Documentation and Archiving Metadata Practices and Needs

May 21, 2019

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 Data

- 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 and/or imputed by source or after integration?
  - If edited and/or imputed by source, how similar were the edits in each source?

Next Steps Toward a Documentation Standard: Processing and Quality

- Are the integrated data cleaned and edited?
- What edits and cleaning procedures are used ?
- Are the integrated data imputed?
- What imputation procedures are used?
- How is data quality evaluated in the integrated data?
- What additional parameters should be documented for unstructured data?

Next Steps Toward a Documentation Standard: Microdata files

- How were data from external sources identified on the data file?
- 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)

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

National Postsecondary Student Aid Study

- **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

Stage 1 sampling: Institution level

- Student counts from lists were compared to counts from IPEDS (universe)

Stage 2 sampling: Student Lists are obtained from institutions to build a sampling frame

- Lists transmitted to Veterans’ Benefits Administration for matching to identify Veterans for oversampling

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

Sampled Students' Records Obtained from Institutions

- Description of collection procedures
- Collection 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

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

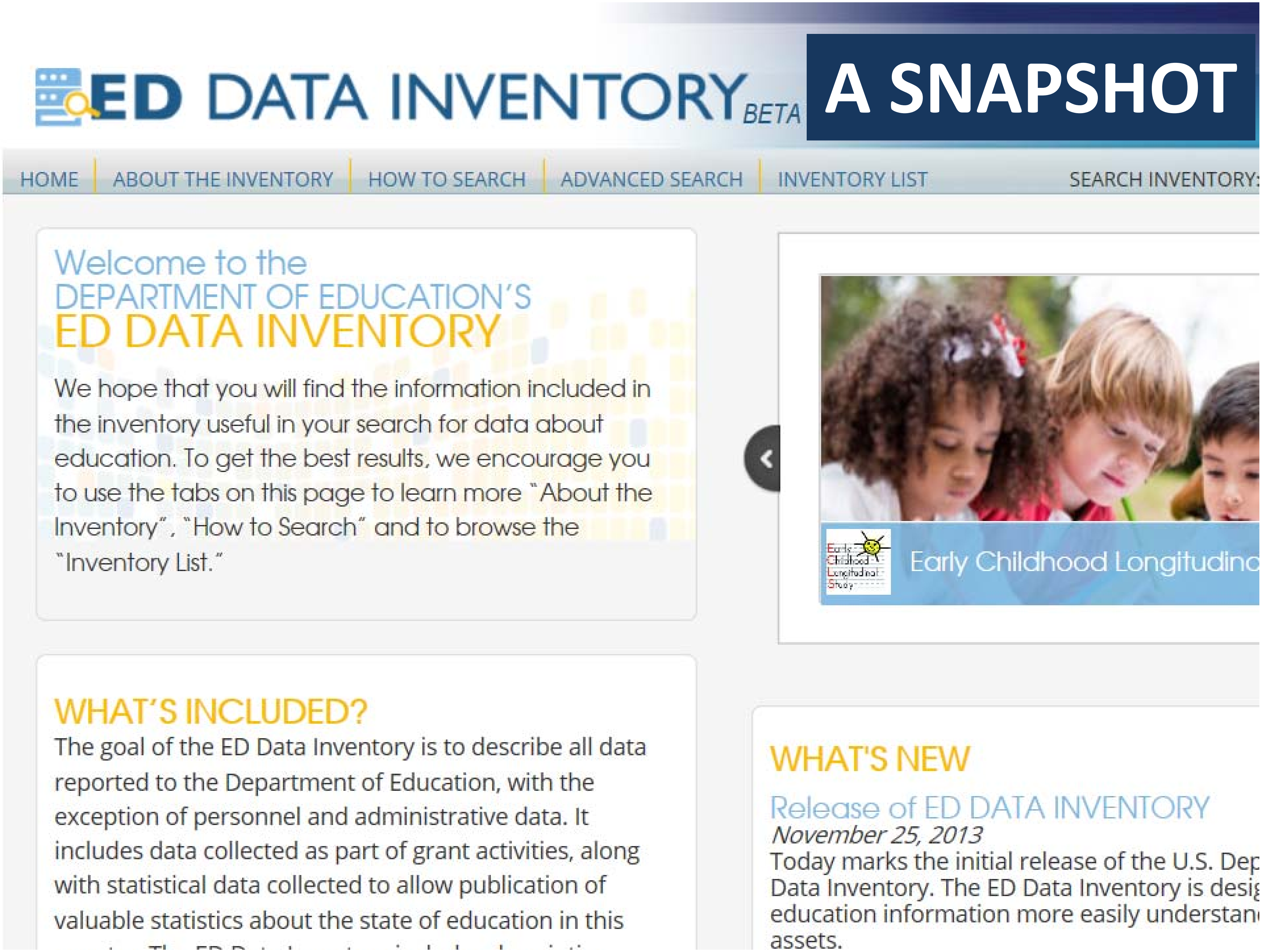
Documentation included details on:

- Matching procedures for each sources
- Outcomes from matching for each source

Student Interview Data Combined with Administrative Data from all Sources

- Editing described at the item level

A SNAPSHOT



Welcome to the
DEPARTMENT OF EDUCATION'S
ED DATA INVENTORY

We hope that you will find the information included in the inventory useful in your search for data about education. To get the best results, we encourage you to use the tabs on this page to learn more "About the Inventory", "How to Search" and to browse the "Inventory List."

WHAT'S INCLUDED?

The goal of the ED Data Inventory is to describe all data reported to the Department of Education, with the exception of personnel and administrative data. It includes data collected as part of grant activities, along with statistical data collected to allow publication of valuable statistics about the state of education in this

WHAT'S NEW

Release of ED DATA INVENTORY

November 25, 2013

Today marks the initial release of the U.S. Dep
Data Inventory. The ED Data Inventory is desig
education information more easily understan
assets.

WHAT IS THE ED DATA INVENTORY?

- It describes administrative and statistical data assembled and maintained by the Department.
- It includes
  - descriptive information (metadata) about each data collection and
  - information on the specific data elements in individual data collections.

WHY WAS THE INVENTORY CREATED?

