colectica

<ddi>

Introduction to DDI and Colectica

Colectica Workshop at EDDI 2014 - London

Agenda: Day 1

Morning

9:00	Introduction to Metadata, DDI and Colectica
10:00	The DDI Information Model
10:30	Break
11:00	In-depth with DDI: Surveys
11:30	In-depth with DDI: Data
12:00	In-depth with DDI: Study Lifecycle
12:30	Lunch

Afternoon

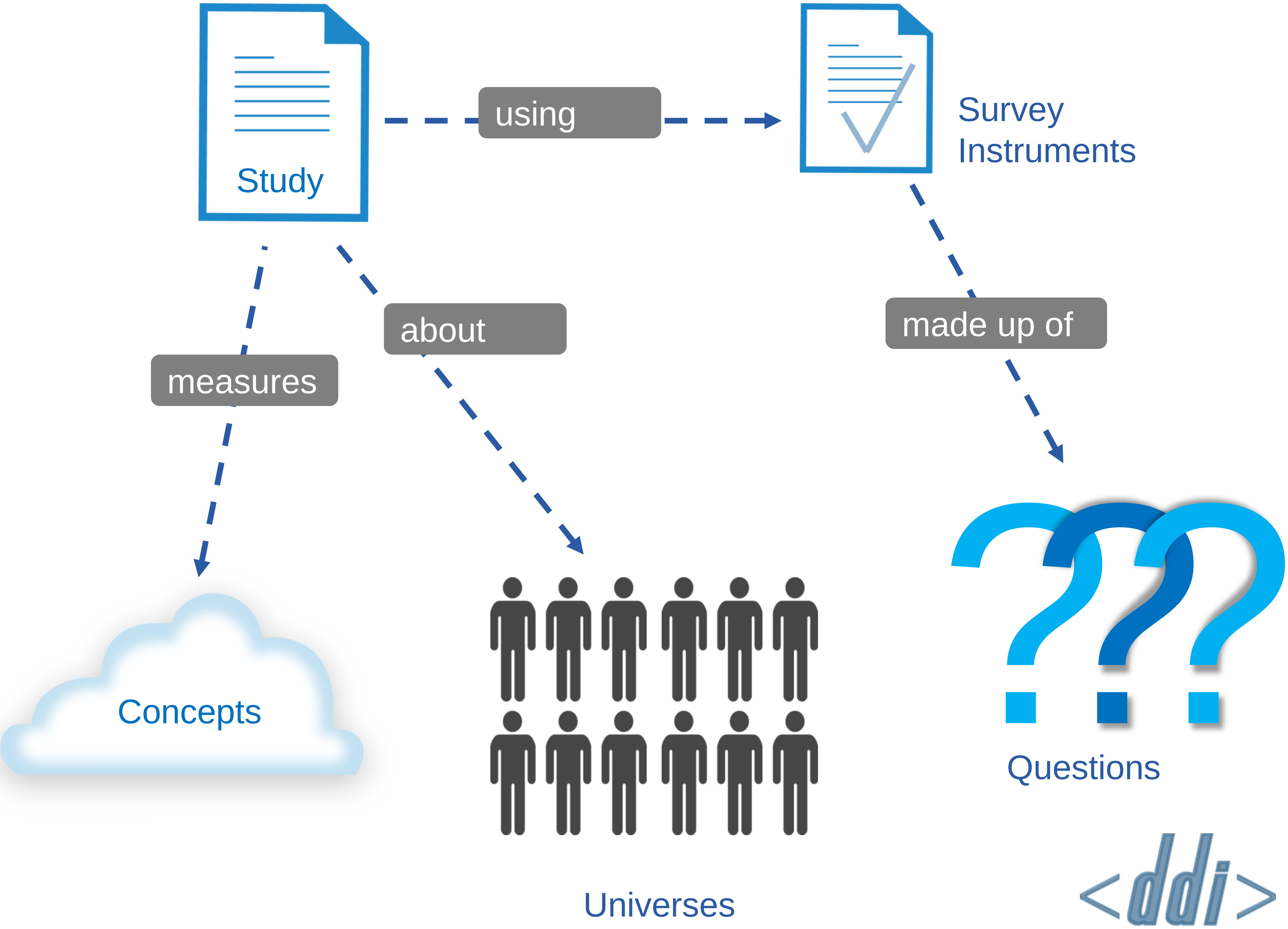
12:30	Lunch
13:30	Metadata publication and versioning
14:15	Processing structured metadata
15:00	Break
15:30	In-depth metadata scenarios and Q & A
16:45	Recap
17:00	The End
18:00	Informal Get-Together Princess Louise, 208 High Holborn, London WC1V 7EP

Goals

- What information can I document with DDI?
- How do I make DDI XML?
- What can I do with DDI?

Introduction to DDI

<ddi>



Study

using

Survey Instruments

measures

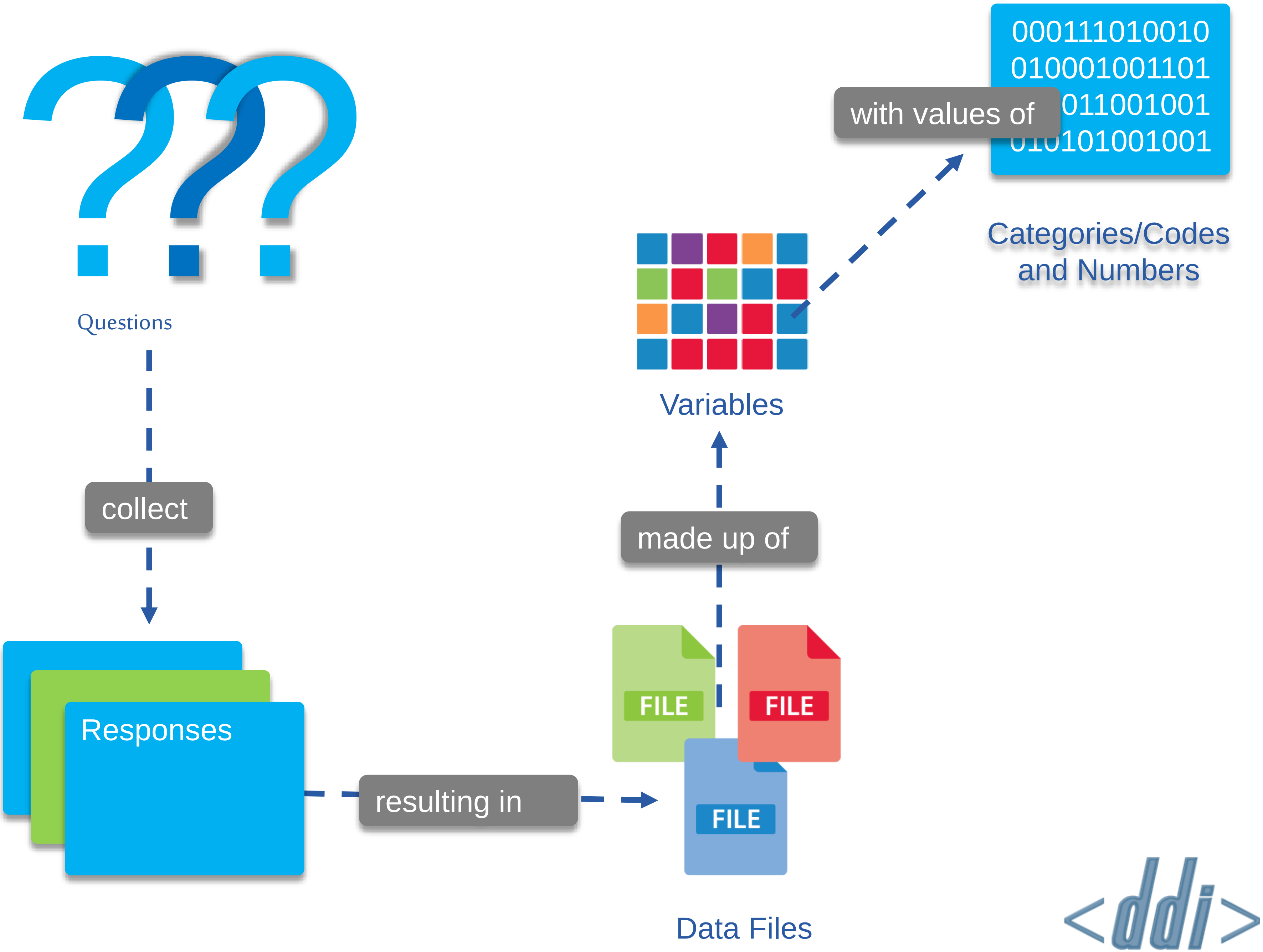
about

made up of

Concepts

Universes

Questions



Questions

collect

Responses

resulting in

Data Files

made up of

Variables

with values of

000111010010
010001001101
011001001
010101001001

Categories/Codes and Numbers

DDI provides a vocabulary for this

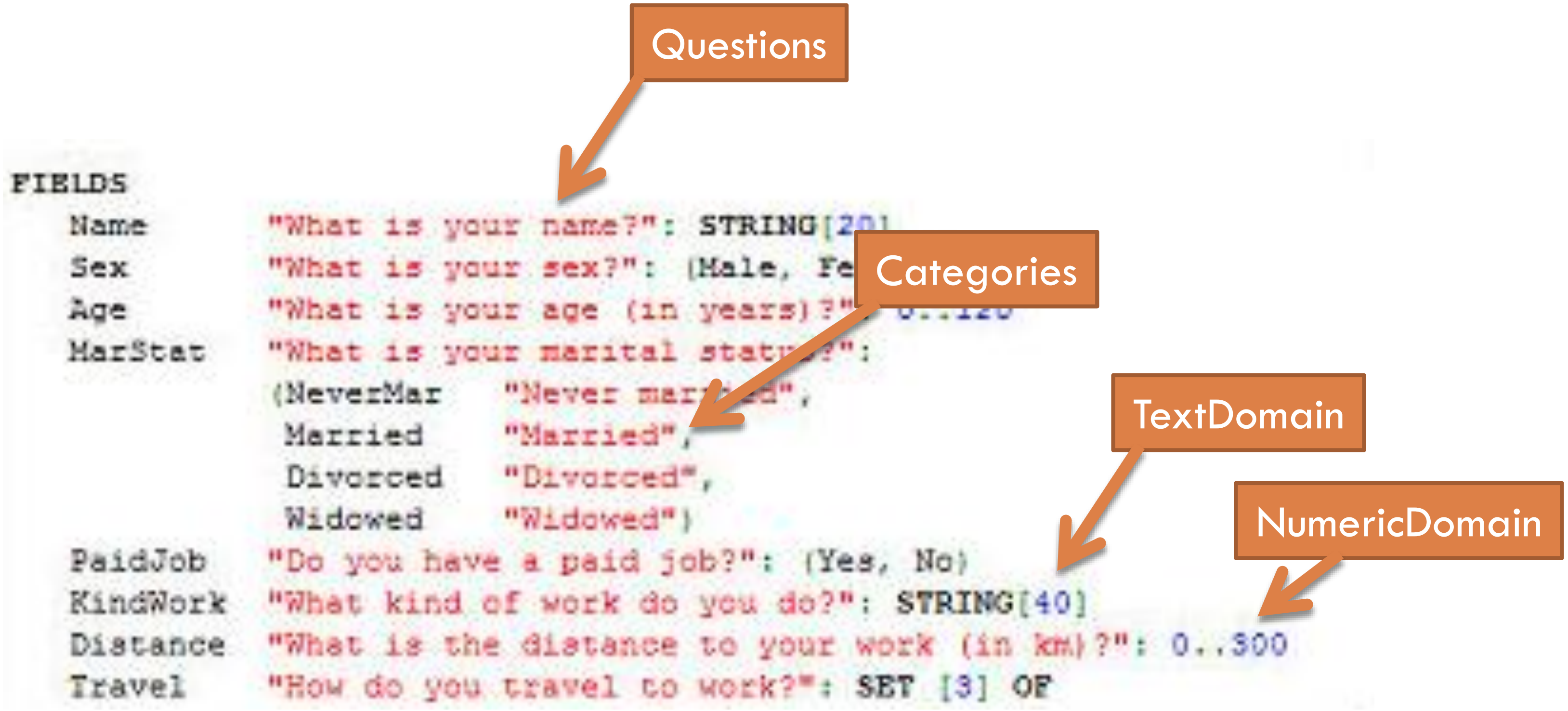
Main Content in DDI

- Study Lifecycle
- Surveys
- Data

Data Documentation Initiative

- Open standard for describing data
  - Focus on social, behavioral, and economic sciences
  - XML
- Users
  - National Statistical Institutes
  - University Research Groups
  - Data Archives
  - Other Data Producers and Publishers
- Since 1995

