Metadata

Taylor J. Wilson

Economist

Division of Consumer Expenditure Surveys

American Society of Hispanic Economists Training Seminar

11 August 2016

BLS

ASHE Seminar: Metadata Roadmap

- What is metadata?
- Why is metadata important?
- BLS and metadata
- International standards
- Pushing the frontier

What is metadata?

- *"…structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource."* – National Information Standards Organization (NISO)

- Data used to describe something

- Metadata is a role for data

BLS

Types of Metadata

- **Descriptive Metadata**
  - Resource for the purpose of identification and discovery.

- **Structural Metadata**
  - How compound objects in a system fit together.

- **Administrative Metadata**
  - Provides information to manage a resource within a system.

Descriptive

Variable Name
Variable Description
Keywords
String v. Numeric

Structural

Skip Patterns
Dependencies
Ordering
Grouping Schemes

Administrative

Permissions
File Location
Time of Creation
Date Modified
Paradata

BLS

ASHE Seminar: Metadata Roadmap

- What is metadata?

- Why is metadata important?

- BLS and metadata

- International standards

- Pushing the frontier

Why is metadata important?

- Archivists, Scientists, Economists, Librarians, and others all work with data in some form.

- All professions are interested in efficient resource management

- Understanding your data makes your data easier to navigate and use.

Source: National Archives, Photo of War Department Records before being transferred to the National Archives in 1935
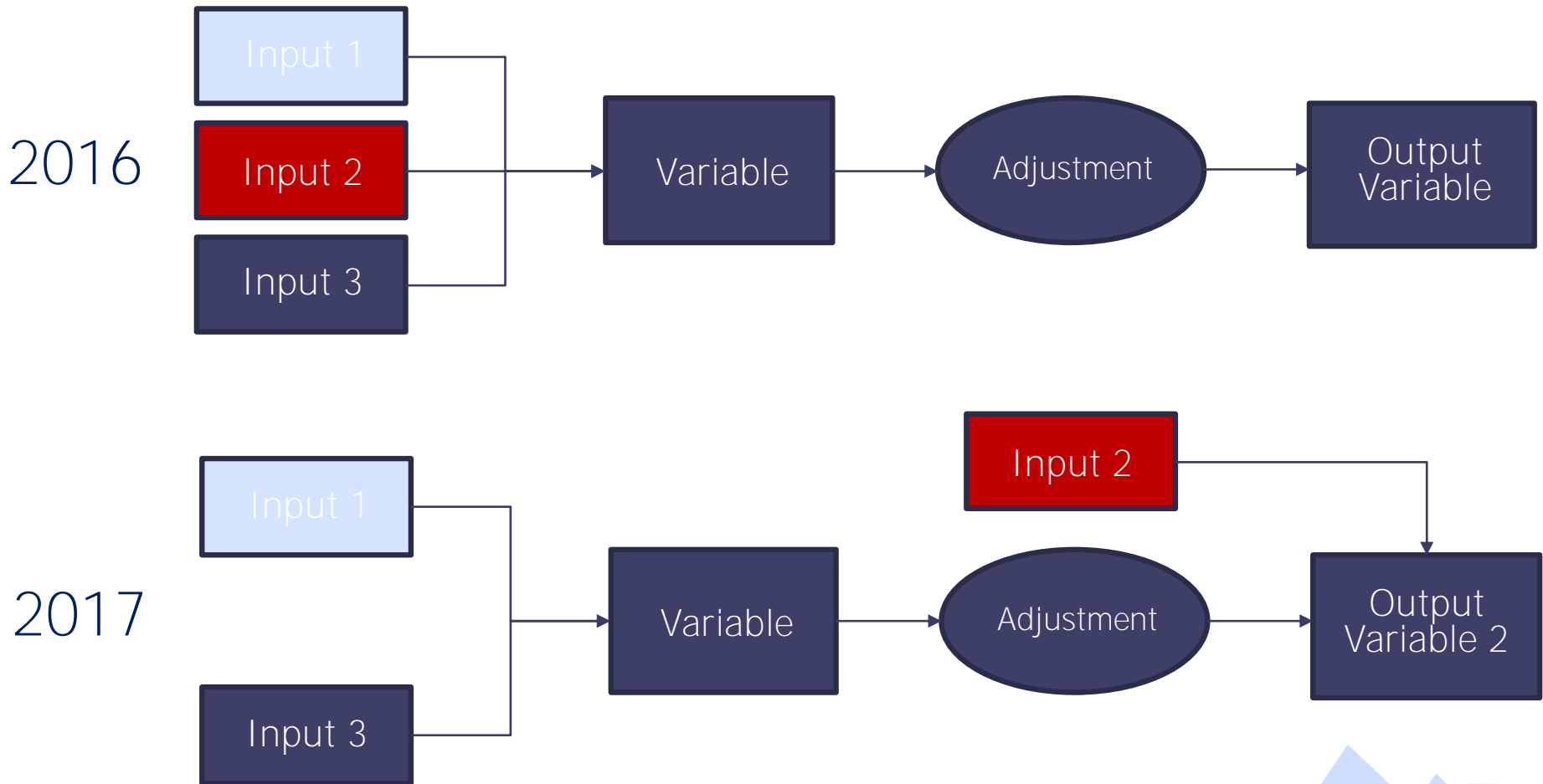
Variable Organization