- ✓ To improve the coordination of data collections across program offices,
- ✓ To minimize respondent burden,
- ✓ To ensure responsible data management at ED,
- ✓ To achieve data transparency with the public about the data that ED collects and maintains.
- ✓ To comply with OMB Information Quality Act Directive M-13-13 (May 9, 2013); and more recently, OMB Directive M-19-15 (April 24, 2019)

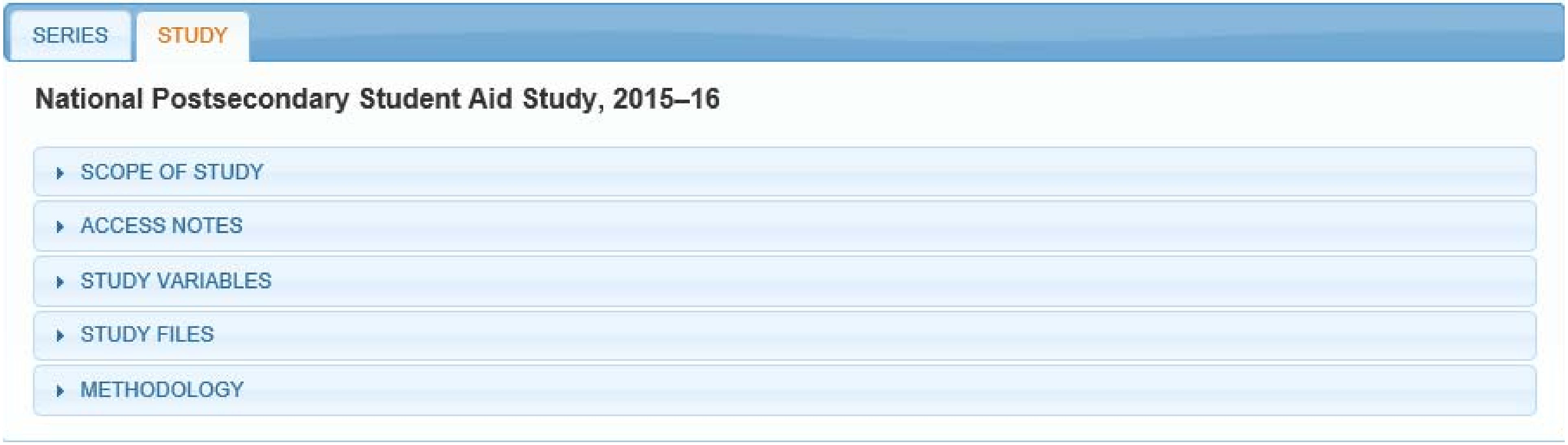
HOW IS THE INVENTORY ORGANIZED?

- A **Series** is a collection of **studies** that are repeated over time (e.g., Series: National Postsecondary Student Aid Study (NPSAS); Study: NPSAS: 16)
- Search for data files or data elements at the series or study level
- Link to study home page and link to data file
- Search for information on **Variables** and files by series or study

WHAT IS INCLUDED FOR VARIABLES?

- Information on **Variables** includes
  - Variable Name, Label, Extended Definition, Value Labels, File Name
  - Select variables of interest or select all variables in a file
  - Export information on selected variables to a csv. file for use in a software package

WHAT IS INCLUDED FOR A STUDY?



National Postsecondary Student Aid Study, 2015–16

- SCOPE OF STUDY
- ACCESS NOTES
- STUDY VARIABLES
- STUDY FILES
- METHODOLOGY

SCOPE OF STUDY

- Abbreviated title, Investigator
- Study Summary, Series Name, URL
- Geography, Date of collection, Periodicity,
- Study Population, Data Type, Purpose
- Age range, education level
- SORN Number, SORN URL, Authorizing Law

ACCESS NOTES

- Initial Reports and URL
- Date of Public Use Data Availability and URL
- Date of Restricted Use Data Availability and URL
- Public Access Level
- Contact Name and e-mail

METHODOLOGY

- Response Rates
- Respondent description
- Data type
- Confidentiality pledge
- Universe/sample size
- Mode of data collection

WHAT IS INCLUDED FOR STUDY INTERVIEWS?

- Methodology by interview
  - Response Rates
  - Respondent description
  - Data type
  - Confidentiality pledge
  - Universe/sample size
  - Mode of data collection

WHAT IS THE PATH FORWARD?

- Institutionalize ongoing data entry as a regular component in the ED data management process
- Drawing information from the OMB Information Clearance Request gets 95% of the metadata in the inventory
- Identify a mechanism to link the remaining 5% of the metadata about the study and the data elements to the data release

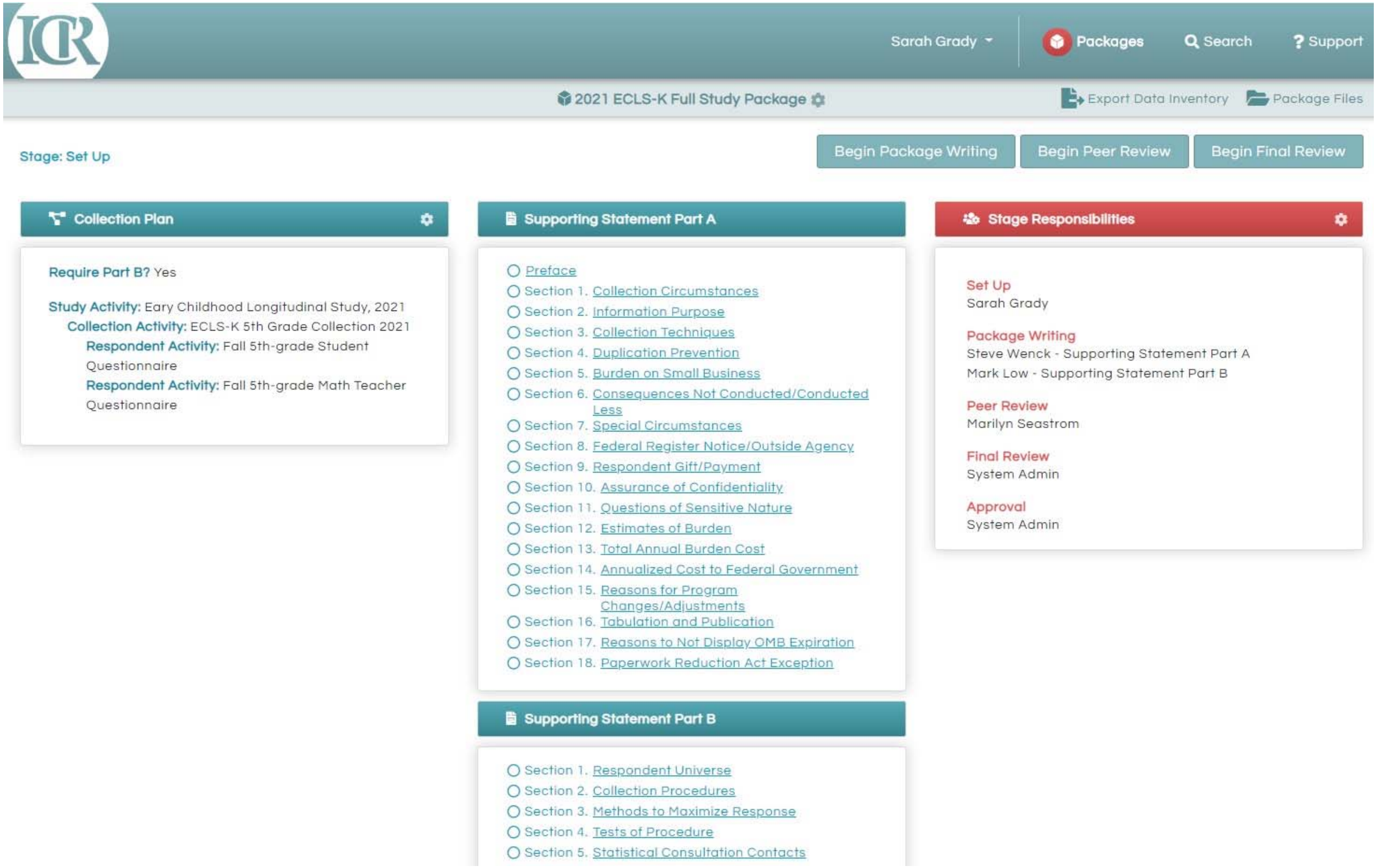
Drawing information from the OMB Information Clearance Request (Phase 1)