Real-world Examples



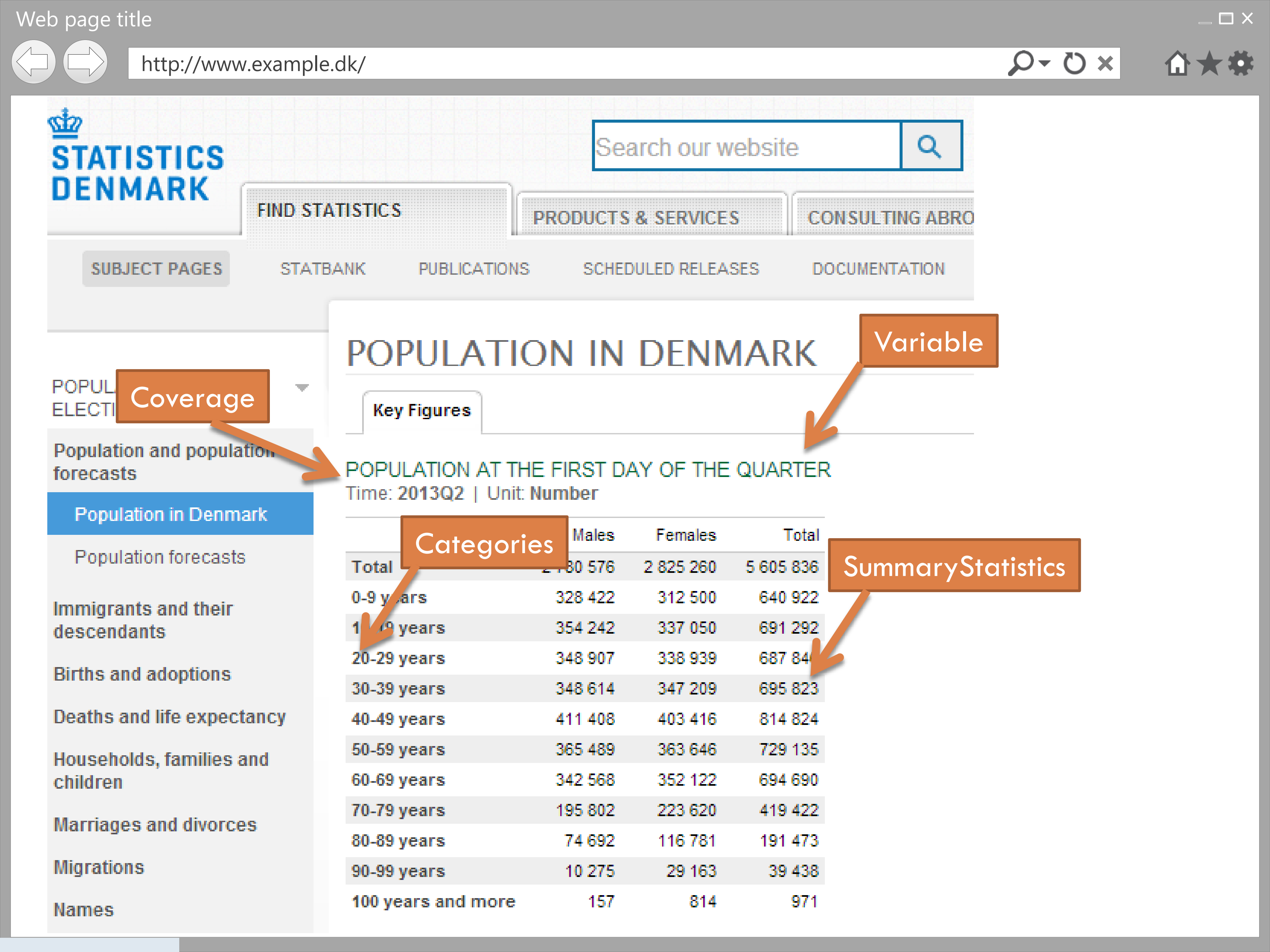
Questions

Categories

TextDomain

NumericDomain

```
FIELDS
   Name       "What is your name?": STRING[20]
   Sex        "What is your sex?": (Male, Fe
   Age        "What is your age (in years)?": 0..120
   MarStat    "What is your marital status?":
              (NeverMar    "Never married",
               Married     "Married",
               Divorced    "Divorced",
               Widowed     "Widowed")
   PaidJob    "Do you have a paid job?": (Yes, No)
   KindWork   "What kind of work do you do?": STRING[40]
   Distance   "What is the distance to your work (in km)?": 0..300
   Travel     "How do you travel to work?": SET [3] OF
```



Coverage

Variable

Categories

SummaryStatistics

POPULATION IN DENMARK

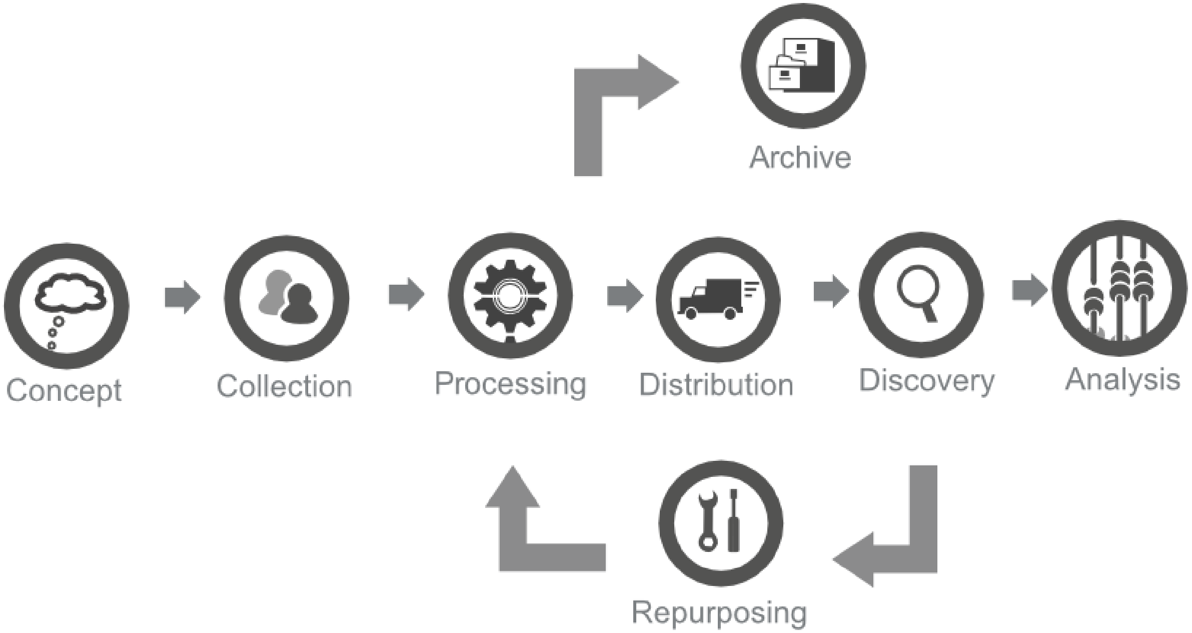
POPULATION AT THE FIRST DAY OF THE QUARTER

Time: 2013Q2 | Unit: Number

	Males	Females	Total
Total	2 780 576	2 825 260	5 605 836
0-9 years	328 422	312 500	640 922
10-19 years	354 242	337 050	691 292
20-29 years	348 907	338 939	687 846
30-39 years	348 614	347 209	695 823
40-49 years	411 408	403 416	814 824
50-59 years	365 489	363 646	729 135
60-69 years	342 568	352 122	694 690
70-79 years	195 802	223 620	419 422
80-89 years	74 692	116 781	191 473
90-99 years	10 275	29 163	39 438
100 years and more	157	814	971

Why DDI?

□ Document the full data lifecycle in a standard manner



Why DDI?

- Reusable metadata definitions
- No copy and paste
- Just point to an item

- Including metadata by reference helps avoid error and confusion
- Reuse is explicit

Metadata Banks

- DDI 3 supports the concept of metadata registries
  - Question banks
  - Variable banks
  - Code lists, concept definitions, or anything else

Metadata-driven Processes

- Generate documentation (Colectica, XSLT, more)
  - PDF
  - Web
- Populate survey systems (from Colectica)
  - Out of the box: Blaise, CASES, CSPro, RedCAP, queXML
  - Custom systems: possible with addins

Multilingual Support

- Most text fields can specify what language the content is in
- These fields can be repeated to represent multiple languages

Colectica Overview

Standards-based metadata management

- Survey design
- Data documentation
- Study lifecycle documentation

The Colectica Platform

Colectica Designer

- Create, ingest, manage, and edit metadata
- Usable DDI 3 for end-users

Colectica Repository

- Centralized, authoritative, metadata store built on DDI 3, ISO 11179, and Web Service standards

Colectica Portal

- Search and browse metadata from Colectica Repository

The Colectica Platform

Colectica SDK

- Allows programmers to work with DDI 3 and interact with Colectica Repository

Colectica Toolkit

- Command line utilities to perform specific tasks

How to create DDI metadata

- Manually enter information
- Import (Colectica or Stat/Transfer)
  - Excel
  - Delimited Files
  - SPSS
  - Stata
  - Blaise
  - CASES
  - RedCAP
  - queXML
  - IBM Data Collection / SPSS Dimensions
  - CSPRo
- Integrate custom data sources

