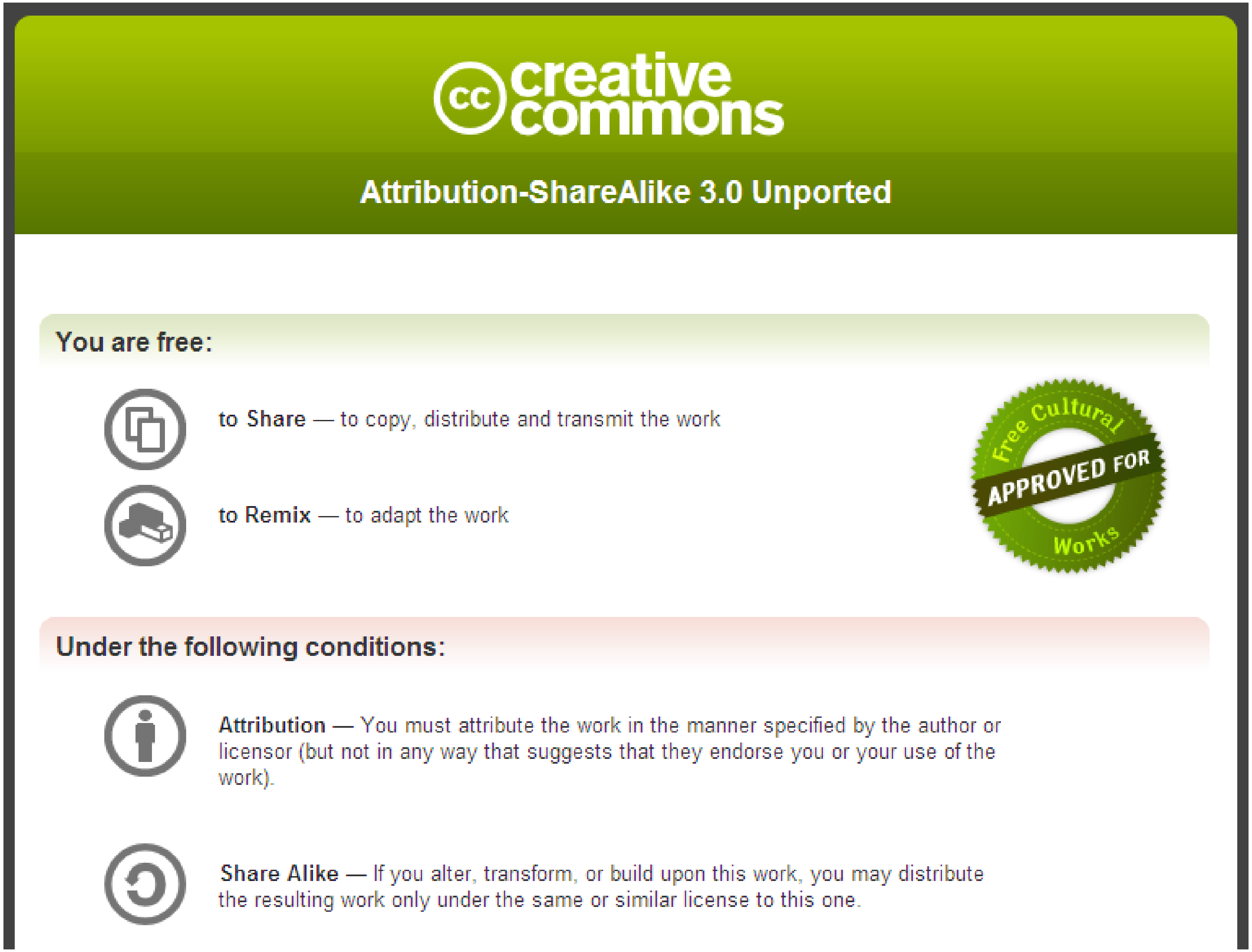
DDI: Capturing metadata throughout the research process for preservation and discovery

Wendy Thomas

NADDI 2012

University of Kansas

License



Attribution-ShareAlike 3.0 Unported

You are free:

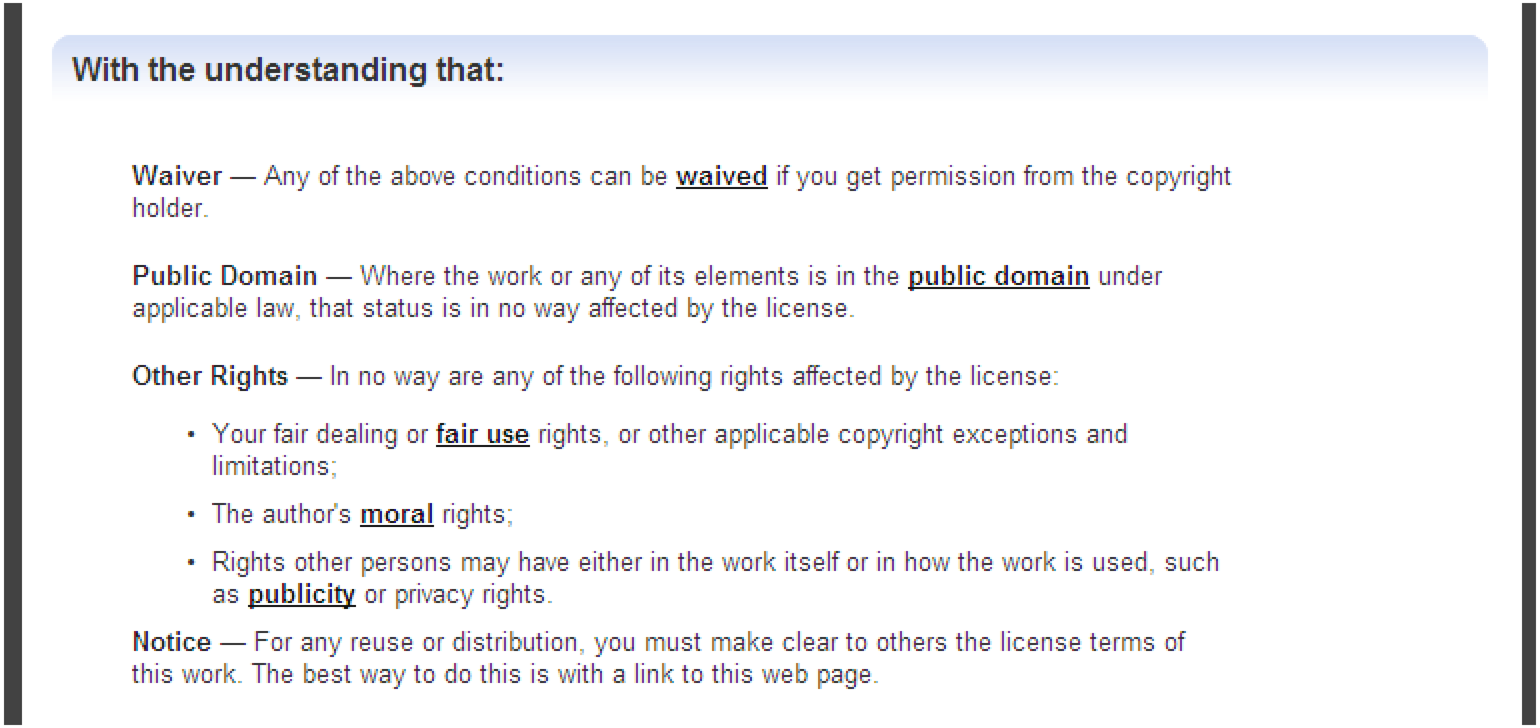
- **to Share** — to copy, distribute and transmit the work
- **to Remix** — to adapt the work

Under the following conditions:

- **Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

Details on next slide.

License (cont.)



With the understanding that:

Waiver — Any of the above conditions can be **waived** if you get permission from the copyright holder.

Public Domain — Where the work or any of its elements is in the **public domain** under applicable law, that status is in no way affected by the license.