- The Cross-Agency Priorities Team provided funds for ED to develop an electronic template of fixed and open fields that:
  - Produces the OMB ICR
  - Internally tags the metadata entries
  - Extracts tagged fields after the ICR is approved
  - Extracted information electronically updates the ED Inventory

Drawing information from the OMB Information Clearance Request (Phases 1-4)

- Next Steps: Test ICR Template in NCES, IES, and another office in ED
- The Cross-Agency Priorities Team is providing funds for ED to
  - Adapt and test electronic ICR template in another Department
  - Provide documentation for government-wide implementation

ICR Template Components: Part A and Part B



Stage: Set Up

Collection Plan

Require Part B? Yes

Study Activity: Eary Childhood Longitudinal Study, 2021
Collection Activity: ECLS-K 5th Grade Collection 2021
Respondent Activity: Fall 5th-grade Student Questionnaire
Respondent Activity: Fall 5th-grade Math Teacher Questionnaire

Supporting Statement Part A

○ Preface
○ Section 1. Collection Circumstances
○ Section 2. Information Purpose
○ Section 3. Collection Techniques
○ Section 4. Duplication Prevention
○ Section 5. Burden on Small Business
○ Section 6. Consequences Not Conducted/Conducted Less
○ Section 7. Special Circumstances
○ Section 8. Federal Register Notice/Outside Agency
○ Section 9. Respondent Gift/Payment
○ Section 10. Assurance of Confidentiality
○ Section 11. Questions of Sensitive Nature
○ Section 12. Estimates of Burden
○ Section 13. Total Annual Burden Cost
○ Section 14. Annualized Cost to Federal Government
○ Section 15. Reasons for Program Changes/Adjustments
○ Section 16. Tabulation and Publication
○ Section 17. Reasons to Not Display OMB Expiration
○ Section 18. Paperwork Reduction Act Exception

Supporting Statement Part B

○ Section 1. Respondent Universe
○ Section 2. Collection Procedures
○ Section 3. Methods to Maximize Response
○ Section 4. Tests of Procedure
○ Section 5. Statistical Consultation Contacts

Stage Responsibilities

Set Up
Sarah Grady

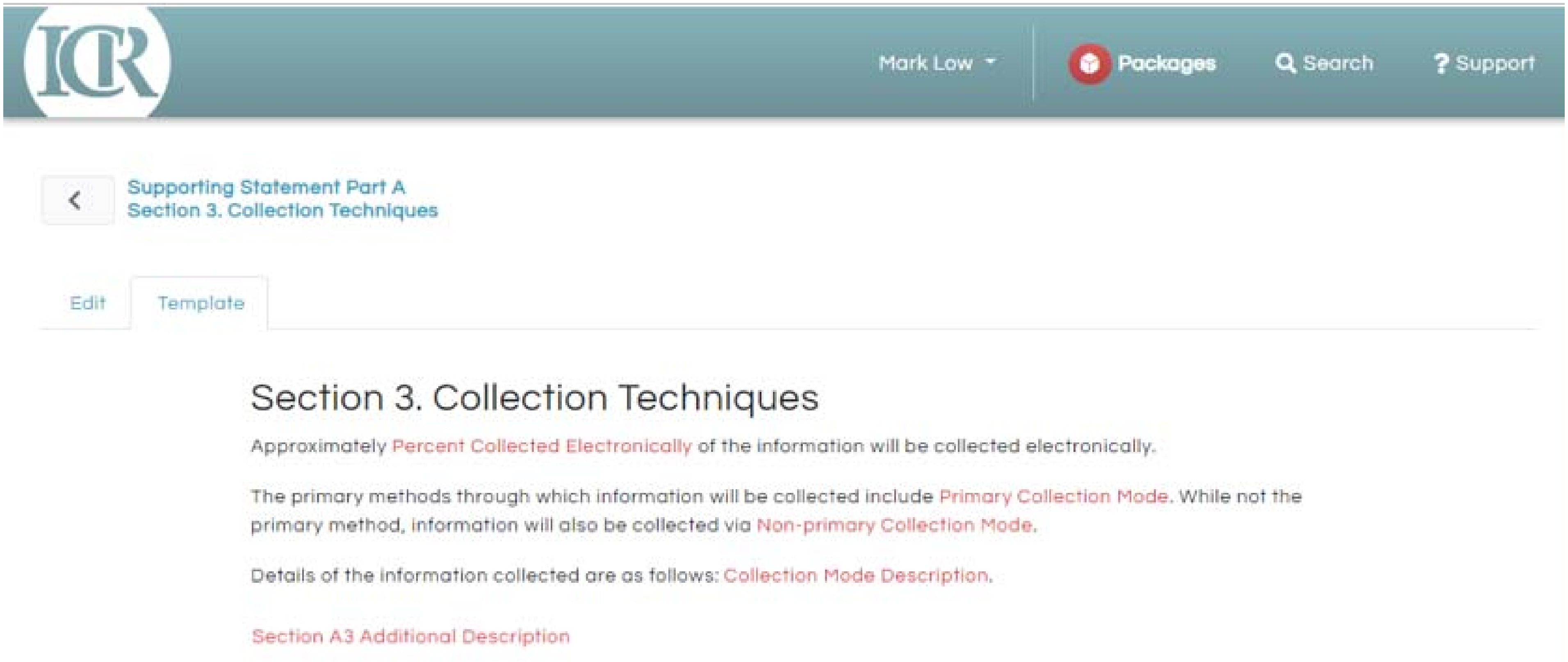
Package Writing
Steve Wenck - Supporting Statement Part A
Mark Low - Supporting Statement Part B

