Using the Data Documentation Initiative to Document the Consumer Expenditure (CE) Survey

Daniel W. Gillman and Reginald Noel

U.S. Bureau of Labor Statistics

ABSTRACT

The Data Documentation Initiative (DDI) is a set of statistical metadata standards for documenting social science data. Currently, there are two series, Codebook - for documenting data sets one at a time; and Lifecycle - for documenting the data sets, questionnaires, and the survey lifecycle for any number of surveys. Each has a number of versions, all of which are in use. Social Science data libraries, data archives, national statistical offices, and international organizations around the world are all making use of the standard. The US Bureau of Labor Statistics (BLS) chose DDI-Lifecycle version 3.2 and the commercial DDI software Colectica Designer to build a pilot system for documenting the Consumer Expenditure Survey (CE). This work was inspired by the realization that publishing PDF files for documenting each year of the CE public use microdata files (PUMD) was inefficient. For instance, using the yearly files, it would very hard to determine whether a variable is used unchanged over every year the PUMD files are published. The DDI-Lifecycle promises a solution to this problem and many others like it. In addition, the DDI-Lifecycle is designed to document the survey lifecycle. This means the design, processing steps, and subsequent data transformations can be described. As CE is undergoing a redesign at this time, the opportunity presents itself for implementing a metadata approach to documentation. BLS has a pilot project to determine the effectiveness of DDI-Lifecycle and the Colectica Designer software. In this poster, we will describe the advantages, accomplishments, pitfalls, and lessons learned in using the approach DDI. We hope to attract other survey groups to try the same approach.

PROJECT OBJECTIVES

- Test the feasibility of using DDI to document CE
  - Data
  - Survey lifecycle
- Test the usefulness of Colectica™ Designer software for this task
- Provide recommendations to management, based on results

TEAM MEMBERS

- CE staff
  - Reggy Noel, Jimmy Choi, Evan Hubener, Taylor Wilson
  - Bryan Rigg, Lucilla Tan, Scott Curtin
- CE management support
- Research staff
  - Dan Gillman

WHAT ARE METADATA STANDARDS?

- Metadata = data used to describe some objects (e.g., survey data)
- Standard = normative document containing
  - Requirements
  - Recommendations
  - Statements
  - Instructions
- Developed under an open, transparent, balanced, fair process

MANAGING DDI

- DDI Alliance
  - Over 30 organizations
    - Data Libraries
    - Data Archives
    - Statistical organizations
    - Small software developers
  - Managed by University of Michigan – ICPSR
  - Legal entity
    - Charter
    - By-laws
- Alliance develops metadata standards (as defined above)
- Member organizations nominate representatives, who
  - Perform technical work
  - Manage process
  - Promote standards

WHAT IS DDI?

- DDI is a suite of statistical metadata standards
- Two main products at this time
  - **DDI Codebook** (2.x)
    - 2.5 is latest version
    - Published as XML-Schema
    - Describes a single Study and its data sets
    - No ability to reuse across studies
  - **DDI Lifecycle** (3.x)
    - 3.2 is latest version
    - Published as XML-Schema
    - Describes many studies, designs, and processing
      - i.e., survey lifecycle
      - Takes advantage of reuse
      - Over time

CE SURVEY

- Conducted by Bureau of Labor Statistics
- Data collected for BLS by the Census Bureau
- Actually, two surveys
  - Diary, Interview
- Combined tabular data released twice yearly
  - Each cover 12 month period
- Public Use Microdata released yearly

- Interview
  - Collect expenses
    - Previous 3 months
    - Easily recalled, large, recurring
    - E.g., rent, utilities
  - Collect demographics, income, assets
  - Sample size – 12K addresses each quarter
    - Rotating panel
      - 1/4th new sample each quarter
    - 6.9K completed each quarter
    - Each address interviewed for 4 consecutive quarters

- Diary
  - Record 2 consecutive one-week diaries
  - Evenly spread throughout year
  - Sample size – 12K addresses per year
    - 6.9K completed (x 2 diaries per household)
  - Collect expenses
    - Small, frequently purchased
    - E.g., food, clothing

- Production
  - Phase 1 (@ Census)
    - Sample selection
    - Interviewing / Data collection
    - Simple editing
    - Data sent to BLS
  - Phase 2 (@ BLS)
    - Editing (Initial Edit Subsystem)
    - Identify completed interviews for use in production

CE SURVEY cont'd

- Production
  - Phase 3 (@ BLS)
    - Edit and Estimation subsystem
    - Impute missing data, allocate combined data
    - Provide universal classification codes to expenditures and items
    - Create quarterly aggregates of spending by household
    - Provide data to CPI
  - Phase 4 (@ BLS)
    - Final edits
      - Top code (PUMD only)
      - Consistency checks with previous years
    - Create tables of annual spending by various demographics
    - Generate PUMD

DEVELOPMENT PROBLEM

- Map the facts of CE
  - Diary
  - Interview
  - Phases
  - Repetition over time
  - Bi-yearly changes
- To Colectica / DDI constructs
- Show ability to
  - Manage variables thru phases
  - Tie variables to questions or derivations
  - Link variables / questions changing over time
  - Link semantically similar variables / questions

SOFTWARE and PILOT

- Colectica
  - Obtained licenses for Colectica/Designer
  - Based on DDI Lifecycle version 3.2
- Development
  - Selected two variables (education and health insurance)
  - Recorded changes over last 3 phases and over time
  - Linked to questions from Interview questionnaire

RESULTS

DDI - Advantages

- Using the international DDI standard.
- Unified metadata across phase.
- Documentation generation for any level of detail.
- Ability top share metadata with other organizations.

Disadvantages

- DDI Lifecycle 3.2 is complex.
- Based on XML.

Colectica Software - Advantages

- A single program which captures and represents change across time.
- Advanced query capability.
- Support from Colectica software developers.

Disadvantages

- High learning curve, due to complexity of DDI.
- Import for Blaise code is not perfected.
- Need for multiple licenses.
- Repository/Portal software necessary for UI and advanced query.
- Advanced query has not been tested, may not exceed MS Access.

CONCLUSIONS

- DDI appears sufficient for documenting CE production
- Colectica Designer
  - Sufficient for metadata input
  - Can't handle versioning or changes over time
  - Limited ability to handle reuse
  - Not robust enough for end-user interface
- Colectica Repository/Portal is needed to complete the pilot

REFERENCES

- DDI Alliance http://www.ddialliance.org/
- Colectica http://www.colectica.com/

CONTACT

- Dan Gillman Gillman.Daniel@bls.gov
- Reggy Noel Noel.Reginald@bls.gov