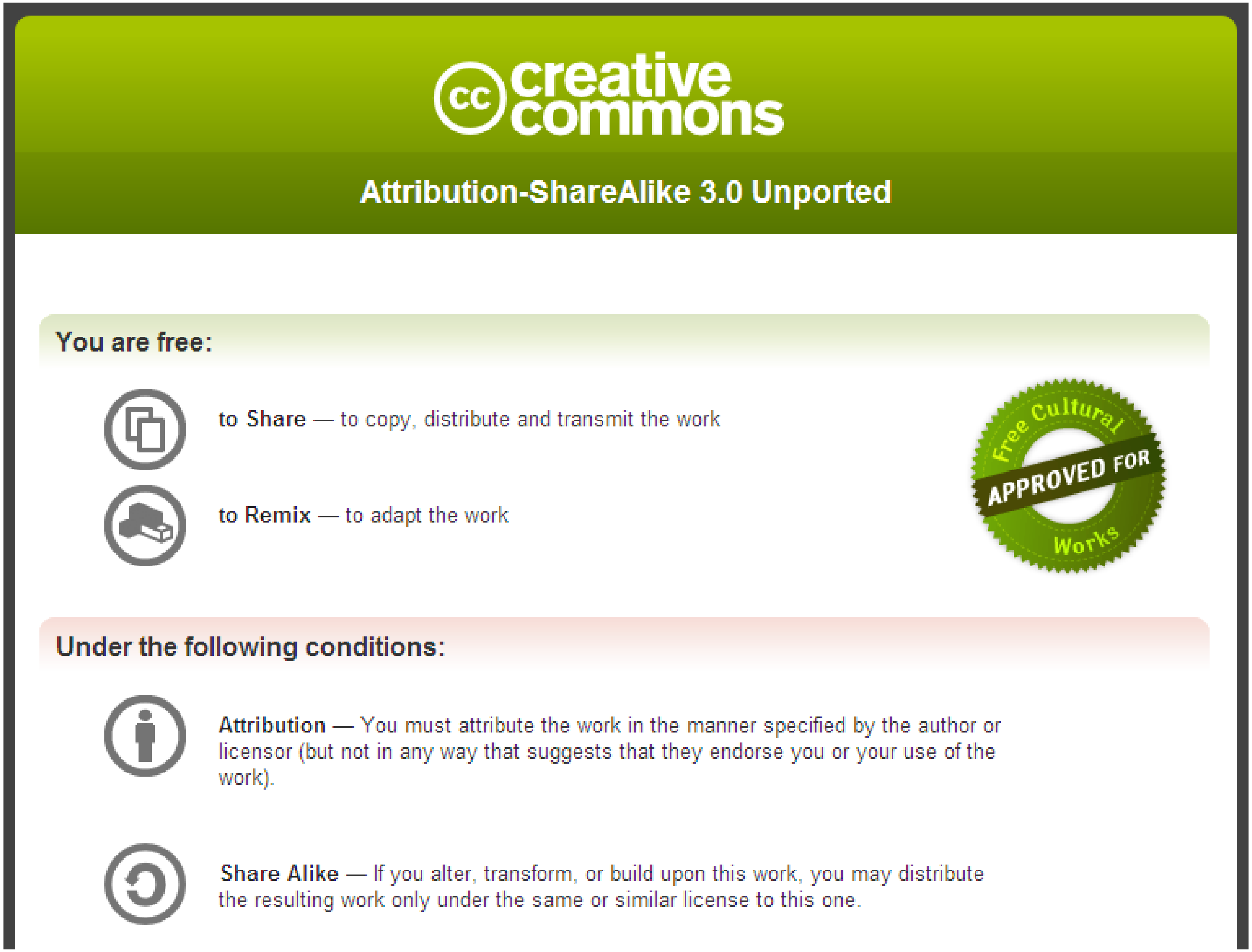
# DDI: Metadata to support collection processes, discovery, and comparability

Wendy Thomas

NADDI 2012

University of Kansas

# License



## Attribution-ShareAlike 3.0 Unported

**You are free:**

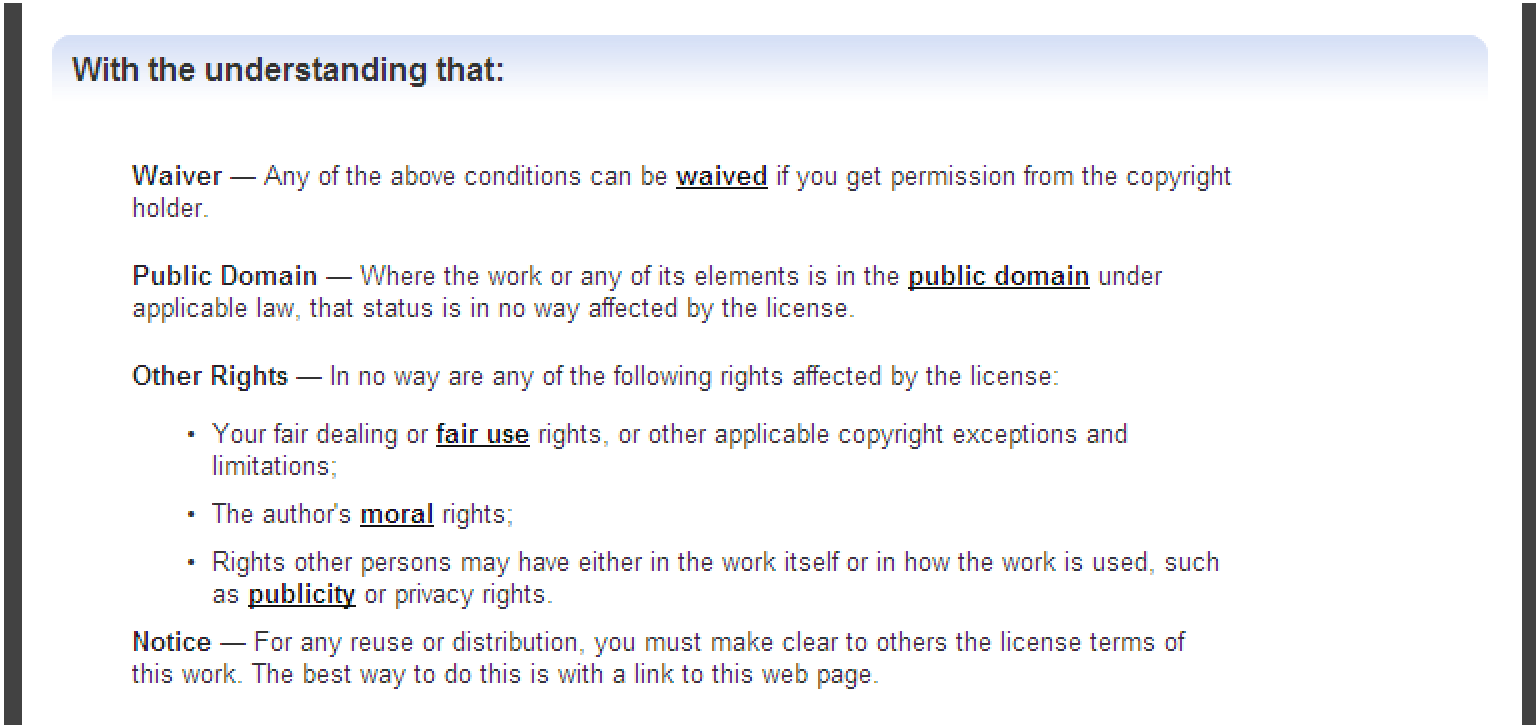
- **to Share** — to copy, distribute and transmit the work
- **to Remix** — to adapt the work

**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

- Intros
- Study Units, Groups, and Resource Packages
- Modules vs. Schemes
- Questions, data collection instruments, variables, concepts, geography, grouping and comparison
- Process control
- Discovery tools and "reusable" metadata
- Support for comparability within and between data sets