Other Rights — In no way are any of the following rights affected by the license:

- Your fair dealing or **fair use** rights, or other applicable copyright exceptions and limitations;
- The author's **moral** rights;
- Rights other persons may have either in the work itself or in how the work is used, such as **publicity** or privacy rights.

Notice — For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.

On-line available at: http://creativecommons.org/licenses/by-sa/3.0/

This is a human-readable summary of the Legal Code at:
http://creativecommons.org/licenses/by-sa/3.0/legalcode

Credits

- Some of these slides were developed for 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

MPC Minnesota Population Center

gesis Leibniz Institute for the Social Sciences

Open Data Foundation

Outline

- Introductions
- History of DDI in terms of development lines
- DDI Codebook (DDI-C)
  - Capturing metadata within DDI-C (NESSTAR)
  - Structural features of DDI-C and implications for use
- DDI Lifecycle (DDI-L)
  - Capturing metadata within DDI-L (DdiEditor)
  - Structural features of DDI-L and implications for use
- DDI resources

Introductions

- Who are you?
- What does your organization do?
    - Data collection
    - Data production
    - Training
    - User access
    - Preservation
- What is the scale of your operations?

HISTORY OF DDI IN TERMS OF DEVELOPMENT LINES

Background

- Concept of DDI and definition of needs grew out of the data archival community
- Established in 1995 as a grant funded project initiated and organized by ICPSR
- Members:
  - Social Science Data Archives (US, Canada, Europe)
  - Statistical data producers (including US Bureau of the Census, the US Bureau of Labor Statistics, Statistics Canada and Health Canada)
- February 2003 – Formation of DDI Alliance
  - Membership based alliance
  - Formalized development procedures

Early DDI: Characteristics of DDI 1/2

- Focuses on the static object of a codebook
- Designed for limited uses
  - End user data discovery via the variable or high level study identification (bibliographic)
  - Only heavily structured content relates to information used to drive statistical analysis
- Coverage is focused on single study, single data file, simple survey and aggregate data files
- Variable contains majority of information (question, categories, data typing, physical storage information, statistics)

Limitations of these Characteristics

- Treated as an “add on” to the data collection process
- Focus is on the data end product and end users (static)
- Limited tools for creation or exploitation
- The Variable must exist before metadata can be created
- Producers hesitant to take up DDI creation because it is a cost and does not support their development or collection process

DDI Development

Codebook		Lifecycle	
Version 1 2000	**Version 2 2003**	**Version 3 2008**	***[Version 4 20??]***
Version 1.01 2001	*Version 2.1 2005*	*Version 3.1 2009*	
Version 1.02 2001	*Version 2.5 2012*	*Version 3.2 [2013]*	
Version 1.3 2002	*[Version 2.6 20??]*	*[Version 3.3 20??]*	

Desired Areas of Coverage

- Simple survey
- Aggregate data
- Complex data file structures
- Series (linkages between studies)
- Programming and software support
- Support for GIS users
- Support for CAI systems
- Support comparability (by-design, harmonization)

Codebook

Version 1

Simple survey

Option for some programming and software support

Version 2

Aggregate data

Some support for GIS users

Lifecycle

Version 3

Simple survey

Aggregate data

Programming and software support

GIS support

CAI support

Complex data files

Series

Comparability

Technical difference between Codebook and Lifecycle structures

- Codebook
  - Codebook based
  - Format originally XML DTD (2.5 XML Schema)
  - After-the-fact
  - Static
  - Metadata replicated
  - Simple study
  - Limited physical storage options

- Lifecycle
  - Lifecycle based
  - Format XML Schema
  - Point of occurrence
  - Dynamic
  - Metadata reused
  - Simple study, series, grouping, inter-study comparison
  - Unlimited physical storage options

DDI CODEBOOK

2.5 Feature Enhancement

- Added sections covering
  - Study authorization
  - Study budget
  - Ex-post study evaluation
  - Collector training
  - Instrument development
- Expanded detail
  - Sample procedure to include sample frame and target sample
  - Response rate
  - Typing of data appraisal
  - Detail for data processing and coding instruction
  - Allows for citation and persistent identifier for individual data files

Compatibility Enhancements

- Ability to capture and retain DDI-Lifecycle identification information (agency, id format, version information)
- Designation of a single note as a master and capture all the related objects it should be attached to
- Use of XHTML for structured content
- Ability to capture DDI-Lifecycle representation or response domain type explicitly
- Capture full range of ISO date types
- Added new sections to DDI-Lifecycle (v.3.2) to reflect added information fields in DDI-Codebook (v.2.5)

Conventionalize the use of Controlled Vocabularies

- Declare the use of a specific controlled vocabulary
- Define the object (element or attribute) that uses the controlled vocabulary
- Define the valid controlled vocabulary value of a non-valid legacy entry

DDI-Codebook Applications

- Simple survey capture
- High level study description with variable information for stand alone studies
- Descriptions of basic nCubes (aggregate / statistical tables)
- Replicating the contents of a codebook including the data dictionary
- Collection management beyond bibliographic records

Continued use of DDI-Codebook

- Current users
- New users whose needs are met by DDI-Codebook
- Software availability
    - NESSTAR
    - IHSN Microdata Toolkit
    - NADA Catalog
    - World Bank Open Data program (Microdata Catalog)

DDI-Codebook

Document Description

- Citation of the codebook document
- Guide to the codebook
- Document status
- Source for the document

Study Description

- Citation for the study
- Study Information
- Methodology
- Data Accessibility
- Other Study Material

File Description

- File Text (record and relationship information)
- Location Map (required for nCubes optional for microdata)

Data Description

- Variable Group and nCube Group
- Variable (variable specification, physical location, question, & statistics)
- nCube

