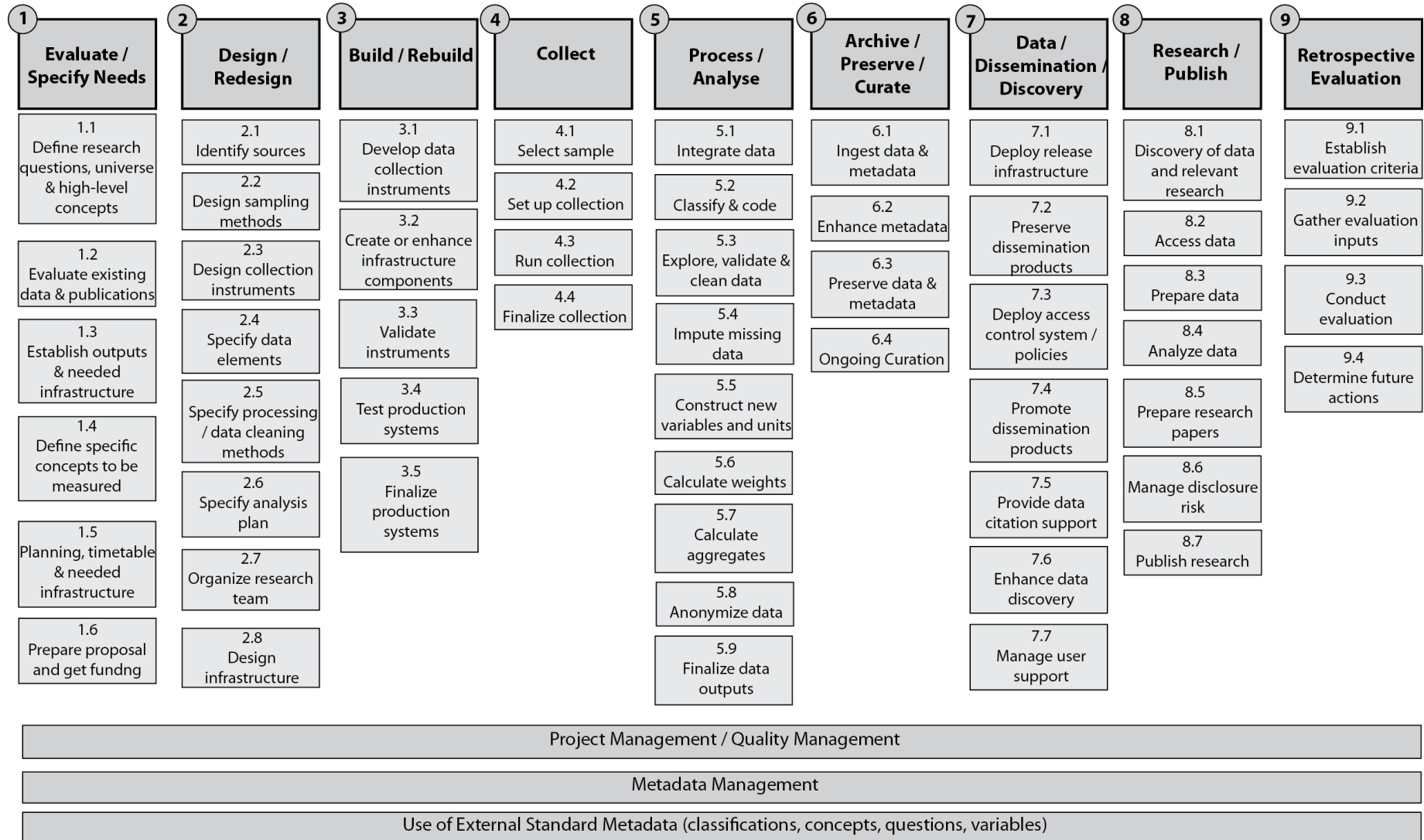
new research
effort

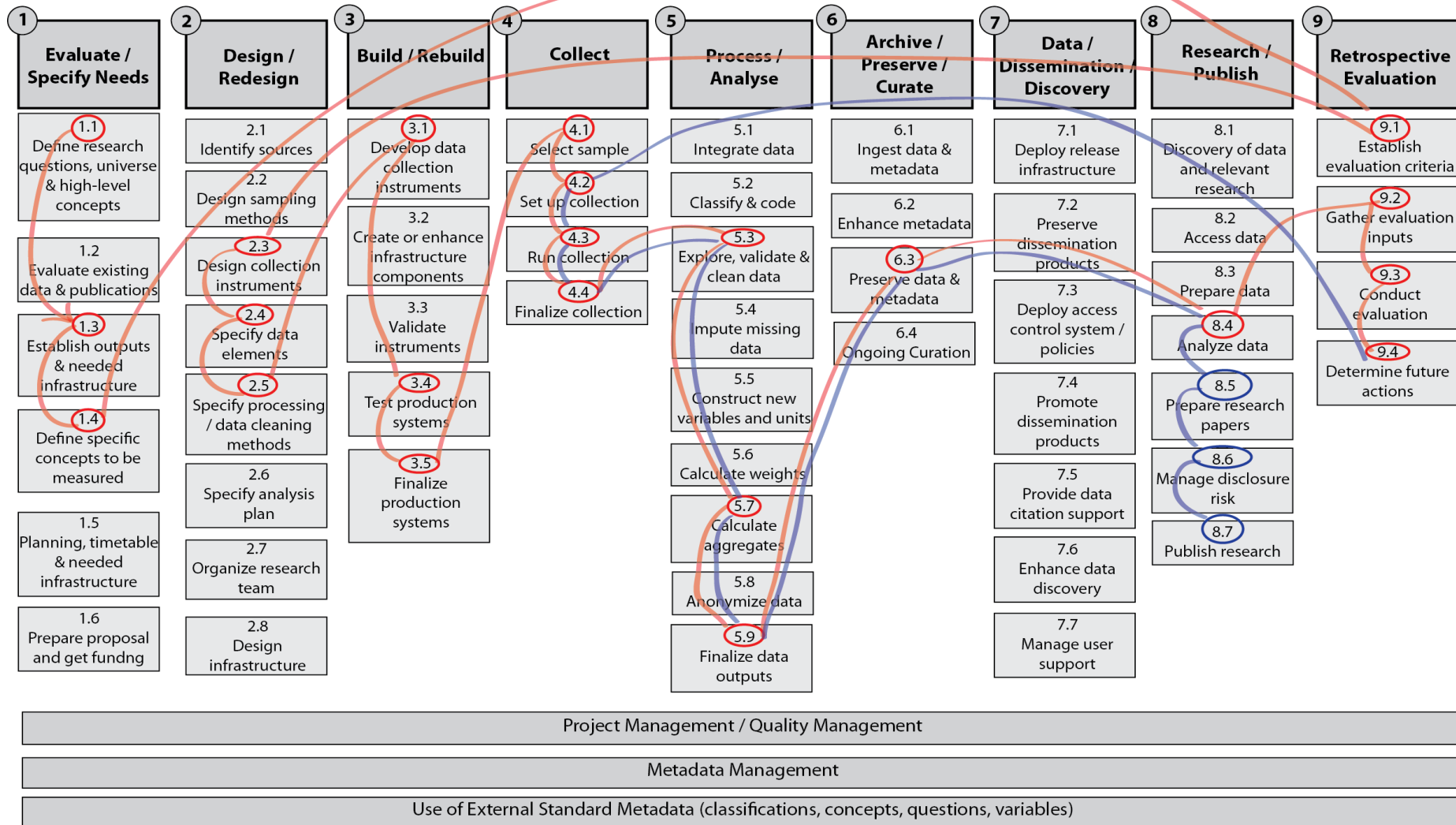
Quality Management / Metadata Management

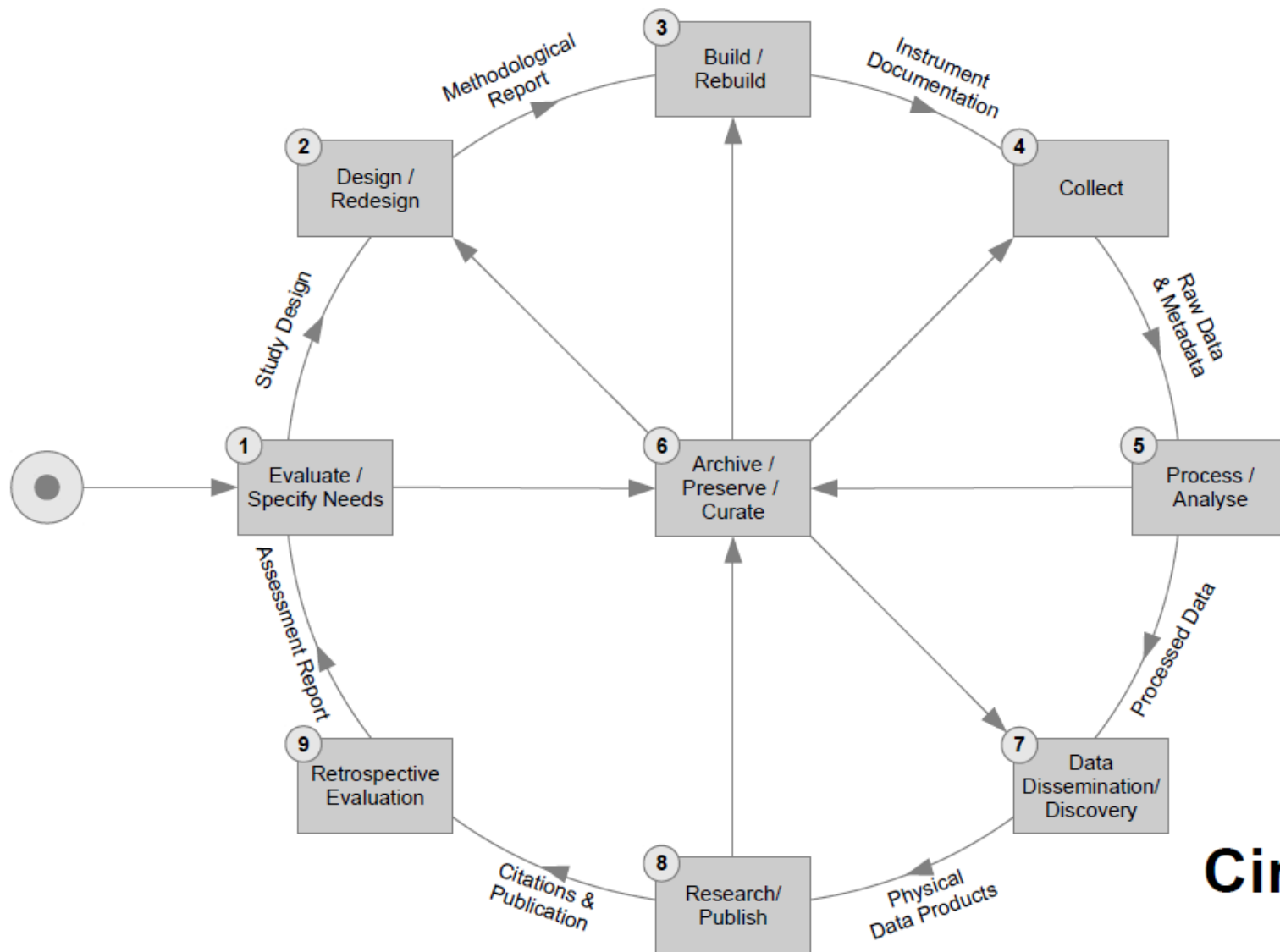
1 Specify Needs	2 Design	3 Build	4 Collect	5 Process	6 Analyse	7 Disseminate	8 Archive	9 Evaluate
1.1 Determine needs for information	2.1 Design outputs	3.1 Build data collection instrument	4.1 Select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Define archive rules	9.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Manage archive repository	9.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design data collection methodology		4.3 Run collection	5.3 Review, Validate & edit				
1.4 Identify concepts	2.4 Design frame & sample methodology	3.3 Configure workflows	4.4 Finalize collection	5.4 Impute	6.3 Scrutinize & explain	7.3 Manage release of dissemination products	8.3 Preserve data and associated metadata	9.3 Agree action plan
1.5 Check data availability	2.5 Design statistical processing methodology	3.4 Test production system		5.5 Derive new variables & statistical units	6.4 Apply disclosure control	7.4 Promote dissemination products	8.4 Dispose of data & associated metadata	
1.6 Prepare business case	2.6 Design production systems & workflow	3.5 Test statistical business process		5.6 Calculate weights	6.5 Finalize outputs			
		3.6 Finalize production system		5.7 Calculate aggregates		7.5 Manage user support		
				5.8 Finalize data files				



Note the similarity to the DDI Combined Lifecycle Model and the top level of the GSBPM







Circle View

Introduction to DDI

THE DDI-STANDARD

What Is DDI I?