# Introductions

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

# 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

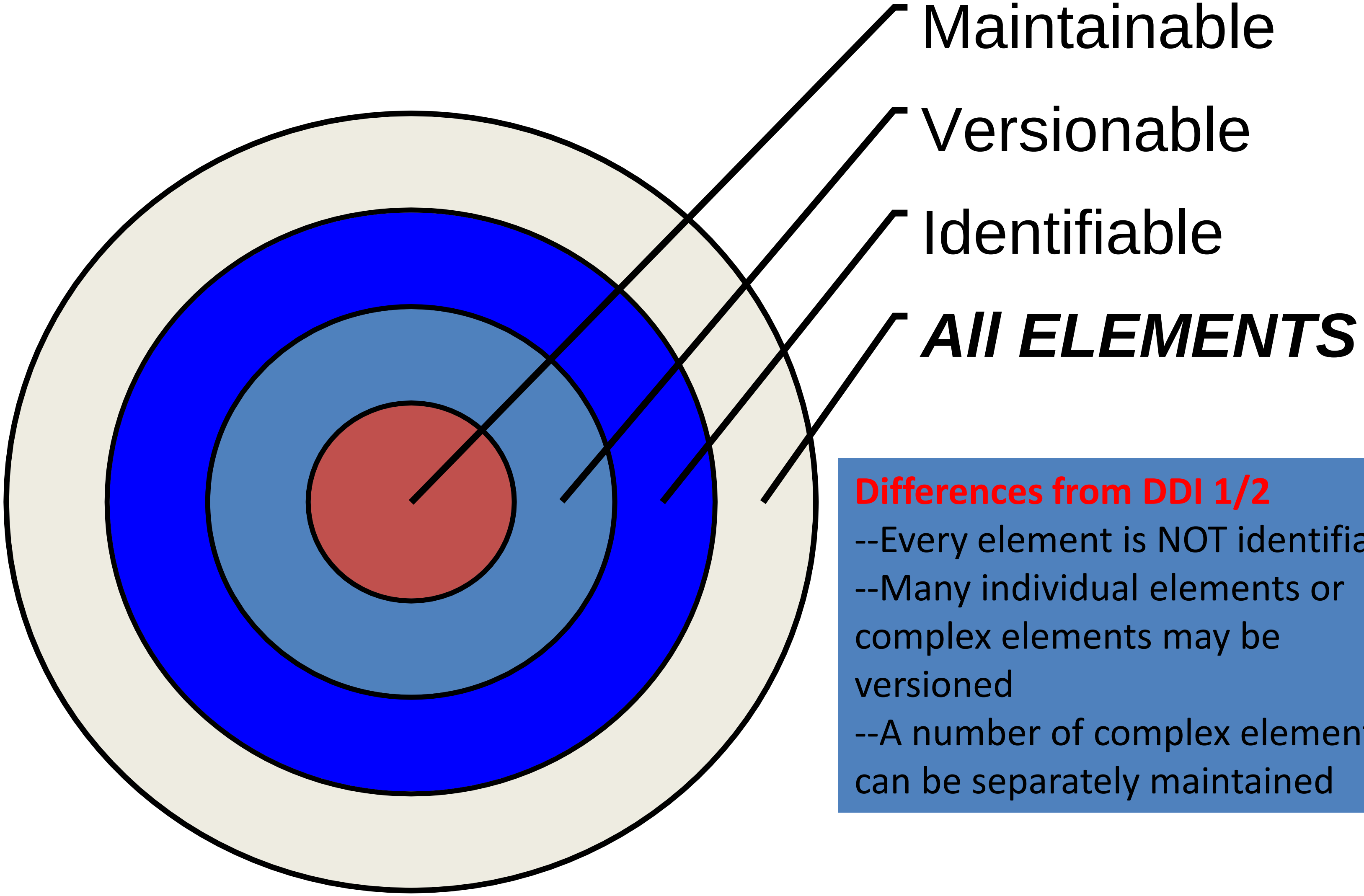
# Change over Time

- 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 can be *versionable*
  - This reflects changes over time
  - The versionable metadata has a version number

# Basic Element Types



Maintainable

Versionable

Identifiable

***All ELEMENTS***

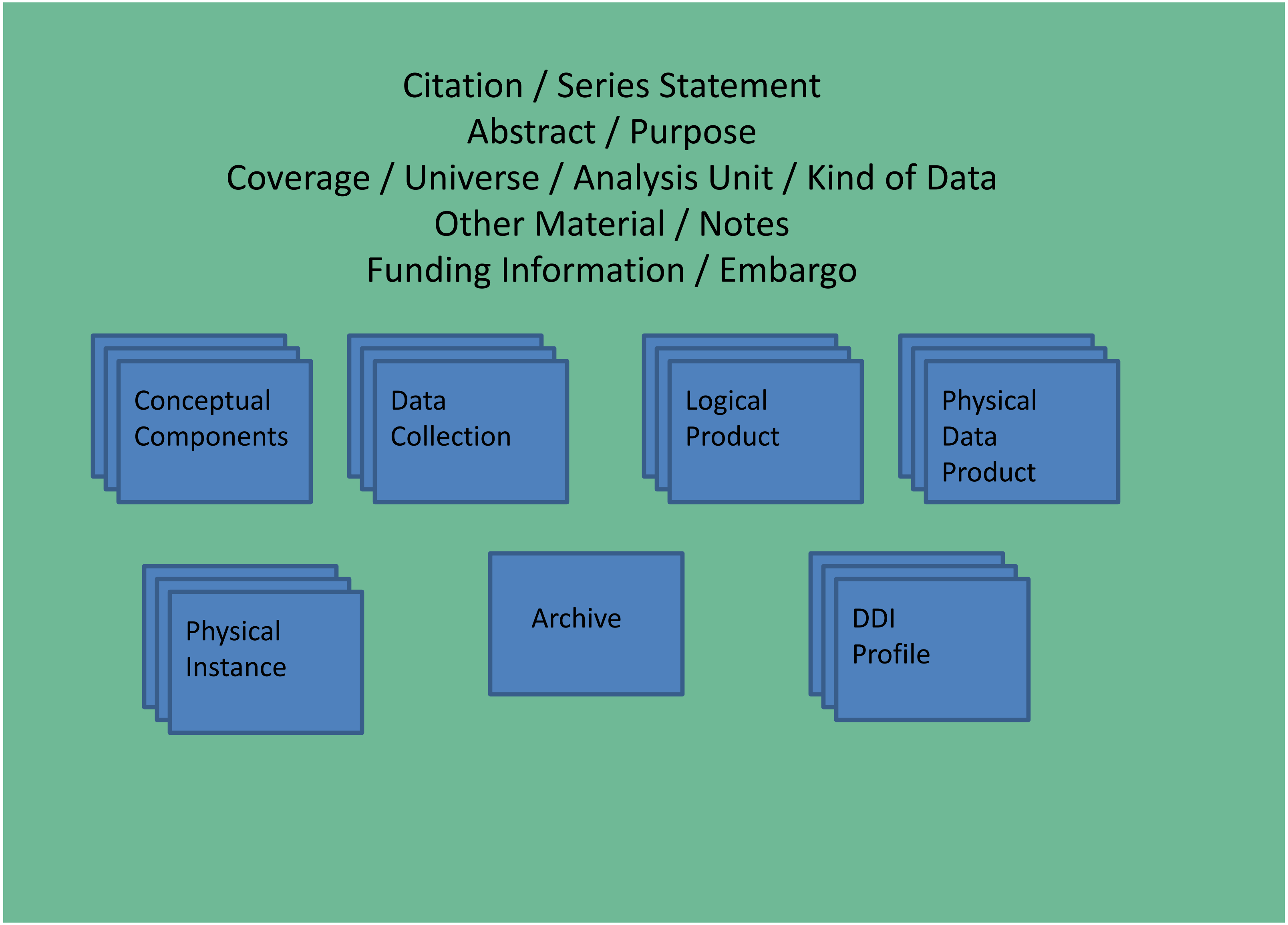
**Differences from DDI 1/2**

--Every element is NOT identifiable

--Many individual elements or complex elements may be versioned

--A number of complex elements can be separately maintained

# 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

# Questions and Instruments

- DDI 3 separates the questions which make up a survey instrument from the survey instrument itself
  - Questions can be re-used!
- There are several different types of question text
  - Many of these are the normal string types found throughout DDI 3

# Questions and Instruments

- DDI 3 separates the questions which make up a survey instrument from the survey instrument itself
    - Questions can be re-used!
- There are several different types of question text
    - Many of these are the normal string types found throughout DDI 3

# 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:

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

# 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)

# Category and Code Domains

- Use CategoryDomain when NO codes are provided for the category response

  [ ] Yes

  [ ] No

- Use CodeDomain when codes are provided on the questionnaire itself

  1. Yes
  2. No

# Category Schemes and Code Schemes

- Use the same structure as variables
- Create the category scheme or schemes first (do not duplicate categories)
- Create the code schemes using the categories
  - A category can be in more than one code scheme
  - A category can have different codes in each code scheme

# Numeric and Text Domains

- Numeric Domain provides information on the range of acceptable numbers that can be entered as a response
- Text domains generally indicate the maximum length of the response and can limit allowed content using a regular expression
- Additional specialized domains such as DateTime are also available
- Structured Mixed Response domain allows for multiple response domains and statements within a single question, when multiple response types are required

# 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

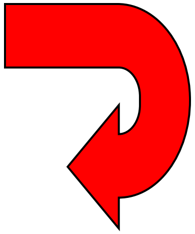
- Questions
- Response Domains
  - Code
  - Numeric
  - Text
- Statements
- Instructions

# Simple Questionnaire

Please answer the following:

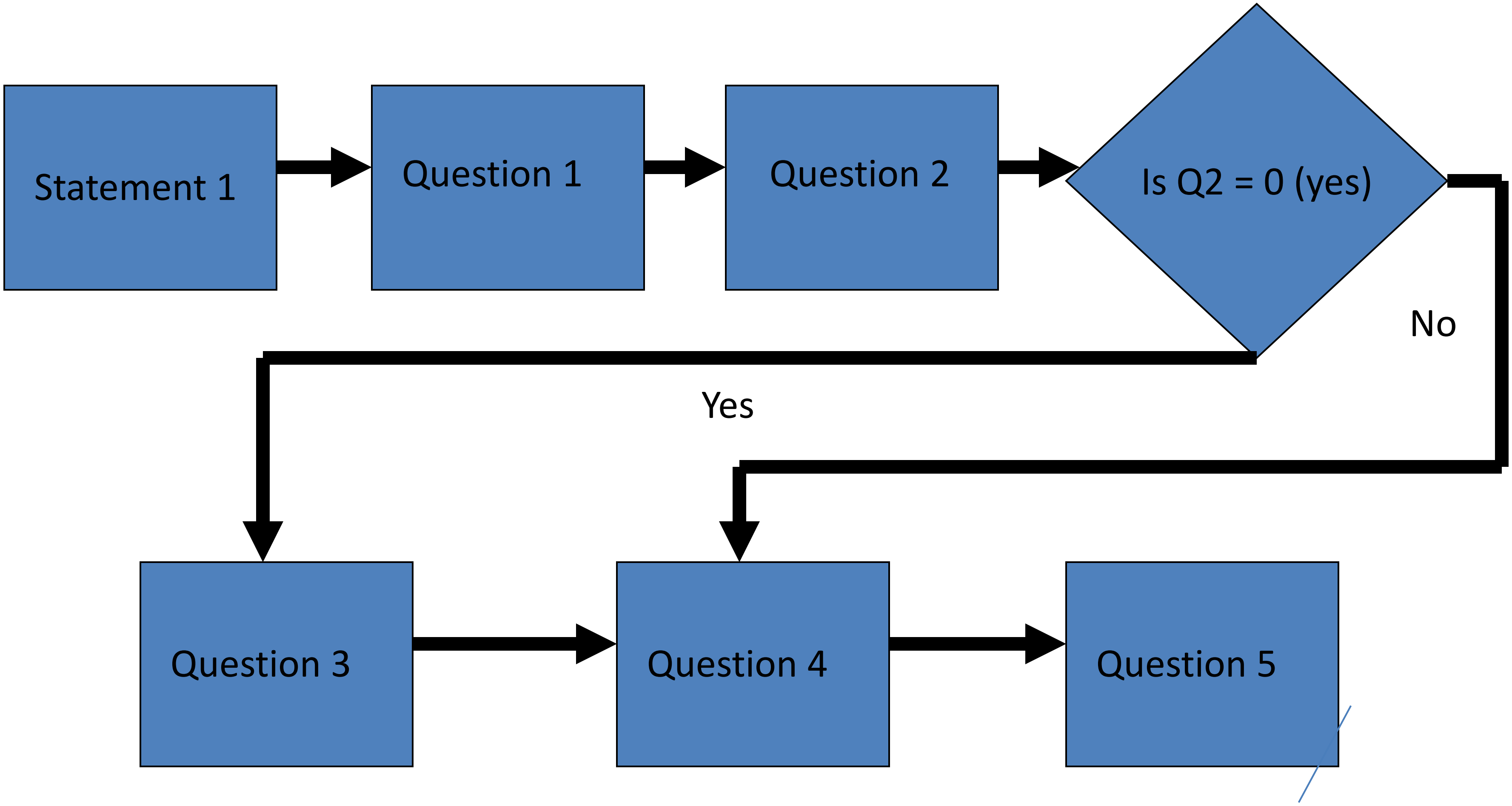
1. Sex
   (1) Male
   (2) Female
2. Are you 18 years or older?
   (0) Yes
   (1) No (Go to Question 4)
3. How old are you? _______
4. Who do you live with?

