Thoughts on the Importance of Metadata for Integrated Data

WSS/FCSM Workshop on Metadata, September 14, 2018

Marilyn Seastrom

Chief Statistician

National Center for Education Statistics

This presentation is intended to promote ideas. The views expressed do not necessarily reflect the position of the U.S. Department of Education.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Going forward

- Federal statistics based on integrated data, regardless of the source, must be **communicated transparently** and understood to ensure that the nation is provided the best available statistical information and that the statistics can be used wisely.
- Metadata is an essential element to these efforts.

European Statistical System Data Quality Framework: Accessibility and Clarity

- "Statistics should be presented in a clear and understandable form . . . with supporting metadata and guidance."
- Metadata should be:
  - Preserved and properly archived
  - Standardized according to systems

One Potential Element of a Quality Framework from the FCSM Working Group

- **"Clarity** is the extent to which easily comprehensible metadata are available, where these metadata are necessary to give a full understanding of statistical data."

Next Steps Toward a Documentation Standard for Integrated Standard

- Standardization of metadata for output microdata files.
  - ➢ **Metadata Content**
  - – Metadata Format

Next Steps Toward a Documentation Standard: Getting Started

- How are data sources for integration identified?
- What is the original intended use of each source?
- When were they collected?
- Where are the source data located?
- Are the source data sample or universe?
- Are the source data structured or unstructured?

Next Steps Toward a Documentation Standard: Getting Started

- What steps are taken to harmonize data from multiple sources?
- How are items drawn from different sources selected (were quality control metrics applied)?
- What controlled classifications (e.g. NAICS, SOC, MSA, Agency glossary) are followed?

Next Steps Toward a Documentation Standard: Modelling

- What data are produced by the model?
- What data sources are used?
- Which variables are used from each source ?
- What analytic techniques are used in the model?
- What statistics are used to evaluate the modelling effort?

Next Steps Toward a Documentation Standard: Matching/Linking

- What data sources are used?
- Which variables are used from each source?
- Which variables are used for matching or linking?
- What software/analytic approaches are used?
- What statistics are used to evaluate the success and quality of the resulting data set (e.g., what is the match or link rate)?

Next Steps Toward a Documentation Standard: Processing and Quality

- What editing or imputation techniques were applied by the original data stewards?
- How is data quality evaluated in source data?
- Are data elements edited by source or after integration?
- Are the data cleaned and edited?
- What edits and cleaning procedures are used conducted?
- How is data quality evaluated in the integrated data?

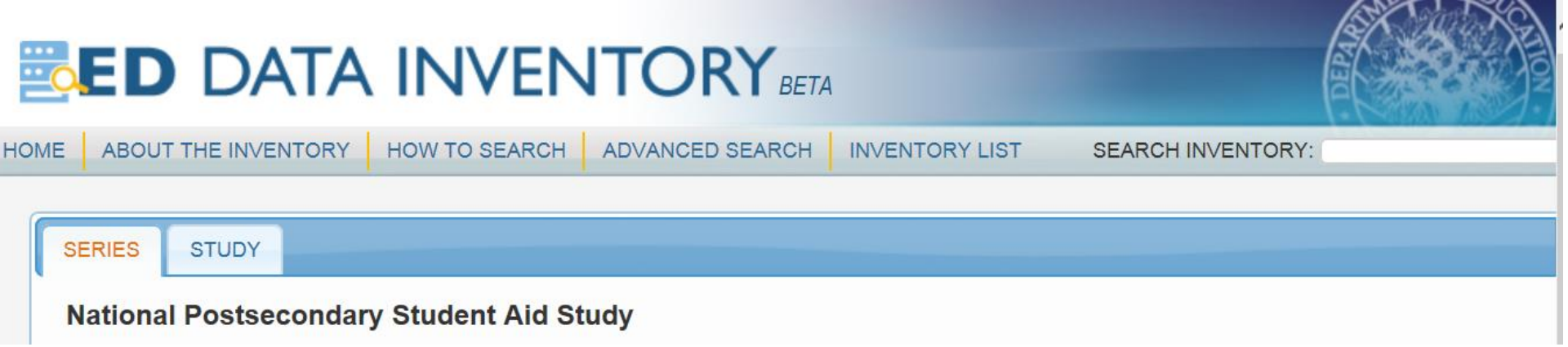
Next Steps Toward a Documentation Standard: Microdata files

- How were data from external sources identified?
- How are modelled data identified?
- Are metadata about data integration processes, outcomes, and assessments of data quality included?
- How are metadata accessed/ disseminated? (e.g. API's, JSON requests)

Thank you!

Marilyn.Seastrom@ed.gov

- Supplemental slides with an example of the content of documentation for one micro data file with integrated data from multiple sources.



- **NPSAS:1989–90—NPSAS:2016**
- Cross-sectional survey based on student-level records of students enrolled in a postsecondary institution
- Uses data from multiple sources
  - institutional records,
  - government databases, and
  - student interviews
- Provides reliable national estimates of characteristics related to financial aid for postsecondary students.

2015–16 National Postsecondary Student Aid Study (NPSAS:16) Data File Documentation

Student Lists from Institutions

- Counts from lists obtained for sampling were compared to counts from IPEDS (universe)
- Lists transmitted to VBA for matching to identify veterans for oversampling

2015–16 National Postsecondary Student Aid Study (NPSAS:16) Data File Documentation

Student Records from Institutions

- Description of collection procedures
- Outcomes:
  - # and % of institutions providing records by mode
  - # and % of institutions and students by control, level, and student types
- Quality: Student records reviewed for completeness

2015–16 National Postsecondary Student Aid Study (NPSAS:16) Data File Documentation

Student Administrative Data from ED and Data from External Sources

- FAFSA data from federal loan applications (ED Central Processing System)
- Data on loans and Pell Grants (National Student Loan Data System)
- Student enrollment in all institutions attended (National Student Clearinghouse)
- SAT/ACT admissions data–scores and survey
- VBA identified veterans and VA education benefits

2015–16 National Postsecondary Student Aid Study (NPSAS:16) Data File Documentation

Student Administrative Data from ED and Data from External Sources

Documentation included details on:

- Matching procedures for each sources
- Outcomes from matching for each source
- Editing described at the item level