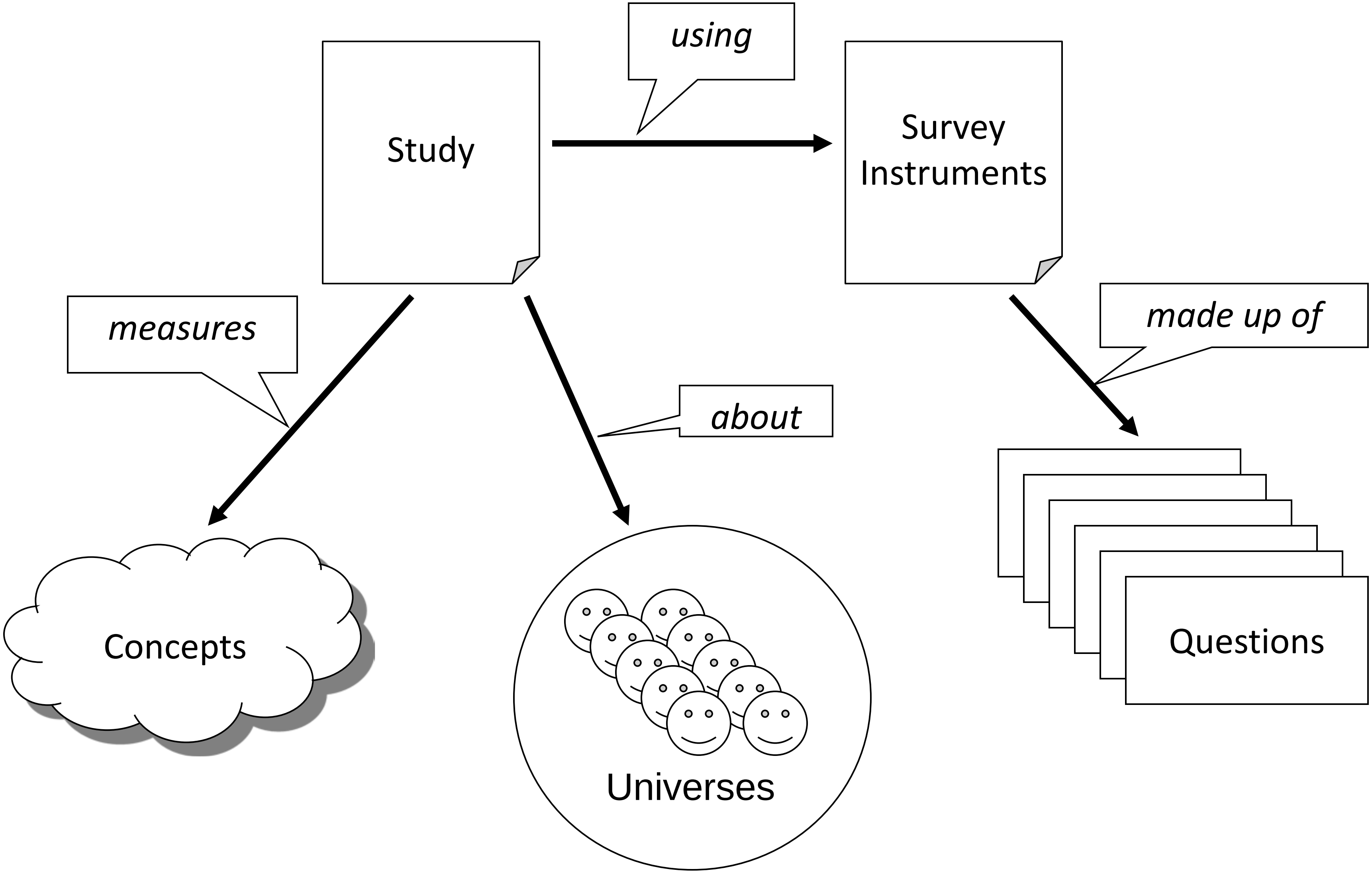
Other Material

DDI Codebook

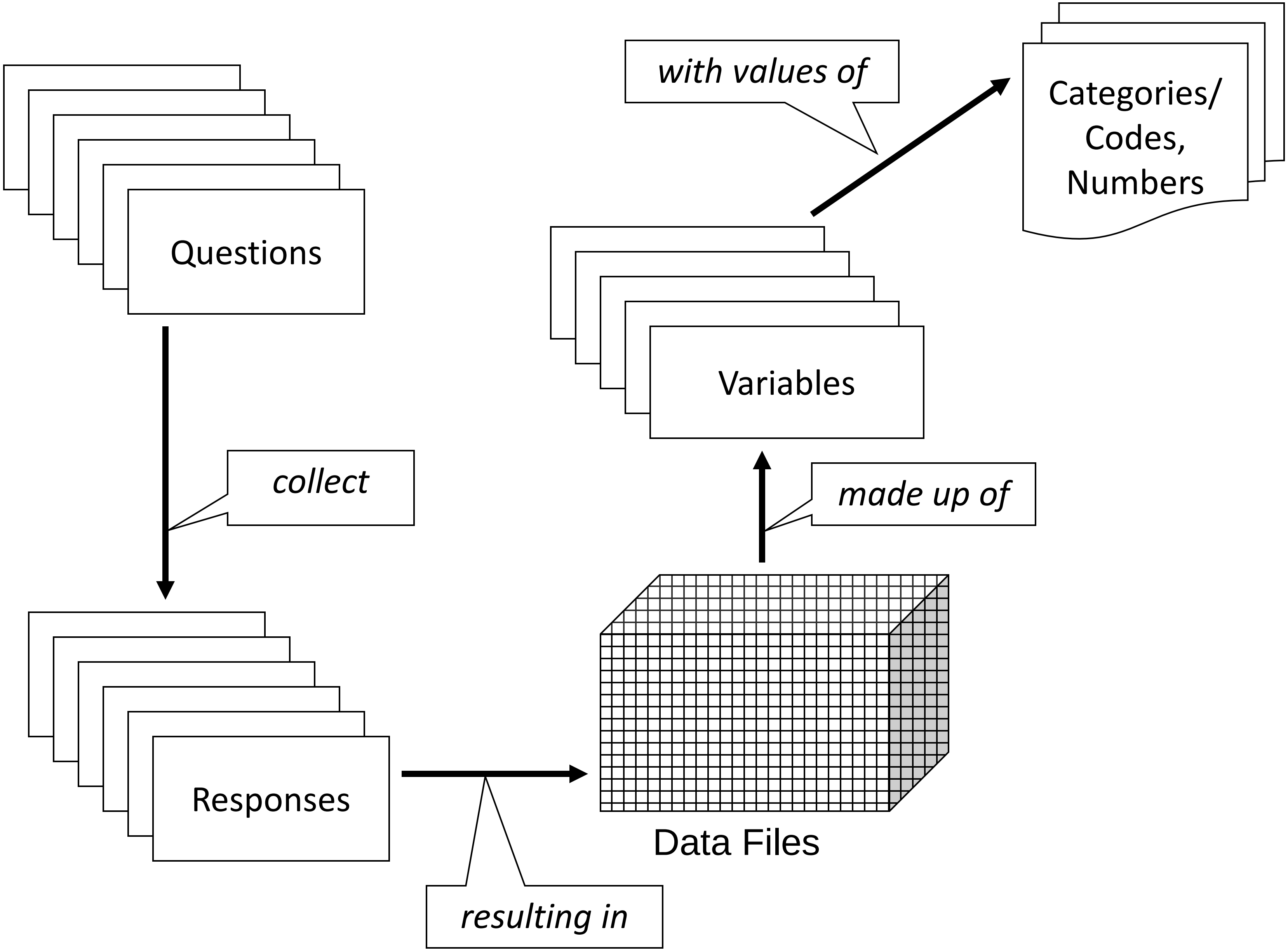
- Looking into NESSTAR

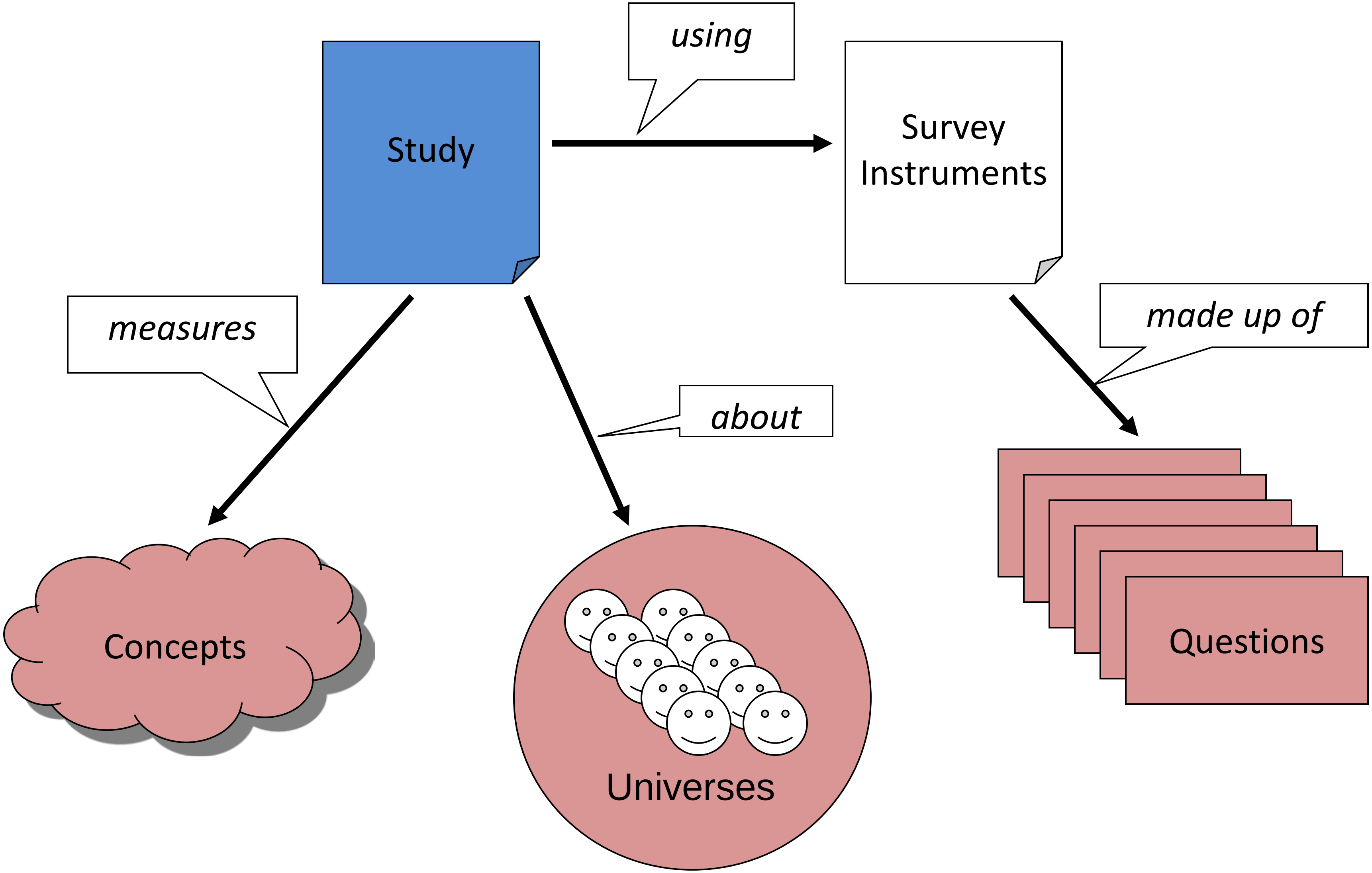
Introduction to DDI

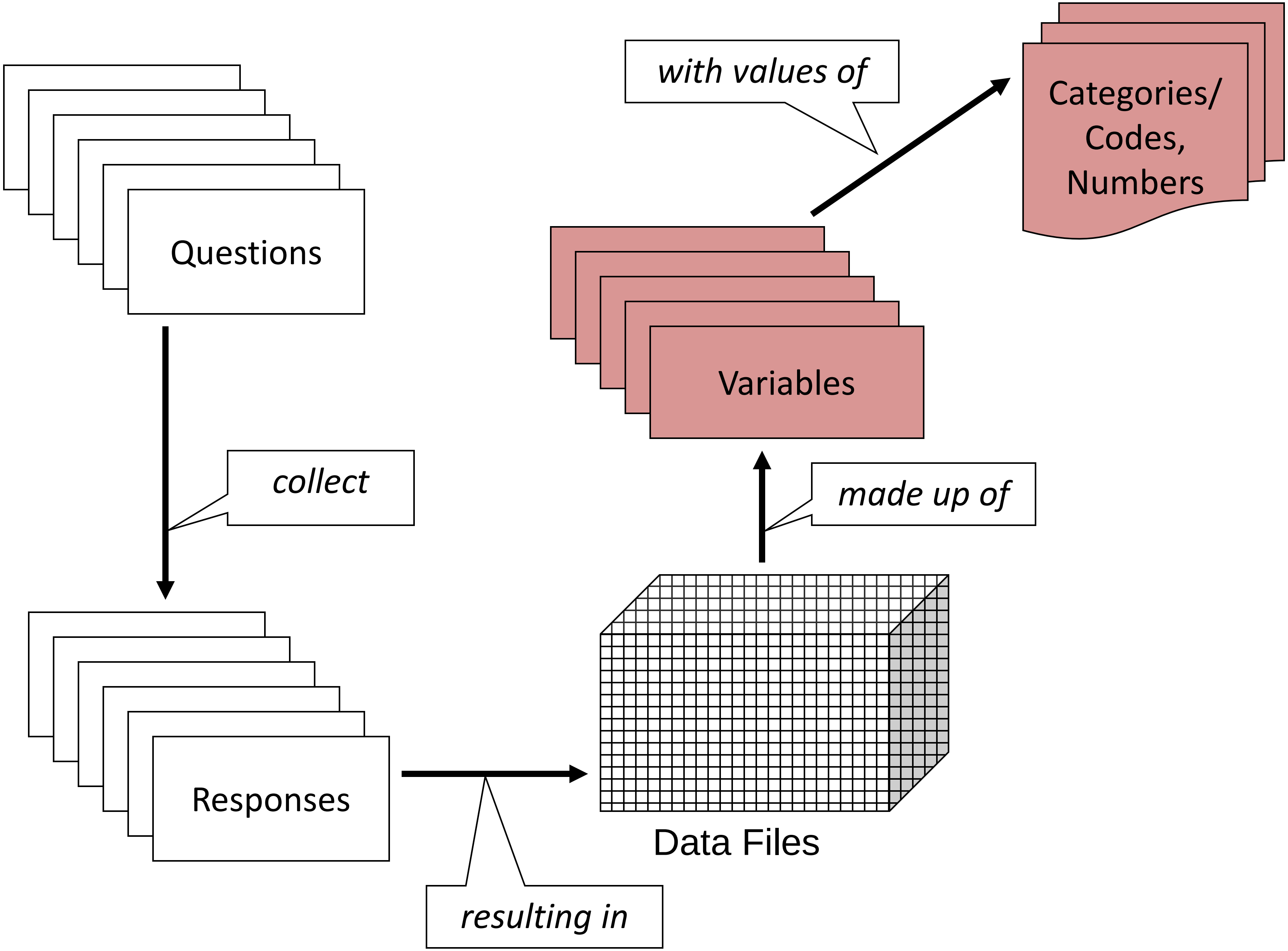
DDI 3 IN 60 SECONDS



using
Study
Survey Instruments
measures
made up of
about
Concepts
Universes
Questions


Questions
collect
Responses
resulting in
Data Files
made up of
Variables
with values of
Categories/
Codes,
Numbers


using
Study
Survey Instruments
measures
about
made up of
Concepts
Universes
Questions


with values of
Categories/ Codes, Numbers
Questions
Variables
collect
made up of
Responses
Data Files
resulting in

DDI LIFECYCLE

Change

- DDI-L is a major change from DDI-C in terms of content and structure. Lets step back and look at:
  - Basic differences between DDI-C and DDI-L
  - Applications for DDI-C and DDI-L
  - Differences that allow DDI-L to do more
  - How these differences provide support for better management of information, data, and metadata

DDI-L Applications

- Describing a series of studies such as a longitudinal survey or cross-cultural survey
- Capturing comparative information between studies
- Sharing and reusing metadata outside the context of a specific study
- Capturing data in the XML
- Capturing process steps from conception of study through data capture to data dissemination and use
- Capturing lifecycle information as it occurs, and in a way that can inform and drive production
- Management of data and metadata within an organization for internal use or external access

