



# **Metadata Standards and Technology Development for the NSF Survey of Earned Doctorates**

Kimberly Noonan (NSF NCSES)

Pascal Heus (MTNA)

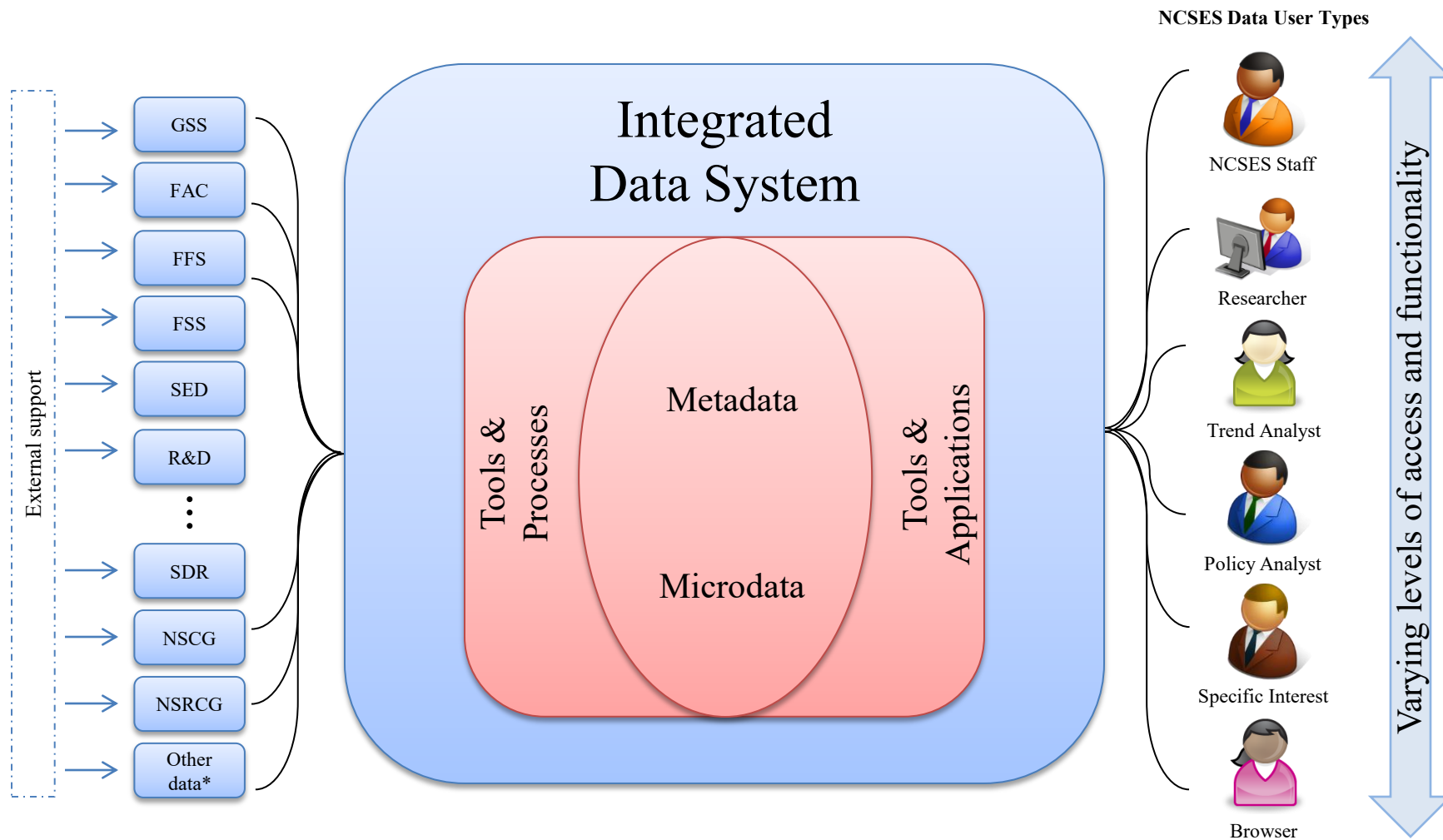
Tim Mulcahy (NORC)

November 5, 2013

# National Center for Science and Engineering Statistics (NCSES)

- **A federal statistical agency within NSF**
- **Charged with the mission to provide a central clearinghouse for the collection, interpretation, and analysis of data on scientific and engineering resources**
- **12 periodic data collections covering science and engineering**
  - **Research and Development**
  - **Education**
  - **Workforce**
- **Over 7 contracts for external support**
- **Building a central data system to store, maintain, and disseminate survey data in a faster, more flexible way**