   _____________________
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


Statement 1
Question 1
Question 2
Is Q2 = 0 (yes)
No
Yes
Question 3
Question 4
Question 5

# Approach to Survey Analysis

- Identify
  - Question Text
  - Statements
  - Instructions or informative materials
  - Response Domains (by type)
- Determine the universe structure and concepts
- Walk through the flow logic

# Approach to Survey Analysis

- Identify
  - Question Text
  - Statements
  - Instructions or informative materials
  - Response Domains (by type)
- Determine the universe structure and concepts
- Walk through the flow logic

# Approach to Survey Analysis

- Identify
  - Question Text
  - Statements
  - Instructions or informative materials
  - Response Domains (by type)
- Determine the universe structure and concepts
- Walk through the flow logic

# Completing Question Items

- Create CodeSchemes reusing common categories
- Determine range for NumericDomains
- Determine maximum length of TextDomains
- Write up control constructs (easiest is to list all QuestionConstruct, all Statement Items)

# Example: Reusing Categories

Full list of all categories:

- Yes
- No
- Don’t know
- Yes
- No
- Yes
- No
- Yes, always
- Sometimes
- Some do, some don’t
- Not to my knowledge
- Never – I don’t let them
- Never – I don’t have a television
- Yes
- No
- Not to my knowledge

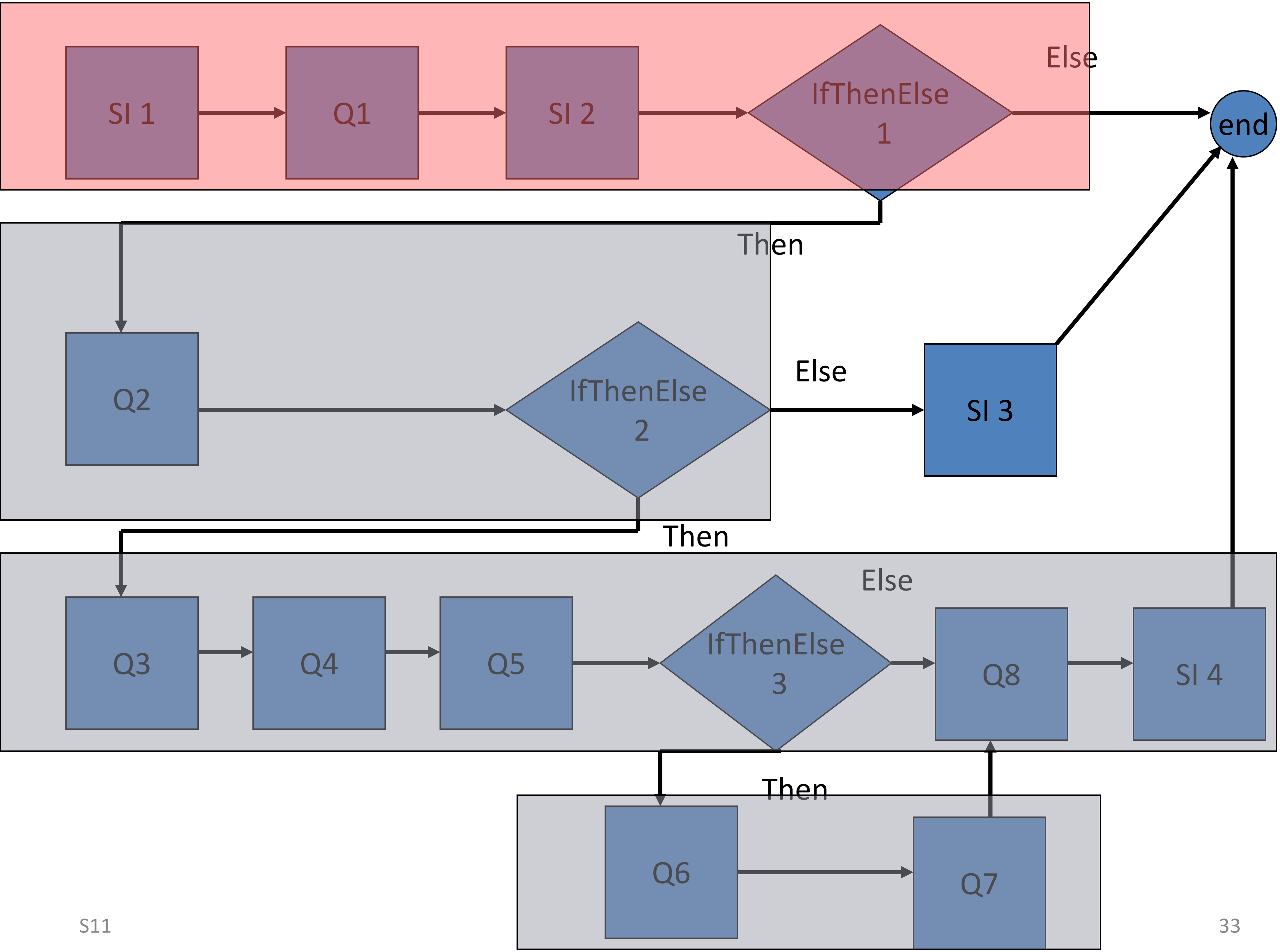
**BECOMES**

Shorter list of reusable categories:

- Yes
- No
- Don’t know
- Yes, always
- Sometimes
- Some do, some don’t
- Not to my knowledge
- Never – I don’t let them
- Never – I don’t have a television

# Flow Logic

- Master Sequence
  - Every instrument has one top-level sequence
- Question and statement order
- Routing – IfThenElse (see next slide)
  - After Statement 2 (all respondents read this)
  - After Q2 Else goes to statement
  - After Q5 Else goes back to a sequence


SI 1
Q1
SI 2
IfThenElse 1
Else
end
Then
Q2
IfThenElse 2
Else
SI 3
Then
Q3
Q4
Q5
IfThenElse 3
Else
Q8
SI 4
Then
Q6
Q7

# Example: Master Sequence

- Statement 1
- Question 1
- Statement 2
- IFThenElse 1
    - Then Sequence 1
        - Question 2
        - IFThenElse 2
            - Then SEQuence 2

                Question 3, Question 4, IFThenElse 3, Question 8, Statement 4<br>
                [Then SEQuence 3 (Question 6,Question 7)]
            - Else Statement 3

# Documentation of Data Processing

- There are two major places where processing is described in DDI 3, termed “Coding instructions”
  - General instructions
  - Generation instructions
  - Standard weight
- In DDI 3, the term “Coding” refers to the programming used to process data
  - Do not confuse it with “Code” or “CodeValue” (3 different things)
- These are used to describe the processing of both survey data and administrative data
  - Can also be used to describe harmonization, aggregation, and other processes later in the life cycle
- For collecting administrative data, the administrative system is described as a DataSource element in CollectionEvent and referenced by Generation Instruction

# Coding Instructions

- General Instruction
  - Handling non-response
  - Imputation or suppression rules
- Generation Instructions
  - Recoding
  - Derivation
  - Data from external sources
- Both human readable and machine-actionable instructions allowed
- If I have a questionnaire and coding instructions I can create the logical product, physical data product, and physical instance automatically with all relational links intact

# General Instruction

- Identification goes on the parent Coding element
- Description – human readable
- Command
  - Command Text – human readable
  - Command File (reference)
  - Structured Command
- IsOveride – reference to Instruction it overrides
- @isOveride (Boolean default=“false”)