- Good data management improves efficiency
- Self describing variables

Consumer Expenditure Survey (CE) Income Variables	Survey of Consumer Finances (SCF) Income Variables
FINCBTAX	INCOME
INTRDVX	NORMINC
ROYESTX	WAGEINC
OTHREGX	BUSSEFARMINC
WELFAREX	INTDIVINC
RETSURVX	KGINC
NETRENTX	SSRETINC
OTHRINCX	TRANSFOTHINC

Tracking Systematic Change

Example: One name changes and a processing input move

2016

Input 1

Input 2

Input 3

→ Variable → Adjustment → Output Variable

2017

Input 1

Input 3

→ Variable → Adjustment → Output Variable 2

Input 2

BLS

Accountability

- BLS produces many datasets which have users ranging from academics to policy makers to the American public.

- Questions involving data use, structures, and changes are frequent.

- Good metadata allows operational questions to be answered quickly and efficiently.

ASHE Seminar: Metadata Roadmap

- What is metadata?

- Why is metadata important?

- BLS and metadata

- International standards

- Pushing the frontier

BLS

BLS and Metadata

- Institutional silos are not conducive to inter-program communication.

- No general standard on metadata construction or storage.

BLS and Metadata cont.

- Each program office is largely self-governing with a hierarchy of management.

- It is currently very difficult for inter-program office communication to occur.

- It is even more difficult for a consensus regarding the organization of metadata to be achieved.

BLS

What do we have in common?

- Social Science Surveys
  - Establishment Surveys
  - Household Surveys
- Economics oriented
  - BLS produces *seven* principal federal economic indicators (PFEI)
- Interested in serving the public good with high quality data.

Economists

Descriptive

Variable Name
Variable Description
Keywords
String v. Numeric

Survey Methodologists

Structural

Skip Patterns
Dependencies
Ordering
Grouping Schemes

Administrative

Permissions
File Location
Time of Creation
Date Modified

Computer Scientists

Where we stand

Administrative

Permissions
File Location
Time of Creation
Date Modified

Computer Scientists

- Easiest metadata to capture

- Auto-generated by computers

- Challenge: Capture and storage

Where we stand

Descriptive

Variable Name
Variable Description
Keywords
String v. Numeric

Economists

- Survey specific

- Naming conventions: low agreement, enforcement, and coverage

- Challenge: Inter-office communication and sharing variable creation methodology.

BLS

Where we stand

Structural

Skip Patterns
Dependencies
Ordering
Grouping Schemes

Survey Methodologists

- Survey specific
- Arguably the most difficult of the three to harmonize
- Challenge: Inter-office communication and harmonization of methodology.

ASHE Seminar: Metadata Roadmap

- What is metadata?

- Why is metadata important?