Why can DDI-L 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

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

Reuse by Reference

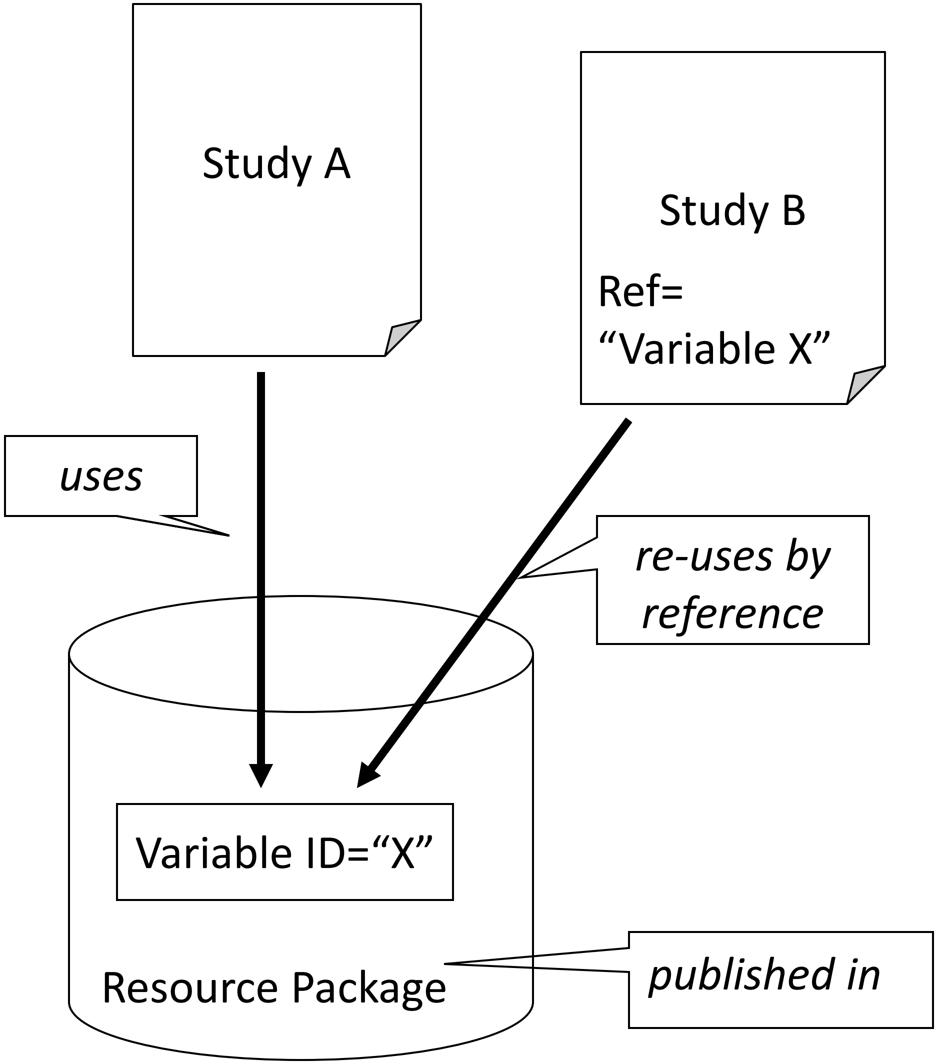
- When a piece of metadata is re-used, a *reference* can be made to the original
- In order to reference the original, you must be able to *identify* it
- You also must be able to *publish* it, so it is visible (and can be referenced)
  - It is published to the user community – those users who are allowed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

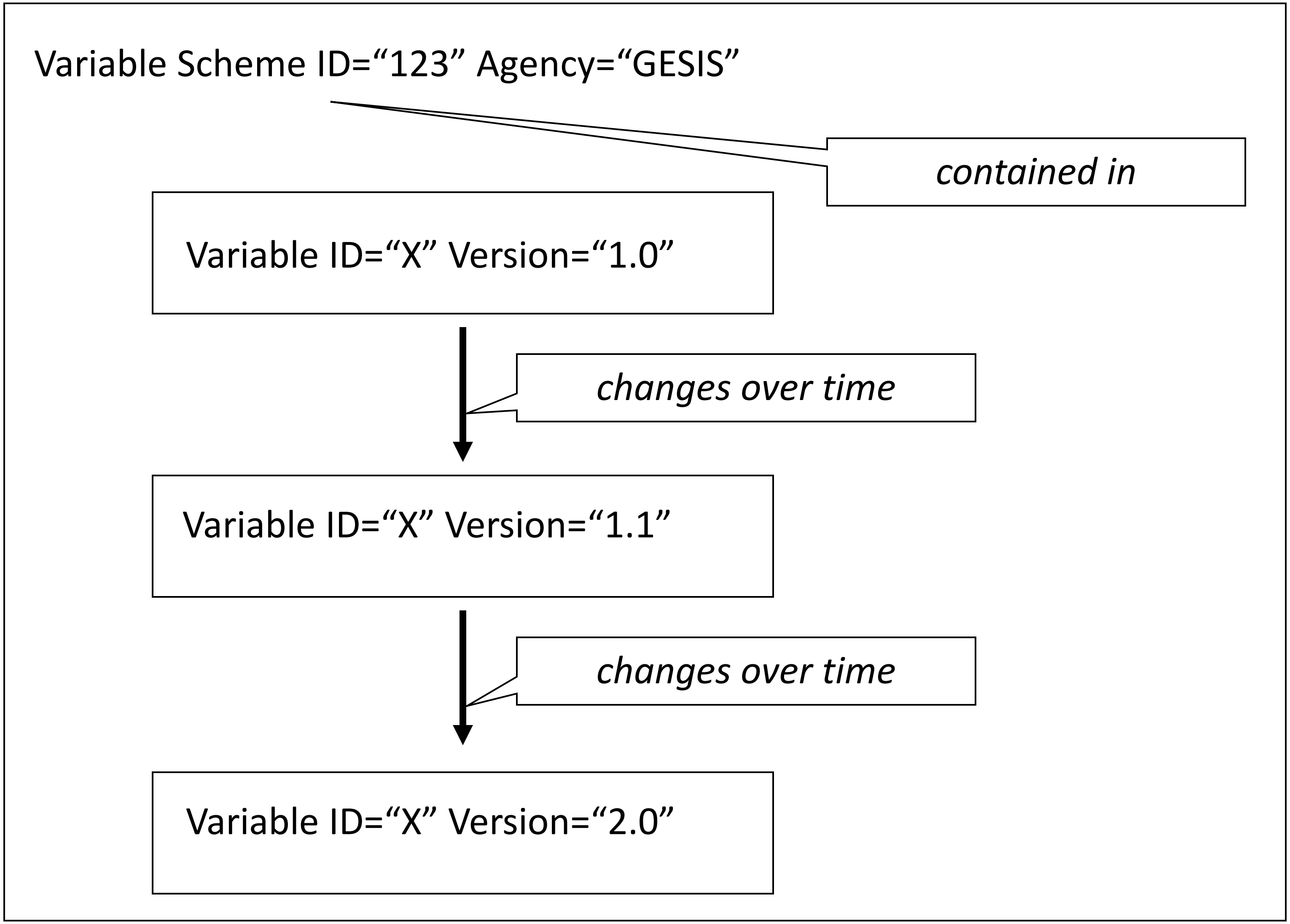
DDI Support for Metadata Reuse

- 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 may be *versionable*
  - Versions reflect changes over time
  - The versionable metadata has a version number


Study A
Study B
Ref=
“Variable X”
uses
re-uses by reference
Variable ID=“X”
Resource Package
published in

Management of Information, Data, and Metadata

- An organization can manage its organizational information, metadata, and data within repositories using DDI-L 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



Variable Scheme ID="123" Agency="GESIS"

contained in

Variable ID="X" Version="1.0"

changes over time

Variable ID="X" Version="1.1"

changes over time

Variable ID="X" Version="2.0"