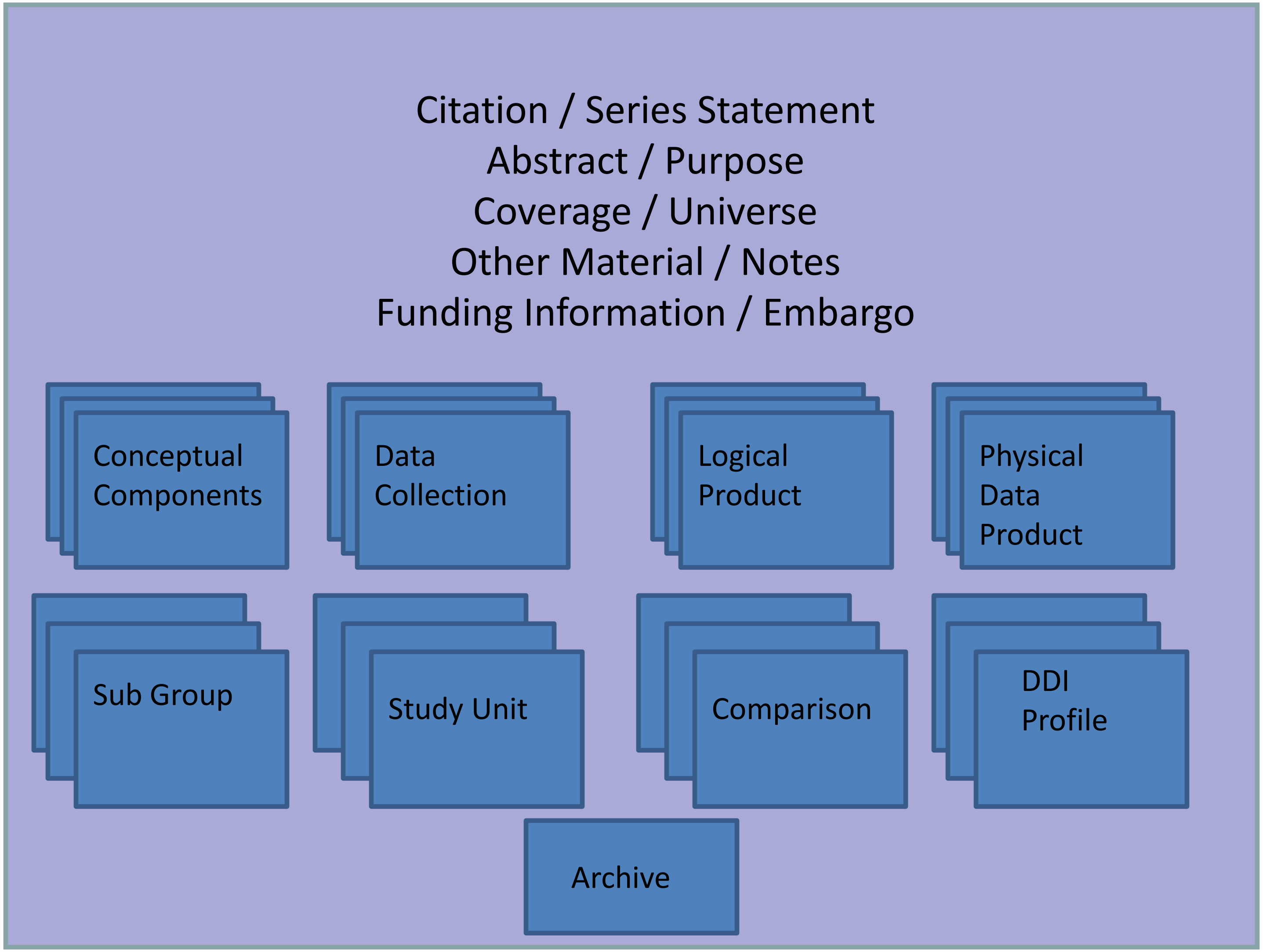
# Generation Instruction

- Source Question and Source Variable
  - can assign a mnemonic that is used in the command for this question or variable
- External Information (reference)
- Description – human readable
- Command
- Control Construct Reference
- Aggregation
  - for NCubes indicating independent and dependent components
- @isDerived

# 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

# Group



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

# 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

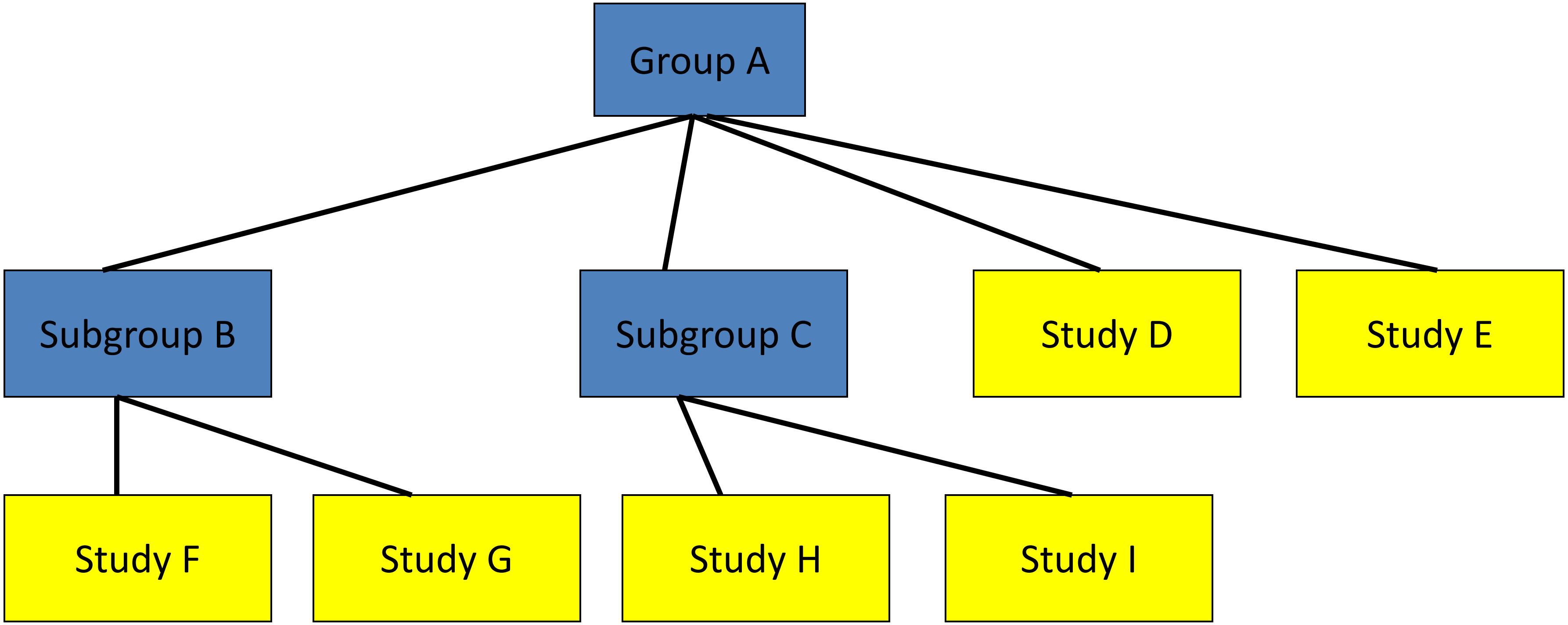
# Group:
# Grouping and Inheritance

- Grouping is the feature which allows DDI 3 to package groups of studies into a single XML instance, and express relationships between them
- To save repetition – and promote re-use – there is an inheritance mechanism, which allows metadata to be automatically shared by studies
- This can be a complicated topic, but it is the basis for many of DDI 3's features, including comparison of studies
- There is a switch which can be used to "turn off" inheritance

# Group Contents

- A group can contain study units, subgroups, and resource packages:
  - Study units document individual studies
  - Subgroups (inline or by reference)
  - Any of the content modules (Logical Product, Data Collection, etc.)
- Groups can nest indefinitely
- They have a set of attributes which explain the purpose of the group (as well as having a human-readable description):
  - Grouping by Time
  - Grouping by Instrument
  - Grouping by Panel
  - Grouping by Geography
  - Grouping by Data Set
  - Grouping by Language
  - Grouping by User-Defined Factor