- BLS and metadata

- International standards

- Pushing the frontier

BLS

International Standards

- Dublin Core Metadata Initiative (DCMI)

- Machine-Readable Cataloging (MARC)

- Metadata Object Description Schema (MODS)

- Text Encoding Initiative (TEI)

- Federal Geographic Data Committee (FGDC)

- Data Documentation Initiative (DDI)

Dublin Core Metadata Initiative

- Developed for digital libraries

- Broad general principals which can be targeted by any metadata scheme

- Interoperability – A system working with other systems without special effort by the customer

Level 1: Shared vocabularies defined in natural language	
Level 2: Shared vocabularies based on formal schematics	
Level 3: Shared formal vocabularies in exchangeable records	
Level 4: Shared formal vocabularies and constraints in records	

Source: Dublin Core Metadata Initiative Website, accessed July 2016, http://dublincore.org/metadata-basics/

International Standards

- Dublin Core Metadata Initiative (DCMI)

- Machine-Readable Cataloging (MARC)

- Metadata Object Description Schema (MODS)

- Text Encoding Initiative (TEI)

- Federal Geographic Data Committee (FGDC)

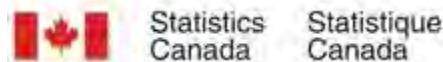- Data Documentation Initiative (DDI)

Data Documentation Initiative (DDI)

- International metadata standard for describing social science datasets.

- Applies Dublin Core Elements in a way that makes sense for social science datasets and surveys.

- Tool box framework allows for an ad hoc solution to the interoperability problem between survey programs.
  - ▶ "Enter once, use often."
  - ▶ Effective Data Sharing
  - ▶ Standardized Terminology

Who uses DDI?

- Center for Comparative European Survey Data
- Consortium of European Social Science Data Archives
- Centralising and Integrating Metadata from European Statistics
- Inter-university Consortium for Political and Social Research
- International Household Survey Network
- Documenting the Survey Cycle using DDI
- The Data Liberation Initiative Metadata Collection
- Canadian Research Data Centres Program
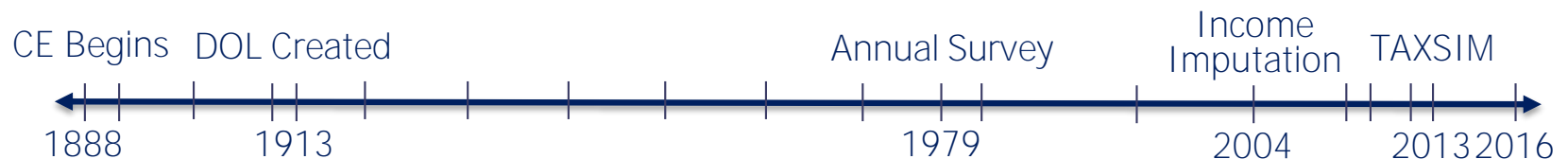
ASHE Seminar: Metadata Roadmap

- What is metadata?

- Why is metadata important?

- BLS and metadata

- International standards

- Pushing the frontier

BLS

Pushing the Frontier

- The Consumer Expenditure Survey (CE) plans to pilot a system implementing the DDI Standards.

- CE is an old and complex survey with a large volume of metadata that is valuable to a user.
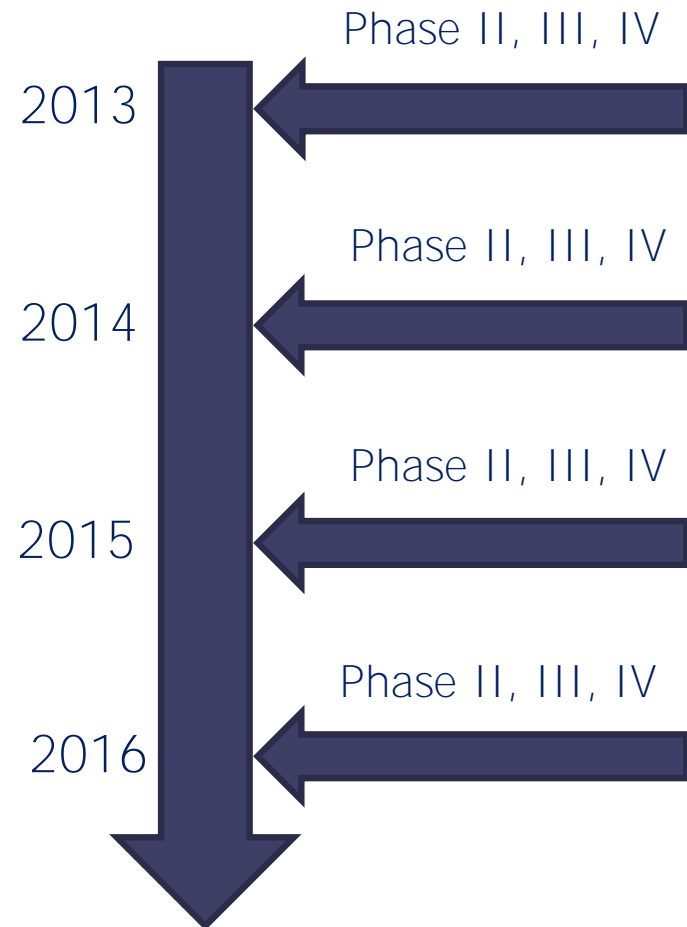
CE Timeline: Major Milestones