- An international specification for structured metadata describing social, behavioral, and economic data
- A standardized framework to maintain and exchange documentation/metadata
- DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.
- A basis on which to build software tools
- Currently expressed in XML – e**X**tensible **M**arkup **L**anguage

History

- 1995 -- First international committee established
- 2000 -- First DDI version published (aligned with codebooks, XML DTD-based)
- 2003 – DDI 2 published (support for aggregate/tabular data and geography added)
- 2003 -- Formation of the DDI Alliance, a self-sustaining membership organization
- 2008 – DDI 3 published (aligned with data lifecycle, XML Schema-based)
- 2010 – DDI rebranding – DDI Codebook (DDI 2 branch) and DDI Lifecycle (DDI 3 branch) development lines

XML Elements

<Book>

<Title> ***The Hitchhiker's Guide to the Galaxy*** </Title>

<Author> Douglas Adams </Author>

<Year> 1979 </Year>

</Book>

XML Attributes

<Book language="English">

<Title> *The Hitchhiker's Guide to the Galaxy* </Title>

<Author> Douglas Adams </Author>

<Year> 1979 </Year>

</Book>

Conflicting Tag Names

```
<MyData>  
  <Table>  
    <Legs>4</Legs>  
    <Length units="feet">5</Length>  
    <Width units="feet">3</Width>  
  </Table>  
  
  <Table>  
    <Rows>4</Rows>  
    <Columns>3</Columns>  
  </Table>  
</MyData>
```

<MyData

xmlns:kitchen="http://www.example.org/kitchen"
xmlns:data="http://www.example.org/data">

<kitchen:Table>

<Legs>4</Legs>

<Length units="feet">5</Length>

<Width units="feet">3</Width>

</kitchen:Table>

<data:Table>

<Rows>4</Rows>

<Columns>3</Columns>

</data:Table>

</MyData>

DDI and XML

<DDIInstance>

 <StudyUnit> ... </StudyUnit>

 <ResourcePackage>

 <QuestionScheme>...</QuestionScheme>

 <VariableScheme>...</VariableScheme>

 <ConceptScheme>...</ConceptScheme>

 <PhysicalInstance>...</PhysicalInstance>

 </ResourcePackage>

</DDIInstance>

Introduction to DDI

TOOLS

MISSY

gesis

Leibniz-Institut
für Sozialwissenschaften

missy

Mikrodaten-
Informationssystem

[Home](#)

[Studie](#)

[Variablen](#)

[Kontakt](#)

[Zur GML-Seite im GESIS-Web](#)

Sie sind hier: [Variablen](#) / [Thematische Gliederung](#)

Thematische Gliederung

- Demographie und Bevölkerung

- + Nationalität und Migration
- + Arbeitsmarkt und Erwerbsbeteiligung
- + Unterhalt und Einkommen
- + Sozialversicherung und Vorsorge
- + Bildung und Qualifikation
- + Pendler
- + Privathaushalt und Familie
- + Wohnverhältnisse
- + Gesundheit
- + Stichprobe

[Gesamtübersicht](#) / [Demographie und Bevölkerung](#) >> [Daten zur Person](#) >> [Familienstand](#) >> [Familienstand, erweitert](#)

- Daten zur Person

+ Alter

+ Geschlecht

- Familienstand

- Familienstand

- Familienstand, erweitert

- Familienstand: Haushaltsbezugsp.

- Familienstand: Familienbezugsp.

- Familienstand: Bezugsp. der Lebensform

- Familienstand: Lebenspartner der Haushaltsbezugsp.

- Familienstand: Lebenspartner der Bezugsp. der Lebensform

- Familienstand: Haupteinkommensbezieher, erweitert

- Familienstand: Ernährer

+ Eheschließungsjahr

+ Geburten

+ Bevölkerungstyp

+ Regionalangaben



[Veröffentlichungen zu diesem Thema](#)

Mikrozensus: Erhebungszeitpunkte

2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1993	1991	1989	1987	1985	1982	1980	1978
EF765	EF765	EF765	EF765	EF35	EF35	EF35	EF35	EF35	EF35	EF35	EF35	EF35	EF38	EF38	EF38	EF38	EF38	EF38	EF21	EF21	EF21

CentERdata – LISS Panel / Questasy

LISS Data Archive > Questionnaire Family&Household - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.lissdata.nl/dataarchive/control_construct_schemes/view/5

(Untitled) x GESIS Intranet - Sozialwisse... x LISS Data Archive > Quest... x

LISS PANEL

CentERdata
Institute for data collection and research

login
Type search text here | Advanced Search

Home Organization About the Panel Proposals Research Access Data Contact

Home > Data Archive > LISS Studies > Family and Household > Wave 1 > Questionnaires > Family&Household

Questionnaire Family&Household

Description This questionnaire was originally conducted in Dutch.

Study Unit [Wave 1](#)

Question	Question Text	Answer Type
1 intro	This questionnaire is about family ties. First we wish to ask you some questions about your father and mother.	
2 cf005	What is the year of birth of your father? We mean your biological father.	Numeric
3 cf006	Can you perhaps indicate the period in which your biological father was born? You can click on the list of response options to select a period. If you don't know the year of birth of your father because you never knew your biological father, you can also indicate this here.	Categories
4 cf007	Is your father still alive?	yes (1), no (2), dk (99)
5 cf008	In what year did your father pass away?	Numeric
6 cf009	What is the year of birth of your mother? We mean your biological mother.	Numeric
7 cf010	Can you perhaps indicate the period in which your biological mother was born? You can click on the list of response options to select a period. If you don't know the year of birth of your mother because you never knew your biological mother, you can also indicate this here.	Categories
8 cf011	Is your mother still alive?	yes (1), no (2), dk (99)
9 cf012	In what year did your mother pass away?	Numeric
10 cf013	Did your own parents ever divorce? With divorce we also mean if your parents were never officially married but did separate. This concerns your biological father and biological mother.	Categories
11 cf014	How old were you when your parents separated?	Numeric
12 cf015	Is your father currently living together with a partner?	yes (1), no (2), dk (99)
13 cf016	Is your mother currently living together with a partner?	yes (1), no (2), dk (99)
14 cf017	We now would like to know the postal code of your parents' address in order to calculate precisely how far apart you and your parents live. This postal code is used for this purpose only and shall absolutely not be provided to third parties or coupled with other information. Do you know the postal code of your parents' address?	Categories
15 cf018	We now would like to know the postal code of your father's address in order to calculate precisely how far apart you and your father live. This postal code is used for this purpose only and shall absolutely not be provided to third parties or coupled with other information. Do you know the postal code of your father's address?	Categories

Done

Midlife in the United States

MIDUS 1

Project 1	In 1994/95, the MacArthur Midlife Research Network carried out a national survey of over 7,000 Americans aged 25 to 74.	Study Details	Browse Data	Download Data	Download Codebook
Project 2	The purpose of the National Study of Daily Experiences is to examine the day-to-day lives, particularly the daily stressful experiences, of a subsample of MIDUS respondents.	Study Details	Browse Data	Download Data	Download Codebook

MIDUS 2

Project 1	Project 1 provided follow-up on the psychosocial, sociodemographic, and health variables assessed in MIDUS I.	Study Details	Browse Data	Download Data	Download Codebook
Project 2	Project 2 provided follow-up on the daily diary study included in MIDUS I.	Study Details	Browse Data	Download Data	Download Codebook
Project 3	Project 3 included new cognitive assessments for the full MIDUS sample, plus longitudinal follow-up for the cognitive subsample from MIDUS I.	Study Details	Browse Data	Download Data	Download Codebook
Project 4	Project 4 included comprehensive biomarker assessments on a subsample of MIDUS respondents, collected at one of 3 General Clinical Research Centers around the country.	Study Details	Browse Data	Download Data	Download Codebook
Project 5	Project 5 included neuroscience assessments on a subsample of respondents in the biomarker study.	Study Details	Browse Data	Download Data	Download Codebook
Milwaukee Sample	As a refinement to MIDUS II, a sample of African Americans from Milwaukee was included to examine health issues in minority populations.	Study Details	Browse Data	Download Data	Download Codebook
Milwaukee Sample (Standalone)	This dataset contains only data from Milwaukee sample.	Study Details	Browse Data	Download Data	Download Codebook

MIDUS Japan

MIDUS Japan	The MIDJA study is a probability sample of Japanese adults (N = 1,027) aged 30 to 79 from the Tokyo metropolitan area.	Study Details	Browse Data	Download Data	Download Codebook
-------------	--	-------------------------------	-----------------------------	-------------------------------	-----------------------------------

- [-] ZACAT
 - [-] ISSP
 - [-] Eurobarometer
 - [-] Standard & Special Eurobarometer
 - [-] EB 72.5 - EB 76.2 (Pre-releases)
 - [-] Eurobarometer 76.2 September-November 2011
 - [-] Metadata
 - [-] Study Description
 - [-] Data Files Description
 - [-] Variable Description
 - [-] Archive and Survey ID variables
 - [-] Standard nation ID variables
 - [-] Nationality (Q1)
 - [-] Employment and social policy (QA)
 - QA1 CRISIS CONCERN: LOSING YOUR JOB
 - QA1 CRISIS CONCERN: PARTNER LOSING JOB
 - QA1 CRISIS CONCERN: CHILDREN LOSING JOB
 - QA2 CRISIS - WILL BE OVER IN 2 YEARS
 - QA3 EU GLOBALISATION ADJ FUND EGF - AWARENESS
 - QA3 EU SOCIAL FUND ESF - AWARENESS
 - QA4 WORKING LIFE - TIMES CHANGED EMPLOYER
 - QA4R1 WORKING LIFE - TIMES CHANGED EMPLOYER
 - QA4R2 WORKING LIFE - TIMES CHANGED EMPLOYER
 - QA5 WORKING LIFE - YEARS FOR LAST EMPLOYER
 - QA5R1 WORKING LIFE - YEARS FOR LAST EMPLOYER
 - QA5R2 WORKING LIFE - YEARS FOR LAST EMPLOYER
 - QA6 JOB FINDING ASSETS: QUALIFICATION

ZA5566: Eurobarometer 76.2 September-November 2011

[ZA5566 Datafiles and Documentation](#) (download via data catalogue)

Variable Group Employment and social policy (QA)

- QA1 CRISIS CONCERN: LOSING YOUR JOB
- QA1 CRISIS CONCERN: PARTNER LOSING JOB
- QA1 CRISIS CONCERN: CHILDREN LOSING JOB
- QA2 CRISIS - WILL BE OVER IN 2 YEARS
- QA3 EU GLOBALISATION ADJ FUND EGF - AWARENESS
- QA3 EU SOCIAL FUND ESF - AWARENESS
- QA4 WORKING LIFE - TIMES CHANGED EMPLOYER
- QA4R1 WORKING LIFE - TIMES CHANGED EMPLOYER
- QA4R2 WORKING LIFE - TIMES CHANGED EMPLOYER
- QA5 WORKING LIFE - YEARS FOR LAST EMPLOYER
- QA5R1 WORKING LIFE - YEARS FOR LAST EMPLOYER
- QA5R2 WORKING LIFE - YEARS FOR LAST EMPLOYER
- QA6 JOB FINDING ASSETS: QUALIFICATION
- QA6 JOB FINDING ASSETS: PROFESS EXPERIENCE
- QA6 JOB FINDING ASSETS: LANGUAGE SKILLS
- QA6 JOB FINDING ASSETS: COMPUTER SKILLS
- QA6 JOB FINDING ASSETS: ABILITY TO ADAPT
- QA6 JOB FINDING ASSETS: WORKING ABROAD
- QA6 JOB FINDING ASSETS: OTHER
- QA6 JOB FINDING ASSETS: NONE OF THESE
- QA6 JOB FINDING ASSETS: DK
- QA7 JOP KEEPING ABILITY - NEXT MONTHS
- QA8 JOB CONFIDENCE - IN 2 YEARS TIME
- QA9 FINDING A JOB IF LAID-OFF - LIKELIHOOD

Find & Analyze Data

[Find ICPSR Data](#)[Bibliography of Data-Related Literature](#)[Variables Database](#)[Analyze Data Online](#)[Thematic Collections](#)[Restricted Data](#)[Publication-Related Archive](#)

V172: Q23 EU COMMON POLICY: EDUCATION

Name: V172

Label: Q23 EU COMMON POLICY: EDUCATION

Question:

Q.23. Some people believe that certain areas of policy should be decided by the (NATIONAL) government, while other areas of policy should be decided jointly within the European Union. Which of the following areas of policy do you think should be decided by the (NATIONAL) government, and which should be decided jointly within the European Union ?

m) Education

Response Categories

Code	Label	Freq.	%
0	<NA>	9	0%
1	NAT GOVERNMENT	10162	63%
2	EUROPEAN UNION	5267	33%
3	DK	716	4%
Total		16154	100%