# Inheritance



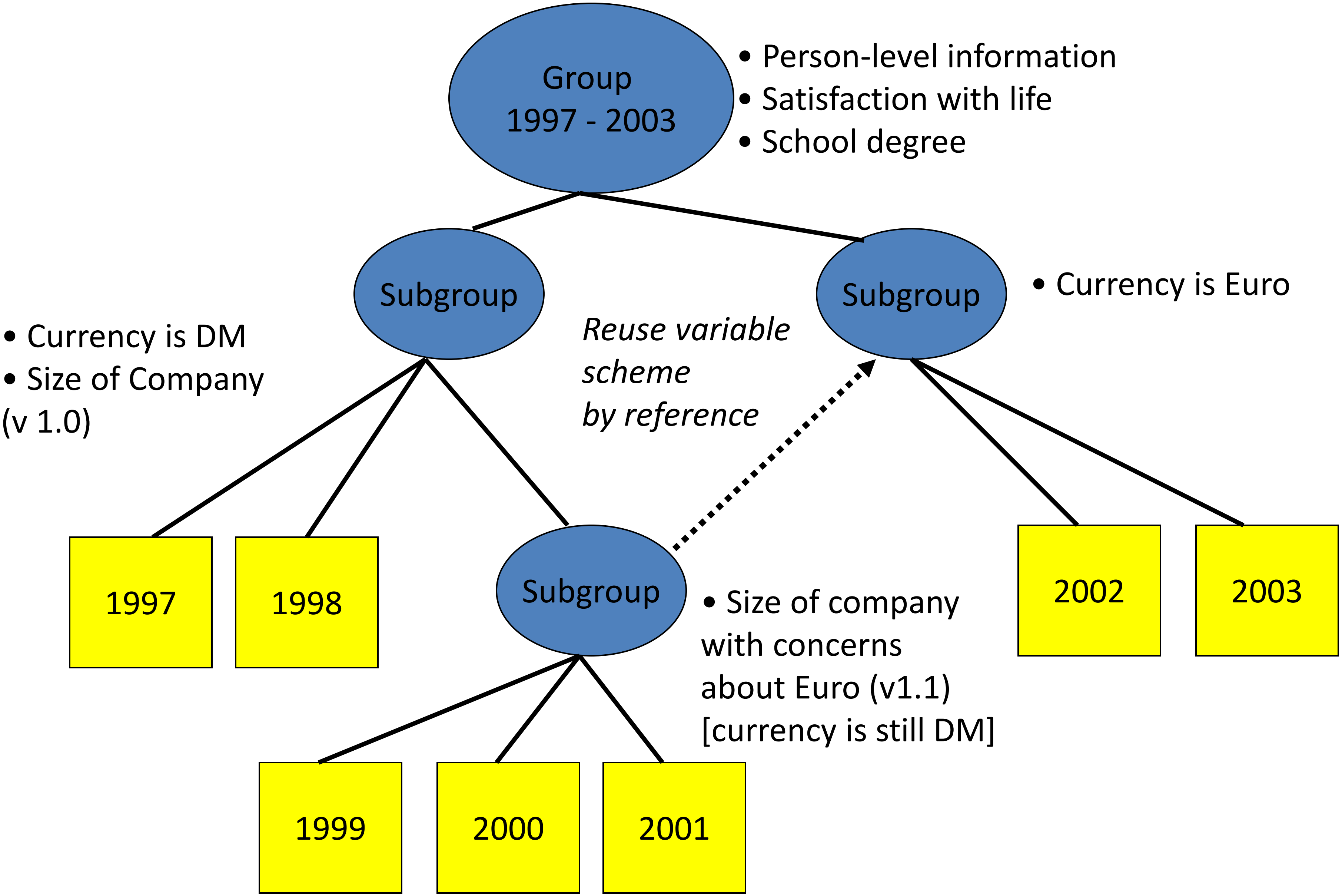
- Modules can be attached at any level
- They are shared – without repetition – by all child study units and subgroups
- If Group A has declared a concept called “X”, it is available to Study Units D – I.
- If Subgroup C has declared a Variable “Gender”, it is available to Study Units H and I without reference or repetition
- Inherited metadata can be changed using local overrides which add, update, or delete inherited properties

# Actions in Identifiers

- In some places – especially in groups where lots of metadata is being inherited – you can Add, Update, and Delete items using identifiers.
  - Using @action attribute = Add/Delete/Update
  - Repeat the identifier of the inherited object being locally modified
- This allows for local re-definition that is *not* reflected in a new version of the scheme
  - It cannot be reused
- For re-use, schemes should be versioned!

# German Social Economic Panel (SOEP) Study Example

- The following slides show how different types of metadata can be shared using grouping and inheritance
- The  SOEP is a panel study, with different panels on different years
  - Variables change over time
  - New questions and data are added



Group
1997 - 2003

- Person-level information
- Satisfaction with life
- School degree

Subgroup

- Currency is DM
- Size of Company
(v 1.0)

*Reuse variable*
*scheme*
*by reference*

Subgroup

- Currency is Euro

1997

1998

Subgroup

- Size of company
with concerns
about Euro (v1.1)
[currency is still DM]

2002

2003

1999

2000

2001

# Comparison

# Comparison

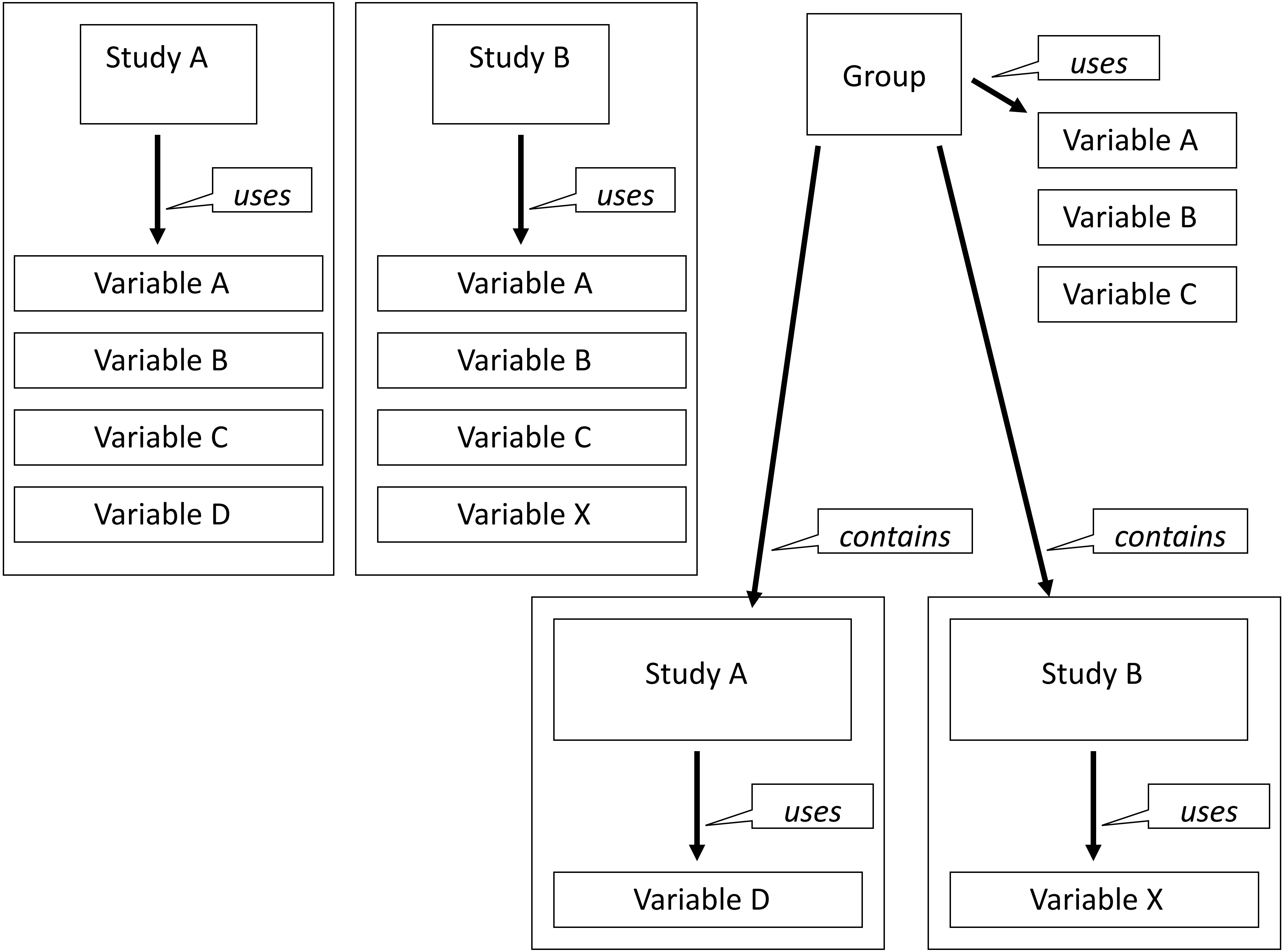
- There are two types of comparison in DDI 3:
  - Comparison by design
  - Ad-hoc (after-the-fact) comparison
- Comparison by design can be expressed using the grouping and inheritance mechanism
- Ad-hoc comparison can be described using the comparison module
- The comparison module is also useful for describing harmonization when performing case selection activities

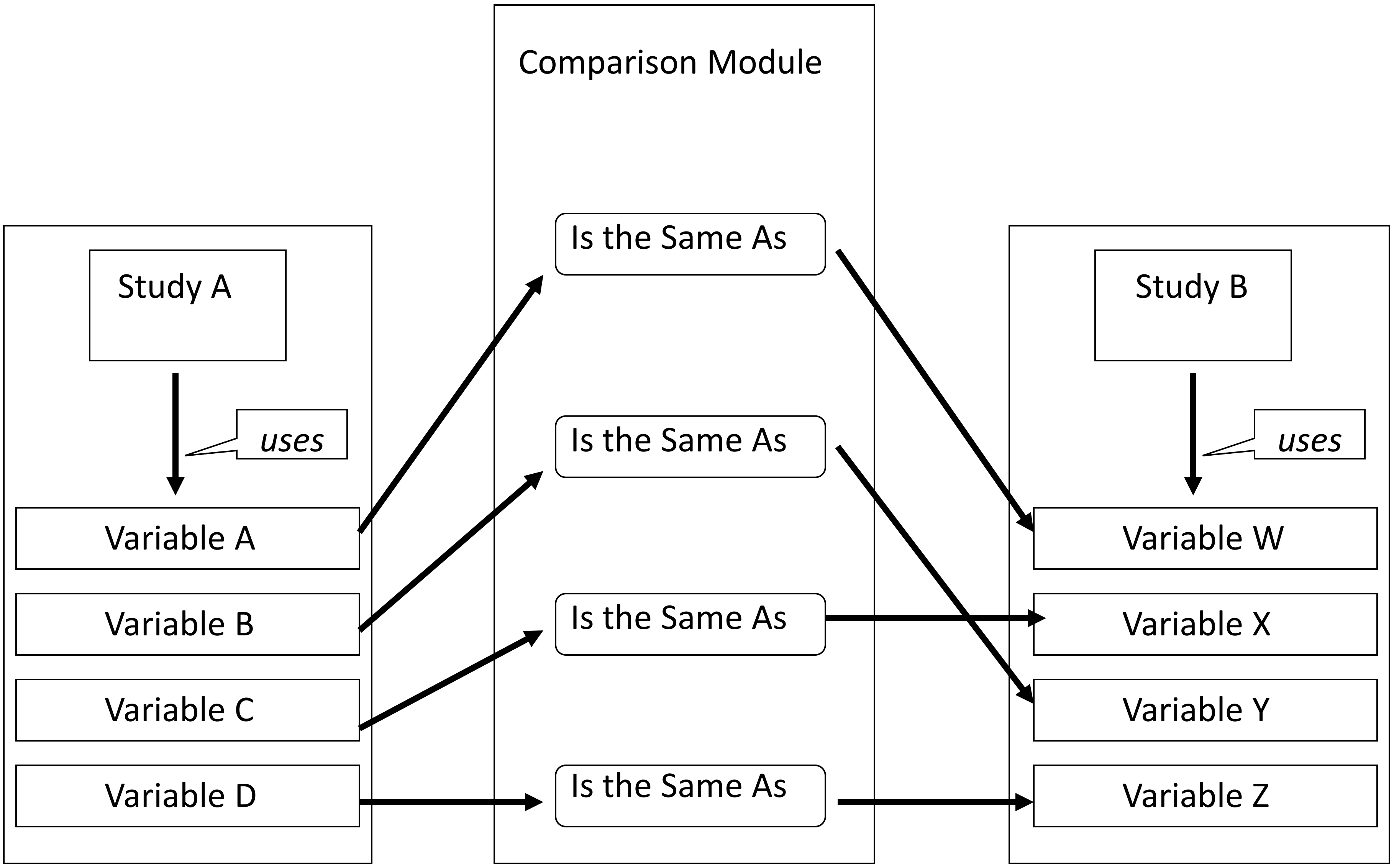
# Data Comparison

- 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)