Peer Review
Marilyn Seastrom

Final Review
System Admin

Approval
System Admin

ICR Template Supporting Statement Part A, Section 3, Collection Techniques



Section 3. Collection Techniques

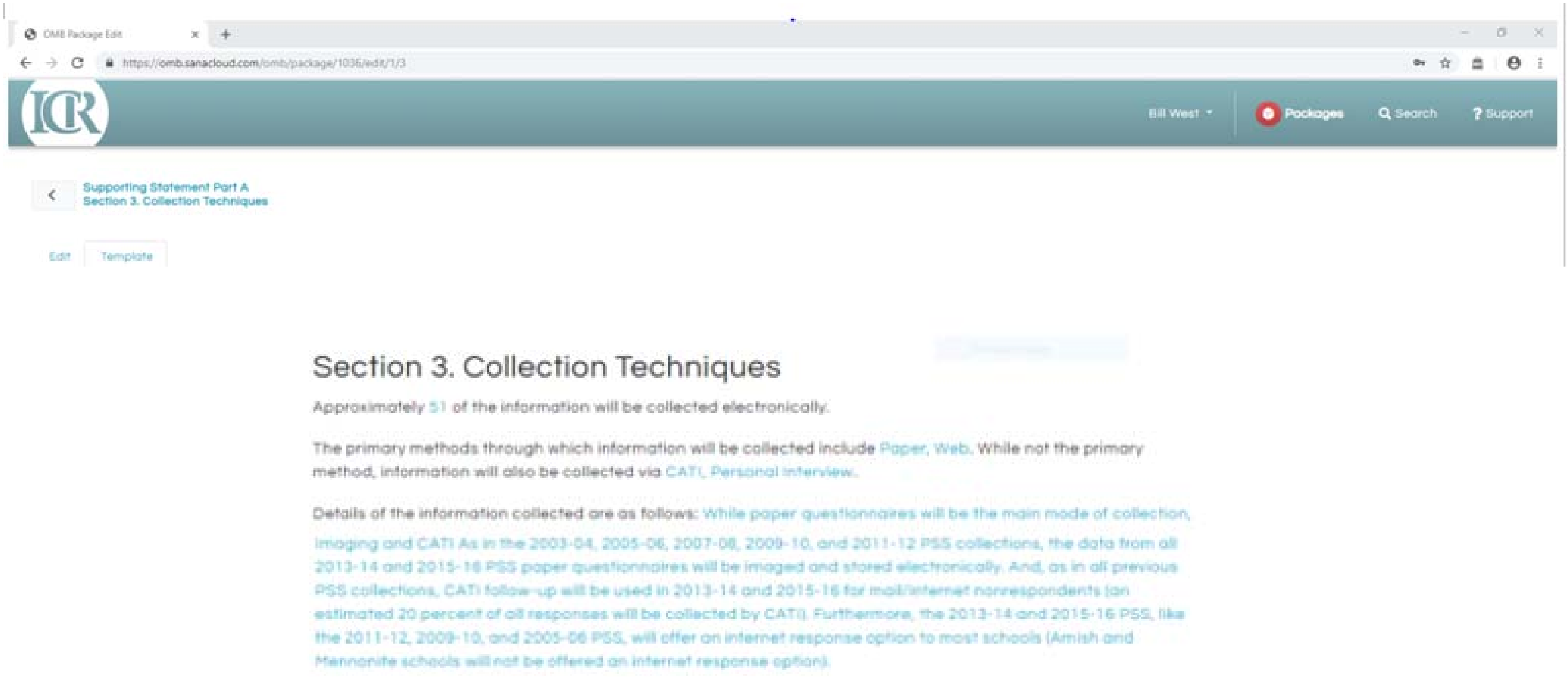
Approximately Percent Collected Electronically of the information will be collected electronically.

The primary methods through which information will be collected include Primary Collection Mode. While not the primary method, information will also be collected via Non-primary Collection Mode.

Details of the information collected are as follows: Collection Mode Description.

Section A3 Additional Description

ICR Template Supporting Statement Part A, Section 3, Collection Techniques


OMB Package Edit
https://omb.sanacloud.com/omb/package/1036/edit/1/3
ICR
Bill West
Packages
Search
? Support
Supporting Statement Part A
Section 3. Collection Techniques
Edit
Template

Section 3. Collection Techniques

Approximately 51 of the information will be collected electronically.