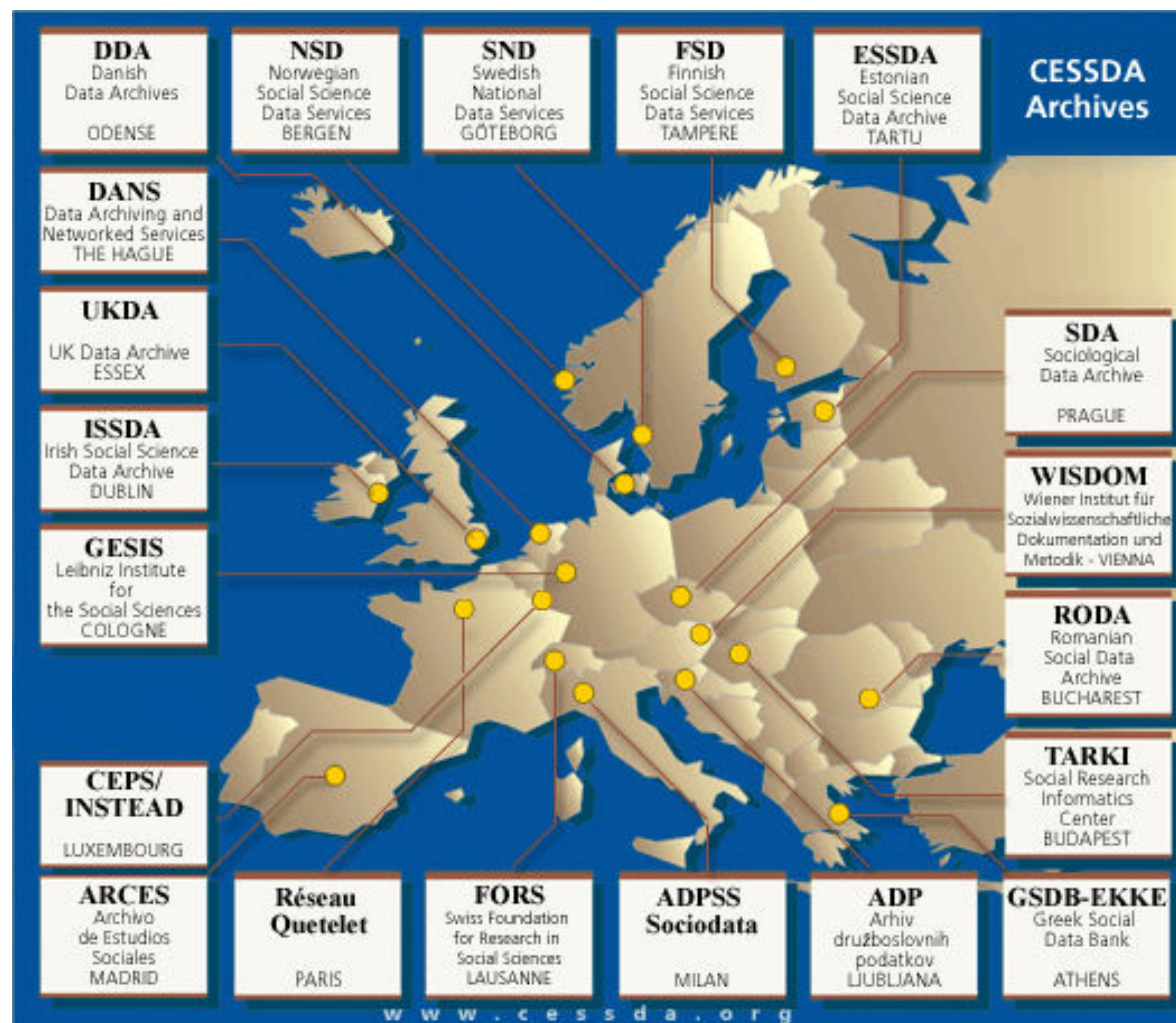
Disclaimer: The frequencies for this variable *may not be weighted*. They are purely descriptive and may not be representative of the study population. Please use with caution and consult the study documentation.

Copyright: ICPSR has an FAQ on [copyright and survey instruments](#).

Summary Statistics

- total responses: 16154
- valid: 16154

CESSDA - Council of European Social Science Data Archives



**IHSN**

International Household Survey Network

[Home](#)
[About](#)
[Activities](#)
[Tools and Guidelines](#)
[Quick Links](#)




IHSN Microdata Management Toolkit

A suite of tools for the documentation and dissemination of your survey and census data according to international standards and good practices

Version 1.1 now available

[1](#)
[2](#)
[3](#)

Technical Support

The Accelerated Data Program (ADP) provides technical and financial support to survey data documentation and dissemination, and improvement of survey methods.



Looking for Data?



The IHSN does not have ownership and is not mandated to disseminate country microdata. But we maintain a [central survey catalog](#), and provide links to national or international databanks.

Highlights



Microdata Management Toolkit

Document your surveys and censuses in compliance with international standards using this specialized metadata editor.

National Data Archive (NADA)

Publish your survey and census metadata in a searchable on-line catalog, and disseminate datasets using this free software compliant with the DDI and the OAI international standards.



Planned surveys and censuses

Information on on-going and planned surveys and censuses in developing countries.

Question Bank

A central repository of survey guidelines, with questionnaire modules, interviewer's instructions, key indicators definitions, and more.

☐ YES
☐ NO
☒ DON'T KNOW


Microdata anonymization

Tools and guidelines for measuring and reducing disclosure risk in microdata.

News & Updates

Second "Regional Training on the Documentation and Archiving of Agricultural Censuses and Surveys" organized in Manila, Philippines, 21-25 February 2011

Tue, 03/22/2011

The second "Regional Training on the Documentation and Archiving of Agricultural Censuses and Surveys" was held in Manila, Philippines, 21-25 February 2011. In line with the objective of the first workshop in Addis Ababa, Ethiopia, the workshop was organized to train participants from... [read more »](#)

"Regional training on the documentation and archiving of Agricultural Censuses and Surveys" held in Addis Ababa, Ethiopia, 7-11 February 2011

Sat, 02/26/2011

The first "Regional training on the documentation and archiving of Agricultural Censuses and Surveys" was held in Addis Ababa, Ethiopia, 7-11 February 2011. The objective of the workshop was to train participants from 15 countries on the use the Microdata Management Toolkit for the documentation... [read more »](#)

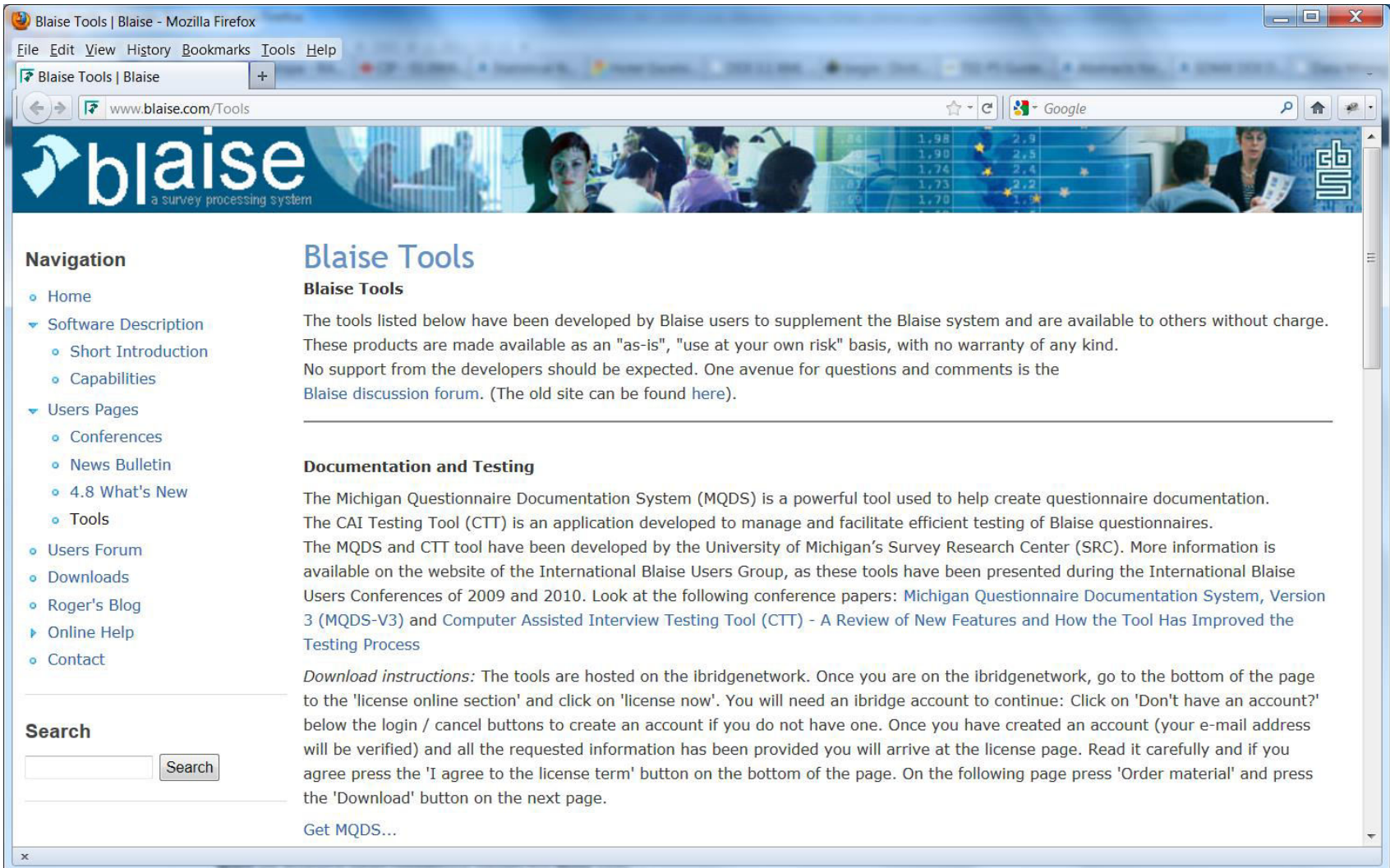
Focus

Dissemination of Microdata Files: Principles, Procedures and Practices



In all countries, data producers are faced by expanding demand for microdata. Determining the best way to disseminate these data is a challenge. The challenge is technical, as data producers have to implement procedures for the documentation, cataloguing and dissemination of the data. The challenge is also legal and ethical.

Michigan Questionnaire Documentation System (MQDS) as Blaise Tool



The screenshot shows a Mozilla Firefox browser window displaying the Blaise Tools website. The browser's address bar shows the URL www.blaise.com/Tools. The website has a blue header with the "blaise" logo and the tagline "a survey processing system". Below the header, there is a navigation menu on the left and a main content area on the right.

Navigation

- Home
- Software Description
 - Short Introduction
 - Capabilities
- Users Pages
 - Conferences
 - News Bulletin
 - 4.8 What's New
 - Tools
- Users Forum
- Downloads
- Roger's Blog
- Online Help
- Contact

Search

Blaise Tools

Blaise Tools

The tools listed below have been developed by Blaise users to supplement the Blaise system and are available to others without charge. These products are made available as an "as-is", "use at your own risk" basis, with no warranty of any kind. No support from the developers should be expected. One avenue for questions and comments is the [Blaise discussion forum](#). (The old site can be found [here](#)).

Documentation and Testing

The Michigan Questionnaire Documentation System (MQDS) is a powerful tool used to help create questionnaire documentation. The CAI Testing Tool (CTT) is an application developed to manage and facilitate efficient testing of Blaise questionnaires. The MQDS and CTT tool have been developed by the University of Michigan's Survey Research Center (SRC). More information is available on the website of the International Blaise Users Group, as these tools have been presented during the International Blaise Users Conferences of 2009 and 2010. Look at the following conference papers: [Michigan Questionnaire Documentation System, Version 3 \(MQDS-V3\)](#) and [Computer Assisted Interview Testing Tool \(CTT\) - A Review of New Features and How the Tool Has Improved the Testing Process](#)

Download instructions: The tools are hosted on the ibridgenetwork. Once you are on the ibridgenetwork, go to the bottom of the page to the 'license online section' and click on 'license now'. You will need an ibridge account to continue: Click on 'Don't have an account?' below the login / cancel buttons to create an account if you do not have one. Once you have created an account (your e-mail address will be verified) and all the requested information has been provided you will arrive at the license page. Read it carefully and if you agree press the 'I agree to the license term' button on the bottom of the page. On the following page press 'Order material' and press the 'Download' button on the next page.

[Get MQDS...](#)

Stat/Transfer Version 11 supports DDI Lifecycle

Contents

- What Stat/Transfer Does
- What's New in Stat/Transfer
- Getting Started
- The Stat/Transfer User Interface
- The Command Processor
- Variable Naming and Limits
- Supported Programs
 - Read-me File
 - Input and Output Variable Types
 - 1-2-3 Worksheet Files
 - Access
 - ASCII Files - Delimited
 - ASCII Files - Fixed Format
 - SCHEMA Files for ASCII
 - Input
 - dBASE Files and Compatibles
 - DDI (Data Documentation Initiative) Schemas
 - Epi Info
 - Excel Worksheets
 - FoxPro Files
 - Gauss Files
 - HTML Tables
 - JMP Files
 - LIMDEP Files
 - Matlab Files
 - Mineset Files
 - Minitab Worksheets
 - Mplus Files
 - NLOGIT Files
 - ODBC Data Sources
 - OpenDocument Spreadsheets
 - OSIRIS Files
 - Paradox Tables
 - Quattro Pro Worksheets
 - R
 - RATS Files

DDI (Data Documentation Initiative) Schemas

The Data Documentation Initiative (DDI) is an open, international effort to provide a a standard way of describing data from the social, behavioral and economic sciences. DDI Schemas are in XML and can describe metadata across the life-cycle from questionnaire design through analysis. See www.ddalliance.org.