Study A
uses
Variable A
Variable B
Variable C
Variable D
Study B
uses
Variable A
Variable B
Variable C
Variable X
Group
uses
Variable A
Variable B
Variable C
contains
contains
Study A
uses
Variable D
Study B
uses
Variable X


Comparison Module
Study A
uses
Variable A
Variable B
Variable C
Variable D
Is the Same As
Is the Same As
Is the Same As
Is the Same As
Study B
uses
Variable W
Variable X
Variable Y
Variable Z

# Comparison Content

- A comparison element is placed on a group or subgroup
- It contains:
  - Description of the comparison
  - Concept maps
  - Variable maps
  - Question maps
  - Category maps
  - Code maps
  - Universe maps
  - Notes
- Each map provides for a description of how two compared items correlate and/or differ, and also allows for a coding to be associated with the correlation

# Ad Hoc Groups

- Creating a course specific group
  - 3 files on aging
  - Create the group and declare the reason for selecting and including these studies
  - Note common or comparable concepts OR clarify why they are similar but NOT the same
  - Map any needed recodes for comparability
  - Provide the links (for example geographic)

# Equivalencies

- FIPS
  - 01 Alabama
  - 02 Alaska
  - 04 Arkansas
  - 06 California
  - 08 Colorado
  - 09 Connecticut
  - 10 Delaware
  - 11 District of Columbia
  - 12 Florida

=

- CENSUS
  - 63 Alabama
  - 94 Alaska
  - 86 Arkansas
  - 71 California
  - 84 Colorado
  - 16 Connecticut
  - 51 Delaware
  - 53 District of Columbia
  - 59 Florida

# Providing Comparative Information

- Create the category and coding schemes
- Use the comparison maps to provide comparability
  - Codes, Categories, Variables, Concepts Questions, Universe
- Example:
  - 6 files using 3 different age variables
  - Single year, five year, and ten year cohorts
- Map each equivalent structure to a single example
- Map the single year to the five year
- Map the five year to the ten year
- Provide the software command to do the conversion

## SINGLE YEARS

< 1 year

1 year

2 years

3 years

4 years

5 years

6 years

7 years

8 years

9 years

10 years

11 years

12 years

13 years

14 years

15 years

16 years

17 years

18 years

19 years

20 years

Etc.

## 5 YEAR COHORTS

< 5 years

5 to 9 years

10 to 14 years

15 to 19 years

20 years plus

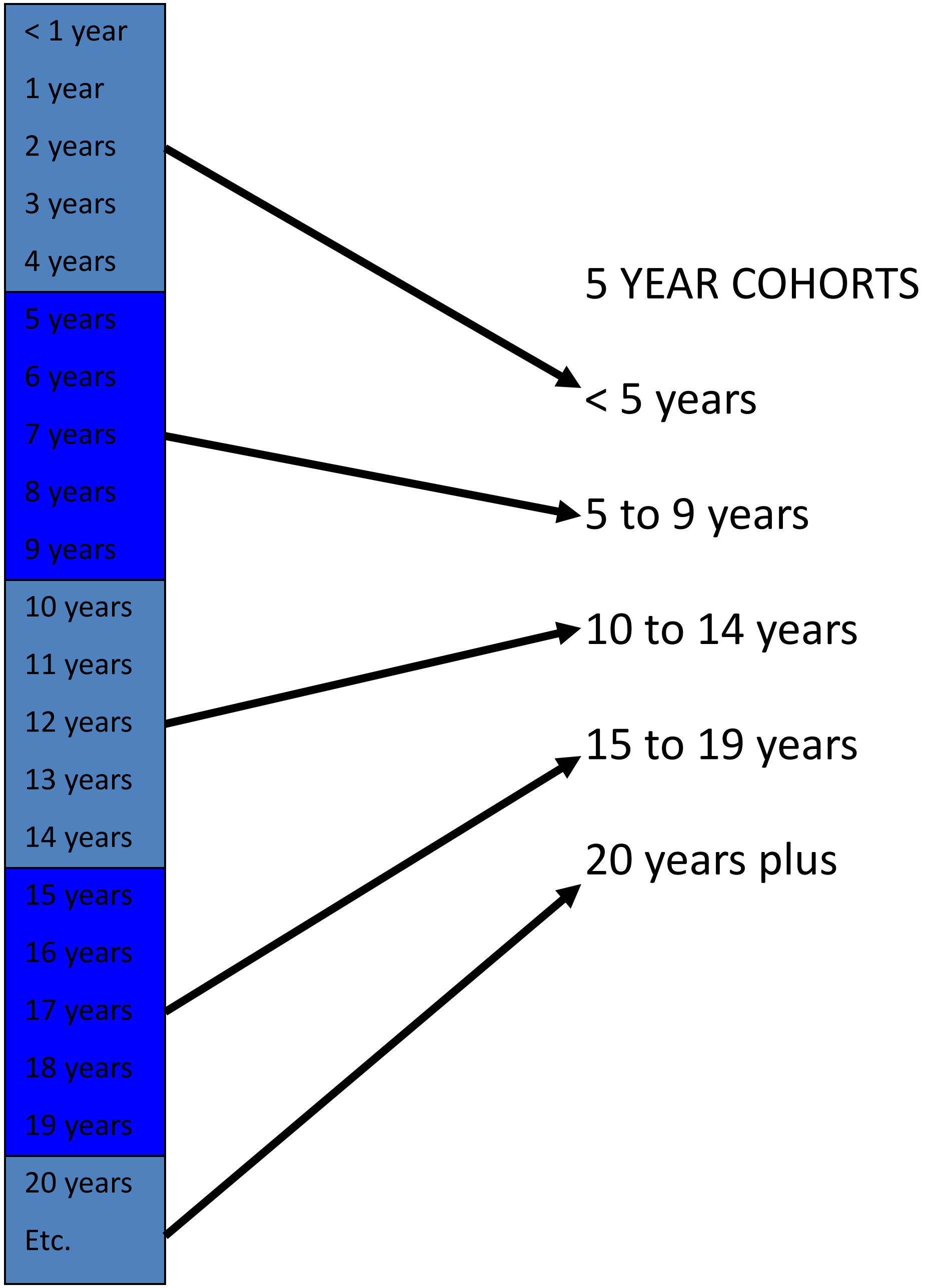
## 10 YEAR COHORTS

< 10 years

10 to 19 years

20 years plus

SINGLE YEARS



< 1 year
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10 years
11 years
12 years
13 years
14 years
15 years
16 years
17 years
18 years
19 years
20 years
Etc.