DDI Information Model

DDI at a Glance

Finding the Details

DDI at a Glance

Study

- Group
- StudyUnit
- Quality

Survey

- DataCollection
- Instrument
- ControlConstructs
- Question

Data

- PhysicalInstance
- DataRelationship
- Variable

Foundational

- Concept
- Universe
- Organization
- CodeList
- CategoryList
- Category

Finding the Details

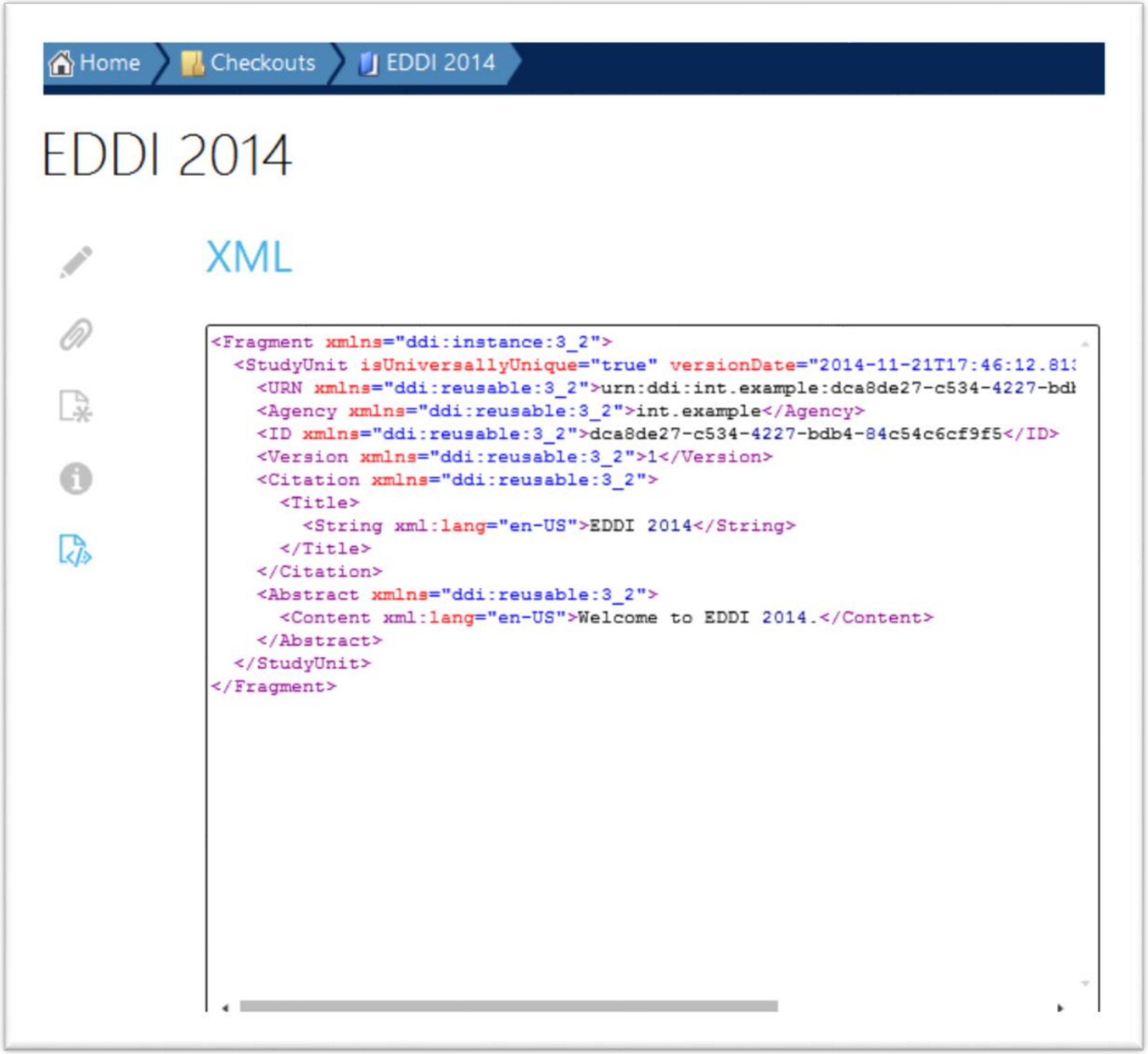
Finding the Details

- Sources
  - DDI Documentation
  - Colectica (or other tools)
  - XML Schemas

DDI Documentation

Available at http://www.ddialliance.org/

DDI in Colectica Reader and Designer



EDDI 2014

XML

```
<Fragment xmlns="ddi:instance:3_2">
  <StudyUnit isUniversallyUnique="true" versionDate="2014-11-21T17:46:12.81
    <URN xmlns="ddi:reusable:3_2">urn:ddi:int.example:dca8de27-c534-4227-bd
    <Agency xmlns="ddi:reusable:3_2">int.example</Agency>
    <ID xmlns="ddi:reusable:3_2">dca8de27-c534-4227-bdb4-84c54c6cf9f5</ID>
    <Version xmlns="ddi:reusable:3_2">1</Version>
    <Citation xmlns="ddi:reusable:3_2">
      <Title>
        <String xml:lang="en-US">EDDI 2014</String>
      </Title>
    </Citation>
    <Abstract xmlns="ddi:reusable:3_2">
      <Content xml:lang="en-US">Welcome to EDDI 2014.</Content>
    </Abstract>
  </StudyUnit>
</Fragment>
```

Common Patterns in DDI

- Identification
- Naming
- Packaging

Common Elements

- Identification
  - Agency ID
  - Identifier
  - Version number

Common Elements: Descriptive

- Either
  - Name
  - Label
  - Description
- Or
  - Citation/Title

DDI 3 Packaging

- **Fragment**
- OR: Modules + Schemes + Packages + Redundancy

Just Enough XML

Just Enough XML

- eXtensible Markup Language (XML)
- XML represents information
- XML is both human & machine readable

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>
```

XML-based Standards

- A standard can define a set of tags and rules for how to use them

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>
```

```xml
<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 Namespaces

Prefix	Namespace
[default] or ddi	ddi:instance:3_2
r	ddi:reusable:3_2
s	ddi:studyunit:3_2
g	ddi:group:3_2
c	ddi:conceptualcomponent:3_2
d	ddi:datacollection:3_2
l	ddi:logicalproduct:3_2
pi	ddi:physicalinstance:3_2

In-depth: Surveys

Study

- Group
- StudyUnit
- Quality

Survey

- DataCollection
- Instrument
- ControlConstructs
- Question

Data

- PhysicalInstance
- DataRelationship
- Variable

Foundational

- Concept
- Universe
- Organization
- CodeList
- CategoryList
- Category

Questions

- What is your name?
- How did you get here?

A Question in DDI

```
<d:QuestionItem>
      <r:Agency>example.org</r:Agency>
      <r:ID>q1</r:ID>
      <r:Version>1</r:Version>

    <d:QuestionItemName xml:lang="en">
        <r:String>name</r:String>
    </d:QuestionItemName>

    <d:QuestionText audienceLanguage="en">
        <d:LiteralText>
            <d:Text>What is your name?</d:Text>
        </d:LiteralText>
    </d:QuestionText>

    <d:TextDomain minLength="1"></d:TextDomain>
</d:QuestionItem>
```

A Question in DDI

```
<d:QuestionItem>
    <r:Agency>example.org</r:Agency>
    <r:ID>q1</r:ID>
    <r:Version>1</r:Version>

    <d:QuestionItemName xml:lang="en">
       <r:String>name</r:String>
    </d:QuestionItemName>

    <d:QuestionText audienceLanguage="en">
        <d:LiteralText>
            <d:Text>What is your name?</d:Text>
        </d:LiteralText>
    </d:QuestionText>