Stat/Transfer supports DDI 3.1 Standard. The data are contained in two files: the dictionary information is stored in a file with the extension **.xml** and the data are stored in a separate delimited ASCII file with the extension **.dat**.

Standard extension: xml, dat

Reading DDI

Stat/Transfer will read and use DDI variable and value labels. Delimited data files are supported for reading and writing. The missing values and the delimiter are taken from the Schema.

Writing DDI

Stat/Transfer writes delimited DDI. The specification requires that elements within the Schema be identified by an "agency". This is typically a url. By default it is "example.org", but you should change it to something more appropriate at **Output Options(1)** of the **Options** dialog box.

Missing Values

On input the missing value is taken from the Schema. On output it is a blank.

Output Variable Types

The output variable type that results from each target variable type is given in the following table:

Target Type	Output Type
byte	Short
int	
float	Float
double	Double
date	Date
time	Time
date/time	DateTime
string	Text

Colectica

Colectica® - DDI Metadata and Survey Design Software Tools - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Colectica® - DDI Metadata and... +

colectica.com

Google

colectica

Blogs News Evaluate Purchase Contact Help

Software Training Services Standards

Colectica® is the fastest way to design, document, and publish your survey research using **open data** standards.

[Learn about Colectica](#)

[See Colectica in Action](#)



For Individuals

- ▶ [Colectica Reader](#)
- ▶ [Colectica Express](#)

For Departments

- ▶ [Colectica Designer](#)
- ▶ [Colectica Repository](#)
- ▶ [Colectica Web](#)

For Institutions

- ▶ [Colectica Fusion](#)

Software

The Colectica Platform is an ideal solution for statistical agencies, survey research groups, public opinion research, data archivists, and other data centric collection operations that are looking to increase the expressiveness and longevity of the data collected through standards based metadata documentation.

Training

In just a few focused days, learn how to adopt open data standards in your organization with our [training courses](#).

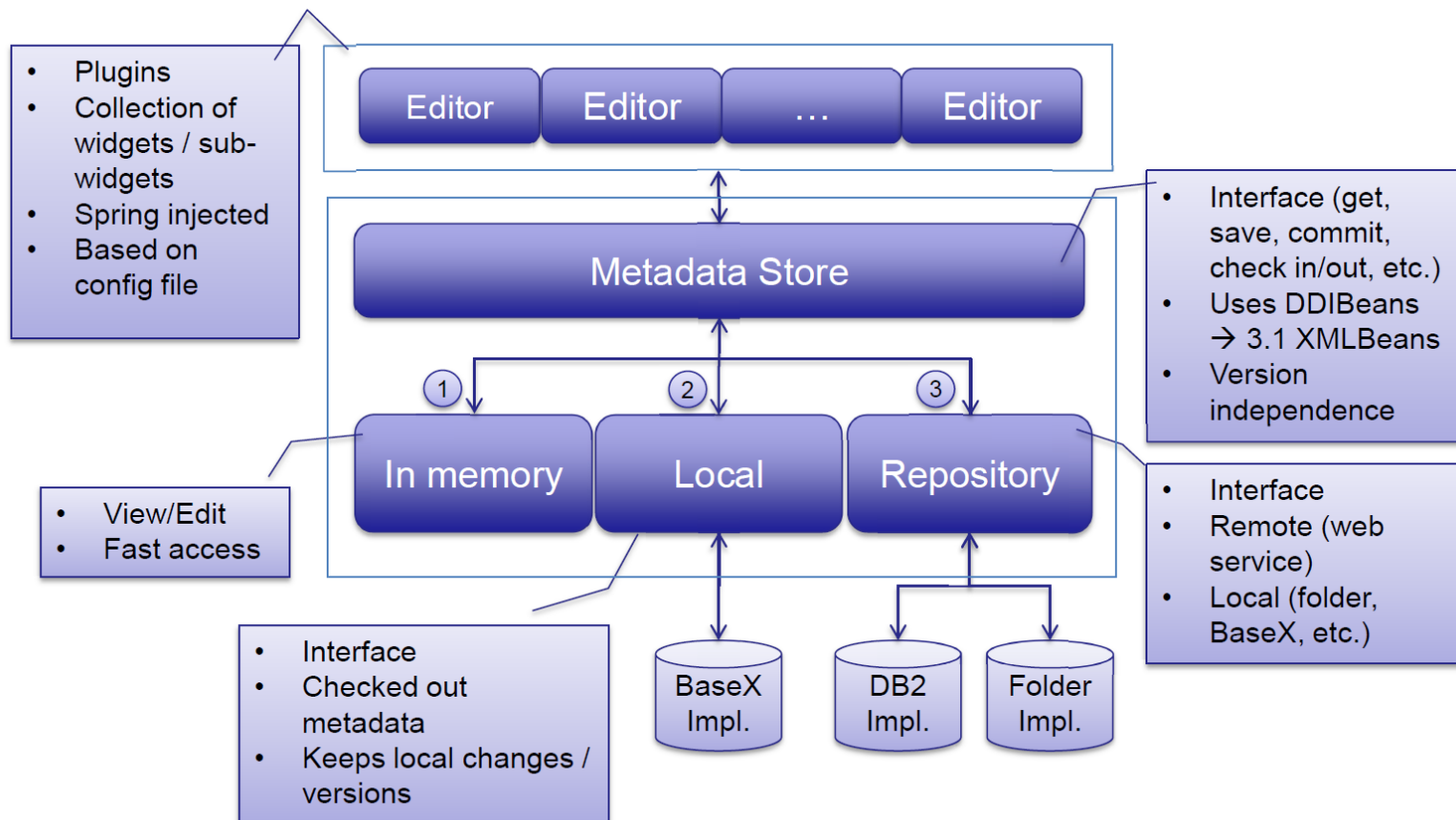
Services

We have years of experience creating software and data documentation based on open data standards. [Put us to work for you.](#)

Canadian Research Data Centre Network (CRDCN)