5 YEAR COHORTS

< 5 years

5 to 9 years

10 to 14 years

15 to 19 years

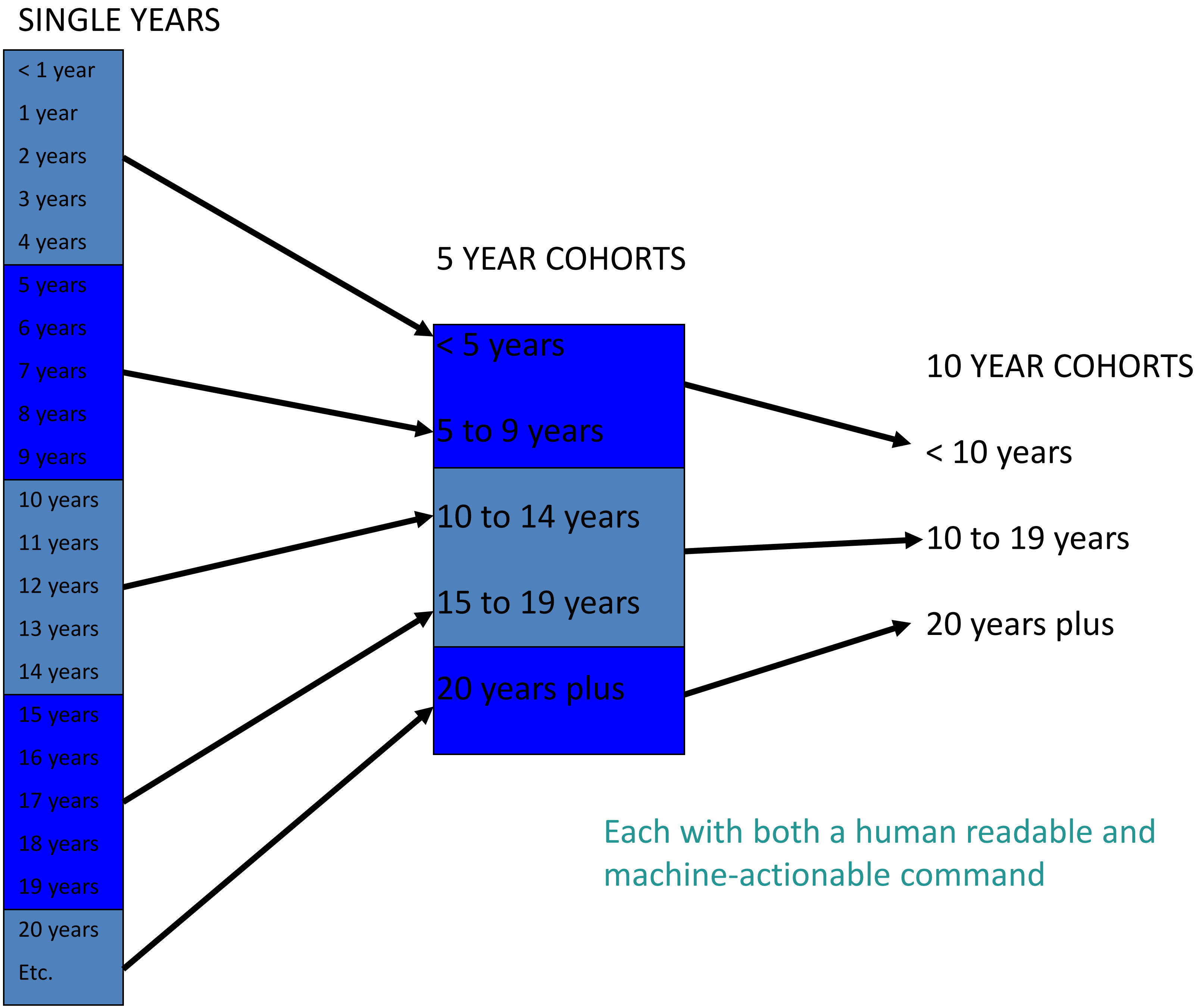
20 years plus

10 YEAR COHORTS

< 10 years

10 to 19 years

20 years plus



SINGLE YEARS

< 1 year
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10 years
11 years
12 years
13 years
14 years
15 years
16 years
17 years
18 years
19 years
20 years
Etc.

5 YEAR COHORTS

< 5 years
5 to 9 years
10 to 14 years
15 to 19 years
20 years plus

10 YEAR COHORTS

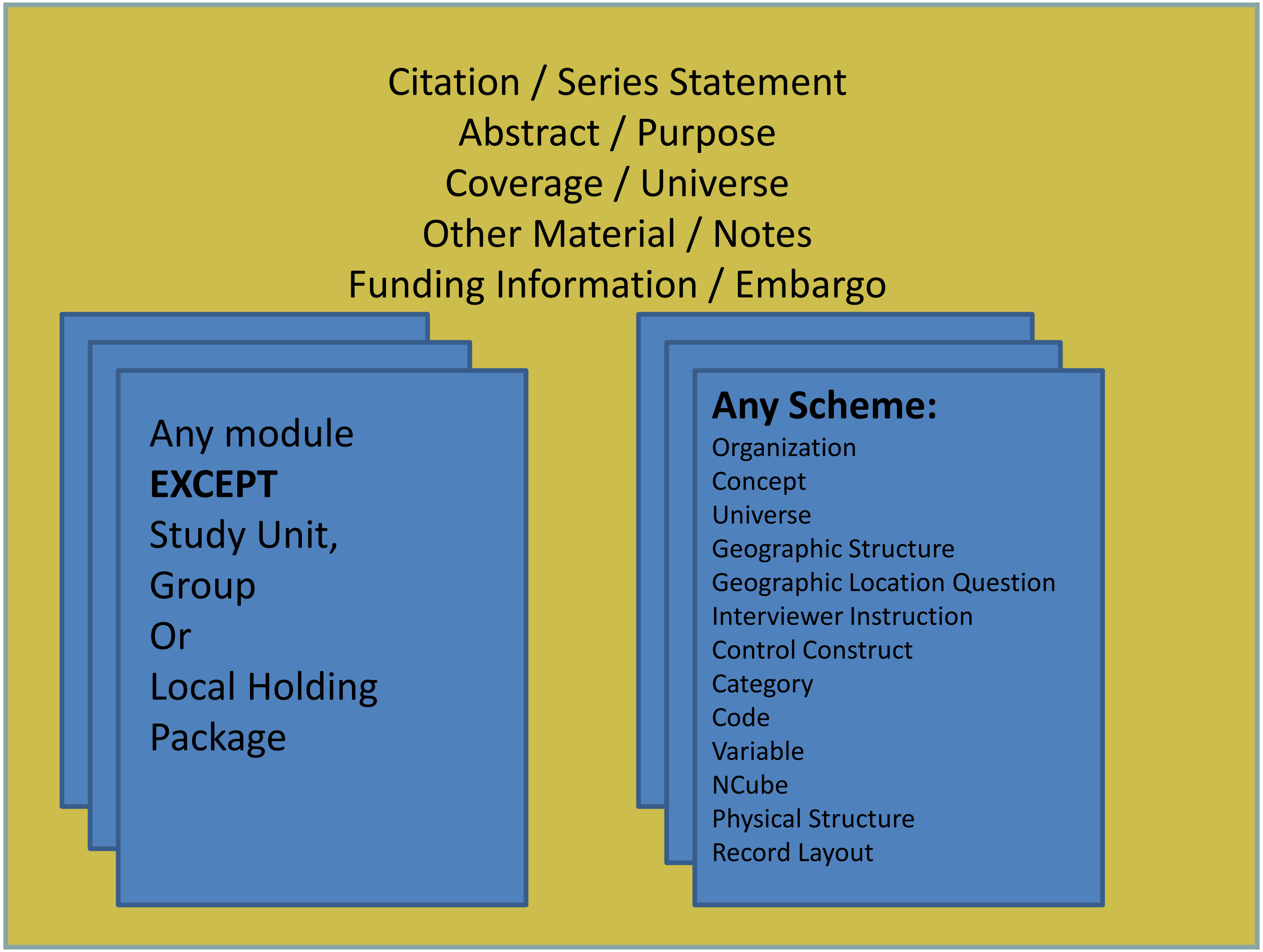
< 10 years
10 to 19 years
20 years plus

Each with both a human readable and machine-actionable command

# 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

# Resource Package



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

Any module
**EXCEPT**
Study Unit,
Group
Or
Local Holding
Package

**Any Scheme:**

Organization
Concept
Universe
Geographic Structure
Geographic Location Question
Interviewer Instruction
Control Construct
Category
Code
Variable
NCube
Physical Structure
Record Layout

# 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

# 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

# Concepts

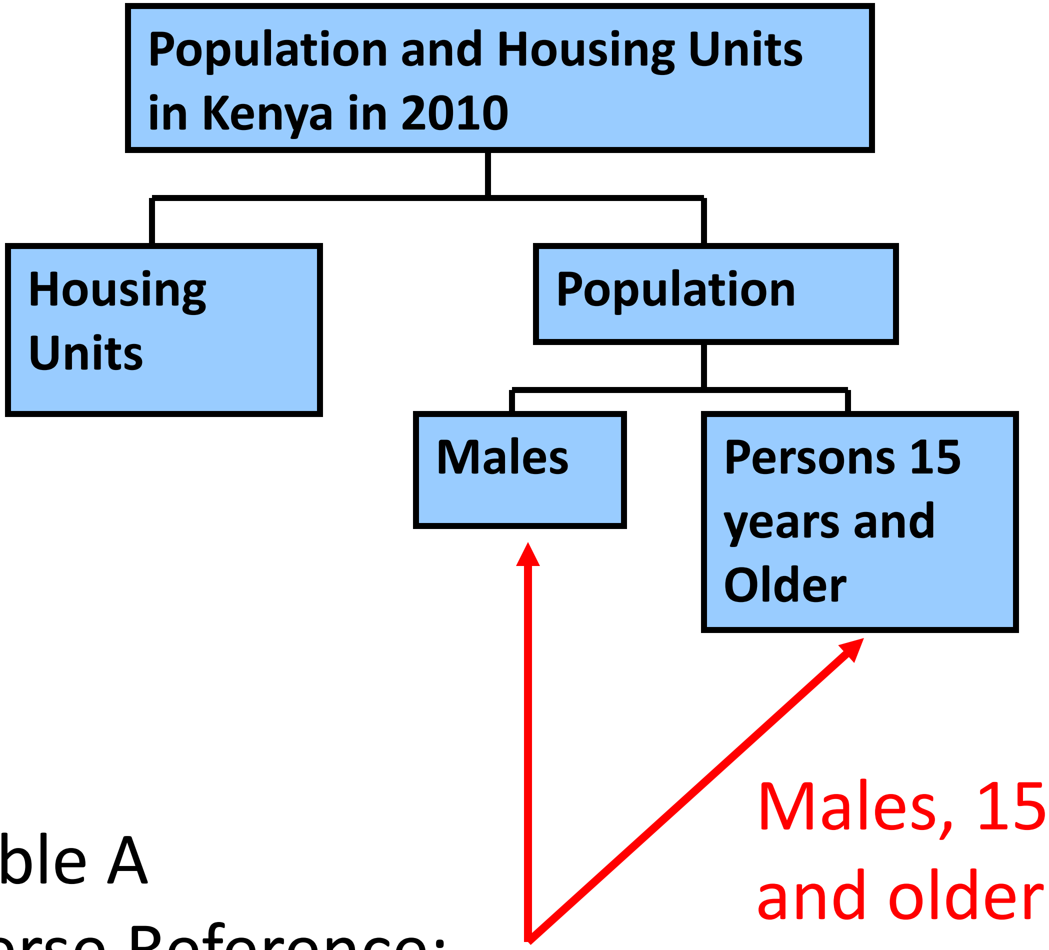
- A concept may be structured or unstructured and consists of a Name, a Label, and a Description. A description is needed if you want to support comparison. Concepts are what questions and variables are designed to measure and are normally assigned by the study (organization or investigator).