# NCSES Data System



\* Other data used regularly in NCSES publications

# NCSES Data Delivery

- **Develop data delivery requirements for all survey microdata and metadata**
- **Ensure comprehensive documentation**
- **Standardize delivery formats**
- **Adopt metadata standards**
  - **Data Documentation Initiative (DDI)**
  - **Globally recommended practices**
  - **Industry standards and technologies**
- **Automate data processing**

# NSF Metadata Project: SED a case study

## Objectives

- Capture comprehensive survey metadata, in DDI format
- Automate generation of essential documentation, standard reports
- Generate delivery package compatible with the NCSES data systems
- Case study with Survey of Earned Doctorates (SED)
  - Extend to other NCSES surveys

## Team

- **NSF NCSES**
  - Building next generation data system and management framework
- **NORC SED team**
  - Survey contractor, years of survey specific knowledge
- **Metadata Technology North America (MTNA)**
  - Domain and technology experts in statistical data management and its challenges

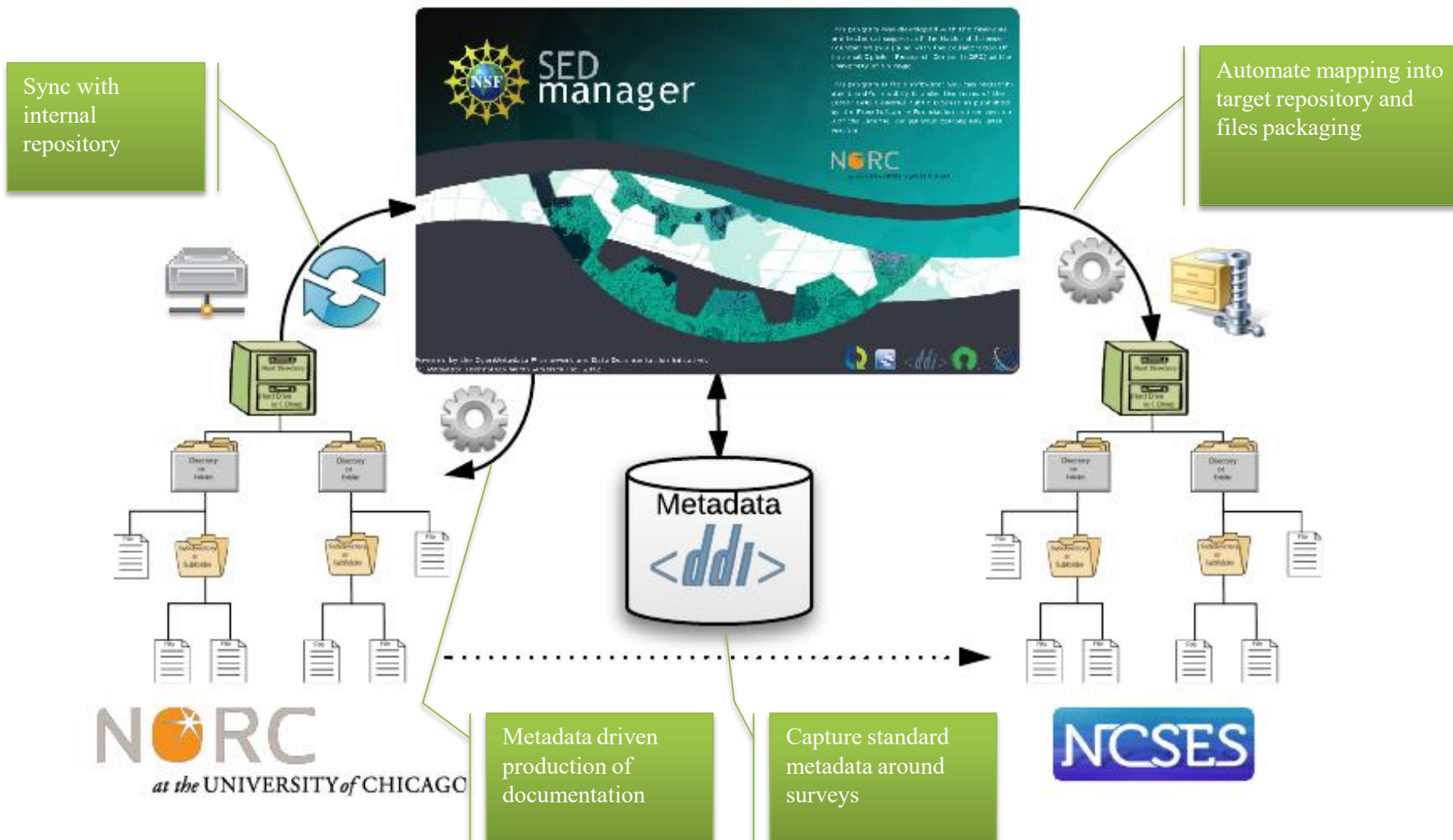
# NSF Survey of Earned Doctorates

- **Began in 1958**
- **Annual survey**