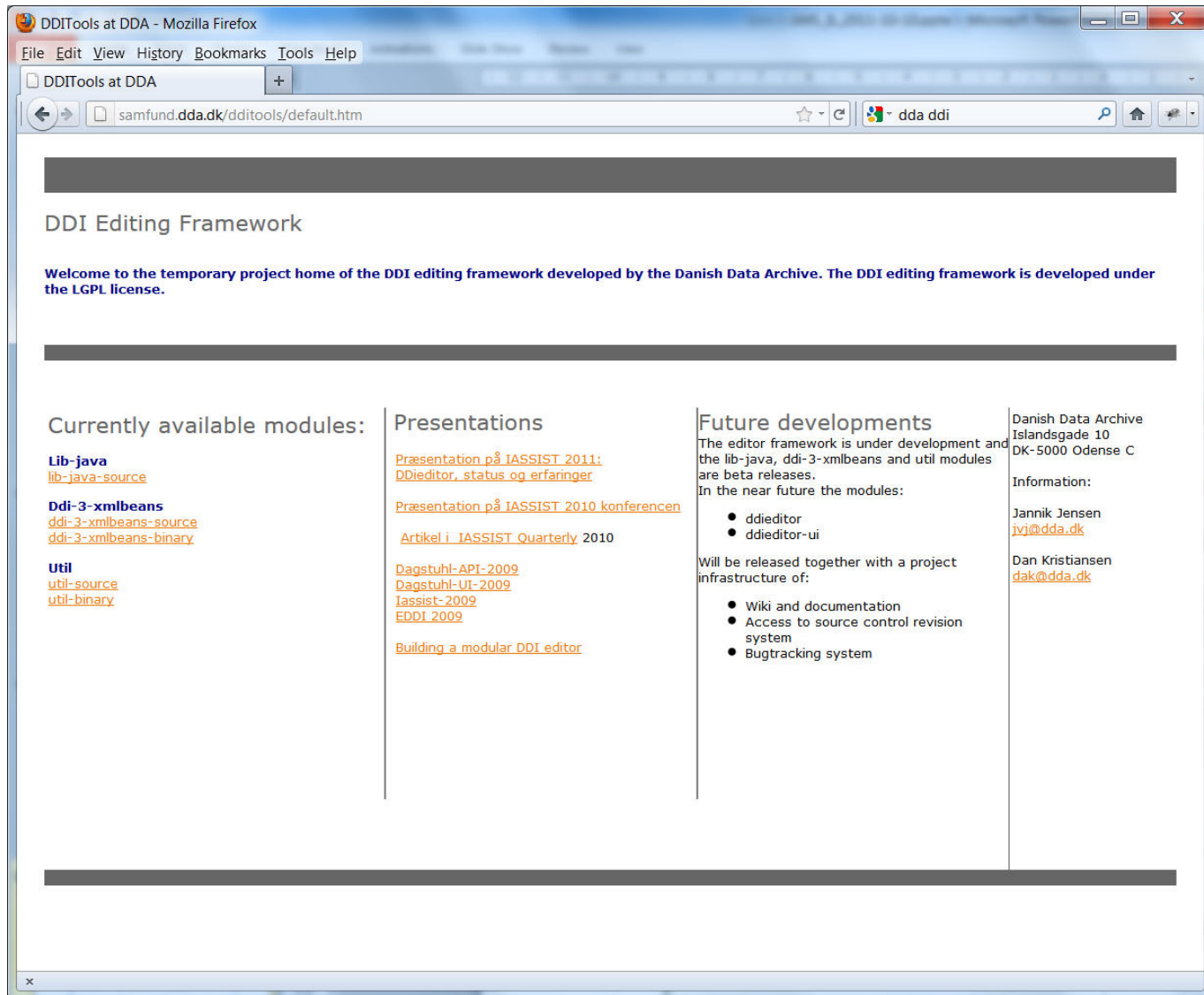
DDI 3 Data/Metadata Management Platform

High Level Editor Architecture

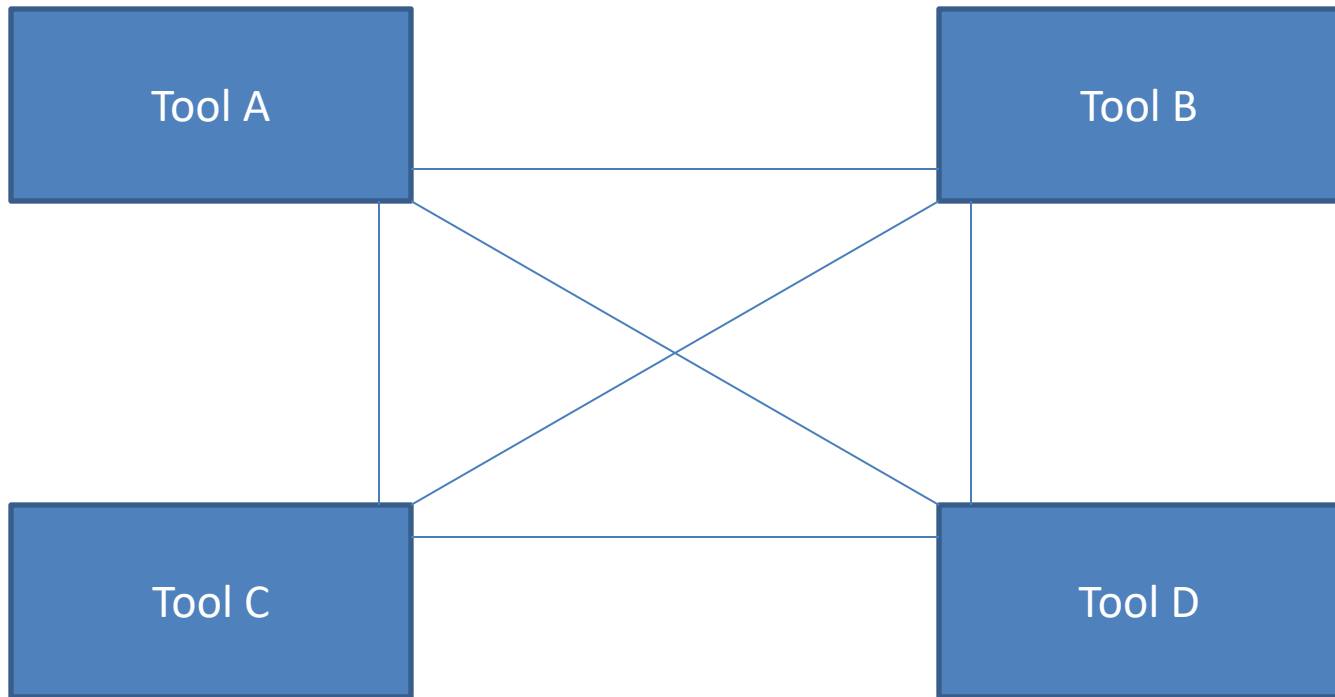


Danish Data Archive

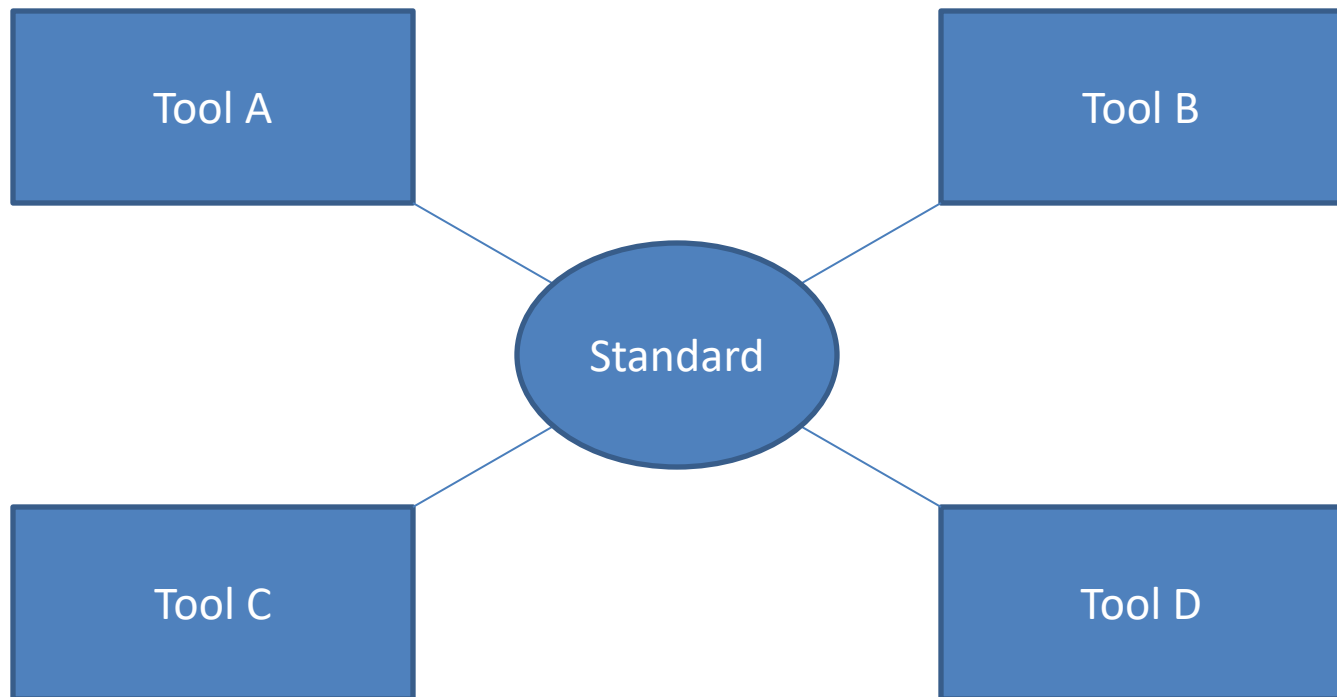
DDI Editing Framework



Tools and Standards



Tools and Standards



Introduction to DDI

CODEBOOK AND LIFECYCLE

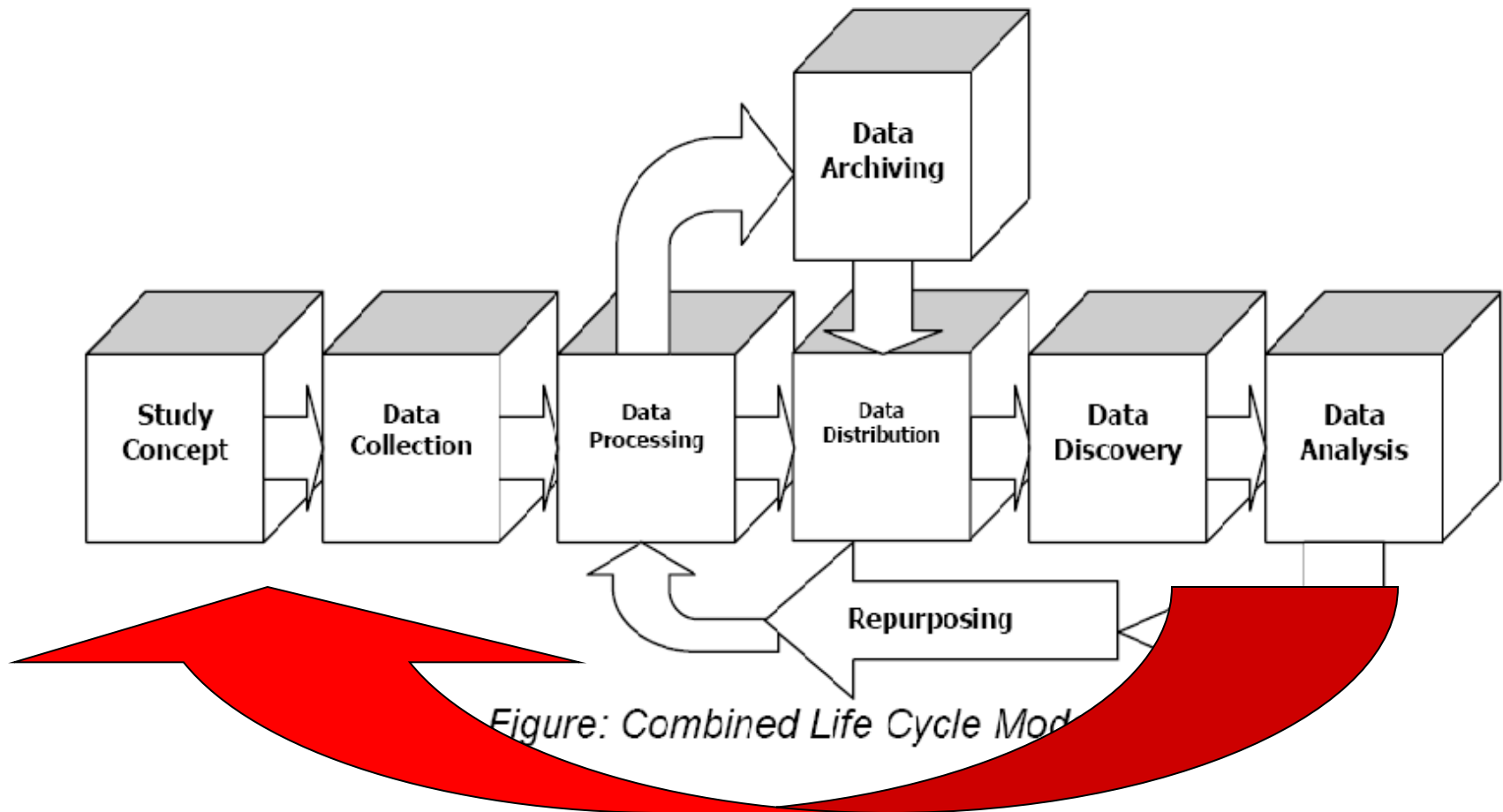
Formerly known as...

- DDI 2 → DDI Codebook → DDI-C
- DDI 3 → DDI Lifecycle → DDI-L

DDI Codebook

- Document Description
- Study Description
- File Description
- Variable Description

DDI Lifecycle Model



Metadata Reuse

DDI Lifecycle Features

- Machine-actionable
- Modular and extensible
- Multi-lingual
- Aligned with other metadata standards
- Can carry data in-line
- Focused on metadata reuse

DDI Lifecycle Features

- Support for CAI instruments
- Support for longitudinal surveys
- Focus on comparison, both by design and after-the-fact (harmonization)
- Robust record and file linkages for complex data files
- Support for geographic content (shape and boundary files)
- Capability for registries and question banks

PART II

Advanced Topics

Advanced Topics

DDI LIFECYCLE IN DETAIL

DDI structures

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

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

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

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)

Advanced Topics

SCHEMES AND RE-USE

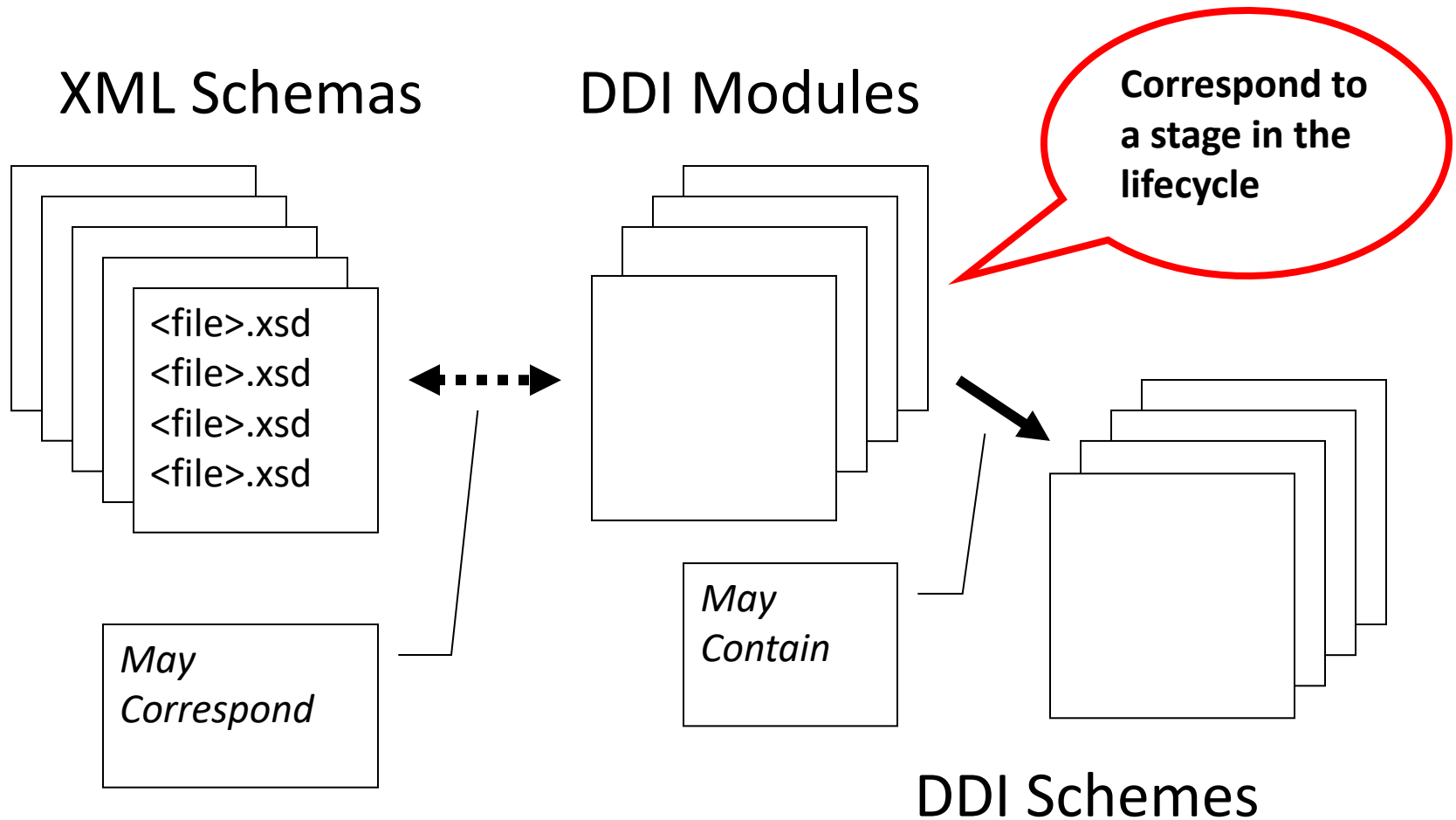
DDI Schemes

- 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



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 (eg, 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 in to 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 by Reference

- When a piece of metadata is re-used, a *reference* can be made to the original
- In order to reference the original, you must be able to *identify* it
- You also must be able to *publish* it, so it is visible (and can be referenced)
 - It is published to the user community – those users who are allowed access