# Universe

- This is the universe of the study which can combines the who, what, when, and where of the data
- Census top level universe: “The population and households within Kenya in 2010”
- Sub-universes: Households, Population, Males, Population between 15 and 64 years of age, …

# Universe Structure

- Hierarchical
  - Makes clear that “Owner Occupied Housing Units” are part of the broader universe “Housing Units”
  - Can be generated from the flow logic of a questionnaire
- Referenced by variables and question constructs
  - Provides implicit comparability when 2 items reference the same universe



Population and Housing Units in Kenya in 2010

Housing Units

Population

Males

Persons 15 years and Older

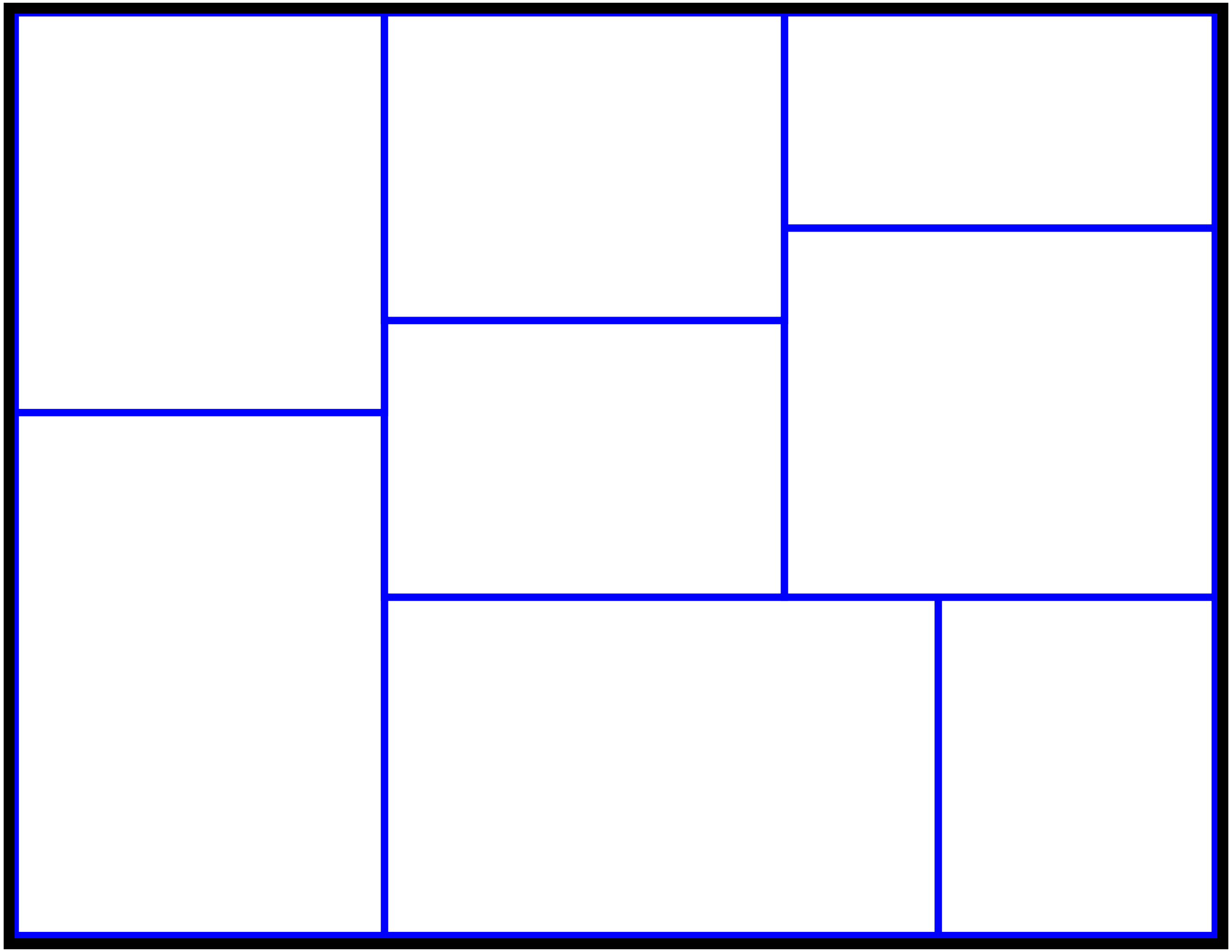
Variable A
Universe Reference:

Males, 15 years of age and older in Kenya in 2010

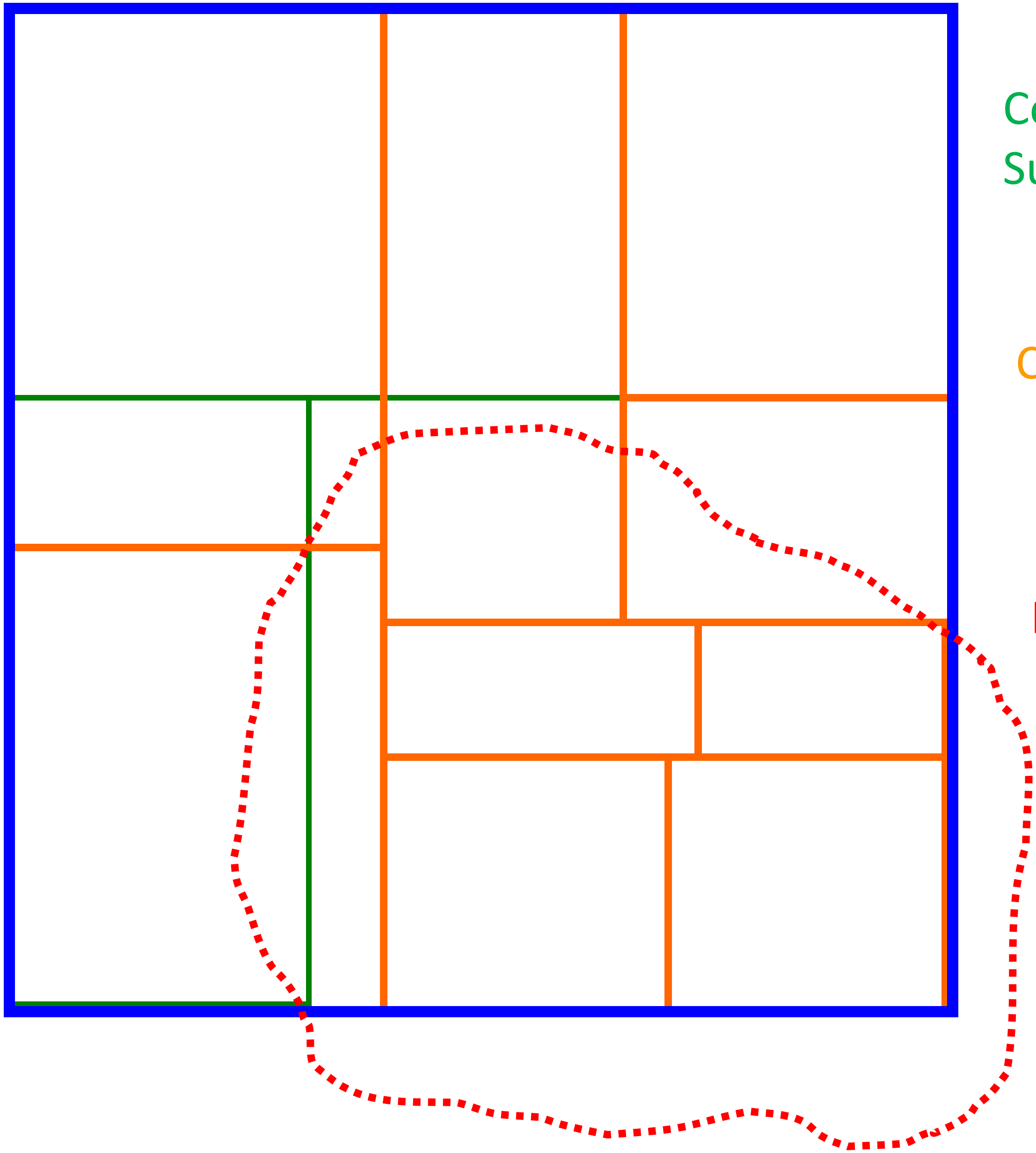
# Geographic Structure

- Level
  - Code, Name, coverage limitation, description
- Parent
  - Reference to a single parent geography
  - This is used to describe single hierarchies
- OR Geographic Layer
  - References multiple base levels where multiple hierarchies are layered to create a resulting polygon

STATE



# COUNTY



County Subdivision

Census Tract

Place

# Hierarchies and Layers

- State (040)
  - County (050)
    - County Subdivision (060)
    - Census Tract (140)
  - Place (160)

- Portion of a Census Tract within a County Subdivision within a Place
- Layer References:
  - 140
  - 060
  - 160

# Geographic Location

- Level description and/or a reference to the level description in the Geographic Structure
- Reference to the variable containing the identifier of the geographic location
- Description of a specific geographic location:
  - Code
  - Name
  - Geographic time
  - Bounding Polygon
  - Excluding Polygon

# Structure and Location

STRUCTURE:

- Level: 040
- Name: State
- U.S. State or state equivalent including Legal Territories and the District of Columbia
- Parent: 010 [country]

LOCATION:

- Level Reference: 040
- Variable Reference: STATEFP
- Name: Minnesota
- Code Value: 27
- Geographic Time: Start: 1857 End: 9999
- Bounding Polygon or Shape File Reference: for each boundary over time

# DDI Resources

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