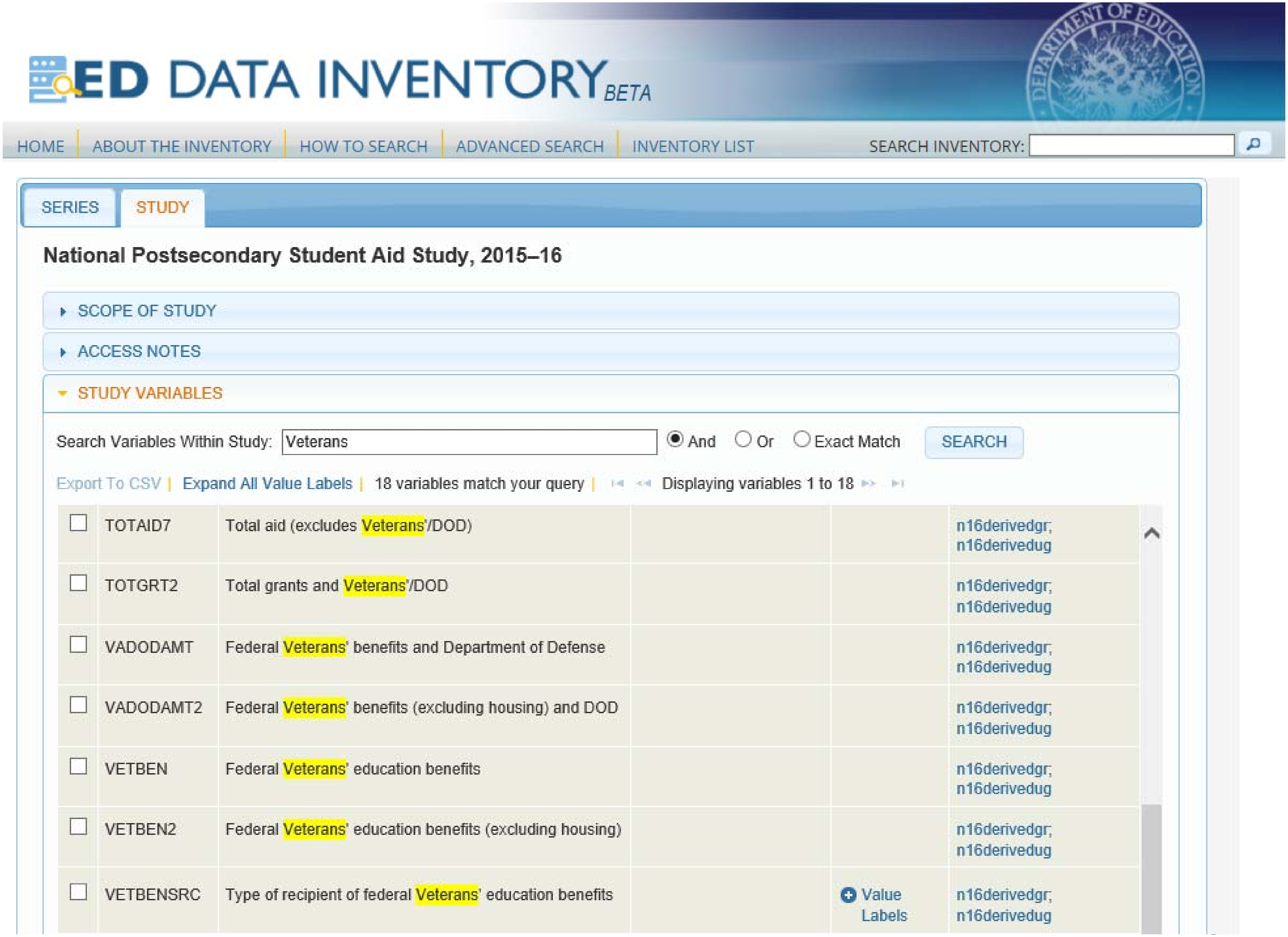
The primary methods through which information will be collected include Paper, Web. While not the primary method, information will also be collected via CATI, Personal interview.

Details of the information collected are as follows: While paper questionnaires will be the main mode of collection, imaging and CATI As in the 2003-04, 2005-06, 2007-08, 2009-10, and 2011-12 PSS collections, the data from all 2013-14 and 2015-16 PSS paper questionnaires will be imaged and stored electronically. And, as in all previous PSS collections, CATI follow-up will be used in 2013-14 and 2015-16 for mail/internet nonrespondents (an estimated 20 percent of all responses will be collected by CATI). Furthermore, the 2013-14 and 2015-16 PSS, like the 2011-12, 2009-10, and 2005-06 PSS, will offer an internet response option to most schools (Amish and Mennonite schools will not be offered an internet response option).

Thank You!