    <d:TextDomain minLength="1"></d:TextDomain>
</d:QuestionItem>
```

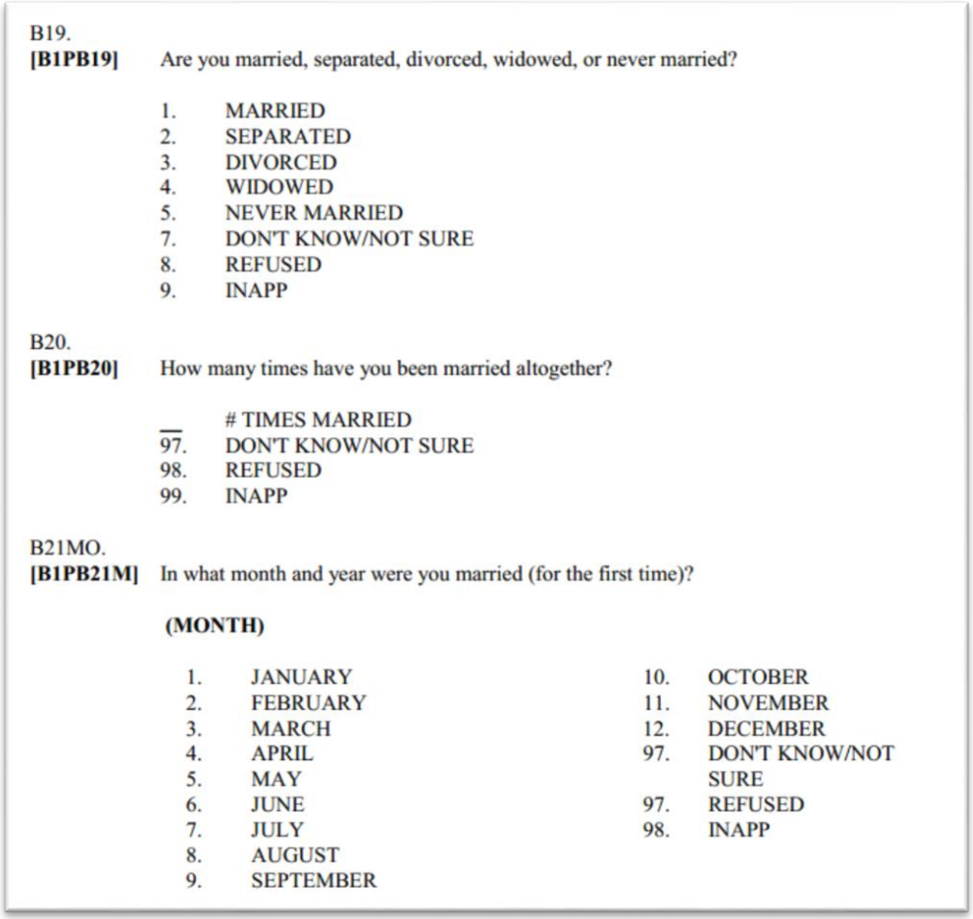
Other Response Types

- Text
- Numeric
- DateTime
- Category
- Code
- Geographic

Question

- Does your organization currently use DDI for any purpose?
  - Yes
  - No

Instrument



B19.
[B1PB19] Are you married, separated, divorced, widowed, or never married?

1. MARRIED
2. SEPARATED
3. DIVORCED
4. WIDOWED
5. NEVER MARRIED
7. DON'T KNOW/NOT SURE
8. REFUSED
9. INAPP

B20.
[B1PB20] How many times have you been married altogether?

__ # TIMES MARRIED
97. DON'T KNOW/NOT SURE
98. REFUSED
99. INAPP

B21MO.
[B1PB21M] In what month and year were you married (for the first time)?

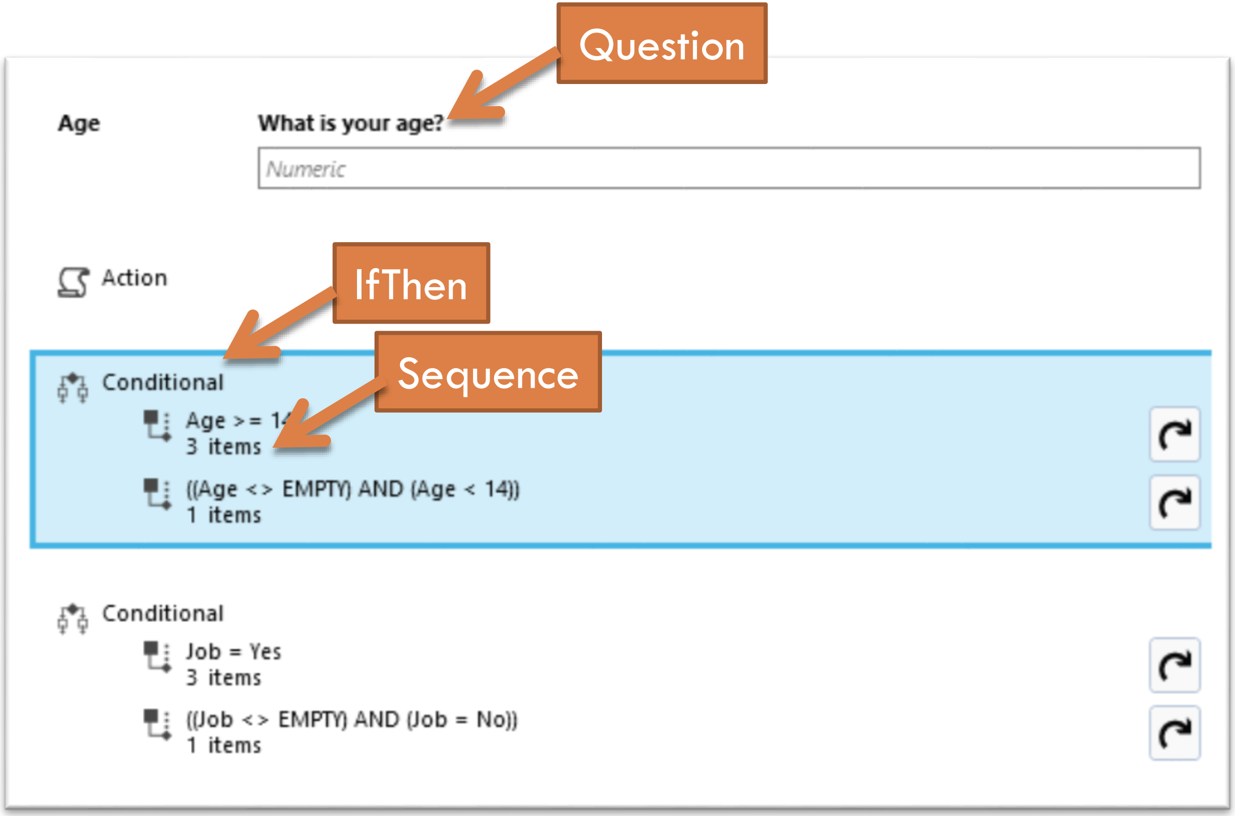
(MONTH)

1. JANUARY
2. FEBRUARY
3. MARCH
4. APRIL
5. MAY
6. JUNE
7. JULY
8. AUGUST
9. SEPTEMBER
10. OCTOBER
11. NOVEMBER
12. DECEMBER
97. DON'T KNOW/NOT SURE
97. REFUSED
98. INAPP

Instrument

- Let’s make a quick survey for the workshop

ControlConstructs



Question

IfThen

Sequence

Age

What is your age?

Numeric

Action

Conditional

Age >= 14
3 items

((Age <> EMPTY) AND (Age < 14))
1 items

Conditional

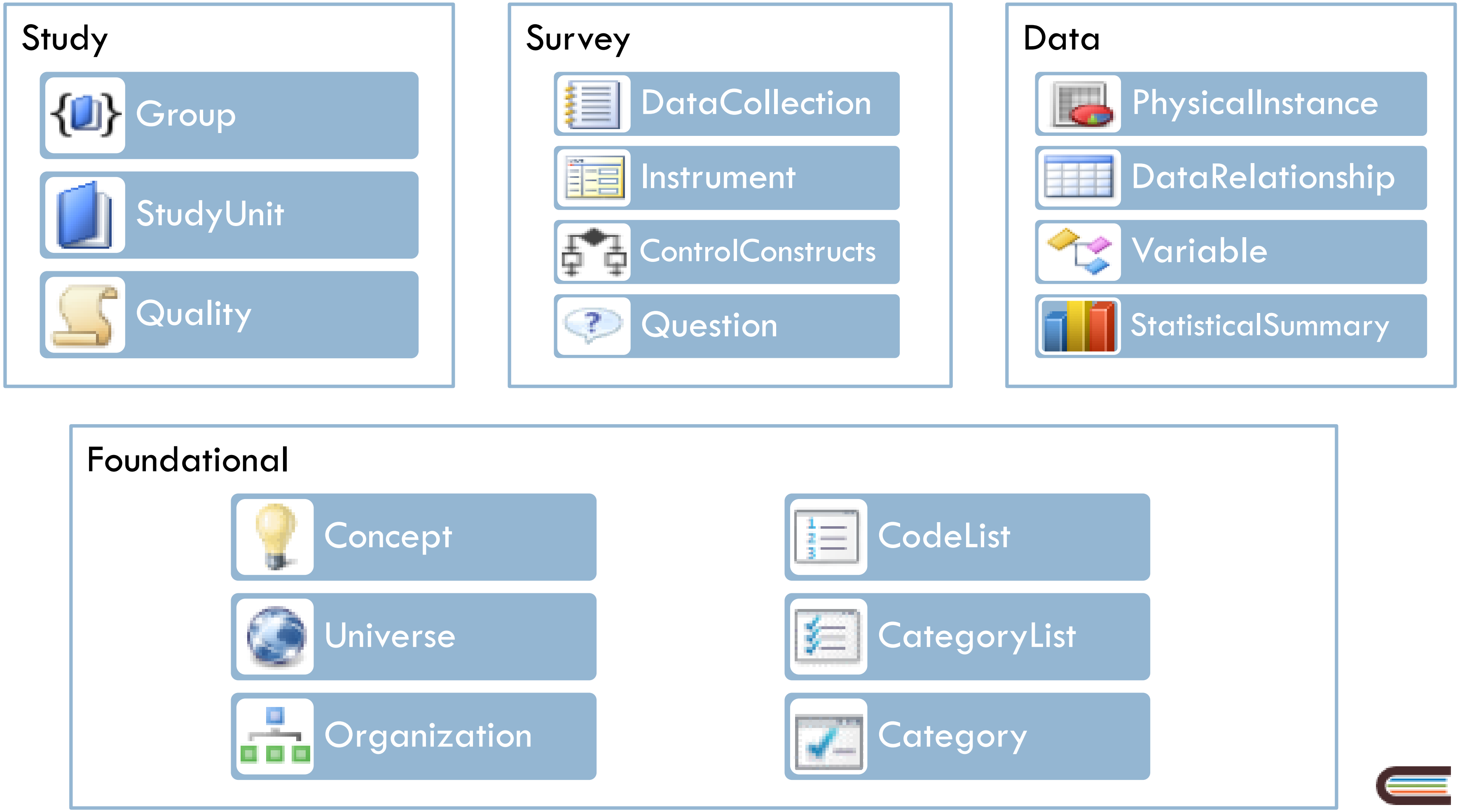
Job = Yes
3 items

((Job <> EMPTY) AND (Job = No))
1 items

ControlConstructs

- If the organization uses DDI, ask:
  - Which parts?
    - Survey
    - Data
    - Study Lifecycle
    - Other

In-depth: Data



Study

- Group
- StudyUnit
- Quality

Survey

- DataCollection
- Instrument
- ControlConstructs
- Question

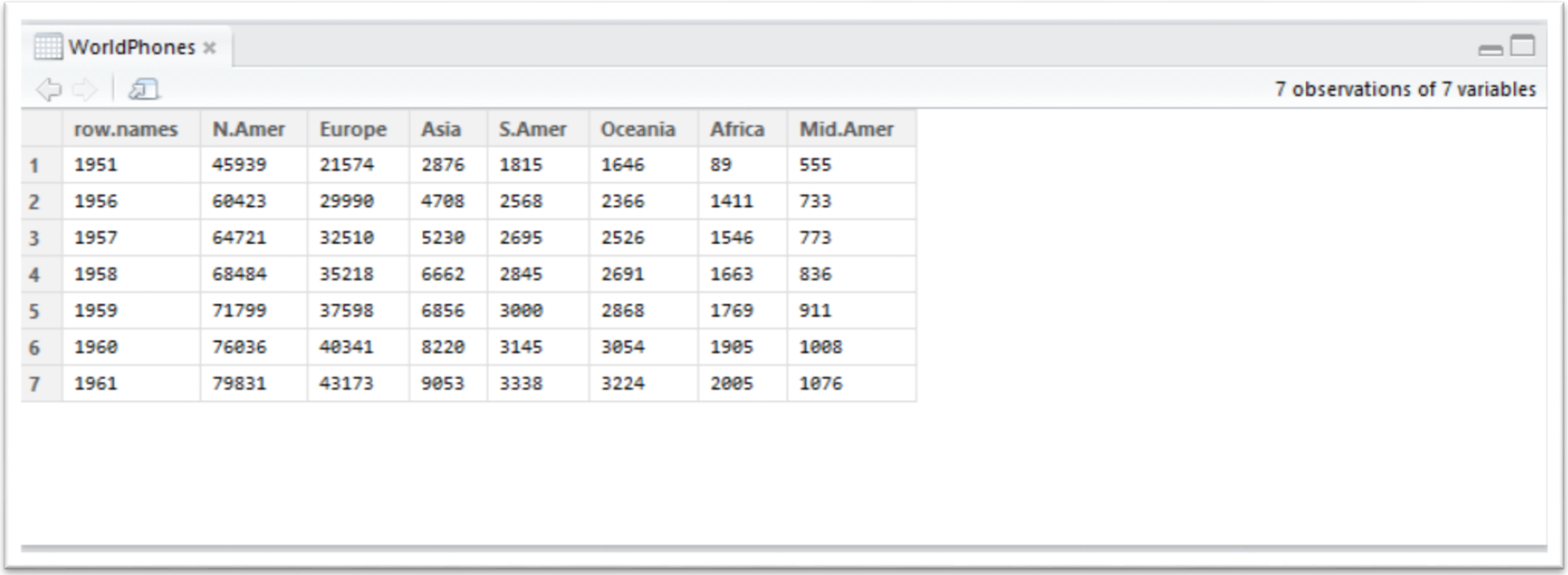