Looking inside the DdiEditor

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

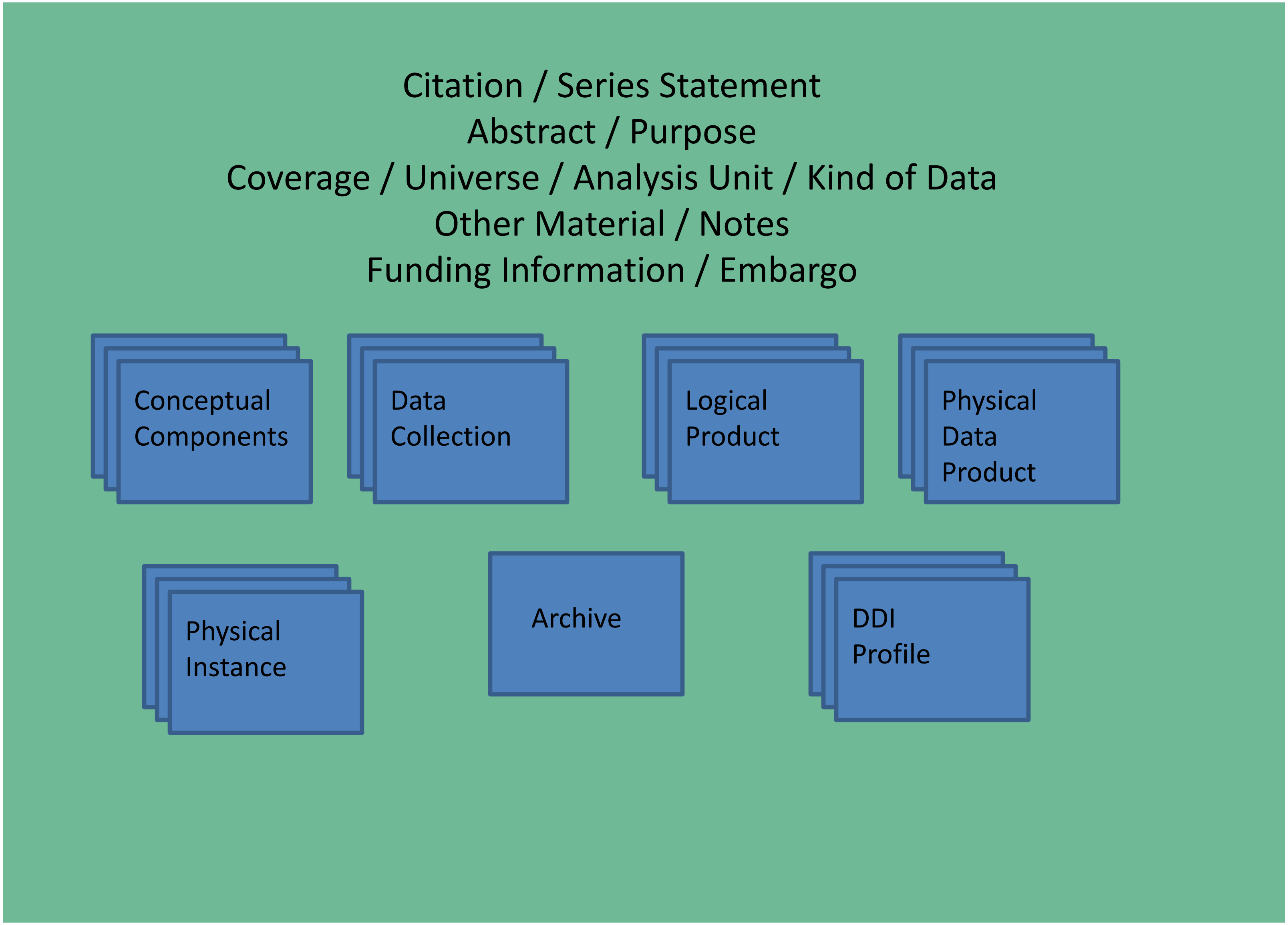
Virtual Data

- When researchers use data, they often combine variables from several sources
  - This can be viewed as a “virtual” data set
  - The re-coding and processing can be captured as useful metadata
  - The researcher’s data set can be re-created from this metadata
  - Comparability of data from several sources can be expressed

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

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

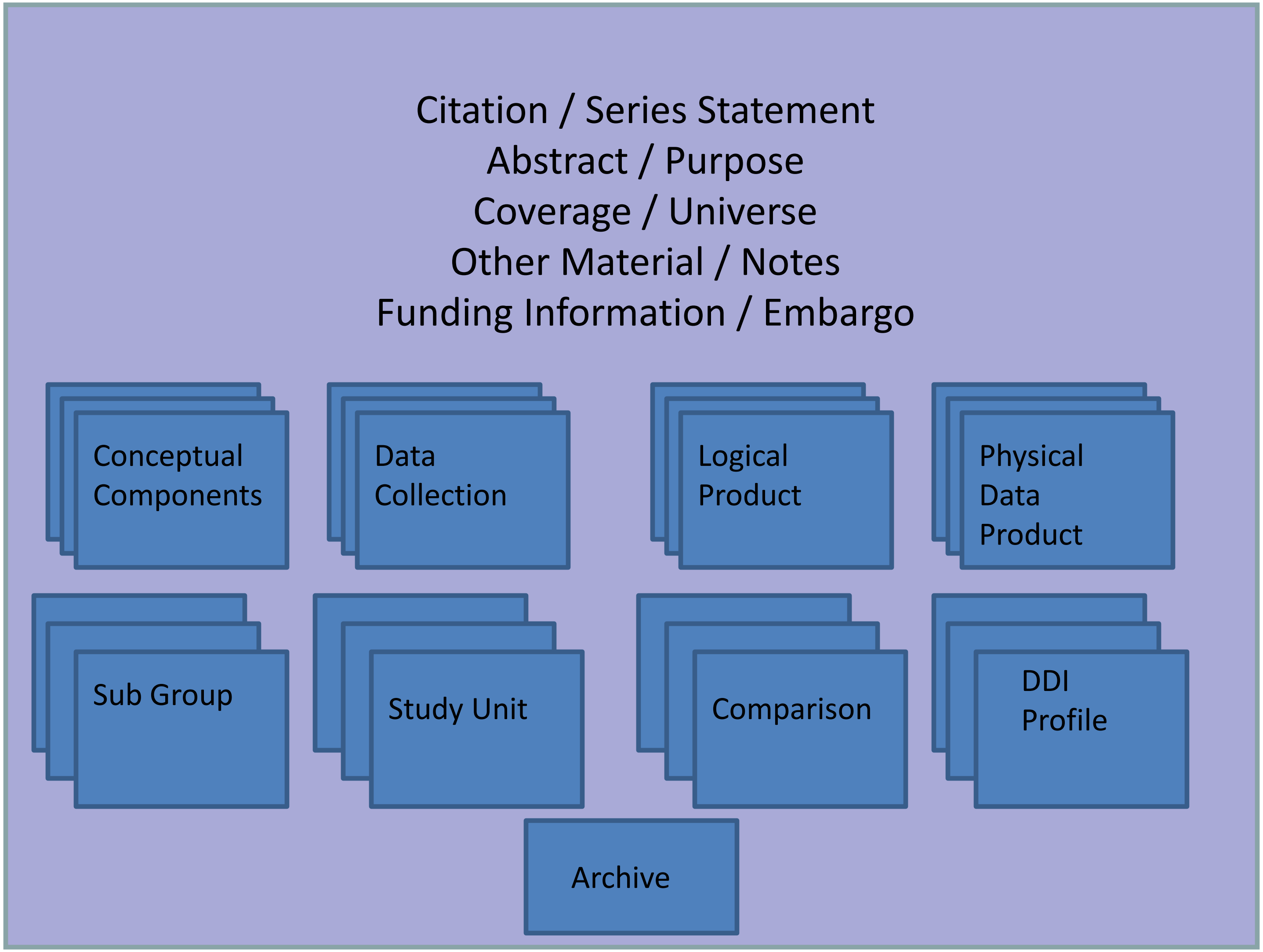
Physical Data Product

Physical Instance

Archive

DDI Profile

Group



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

Conceptual Components

Data Collection

Logical Product

Physical Data Product

Sub Group

Study Unit

Comparison

DDI Profile

Archive

DDI 3 Lifecycle Model and Related Modules

Groups and Resource Packages are a means of publishing any portion or combination of sections of the life cycle