- **All individuals receiving research doctoral degrees from accredited U.S. Institutions**
- **Results used to assess characteristics and trends in doctorate education and degrees**
- **Survey is currently conducted by NORC**
- **NCSES disseminates data, reports and documentation**
- **SED sponsors**
  - National Science Foundation
  - National Institutes of Health
  - US Department of Agriculture
  - Department of Education
  - National Endowment for the Humanities
  - National Aeronautics & Space Administration

# SED Metadata Project Plan

- **Define SED metadata**
  - Assess current situation, file inventory
  - Capture comprehensive survey metadata
- **Develop SED metadata schema**
  - Develop metadata model
    - Based on metadata standards
    - Aligned with NCSES data system and dissemination needs
- **Prepare metadata for SED 2011**
  - Develop software tool, extend existing MTNA open source application
- **Automate SED metadata preparation for 2008, 2009, 2010**
  - Extend to additional survey cycles
- **Recommend maintenance and future steps**
  - Lessons learned
  - Next steps

# SED NSF Manager





# NSF SED Manager

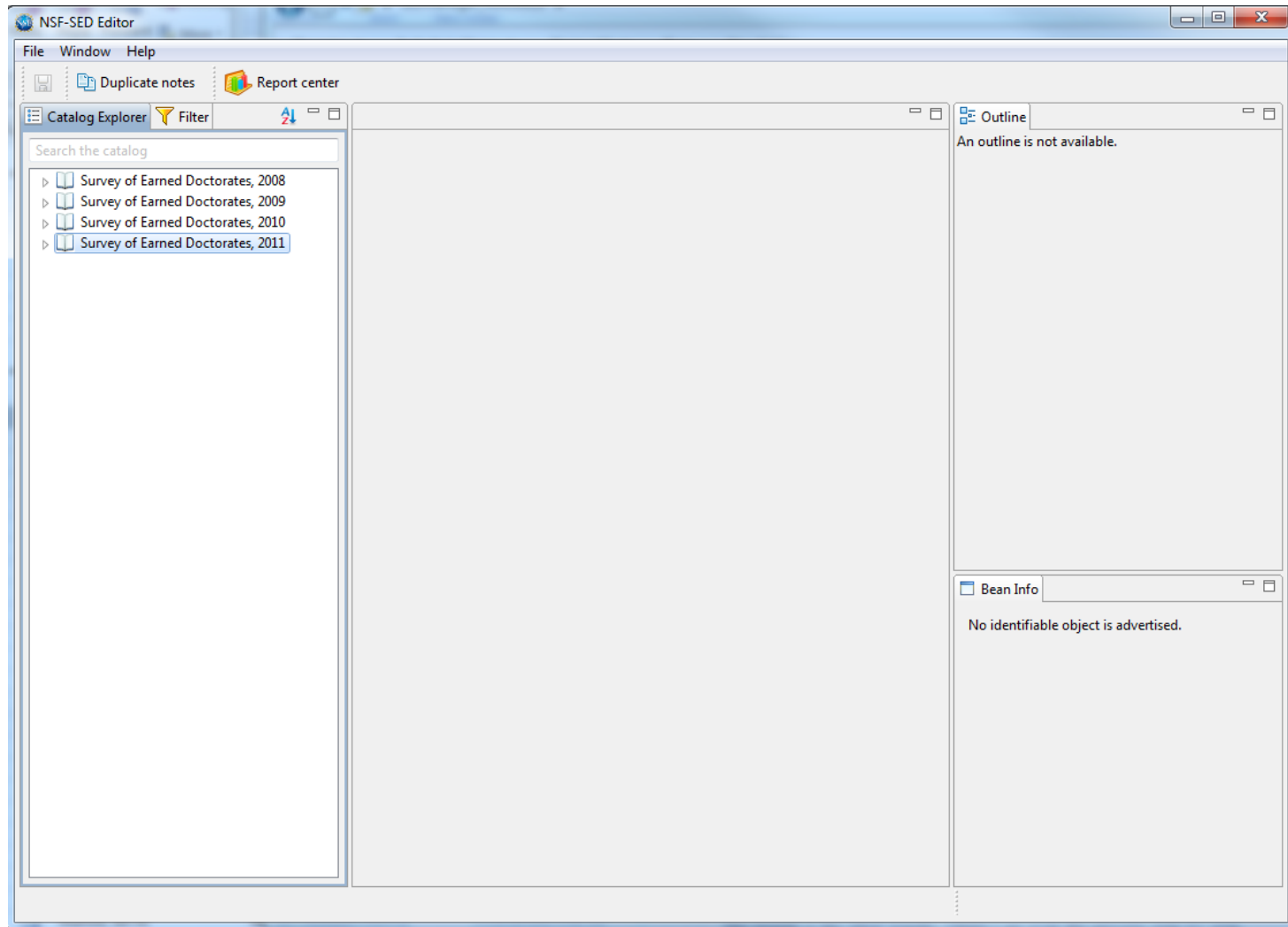
## Features:

- **Catalog Explorer**
  - Import/create survey
- **Metadata Editors**
  - Survey, Questionnaire, Classification, Variables, Data, Documentation, Notes,
  - Repository packaging module
- **Report Center**
  - Codebook
  - Comparison reports, e.g. codebook comparison
  - Custom reports

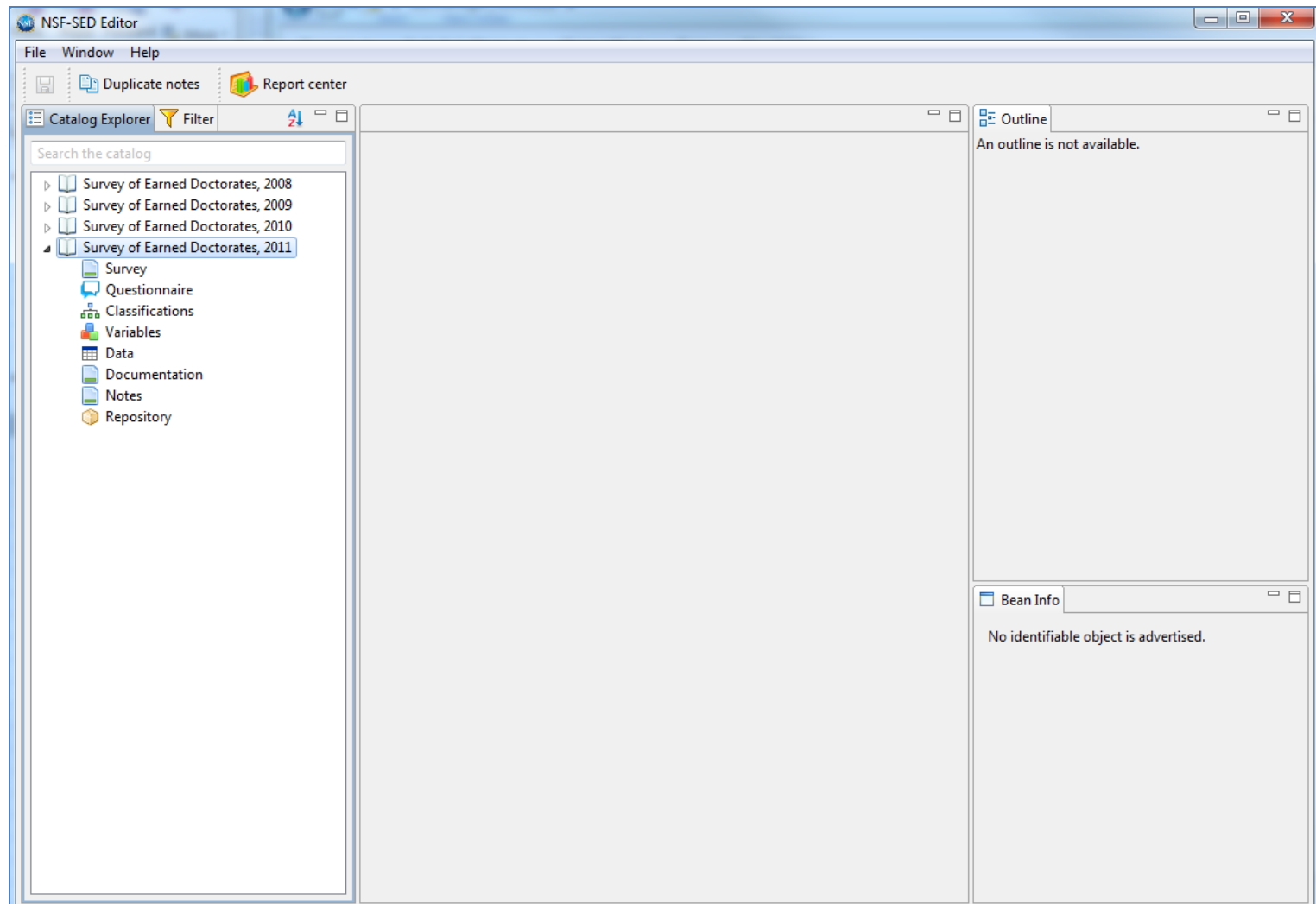
## Benefits:

- Metadata are standardized
- Metadata are DDI compliant
- Metadata are automatically captured

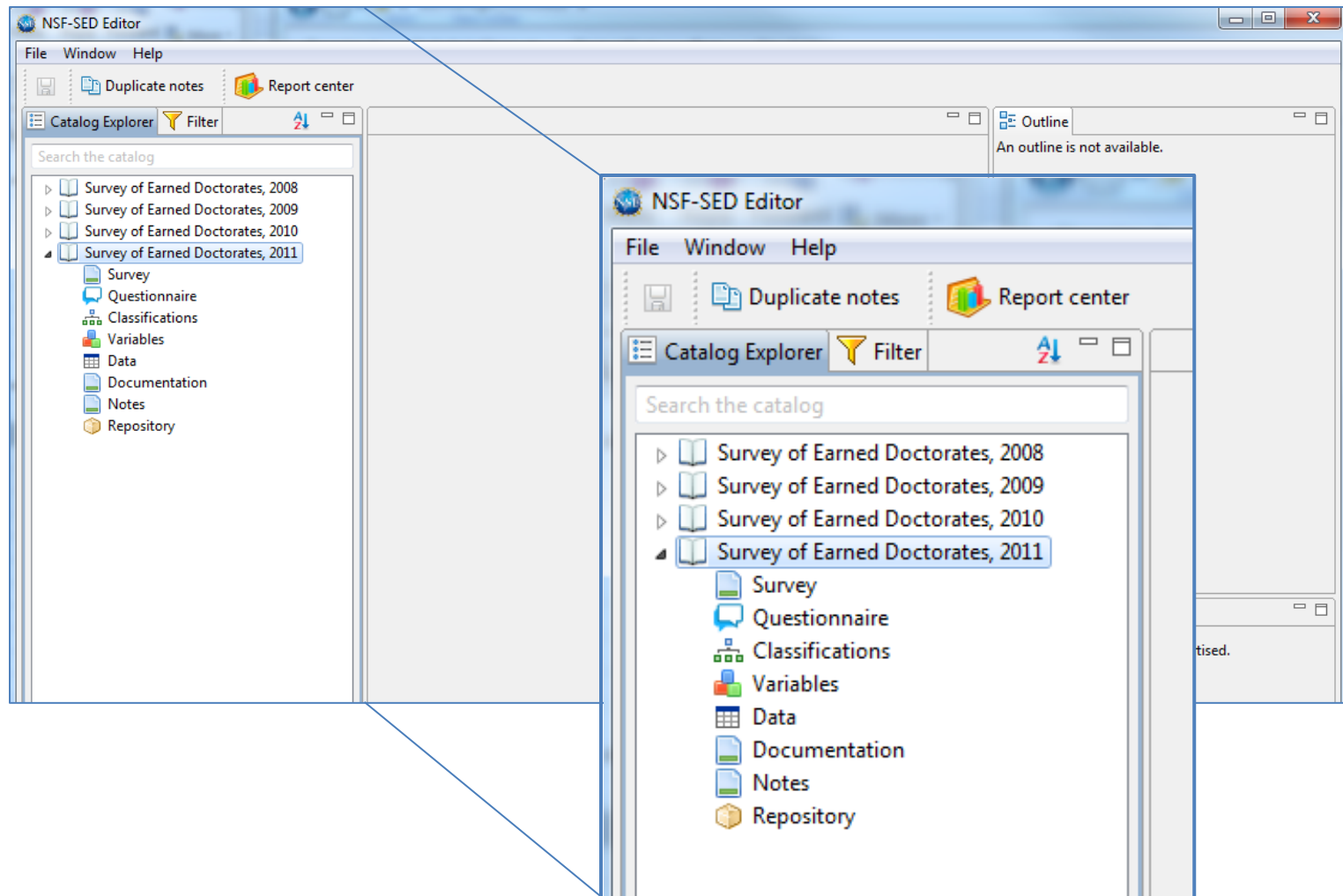
# Catalog Explorer



# Metadata Editors



# Metadata Editors



# Survey Editor

NSF-SED Editor

File Window Help

Duplicate notes Report center

Catalog Explorer Filter

Search the catalog

- ☐ Survey of Earned Doctorates, 2008
- ☐ Survey of Earned Doctorates, 2009
- ☐ Survey of Earned Doctorates, 2010
- ☒ Survey of Earned Doctorates, 2011
  - ☒ Survey
  - ☐ Questionnaire
  - ☐ Classifications
  - ☐ Variables
  - ☐ Data
  - ☐ Documentation
  - ☐ Notes
  - ☐ Repository