Data

- PhysicalInstance
- DataRelationship
- Variable
- StatisticalSummary

Foundational

- Concept
- Universe
- Organization
- CodeList
- CategoryList
- Category

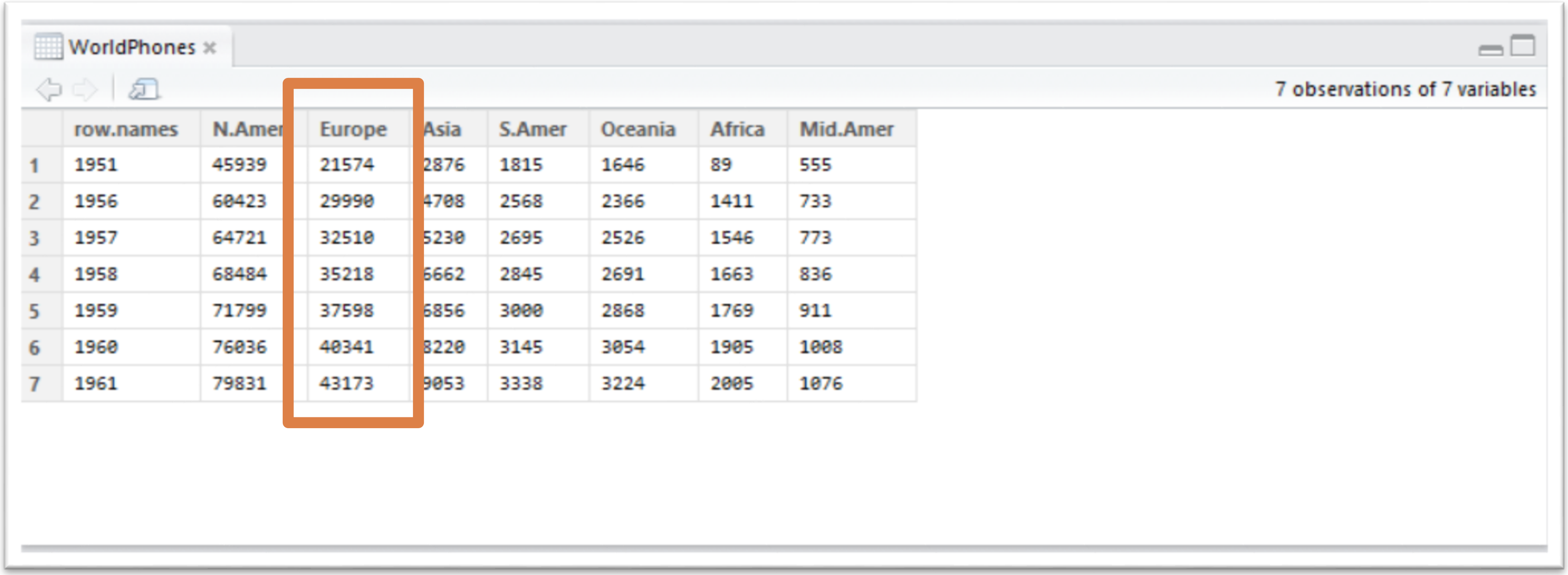
PhysicalInstance (Dataset)



7 observations of 7 variables

	row.names	N.Amer	Europe	Asia	S.Amer	Oceania	Africa	Mid.Amer
1	1951	45939	21574	2876	1815	1646	89	555
2	1956	60423	29990	4708	2568	2366	1411	733
3	1957	64721	32510	5230	2695	2526	1546	773
4	1958	68484	35218	6662	2845	2691	1663	836
5	1959	71799	37598	6856	3000	2868	1769	911
6	1960	76036	40341	8220	3145	3054	1905	1008
7	1961	79831	43173	9053	3338	3224	2005	1076

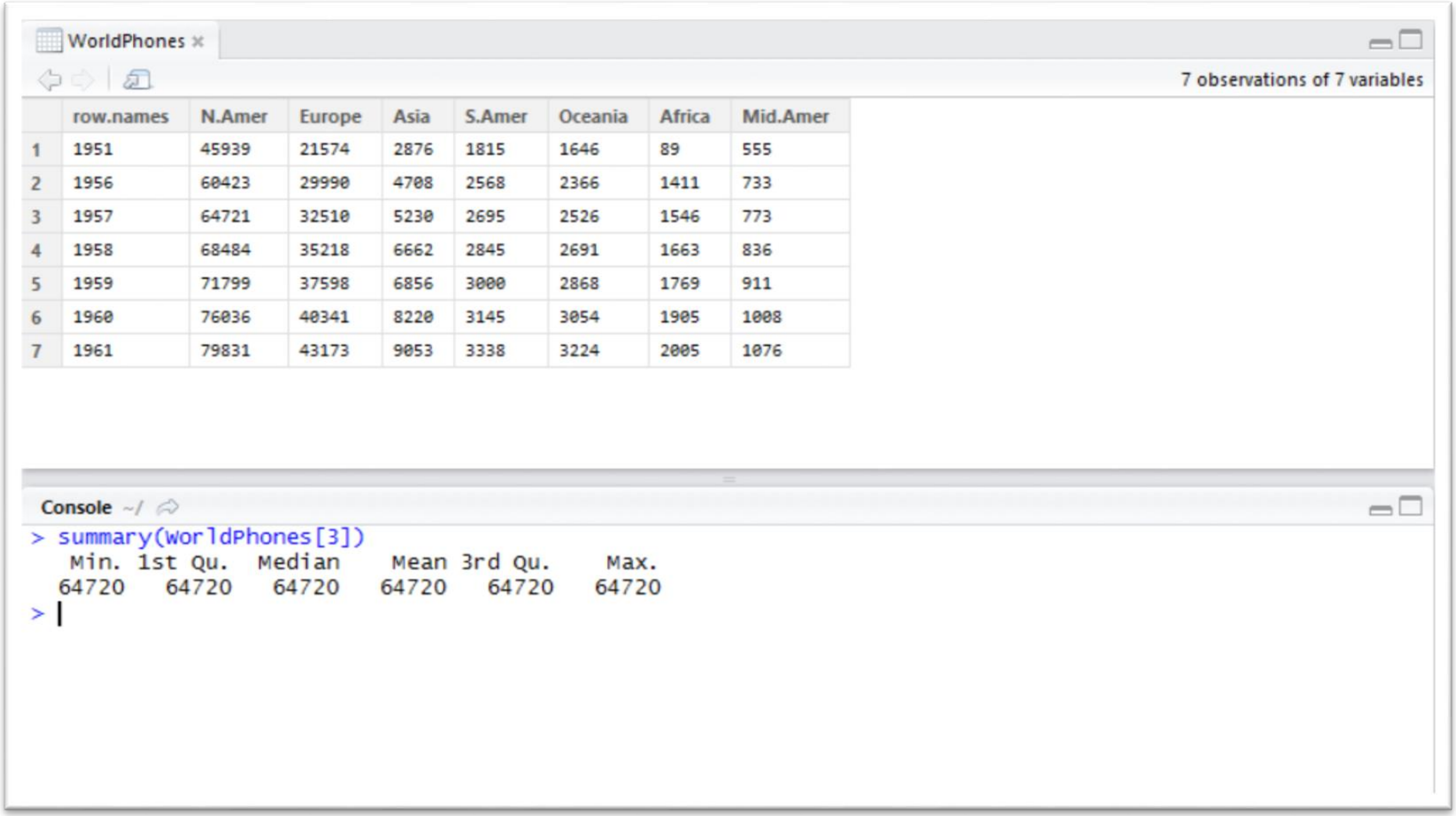
Variable



7 observations of 7 variables

	row.names	N.Amer	Europe	Asia	S.Amer	Oceania	Africa	Mid.Amer
1	1951	45939	21574	2876	1815	1646	89	555
2	1956	60423	29990	4708	2568	2366	1411	733
3	1957	64721	32510	5230	2695	2526	1546	773
4	1958	68484	35218	6662	2845	2691	1663	836
5	1959	71799	37598	6856	3000	2868	1769	911
6	1960	76036	40341	8220	3145	3054	1905	1008
7	1961	79831	43173	9053	3338	3224	2005	1076

StatisticalSummary



WorldPhones

7 observations of 7 variables

	row.names	N.Amer	Europe	Asia	S.Amer	Oceania	Africa	Mid.Amer
1	1951	45939	21574	2876	1815	1646	89	555
2	1956	60423	29990	4708	2568	2366	1411	733
3	1957	64721	32510	5230	2695	2526	1546	773
4	1958	68484	35218	6662	2845	2691	1663	836
5	1959	71799	37598	6856	3000	2868	1769	911
6	1960	76036	40341	8220	3145	3054	1905	1008
7	1961	79831	43173	9053	3338	3224	2005	1076

Console ~/