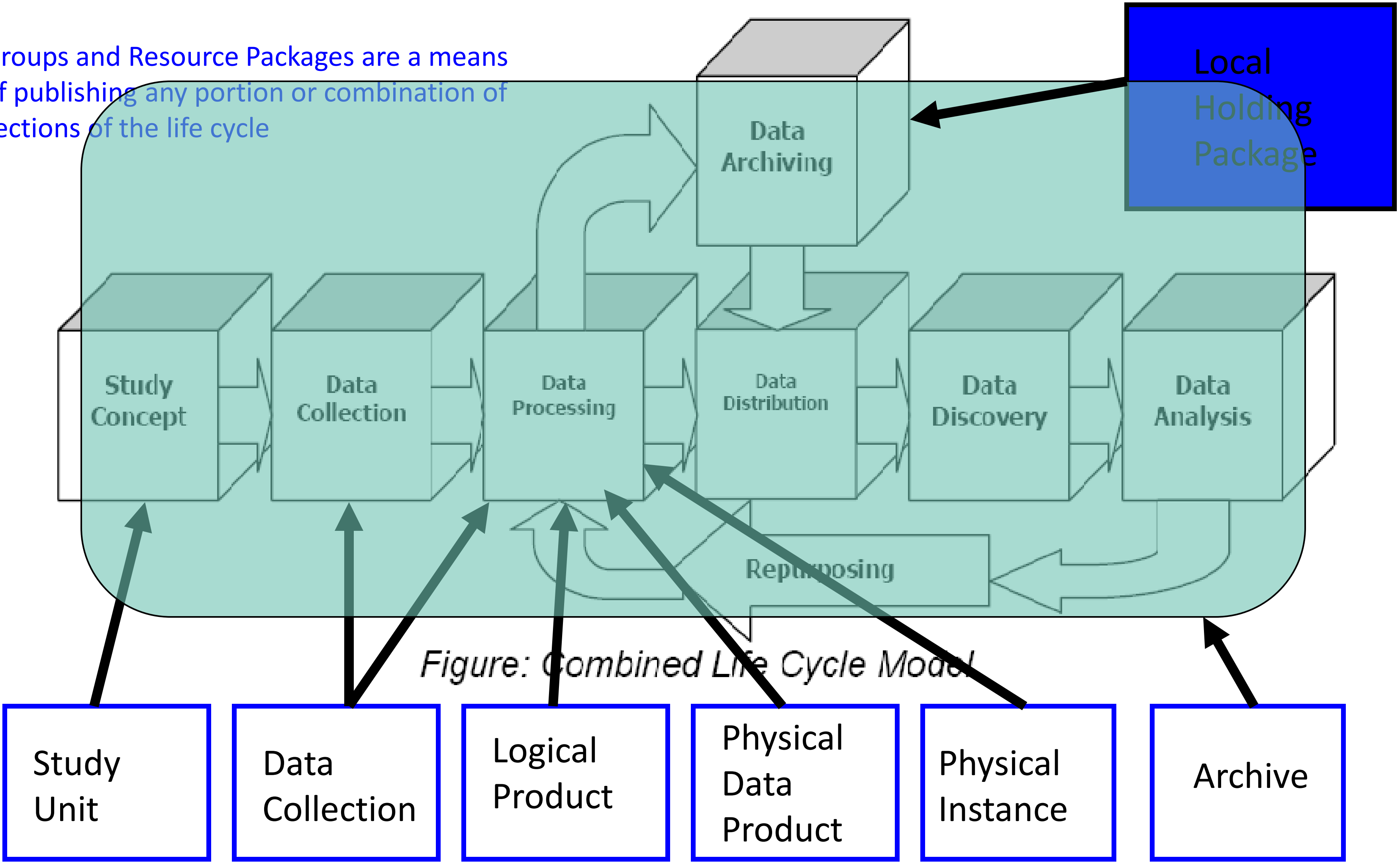


Figure: Combined Life Cycle Model

Local Holding Package

Study Unit

Data Collection

Logical Product

Physical Data Product

Physical Instance

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

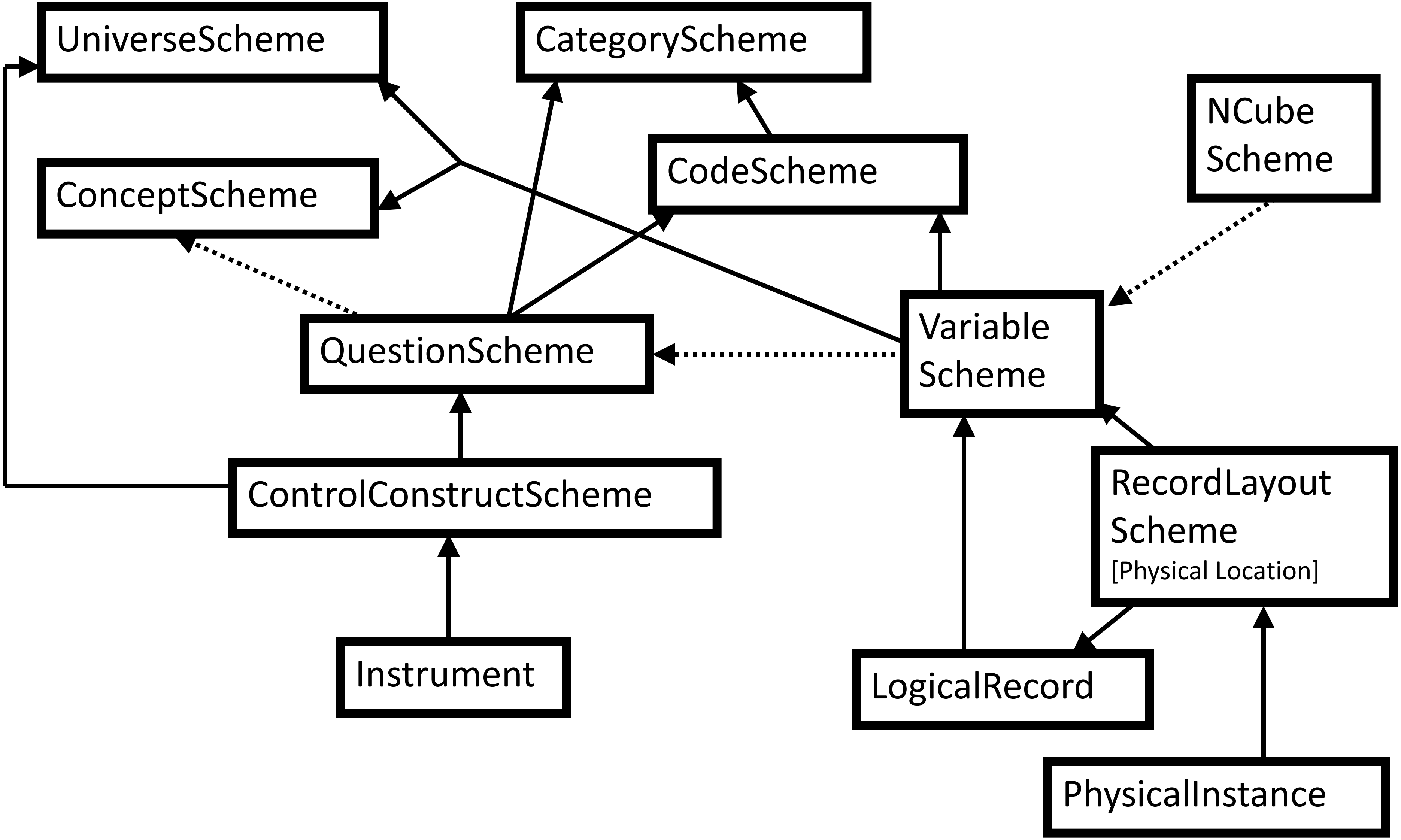
Physical storage

- 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

- An archive is whatever organization or individual has current control over the metadata
- Contains persistent lifecycle events
- Contains archive specific information
  - local identification
  - local access constraints