http://datainventory.ed.gov/

Marilyn.Seastrom@ed.gov

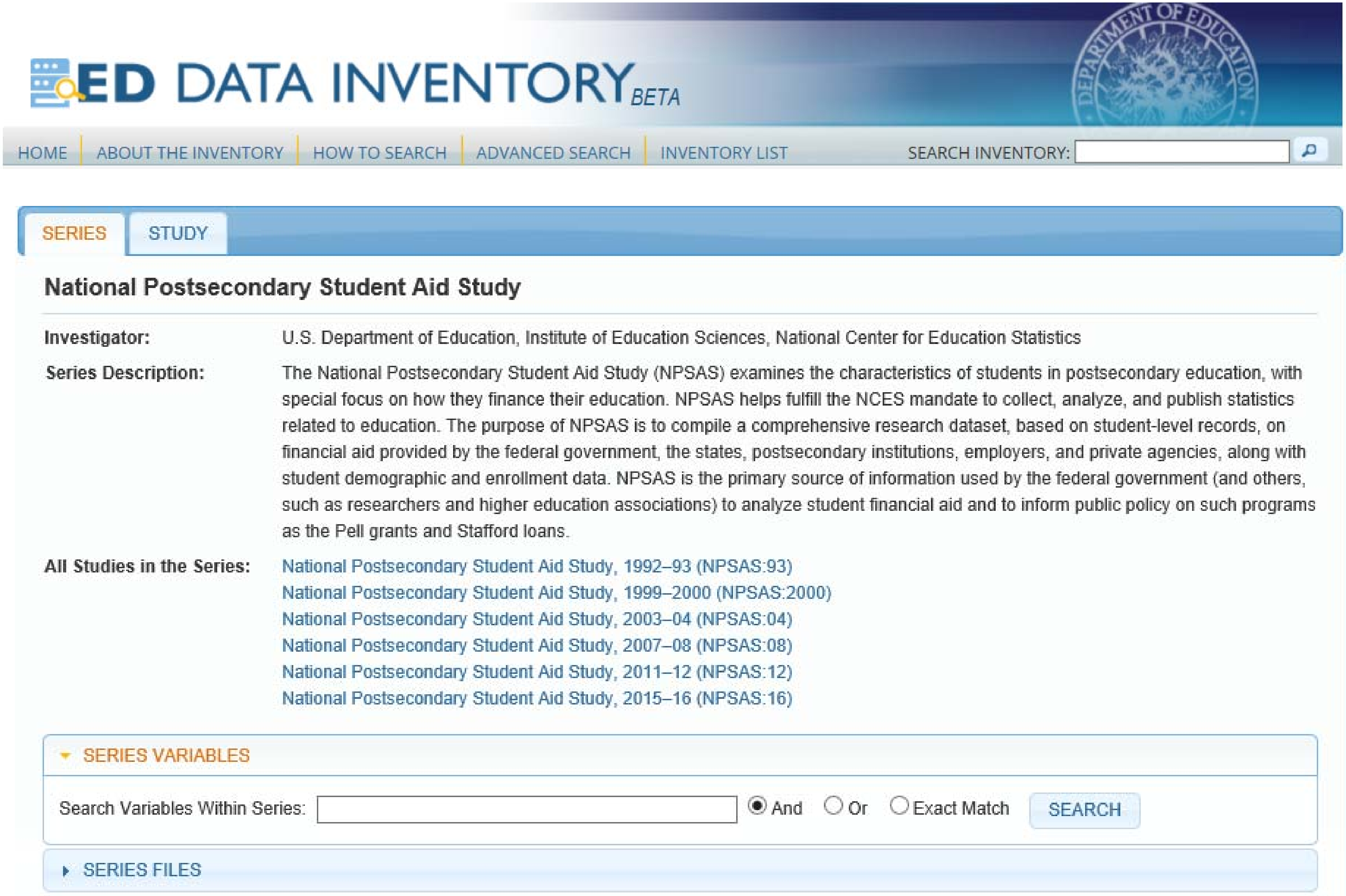


National Postsecondary Student Aid Study, 2015–16

Search Variables Within Study: Veterans ● And ○ Or ○ Exact Match SEARCH

Export To CSV | Expand All Value Labels | 18 variables match your query | Displaying variables 1 to 18

□	TOTAID7	Total aid (excludes Veterans'/DOD)			n16derivedgr; n16derivedug
□	TOTGRT2	Total grants and Veterans'/DOD			n16derivedgr; n16derivedug
□	VADODAMT	Federal Veterans' benefits and Department of Defense			n16derivedgr; n16derivedug
□	VADODAMT2	Federal Veterans' benefits (excluding housing) and DOD			n16derivedgr; n16derivedug
□	VETBEN	Federal Veterans' education benefits			n16derivedgr; n16derivedug
□	VETBEN2	Federal Veterans' education benefits (excluding housing)			n16derivedgr; n16derivedug
□	VETBENSRC	Type of recipient of federal Veterans' education benefits		Value Labels	n16derivedgr; n16derivedug



National Postsecondary Student Aid Study

Investigator: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics

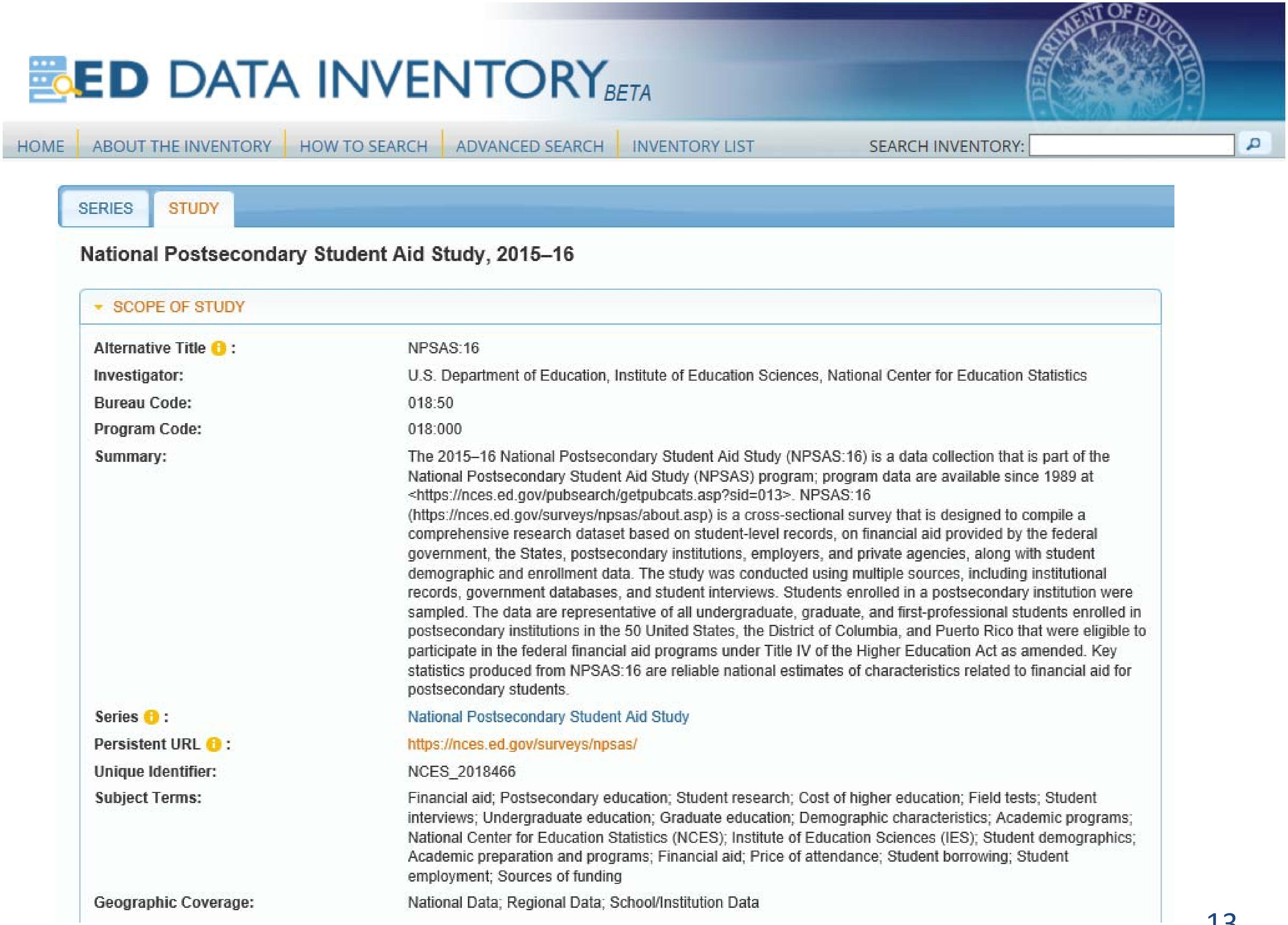
Series Description: The National Postsecondary Student Aid Study (NPSAS) examines the characteristics of students in postsecondary education, with special focus on how they finance their education. NPSAS helps fulfill the NCES mandate to collect, analyze, and publish statistics related to education. The purpose of NPSAS is to compile a comprehensive research dataset, based on student-level records, on financial aid provided by the federal government, the states, postsecondary institutions, employers, and private agencies, along with student demographic and enrollment data. NPSAS is the primary source of information used by the federal government (and others, such as researchers and higher education associations) to analyze student financial aid and to inform public policy on such programs as the Pell grants and Stafford loans.