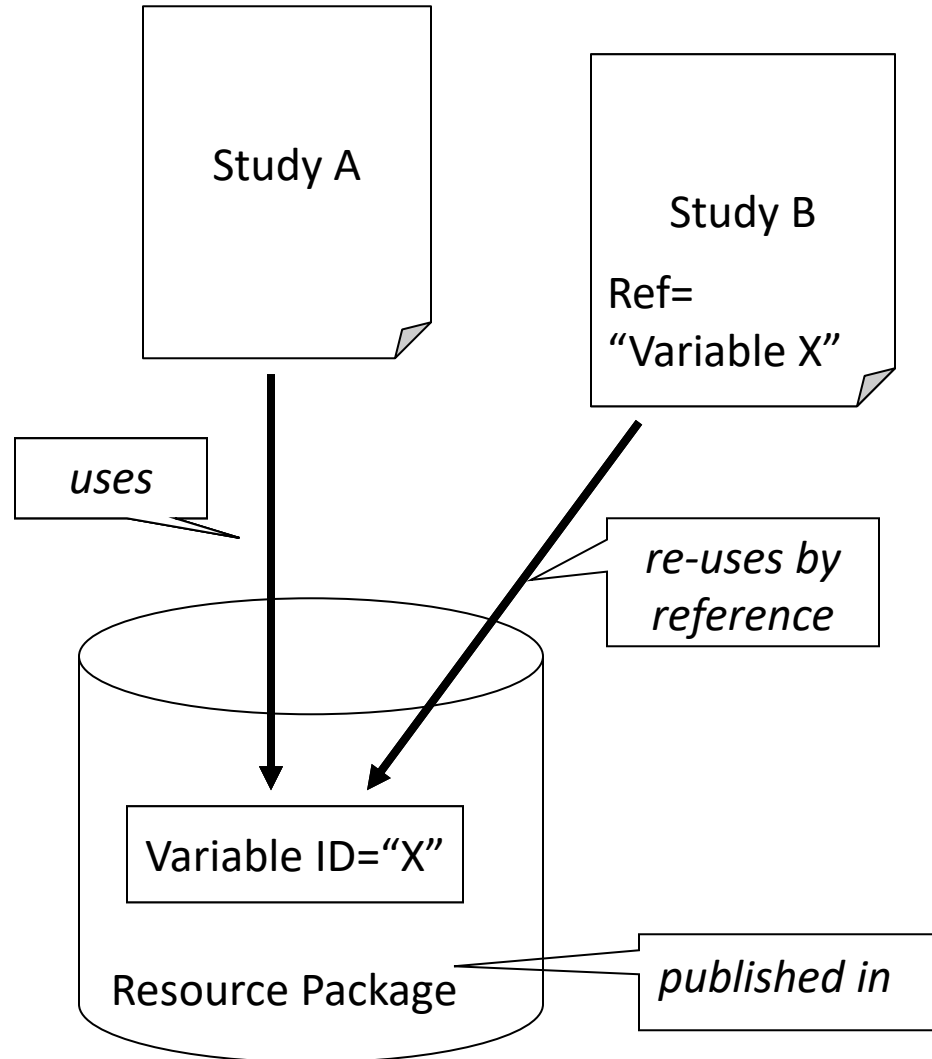
Change over Time

- Metadata items change over time, as they move through the data lifecycle
 - This is especially true of longitudinal/repeat cross-sectional studies
- This produces different *versions* of the metadata
- The metadata versions have to be *maintained* as they change over time
 - If you reference an item, it should not change: you reference a specific version of the metadata item

DDI Support for Metadata Reuse

- DDI allows for metadata items to be *identifiable*
 - They have unique IDs
 - They can be re-used by *referencing* those IDs
- DDI allows for metadata items to be *published*
 - The items are published in *resource packages*
- Metadata items are *maintainable*
 - They live in “schemes” (lists of items of a single type) or in “modules” (metadata for a specific purpose or stage of the lifecycle)
 - All maintainable metadata has a known owner or *agency*
- Maintainable metadata can be *versionable*
 - This reflects changes over time
 - The versionable metadata has a version number

Reusable Study-independent Information in Resource Package



Advanced Topics

QUESTIONNAIRE EXAMPLE

Questionnaires

- Questions
 - Question Text
 - Response Domains
- Statements
 - Pre- Post-question text
- Instructions
 - Routing information
 - Explanatory materials
- Question Flow

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:

- Questions

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

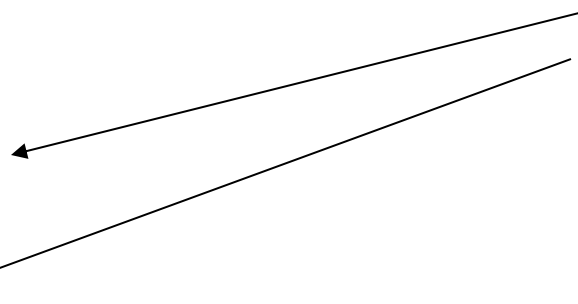
- Questions

- Response Domains

- Code

- Numeric

- Text



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

- Questions
- Response Domains
 - Code
 - Numeric
 - Text
- Statements
- [Instructions](#)

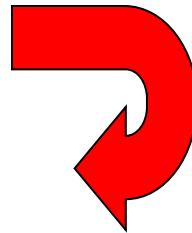
Simple Questionnaire

Please answer the following:

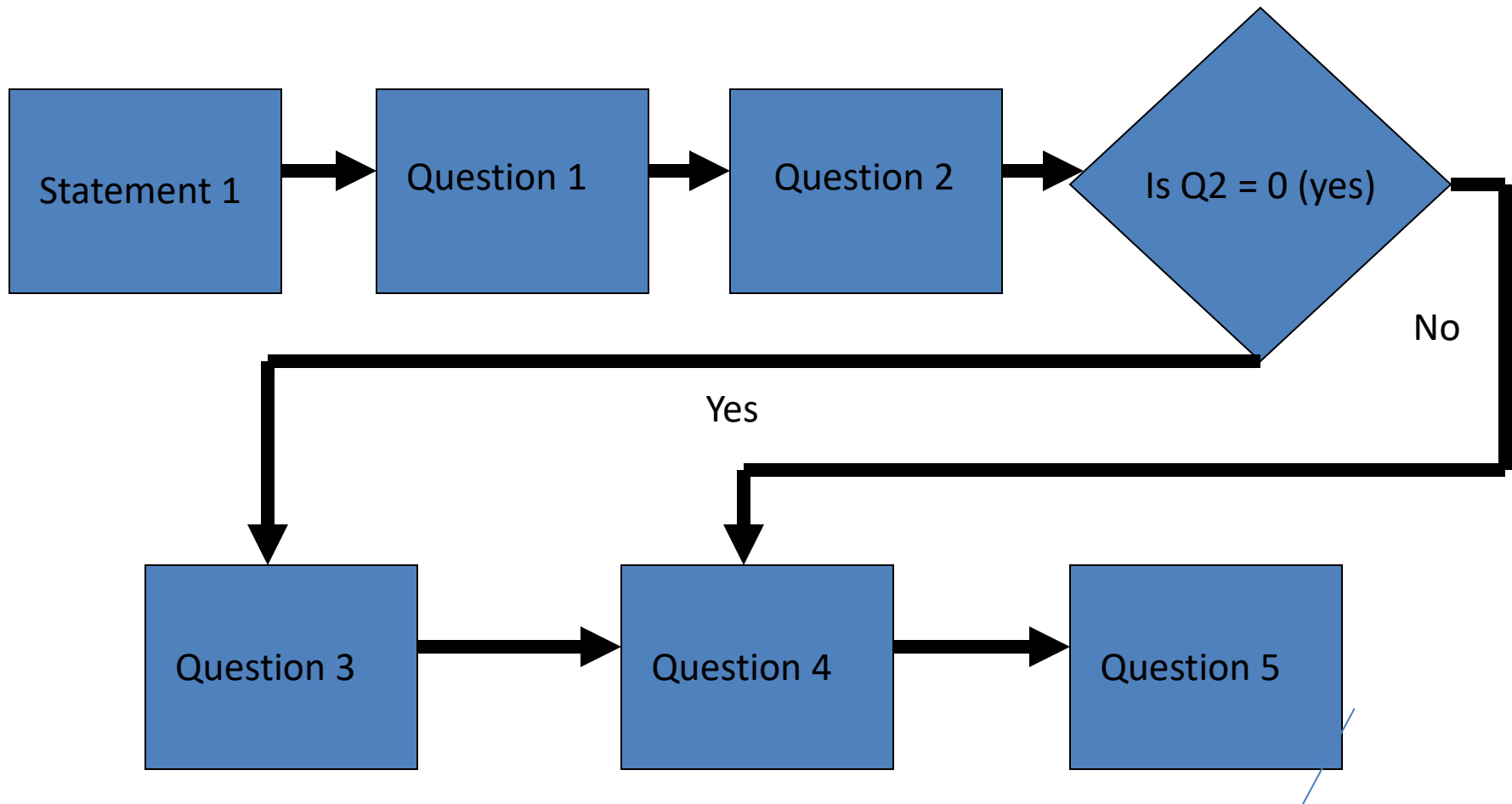
1. Sex
 - (1) Male
 - (2) Female
2. Are you 18 years or older?
 - (0) Yes
 - (1) No (Go to Question 4)
3. How old are you? _____
4. Who do you live with?

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

Skip Q3



- Questions
- Response Domains
 - Code
 - Numeric
 - Text
- Statements
- Instructions
- Flow



Advanced Topics

COMPARISON

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

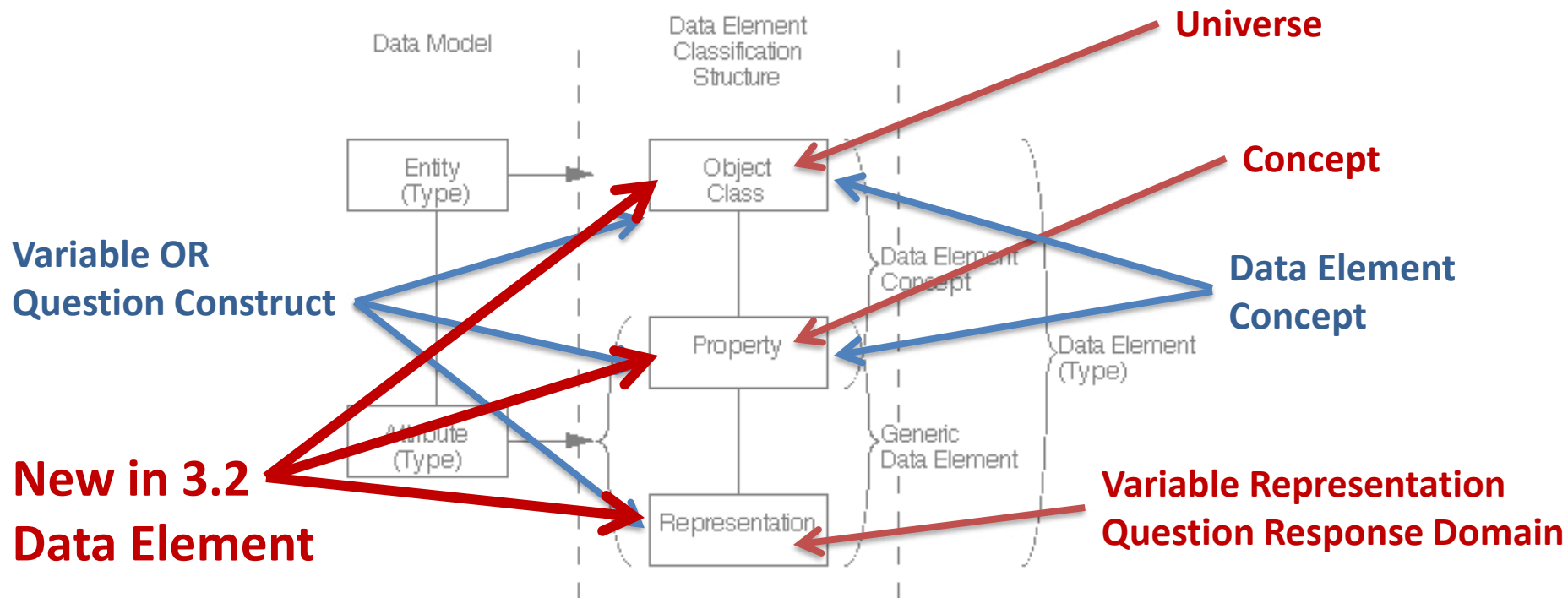
- For data which is completely the same, DDI provides a way of showing comparability: Grouping
 - These things are comparable “by design”
 - This typically includes longitudinal/repeat cross-sectional studies
- For data which *may* be comparable, DDI allows for a statement of what the comparable metadata 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

Comparison Map	Textual Content of Main Body		Category		Code Scheme	
	Same	Similar	Same	Similar	Same	Different
Question	X		X		X	
	X		X			X
	X			X	X	
	X			X		X
		X	X		X	
		X	X			X
		X		X	X	
		X		X		X