Study - NSF-SED 2011

Survey of Earned Doctorates, 2011

▼ Identification

Study Year: 2011

URN: urn:ddi:us.norc:StudyUnit.NSF\_SED\_2011.1.0.0

▼ Citation

English Title: Survey of Earned Doctorates, 2011

Abbreviation: NSF-SED 2011

Date: 2011

Creators:

Contributors:

Copyright:

Description:

▼ Abstract

Outline

- ☒ Identification
- ☒ Citation
- ☒ Abstract
- ☒ Purpose

Bean Info

No identifiable object is advertised.

# Variable Editor

NSF-SED Editor

File Window Help

Duplicate notes Report center

Catalog E... Filter

Variables - NSF-SED 2011

Survey of Earned Doctorates, 2011

Search: Columns LIST New variable

Showing 1 to 100 of 134 entries

Name	Label	DataType	Format	Notes	Question	Universe
DRF_ID	ID Number	Text		1		
PHDFY	Fiscal year of Doctorate	Double		3		
FORMIND	Form type indicator	Code	Code Format	3		
DOCCODE	Type of Doctorate	Code	Code Format	3	Type of Research Docto...	
PHDDISS	Dissertation field	Code	Code Format	2	Using the list on pages ...	
PHDDISS2	Secondary dissertation f...	Code	Code Format	3	If your dissertation rese...	
TUITREMS	Tuition remission - full ...	Code	Code Format	1	Did you receive full or p...	
SRCEPRIM	Primary source of supp...	Code	Code Format	2	Which TWO sources list...	
SRCE1ED	Edited primary source o...	Code	Code Format	2	Which TWO sources list...	
SRCESEC	Secondary source of su...	Code	Code Format	1	Which TWO sources list...	
SRCEA	Fellowship, scholarship	Code	Code Format	2	Which of the following ...	
SRCEB	Grant	Code	Code Format	2	Which of the following ...	

Page 1 of 2

First < 5 << Previous 1 Next >> 5 > Last

Show 100 / Page

Variable Representation Question Notes Concept/Universe Generation Instruction Files Summary Statistics Used By

Name: PHDDISS

Label: Dissertation field

☐ Is Time ☐ Is Geographic ☐ Is Weight

Description: Using the list...choose the code that best describes the primary field of your dissertation research. The three digit codes from the Specialties List (Exhibit E) are used to identify primary field of doctoral dissertation.

Reponse Unit:

Outline

Select all Unselect all

- ☐ Ungrouped (0)
- ☐ Section I: Identification (3 variables, 0 groups)
- ☐ Section II: Doctoral Degree (3 variables, 0 groups)
- ☐ Section III: Financial Support for Education (2 variables, 0 groups)
- ☐ Section IV: Postsecondary Educational History (2 variables, 0 groups)
- ☐ Section V: Postgraduate Plans (19 variables, 0 groups)
- ☐ Section VI: Background Information (demographic, 10 variables, 0 groups)
- ☐ Section VII: Response Information (3 variables, 0 groups)

Bean Info

URN:

# Variable Editor

Variable	Representation	Question	Notes	Concept/Universe	Generation Instruction	Files	Summary Statistics	Used By
Name:	<input type="text" value="PHDDISS"/>							
Label:	<input type="text" value="Dissertation field"/>							
	<input type="checkbox"/> Is Time <input type="checkbox"/> Is Geographic <input type="checkbox"/> Is Weight							
Description:	<input type="text" value="Using the list...choose the code that best describes the primary field of your dissertation research. The three digit codes from the Specialties List (Exhibit E) are used to identify primary field of doctoral dissertation."/>							
Reponse Unit:	<input type="text"/>							

# Questionnaire Editor

The screenshot displays the NSF-SED Editor software interface, specifically the Questionnaire Editor window. The interface is divided into several panes:

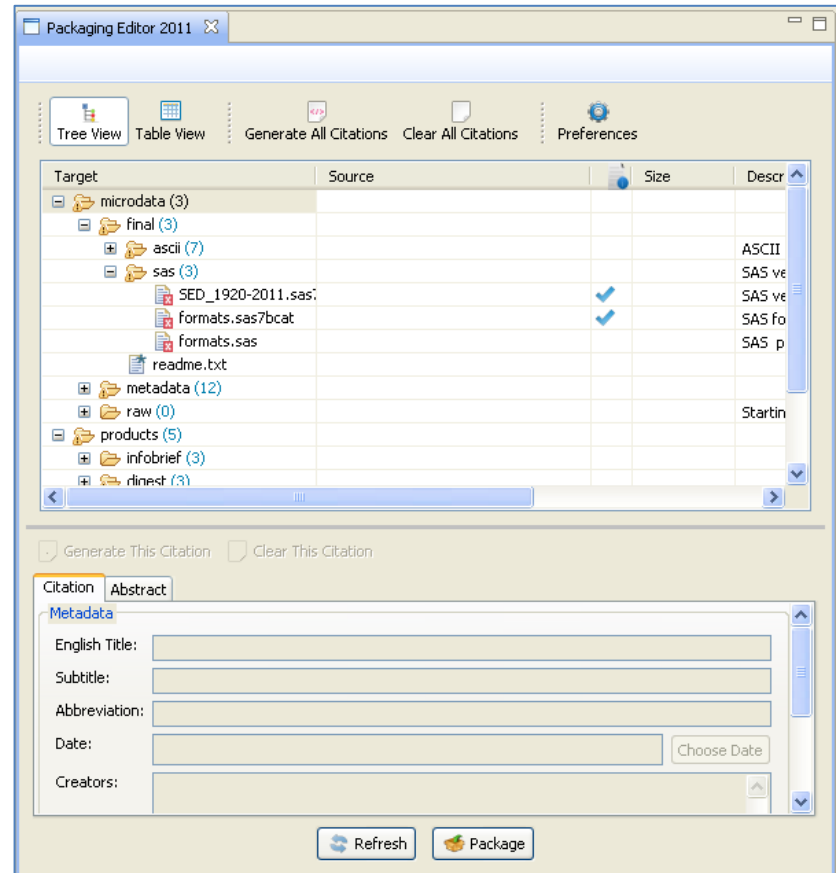
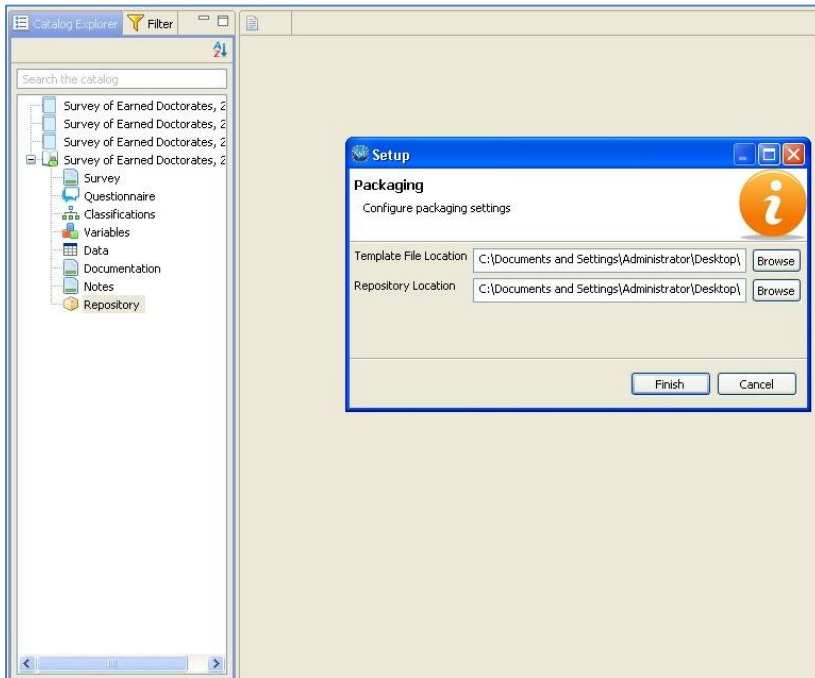
- Left Pane (Catalog):** Lists various data sources including "Survey of Earned Doctorates, 2011", "Survey of Earned Doctorates", "Survey of Earned Doctorates", "Survey of Earned Doctorates", "Survey", "Questionnaire", "Classifications", "Variables", "Data", "Documentation", "Notes", and "Repository".
- Top Pane (Questions - NSF-SED 2011):** Displays a list of questions for the "Survey of Earned Doctorates, 2011". The questions are grouped under "Questions" and "Groups". The "Questions" group is selected, showing a list of 88 entries. The "Groups" group is also visible, showing a list of 16 entries.
- Right Pane (Outline):** Provides a hierarchical outline of the questionnaire structure, including sections like "Part A - Education" and "Part B - Financial Support".
- Bottom Pane (Questionnaire Editor):** The main workspace for editing the questionnaire. It shows a table of questions and their associated variables. The table has columns for "Name", "Type", and "Label".