All Studies in the Series:
National Postsecondary Student Aid Study, 1992–93 (NPSAS:93)
National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000)
National Postsecondary Student Aid Study, 2003–04 (NPSAS:04)
National Postsecondary Student Aid Study, 2007–08 (NPSAS:08)
National Postsecondary Student Aid Study, 2011–12 (NPSAS:12)
National Postsecondary Student Aid Study, 2015–16 (NPSAS:16)

▾ SERIES VARIABLES

Search Variables Within Series: ● And ○ Or ○ Exact Match

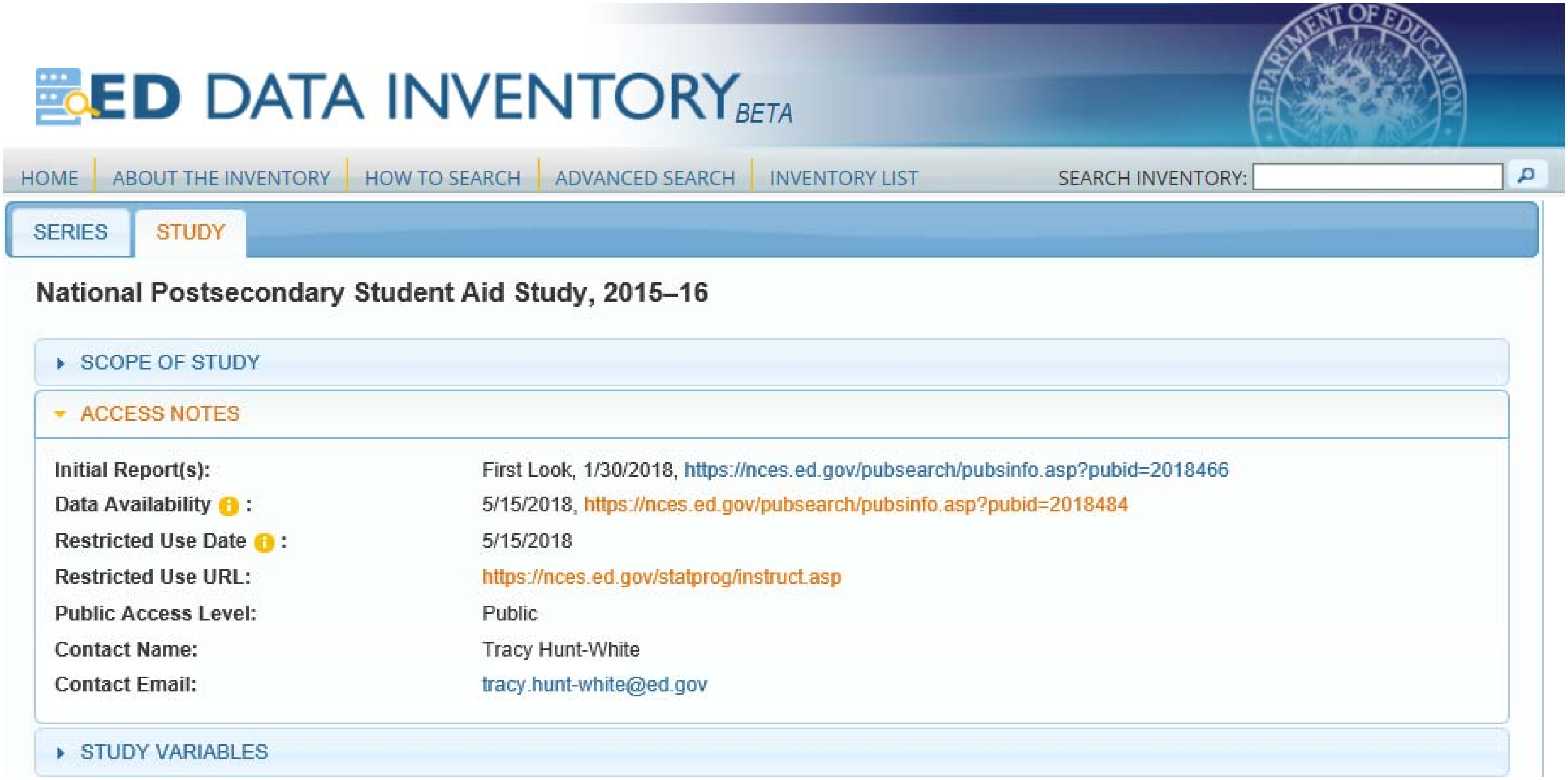
▸ SERIES FILES



National Postsecondary Student Aid Study, 2015–16

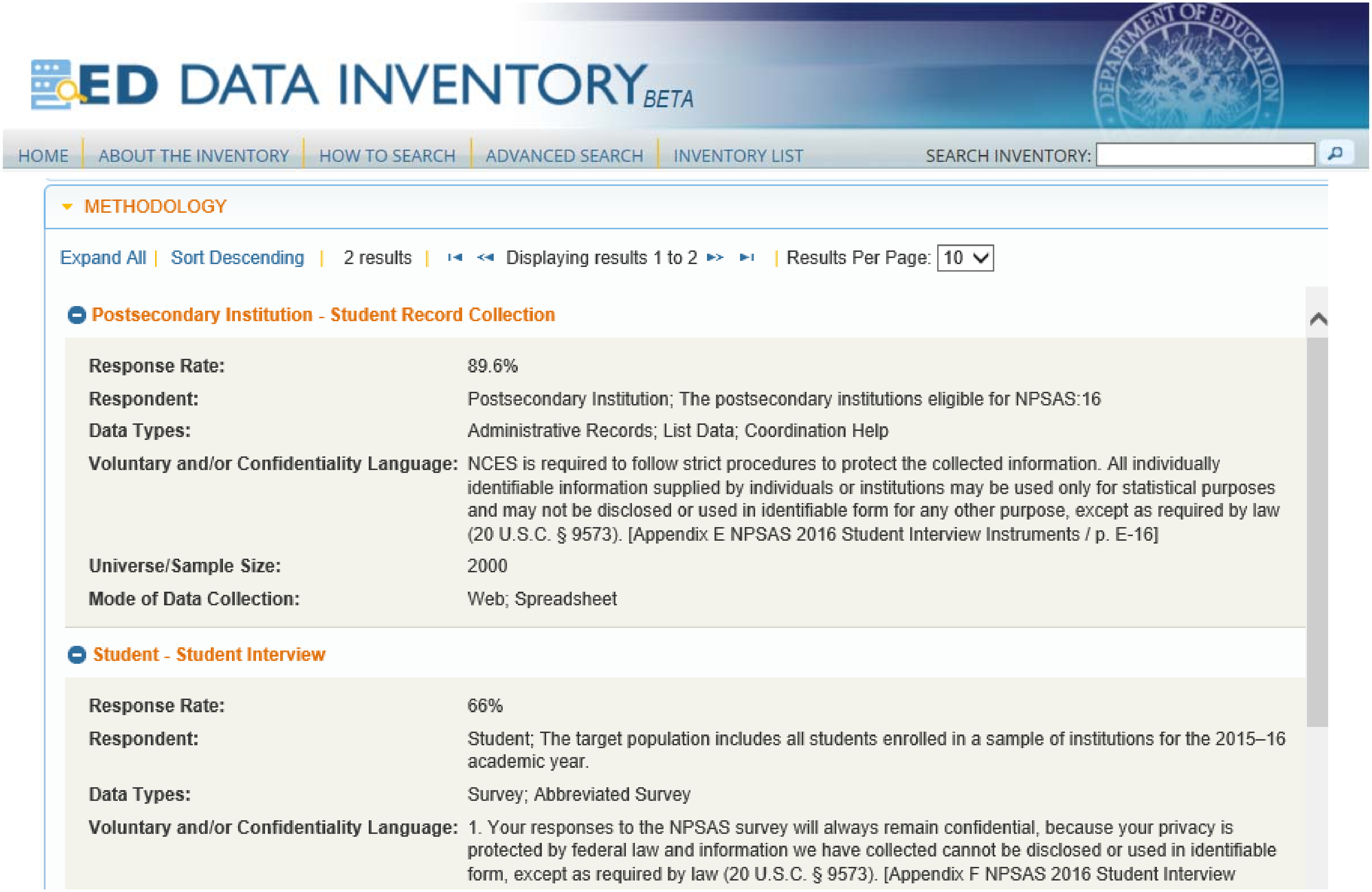
SCOPE OF STUDY

Alternative Title :	NPSAS:16
Investigator:	U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics
Bureau Code:	018:50
Program Code:	018:000
Summary:	The 2015–16 National Postsecondary Student Aid Study (NPSAS:16) is a data collection that is part of the National Postsecondary Student Aid Study (NPSAS) program; program data are available since 1989 at <https://nces.ed.gov/pubsearch/getpubcats.asp?sid=013>. NPSAS:16 (https://nces.ed.gov/surveys/npsas/about.asp) is a cross-sectional survey that is designed to compile a comprehensive research dataset based on student-level records, on financial aid provided by the federal government, the States, postsecondary institutions, employers, and private agencies, along with student demographic and enrollment data. The study was conducted using multiple sources, including institutional records, government databases, and student interviews. Students enrolled in a postsecondary institution were sampled. The data are representative of all undergraduate, graduate, and first-professional students enrolled in postsecondary institutions in the 50 United States, the District of Columbia, and Puerto Rico that were eligible to participate in the federal financial aid programs under Title IV of the Higher Education Act as amended. Key statistics produced from NPSAS:16 are reliable national estimates of characteristics related to financial aid for postsecondary students.
Series :	National Postsecondary Student Aid Study
Persistent URL :	https://nces.ed.gov/surveys/npsas/
Unique Identifier:	NCES_2018466