ISO/IEC 11179-1

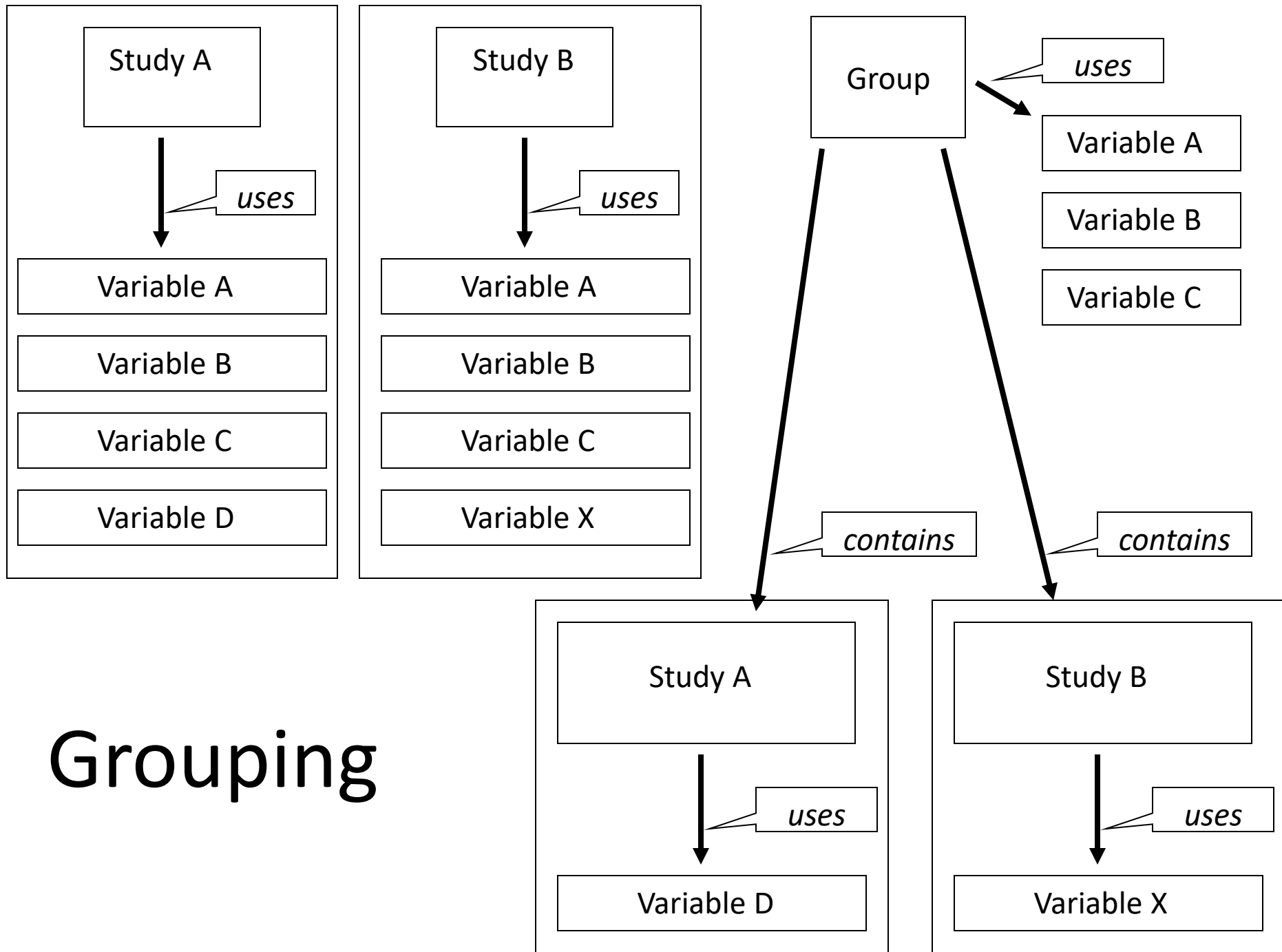
International Standard ISO/IEC 11179-1: Information technology – Specification and standardization of data elements – Part 1: Framework for the specification and standardization of data elements
 Technologies de l'informatique – Spécification et normalisation des éléments de données – Partie 1: Cadre pour la spécification et la normalisation des éléments de données. First edition 1999-12-01 (p26)
http://metadata-standards.org/11179-1/ISO-IEC_11179-1_1999_IS_E.pdf

Data Comparison

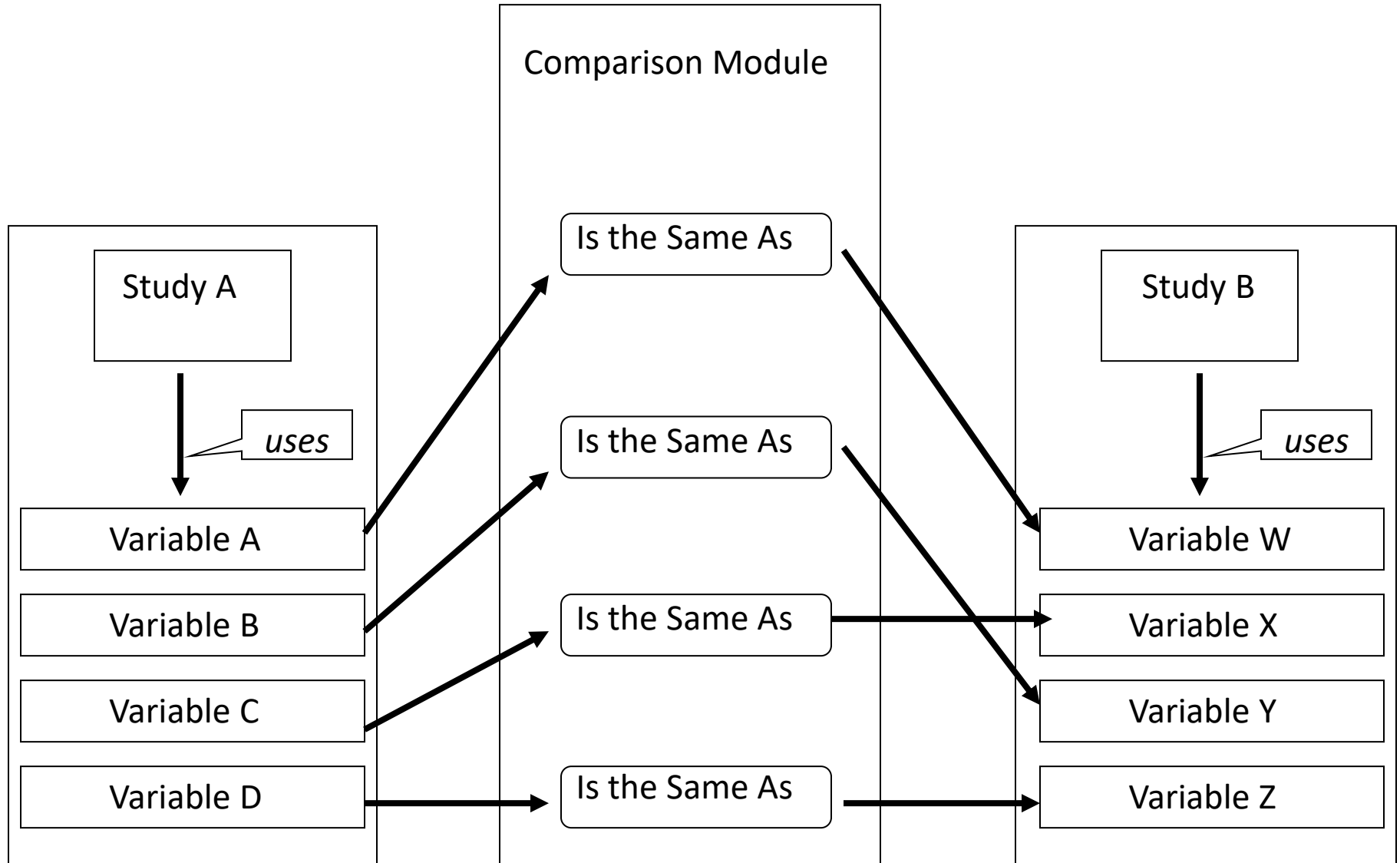
- To compare data from different studies (or even waves of the same study) we use the *metadata*
 - The metadata explains which things are comparable in data sets
- When we compare two variables, they are comparable if they have the same set of properties
 - They measure the same concept for the same 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/repeated 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)



Comparison



Advanced Topics

YOUR QUESTIONS

Advanced Topics

CONCLUSION

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?

BACKUP SLIDES

DDI structures

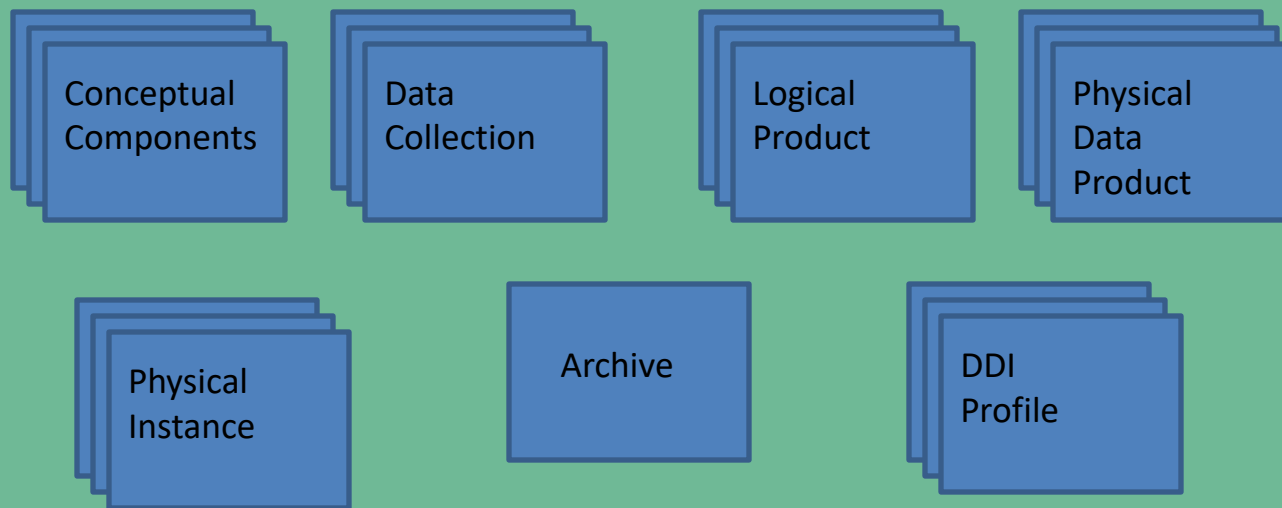
- Schemes
 - Data Element
 - Concepts
 - Geography
 - Questions
 - Variables
- Group
- Comparison
- Control Construct
- Processing Events
- Processing Instructions

DDI Structures

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

Study Unit

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



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]

MANAGEMENT

Administration

Data Management

Descriptive Information

Descriptive Information

Access

DIP

Order response

Order

Query/
Response

Consumer

Archival Storage

AIP

AIP

Ingest

SIP

Producer

Preservation Planning



Functional Entity



Terminator (Actor/Agent in this case)



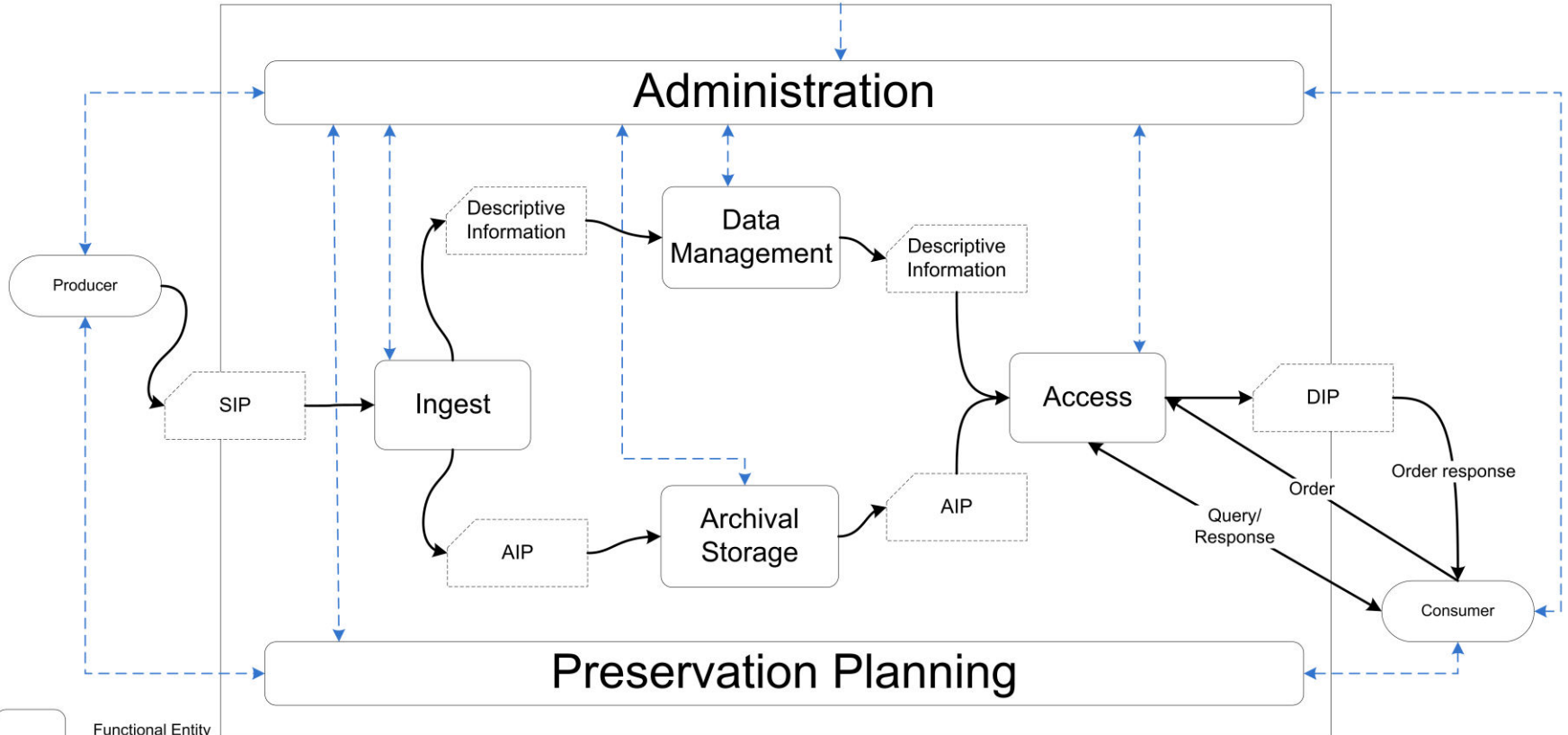
Package Data Object (making use of the old punch card symbol)