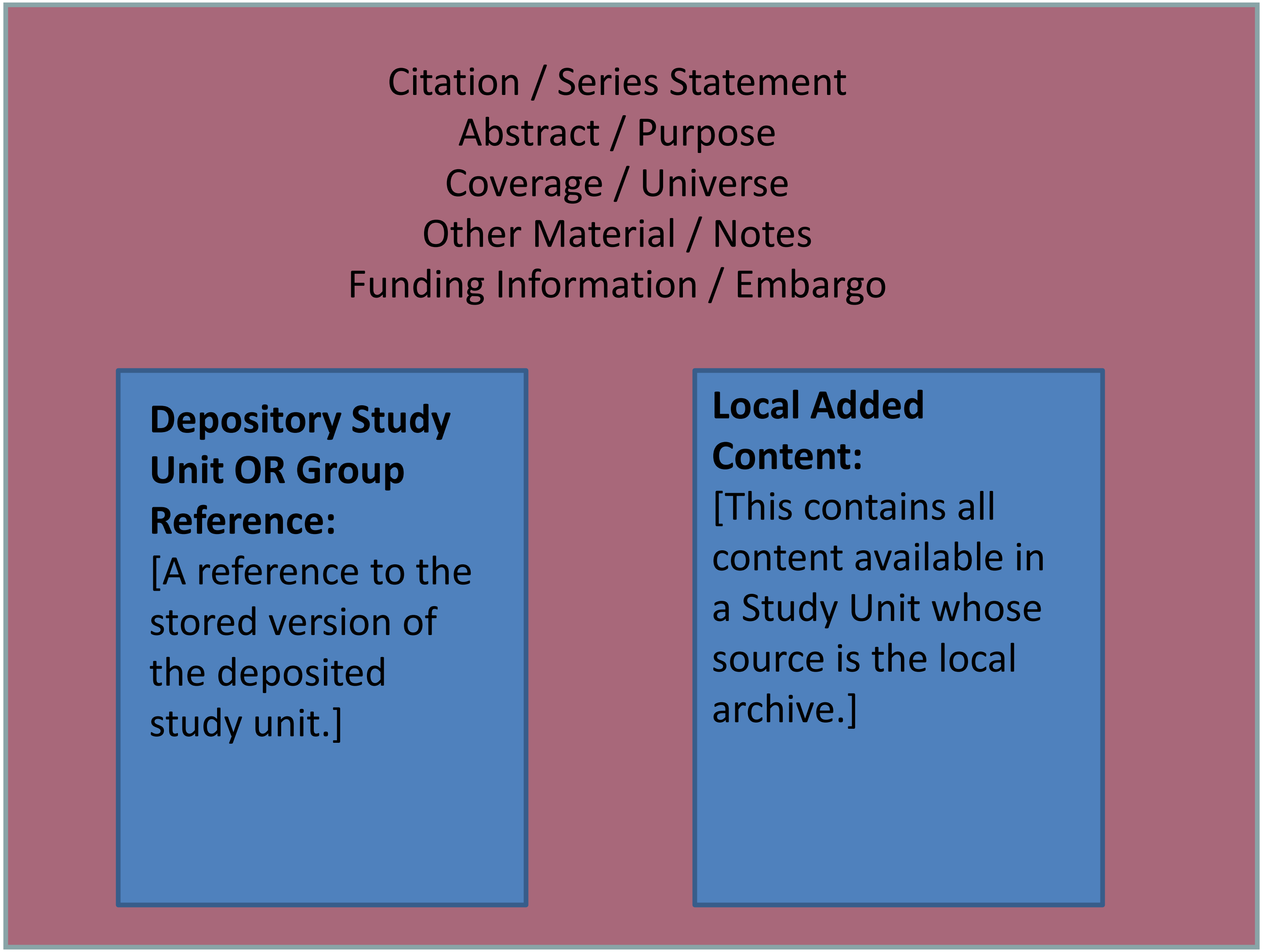
Building from Component Parts



Group

- Resource Package
  - Allows packaging of any maintainable item as a resource item
- Group
  - Up-front design of groups – allows inheritance
  - Ad hoc (“after-the-fact”) groups – explicit comparison using comparison maps for Universe, Concept, Question, Variable, Category, and Code
- Local Holding Package
  - Allows attachment of local information to a deposited study without changing the version of the study unit itself

3.1 Local Holding Package

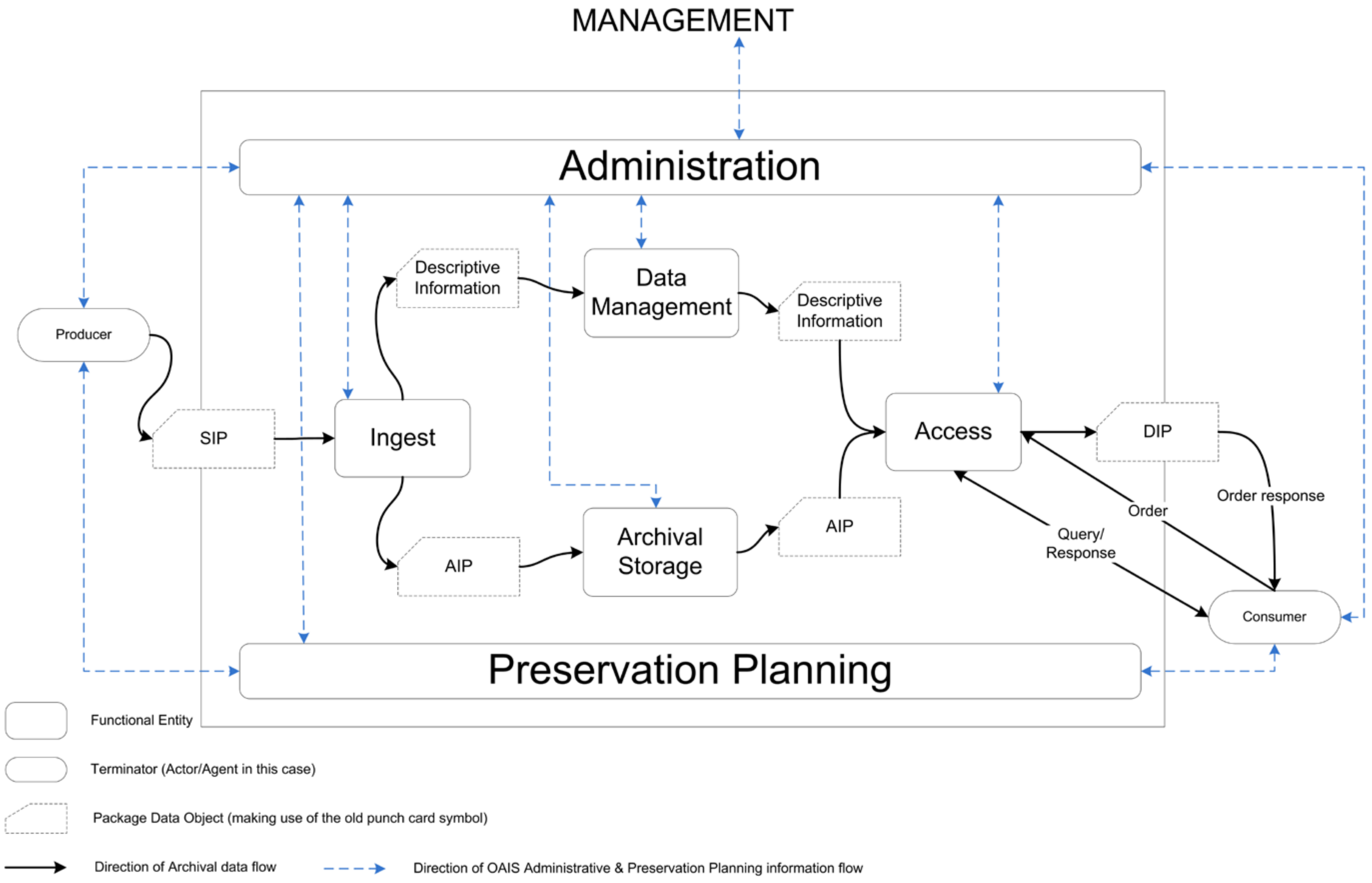


Citation / Series Statement
Abstract / Purpose
Coverage / Universe
Other Material / Notes
Funding Information / Embargo

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.]

Based on “Figure 4-1: OAIS Functional Entities” OAIS Pink Book 2009



Support for OAIS content

- SIP
  - Study level material
  - Variable and Question content
  - Methodology and process
- AIP
  - Lifecycle events (Archive processing)
  - Value added
  - Provenance
- DIP
  - Selections of all of the above

DDI Resources

- http://ddialliance.org
  - Specifications
  - Resources
    - Tools
    - Best Practices
    - Use cases
  - DDI Users list