Subject Terms:	Financial aid; Postsecondary education; Student research; Cost of higher education; Field tests; Student interviews; Undergraduate education; Graduate education; Demographic characteristics; Academic programs; National Center for Education Statistics (NCES); Institute of Education Sciences (IES); Student demographics; Academic preparation and programs; Financial aid; Price of attendance; Student borrowing; Student employment; Sources of funding
Geographic Coverage:	National Data; Regional Data; School/Institution Data



National Postsecondary Student Aid Study, 2015–16

Initial Report(s):	First Look, 1/30/2018, https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018466
Data Availability ⓘ :	5/15/2018, https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018484
Restricted Use Date ⓘ :	5/15/2018
Restricted Use URL:	https://nces.ed.gov/statprog/instruct.asp
Public Access Level:	Public
Contact Name:	Tracy Hunt-White
Contact Email:	tracy.hunt-white@ed.gov



Postsecondary Institution - Student Record Collection

Response Rate: 89.6%

Respondent: Postsecondary Institution; The postsecondary institutions eligible for NPSAS:16

Data Types: Administrative Records; List Data; Coordination Help

Voluntary and/or Confidentiality Language: NCES is required to follow strict procedures to protect the collected information. All individually identifiable information supplied by individuals or institutions may be used only for statistical purposes and may not be disclosed or used in identifiable form for any other purpose, except as required by law (20 U.S.C. § 9573). [Appendix E NPSAS 2016 Student Interview Instruments / p. E-16]

Universe/Sample Size: 2000

Mode of Data Collection: Web; Spreadsheet

Student - Student Interview

Response Rate: 66%

Respondent: Student; The target population includes all students enrolled in a sample of institutions for the 2015–16 academic year.

Data Types: Survey; Abbreviated Survey

Voluntary and/or Confidentiality Language: 1. Your responses to the NPSAS survey will always remain confidential, because your privacy is protected by federal law and information we have collected cannot be disclosed or used in identifiable form, except as required by law (20 U.S.C. § 9573). [Appendix F NPSAS 2016 Student Interview

Capturing the variables and remaining metadata

- Release of data file and documentation contingent upon
  - Submission of achieved response rates
  - Submission of information on variables
  - Date of data release