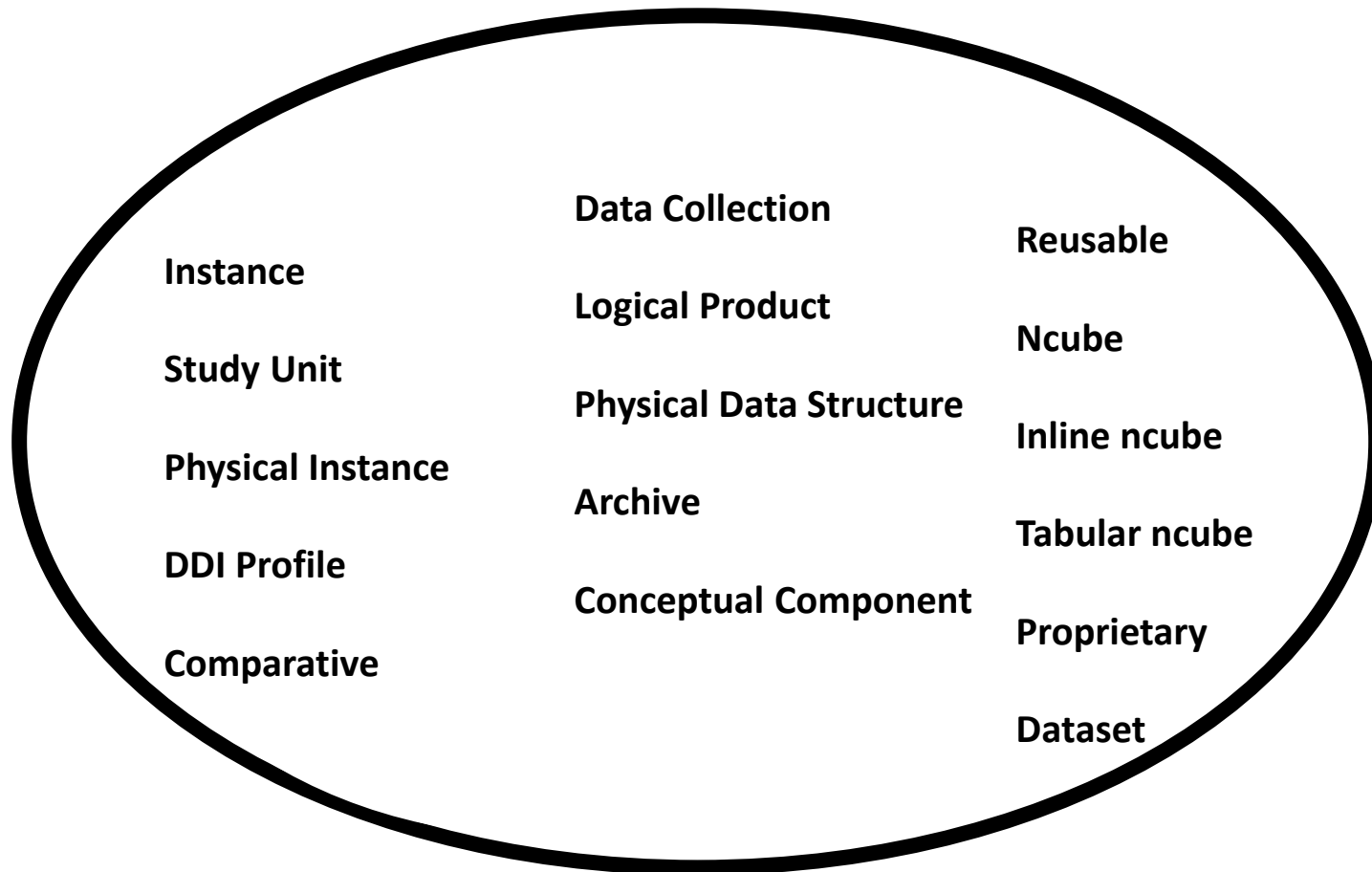
Direction of Archival data flow



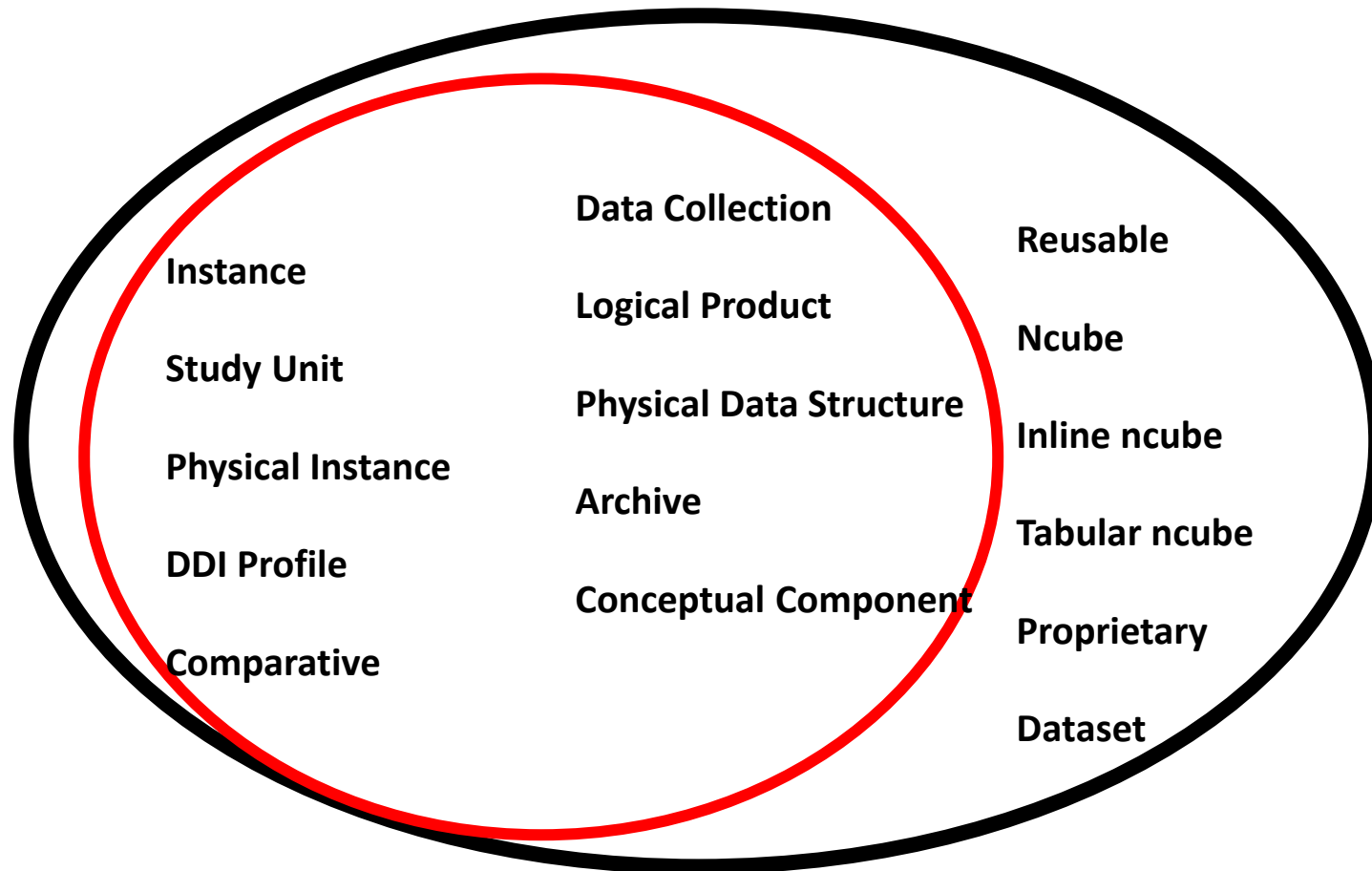
Direction of OAIS Administrative & Preservation Planning information flow



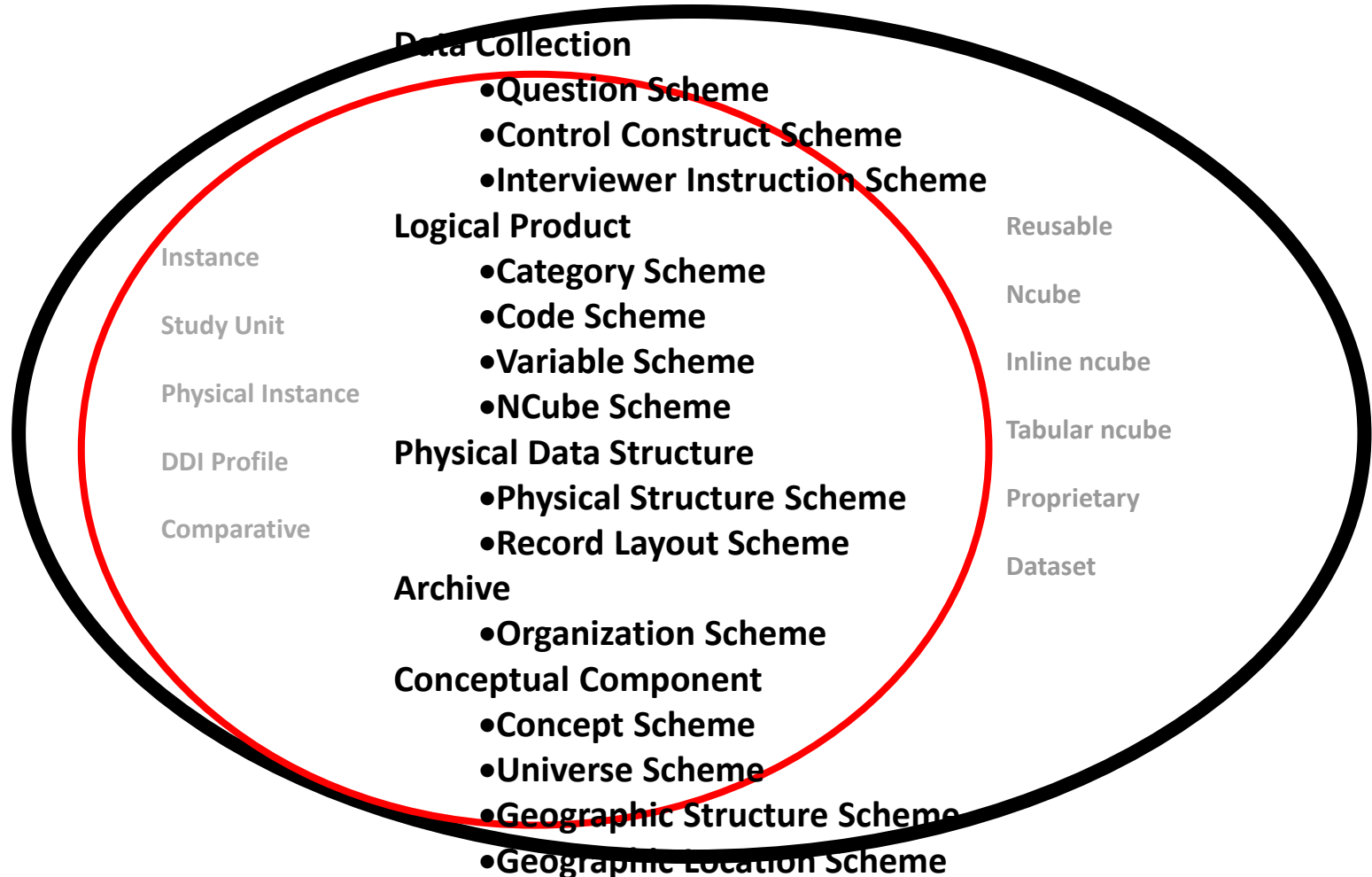
XML Schemas, DDI Modules, and DDI Schemes



, XML Schemas DDI Modules, and DDI Schemes



XML Schemas, DDI Modules, and DDI Schemes



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

Reuse Across the Lifecycle

- This basic metadata is reused across the lifecycle
 - Responses may use the same categories and codes which the variables use
 - Multiple waves of a study may re-use concepts, questions, responses, variables, categories, codes, survey instruments, etc. from earlier waves