```
> summary(WorldPhones[3])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  64720   64720   64720   64720   64720   64720 
> 
```

PhysicalInstance (Dataset)

- Let’s download it

In-depth: Study Lifecycle

Study

- Group
- StudyUnit
- Quality

Survey

- DataCollection
- Instrument
- ControlConstructs
- Question

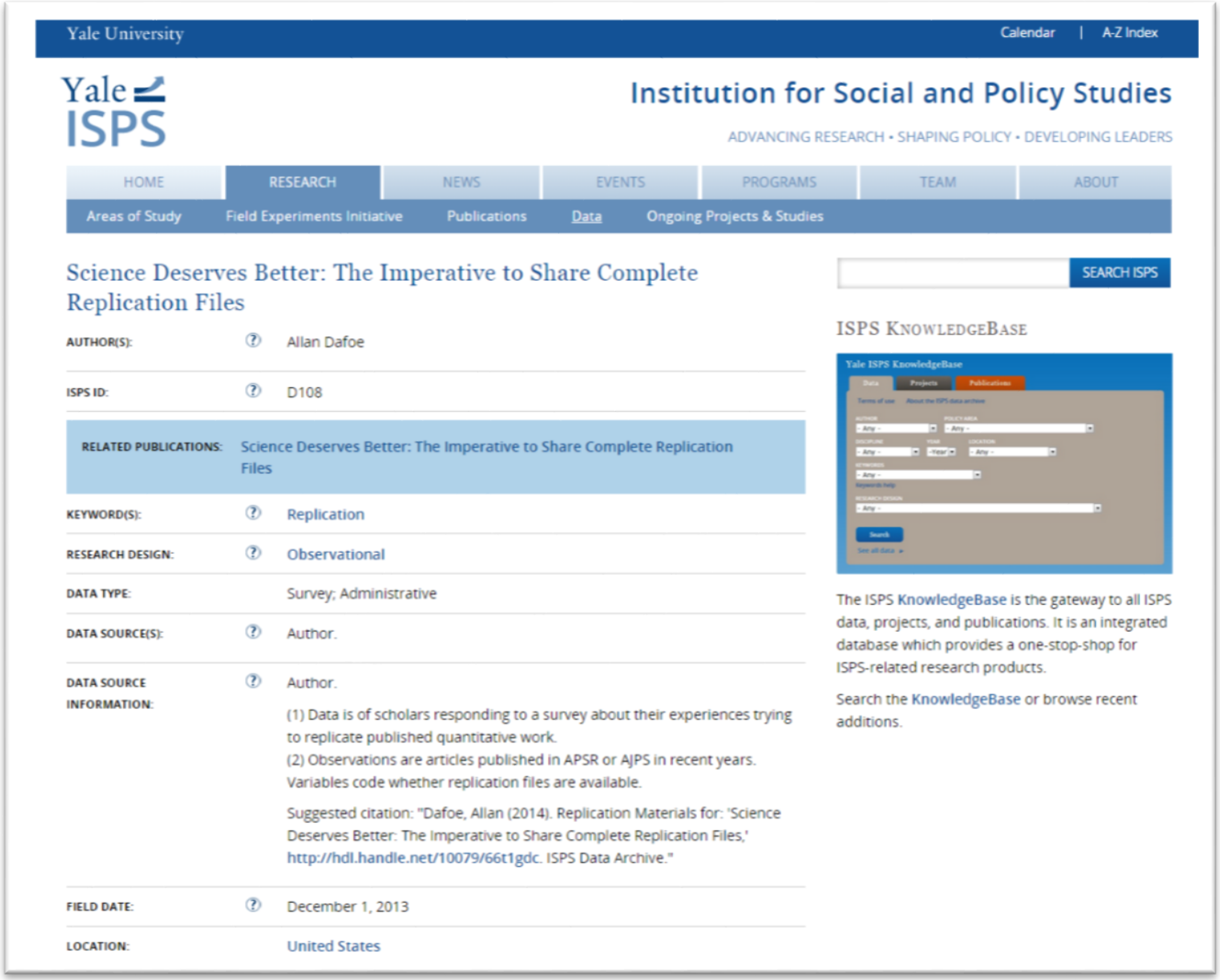
Data

- PhysicalInstance
- DataRelationship
- Variable

Foundational

- Concept
- Universe
- Organization
- CodeList
- CategoryList
- Category

StudyUnit



Yale University
Calendar | A-Z Index
Yale ISPS
Institution for Social and Policy Studies
ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS
HOME
RESEARCH
NEWS
EVENTS
PROGRAMS
TEAM
ABOUT
Areas of Study
Field Experiments Initiative
Publications
Data
Ongoing Projects & Studies
SEARCH ISPS

Science Deserves Better: The Imperative to Share Complete Replication Files

AUTHOR(S):	Allan Dafoe
ISPS ID:	D108
RELATED PUBLICATIONS:	Science Deserves Better: The Imperative to Share Complete Replication Files
KEYWORD(S):	Replication
RESEARCH DESIGN:	Observational
DATA TYPE:	Survey; Administrative
DATA SOURCE(S):	Author.
DATA SOURCE INFORMATION:	Author. (1) Data is of scholars responding to a survey about their experiences trying to replicate published quantitative work. (2) Observations are articles published in APSR or AJPS in recent years. Variables code whether replication files are available. Suggested citation: "Dafoe, Allan (2014). Replication Materials for: 'Science Deserves Better: The Imperative to Share Complete Replication Files,' http://hdl.handle.net/10079/66t1gdc. ISPS Data Archive."
FIELD DATE:	December 1, 2013
LOCATION:	United States

ISPS KnowledgeBase

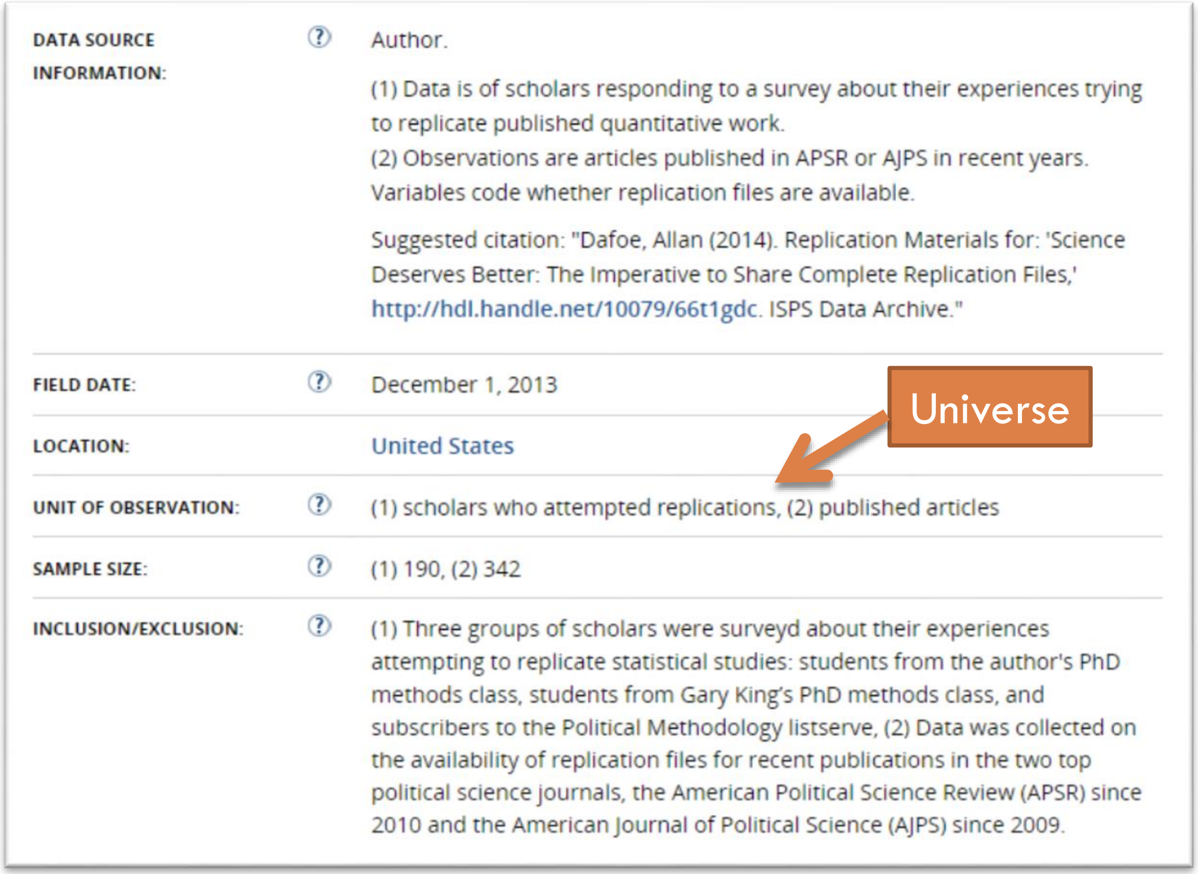
The ISPS KnowledgeBase is the gateway to all ISPS data, projects, and publications. It is an integrated database which provides a one-stop-shop for ISPS-related research products.

Search the KnowledgeBase or browse recent additions.

StudyUnit

- EDDI 2014
  - A study of EDDI workshop attendees in London

Universe (Population)

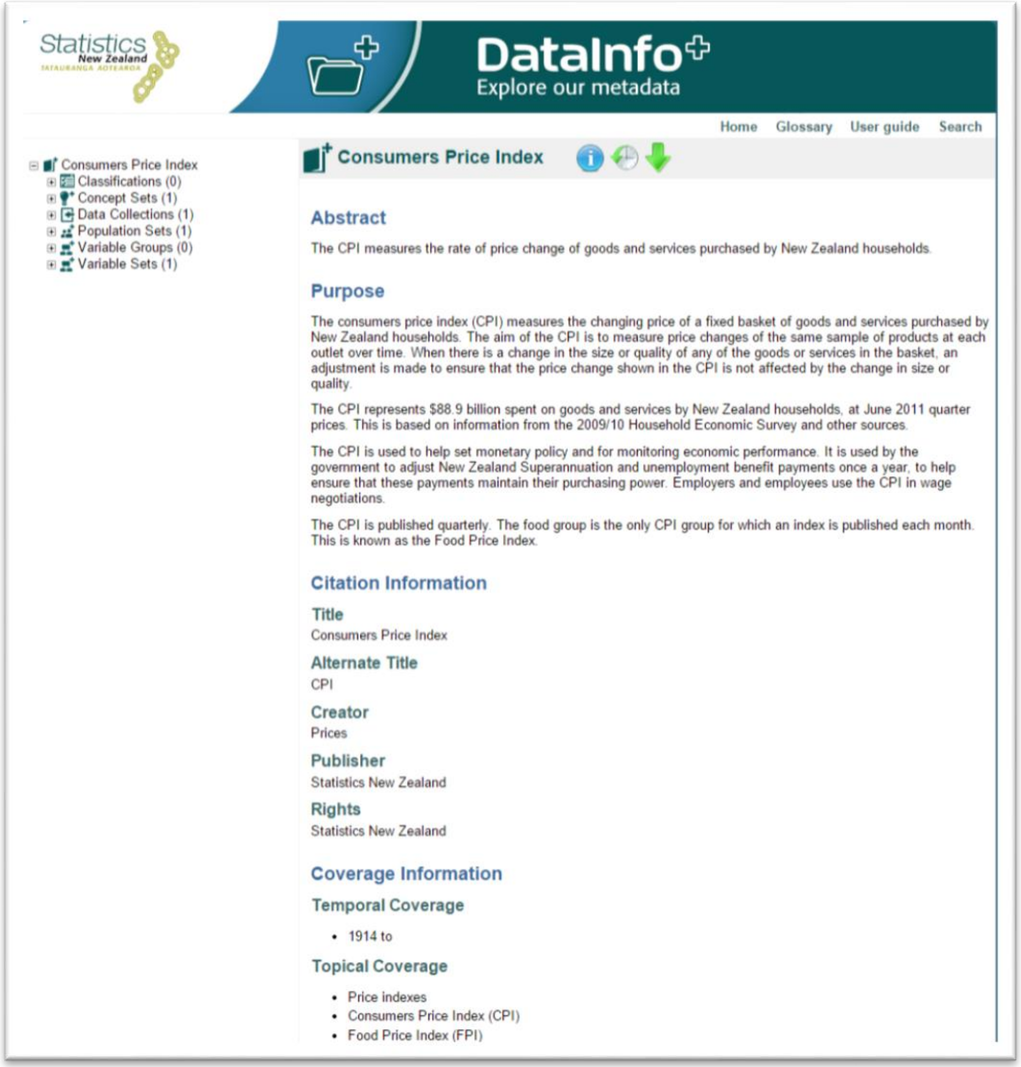


DATA SOURCE INFORMATION:	Author. (1) Data is of scholars responding to a survey about their experiences trying to replicate published quantitative work. (2) Observations are articles published in APSR or AJPS in recent years. Variables code whether replication files are available. Suggested citation: "Dafoe, Allan (2014). Replication Materials for: 'Science Deserves Better: The Imperative to Share Complete Replication Files,' http://hdl.handle.net/10079/66t1gdc. ISPS Data Archive."
FIELD DATE:	December 1, 2013
LOCATION:	United States
UNIT OF OBSERVATION:	(1) scholars who attempted replications, (2) published articles
SAMPLE SIZE:	(1) 190, (2) 342