| CE Begins | DOL Created | | | | | | | Annual Survey | | | Income Imputation | | TAXSIM | |
| 1888 | | 1913 | | | | | | | 1979 | | | 2004 | | 2013 | 2016 |

Not to scale; not all changes are marked or labeled

One Survey, Two Dimensions

- Most end users are interested in changes across time.

- Changes to a variable through the processing phases helps users understand how the data is generated.

Phase II, III, IV
2013

Phase II, III, IV
2014

Phase II, III, IV
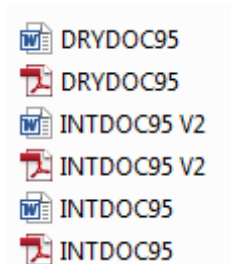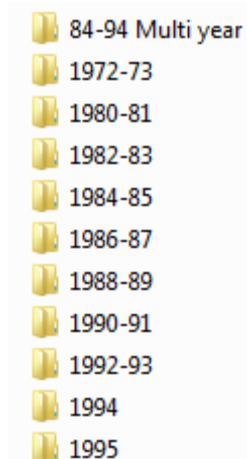2015

Phase II, III, IV
2016

BLS

Dear Sir/Madam,

I have accessed a derived dataset with information on the educational attainment of the household's reference person before the year 1996. This information was codified on a scale of 1 to 7, but I don't know the meaning of the codes.

Could you tell me their meaning?

Thank you!

Answering Inquiries

| 84-94 Multi year |
| 1972-73 |
| 1980-81 |
| 1982-83 |
| 1984-85 |
| 1986-87 |
| 1988-89 |
| 1990-91 |
| 1992-93 |
| 1994 |
| 1995 |

→

DRYDOC95
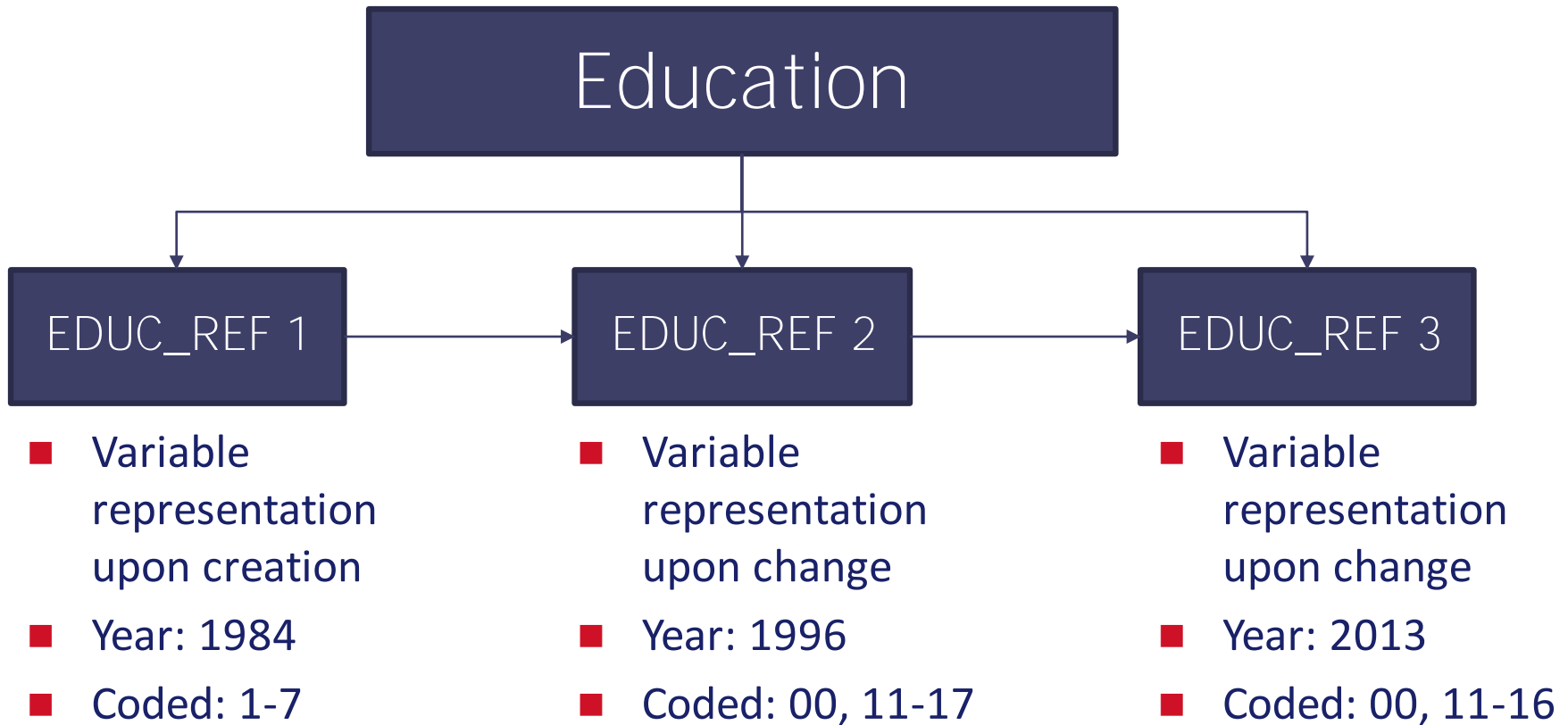DRYDOC95
INTDOC95 V2
INTDOC95 V2
INTDOC95
INTDOC95

→

EDUC_REF   Education of reference person
CODED
1 Elementary (1-8 yrs.)
2 High school, less than H.S. graduate
3 High school graduate
4 College, less than college graduate
5 College graduate
6 Graduate school
7 Never attended school

BLS derived

■ Many copies of the same information is not conducive to:

▶ Building time series

▶ Efficiently locating information

▶ Efficiently utilizing government resources

Example DDI Solution

```
                    ┌─────────────────────────┐
                    │       Education         │
                    └─────────────────────────┘
                               │
          ┌────────────────────┼────────────────────┐
          ▼                    ▼                    ▼
   ┌──────────────┐     ┌──────────────┐     ┌──────────────┐
   │  EDUC_REF 1  │ ──▶ │  EDUC_REF 2  │ ──▶ │  EDUC_REF 3  │
   └──────────────┘     └──────────────┘     └──────────────┘
```

- Variable representation upon creation
- Year: 1984
- Coded: 1-7

- Variable representation upon change
- Year: 1996
- Coded: 00, 11-17

- Variable representation upon change
- Year: 2013
- Coded: 00, 11-16

BLS

Benefits of Metadata Management

- Minimizes lag time between question and answer.

- Improves processing efficiency

- Users will be able to answer their own questions without ambiguity.

- Reduces analytical errors from data anomalies and inconsistencies.

- Creates better, more user-friendly survey data.

Pushing the Frontier

- CE is motivated to provide users with effective metadata documentation.

- Focusing on reuse allows us to create more user-friendly documentation for the CE Survey.

- Adhering to the international standards will encourage moves toward interoperability.

Contact Information

Taylor J. Wilson
Economist
Division of Consumer Expenditure Surveys
www.bls.gov/cex/
202-691-6550
wilson.taylor@bls.gov

BLS