INCLUSION/EXCLUSION:	(1) Three groups of scholars were surveyd about their experiences attempting to replicate statistical studies: students from the author's PhD methods class, students from Gary King's PhD methods class, and subscribers to the Political Methodology listserve, (2) Data was collected on the availability of replication files for recent publications in the two top political science journals, the American Political Science Review (APSR) since 2010 and the American Journal of Political Science (AJPS) since 2009.

Universe (Population)

- People in this room, right now

Group (Series)



Consumers Price Index

Abstract

The CPI measures the rate of price change of goods and services purchased by New Zealand households.

Purpose

The consumers price index (CPI) measures the changing price of a fixed basket of goods and services purchased by New Zealand households. The aim of the CPI is to measure price changes of the same sample of products at each outlet over time. When there is a change in the size or quality of any of the goods or services in the basket, an adjustment is made to ensure that the price change shown in the CPI is not affected by the change in size or quality.

The CPI represents $88.9 billion spent on goods and services by New Zealand households, at June 2011 quarter prices. This is based on information from the 2009/10 Household Economic Survey and other sources.

The CPI is used to help set monetary policy and for monitoring economic performance. It is used by the government to adjust New Zealand Superannuation and unemployment benefit payments once a year, to help ensure that these payments maintain their purchasing power. Employers and employees use the CPI in wage negotiations.

The CPI is published quarterly. The food group is the only CPI group for which an index is published each month. This is known as the Food Price Index.

Citation Information

Title
Consumers Price Index

Alternate Title
CPI

Creator
Prices

Publisher
Statistics New Zealand

Rights
Statistics New Zealand

Coverage Information

Temporal Coverage

- 1914 to

Topical Coverage

- Price indexes
- Consumers Price Index (CPI)
- Food Price Index (FPI)

Group (Series)

- European DDI User Conference
  - EDDI 2012
  - EDDI 2013
  - EDDI 2014
  - EDDI 2015
  - EDDI 2016

Metadata Publication

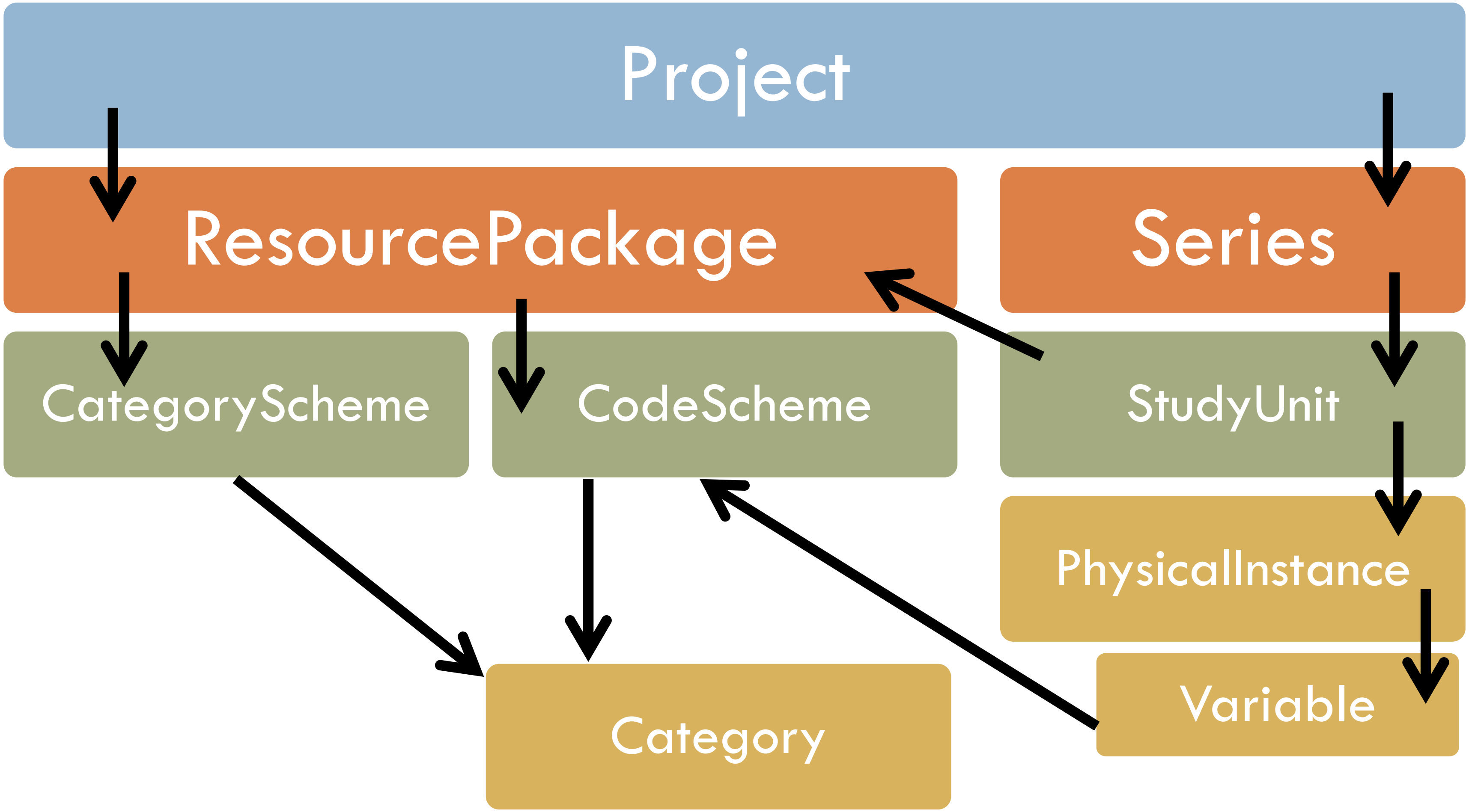
Publish to PDF

Publish to a Repository

Publish to the Web

DDI Versioning

Version Propagation



Colectica takes care of this for you

Processing Structured Metadata

Addin example

- Export survey instrument information to CSV for further processing or analysis

Use Cases and Q & A

Use Cases

- What are yours?

Colectica/DDI Users

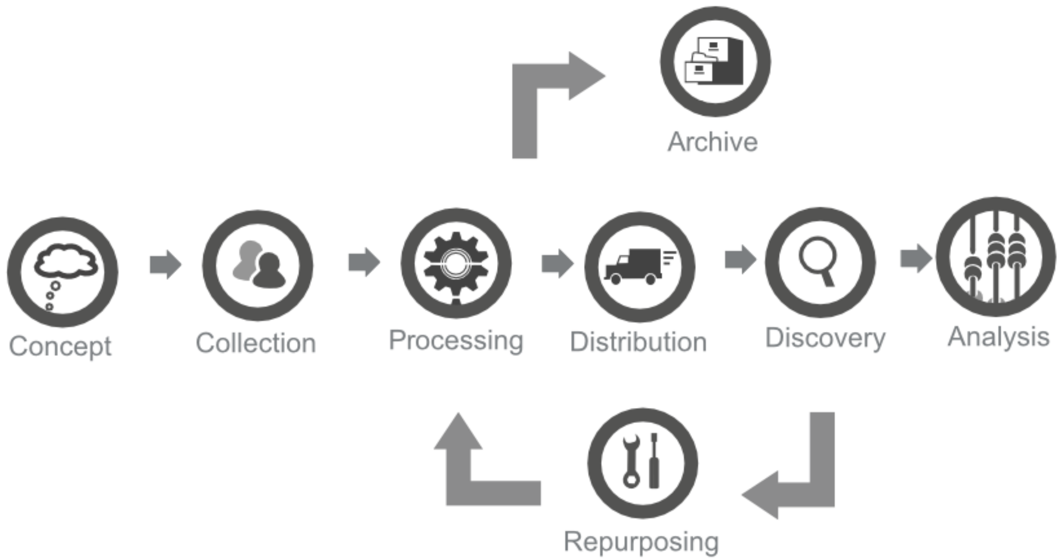
- Official Statistics
  - Statistics New Zealand
  - INSEE
  - Statistics Denmark
- Long-term Longitudinal Studies
  - National Children's Study (NIH and BAH)
  - Midlife in the United States
  - Wisconsin Longitudinal Study
- Archives
  - UCLA Social Science Data Archive
  - Yale Institution for Social and Policy Studies

Statistics New Zealand

> **"Turning data into relevant knowledge, efficiently."**

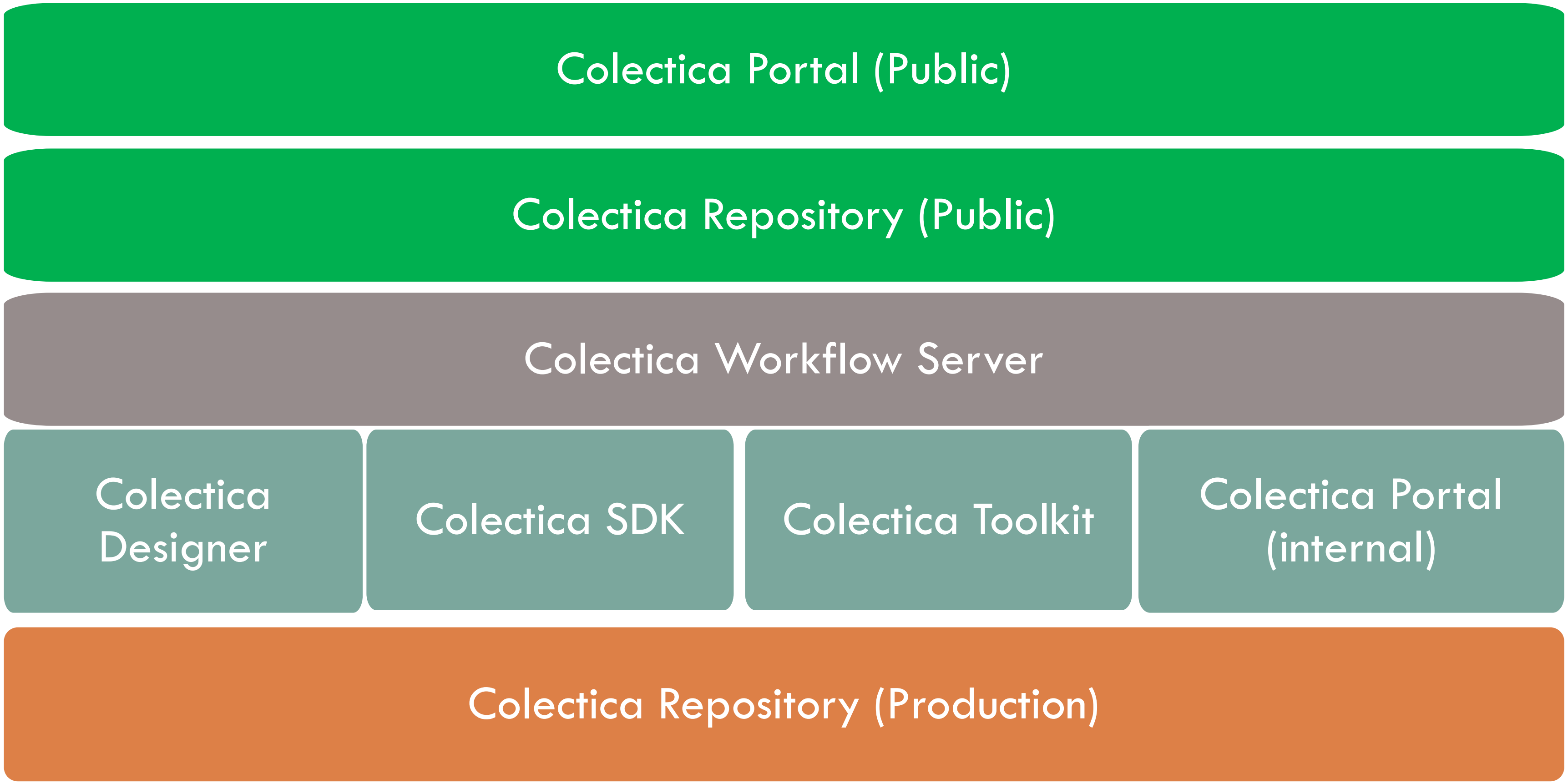
- Ensure New Zealand has the statistical information it needs to grown and prosper
- We do this by:
  - Make sure the right statistics are produced
  - Make sure as many people as possible use the statistics to support informed decision making

Statistics New Zealand Metadata Infrastructure Project



- Create a single, canonical source for all this information
- Solution: central repository

Architecture: Repository



Colectica Portal (Public)

Colectica Repository (Public)

Colectica Workflow Server

Colectica Designer

Colectica SDK

Colectica Toolkit

Colectica Portal (internal)

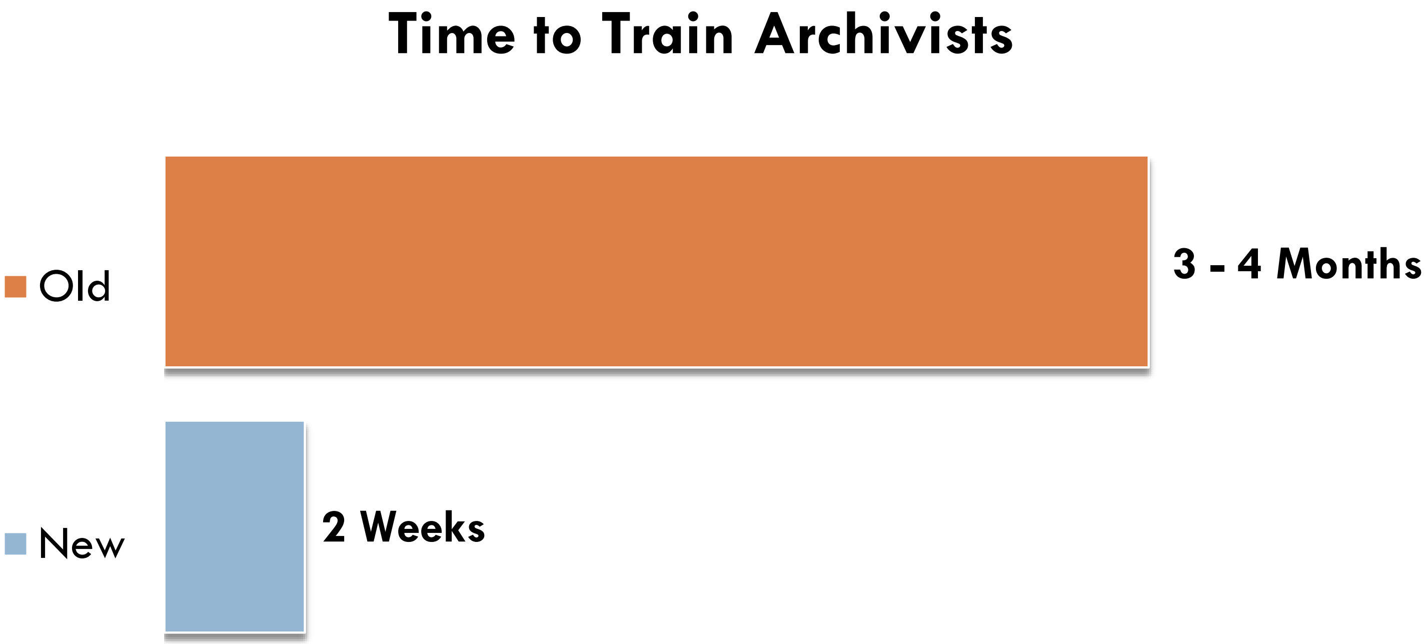
Colectica Repository (Production)

Key Result 1 – Metadata Capture

“We used to record all metadata at the end of the lifecycle.”

“Now, curators capture the information when they first think of it.”

Key Result 2 - Archiving



Facts and Figures

1,008

Datasets

200

Series

20 - 40

Metadata Curators

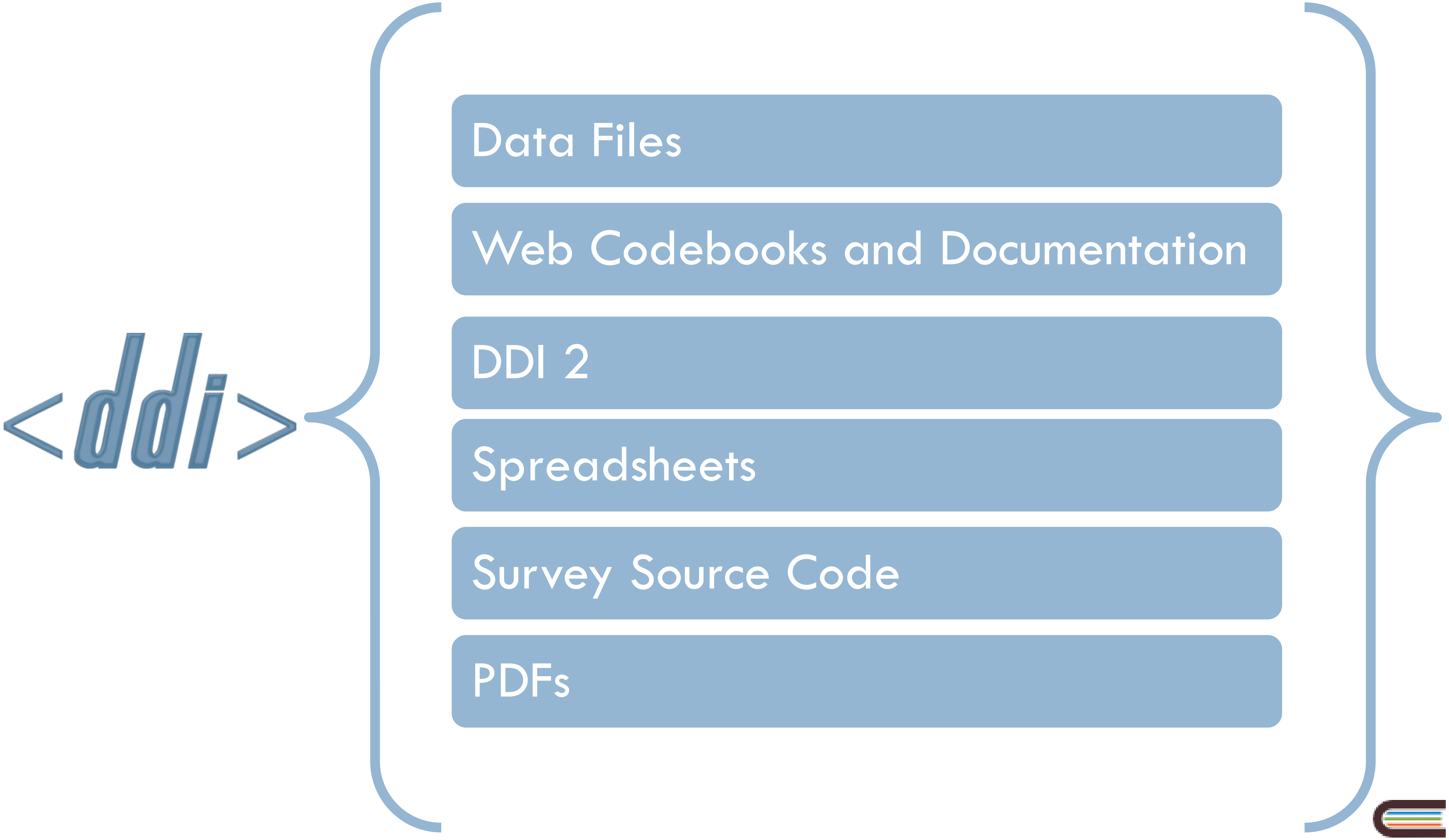
219

Unique Portal Users

Statistics Denmark

- Managing Series and Study-level metadata
- Eurostat Quality reporting requirements
  - From DDI to ESMS, ESQRS SDMX formats

MIDUS: Data Integration Project

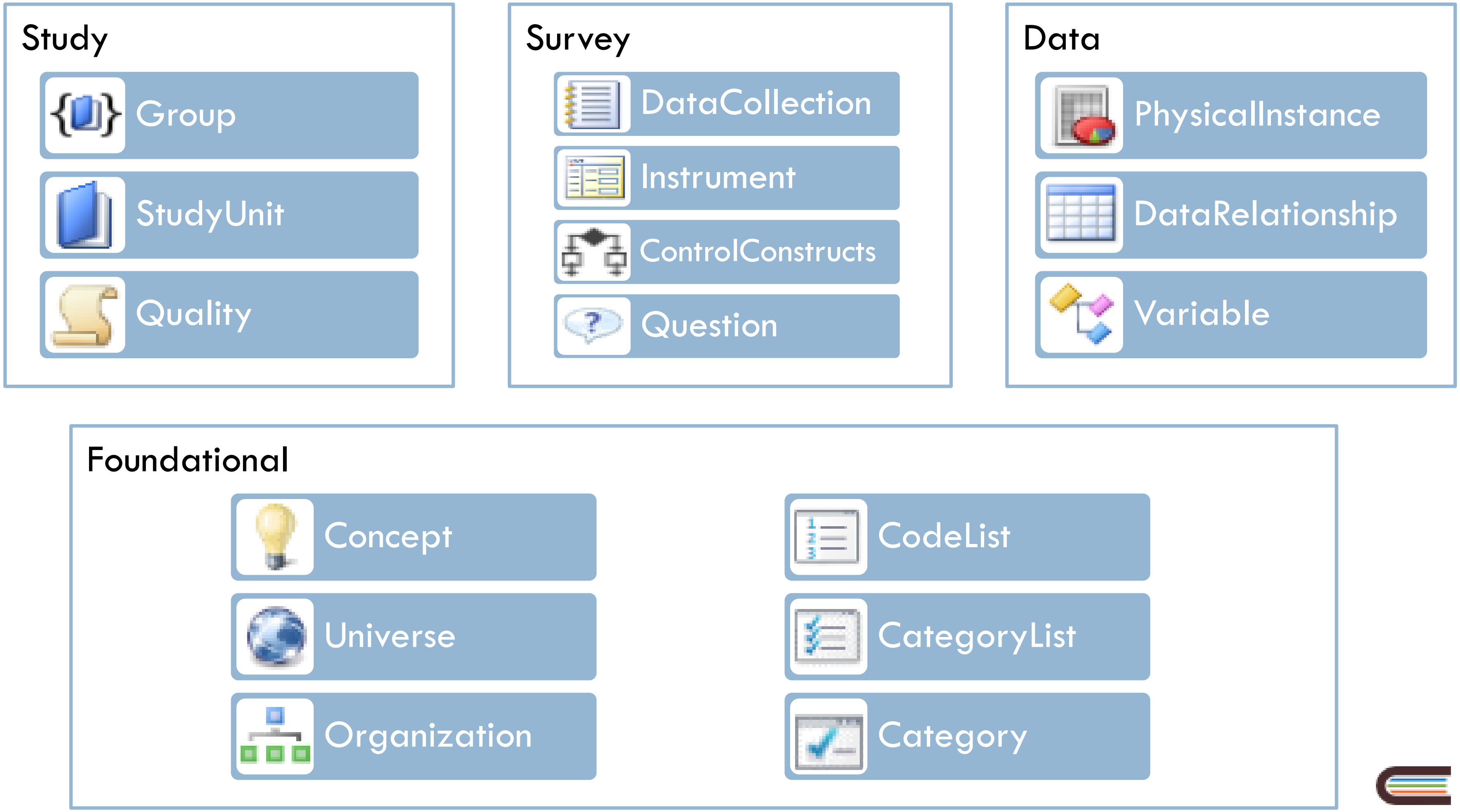


- Data Files
- Web Codebooks and Documentation
- DDI 2
- Spreadsheets
- Survey Source Code
- PDFs

Mapping MIDUS to DDI Lifecycle

- Joint project between MIDUS and Colectica
- Main tool: Colectica
  - Repository
  - Designer
  - SDK

DDI at a Glance



Study

- Group
- StudyUnit
- Quality

Survey

- DataCollection
- Instrument
- ControlConstructs
- Question

Data

- PhysicalInstance
- DataRelationship
- Variable

Foundational

- Concept
- Universe
- Organization
- CodeList
- CategoryList
- Category

“Research Data Management: Facilitating Discoverability using Open Metadata Standards”

April 8 – 10, 2015

University of Wisconsin - Madison

Feedback

bit.ly/eddi-workshop

colectica

Thank you

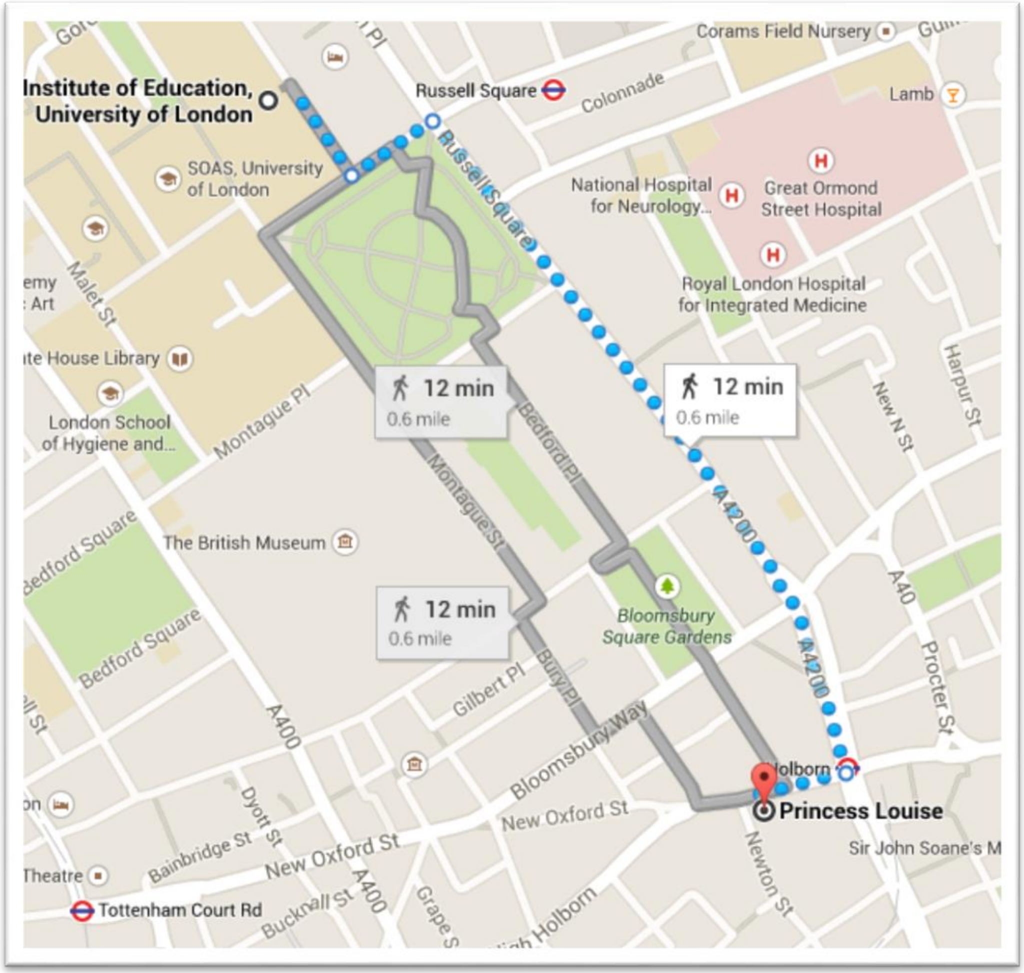
Jeremy Iverson
jeremy@colectica.com

Dan Smith
dan@colectica.com

colectica.com

“Informal Gathering”

“Informal gathering”


Institute of Education,
University of London
Russell Square
Corams Field Nursery
Colonnade
Lamb
SOAS, University of London
National Hospital for Neurology...
Great Ormond Street Hospital
Royal London Hospital for Integrated Medicine
Malet St
Russell Square
London School of Hygiene and...
Montague Pl
12 min
0.6 mile
12 min
0.6 mile
Bedford Pl
Montague St
A4200
Harpur St
New N St
Bedford Square
The British Museum
12 min
0.6 mile
Bloomsbury Square Gardens
A40
Procter St
Gilbert Pl
Bury Pl
Bloomsbury Way
A400
Holborn
Princess Louise
New Oxford St
Sir John Soane's M
Dyott St
Bainbridge St
New Oxford St
Theatre
Tottenham Court Rd
Bucknall St
A400
Grape St
High Holborn
Newton St