The "Questionnaire Editor" table lists the following variables and their labels:

Type	Name	Label
Variable	SRCEK	Spouse's, partner's, or family's ea.
Variable	SRCEE	Other assistantship
Variable	SRCEJ	Other source of support
Variable	SRCEI	Personal earnings during graduat.
Variable	SRCEL	Employer reimbursement/assista.
Variable	SRCEB	Personal savings
Variable	SRCEF	Grant
Variable	SRCEM	Traineeship
Variable	SRCEM	Foreign (non-U.S.) support



# Repository Editor



# Report Center

**Intro Page**  
Select study and output location for report(s)

Primary Study  
Survey of Earned Doctorates, 2010

Comparative Study(ies)

Study	Order
<input type="checkbox"/> Survey of Earned Doctorates, 2011	
<input type="checkbox"/> Survey of Earned Doctorates, 2009	
<input type="checkbox"/> Survey of Earned Doctorates, 2008	

Output Directory C:\Documents and Settings\Administrator\Desktop\SED\_REPO\2010\Reports

< Back **Next >** Finish Cancel

**Report Selection**  
Select reports to generate

- ☒ Survey Documentation
  - ☒ Codebook (HTML)
  - ☒ Variable Groups (HTML)
  - ☒ Record Layout (HTML)
  - ☒ Record Layout (PDF)
  - ☒ Special Data Flags and Values (HTML)
  - ☒ Classifications (HTML)
  - ☒ Questionnaire Outline (HTML)
- ☒ Quality Assurance
  - ☒ Notes: Valid Years (HTML)
  - ☒ Notes: Historical (HTML)
  - ☒ Questions (HTML)
- ☒ Comparison
  - ☒ Variable Comparison (HTML)

Select All Clear Selection

< Back **Next >** Finish Cancel

# Example Reports

## Codebook

DOCUMENTATION  
of the  
DOCTORATE RECORDS FILE  
1920 - 2011

November, 2012

Survey of Earned Doctorates  
National Science Foundation

Maintained by:  
NORC at the University of Chicago  
55 East Monroe, Suite 2000  
Chicago, IL 60603

NORC  
at the UNIVERSITY of CHICAGO

See Appendix C and the electronic crosswalk and institution labeling file provided with the dataset for more information.

**Valid Values:**  
Institution code (see Appendix D)  
(Blank) = "No" (no data/no information was reported or no baccalaureate degree was received (see BANCHE to distinguish))

Variable	Column	Column	Start	End
1	236	238		
2	239	241		
3	242	244		
4	245	247		
5	248	250		
6	251	252		
7	253	255		
8	256	258		
9	259	261		
10	262	264		
11	265	267		
12	268	269		
13	270	271		
14	272	274		
15	275	277		
16	278	279		
17	280	282		
18	283	285		
19	286	287		
20	288	289		
21	290	291		
22	292	293		
23	294	295		
24	296	298		
25	299	308		
26	307	309		
27	310	312		
28	313	315		
29	316	318		
30	319	321		
31	322	324		
32	325	327		
33	328	329		
34	330	331		
35	332	334		
36	335	337		
37	338	340		
38	341	343		
39	344	345		

## Comparison report

Variable Comparison - Windows Internet Explorer

C:\Documents and Settings\Administrator\Desktop\SED\_Repository\Report\_Center\Variable

File Edit View Favorites Tools Help

Web Search

Variable Comparison

### Variable Comparison

2011	DRF_ID	2010	DRF_ID
ID Number		ID Number	
index: 1 start: 1 end: 7 width: 7		index: 1 start: 1 end: 7 width: 7	
n/a			
Type: Character			

2011	PHDFY	2010	SRCEB
Fiscal year of Doctorate		Grant	
index: 2 start: 8 end: 11 width: 4		index: 12 start: 34 end: 36 width: 3	
n/a		n/a	
Type: Numeric		Which of the following were sources of financial support during graduate school?	
		Type: Numeric	
		1 Yes	1 Yes
		2 No	2 No
		Missing	Missing

2011	FORMIND	2010	SRCEC
Form type indicator		Teaching assistantship	
index: 3 start: 12 end: 13 width: 2		index: 13 start: 37 end: 39 width: 3	
n/a		n/a	
Type: Character		Which of the following were sources of financial support during graduate school?	
		Type: Numeric	
		1 Yes	1 Yes
		2 No	2 No
		Missing	Missing

2011	SRCED	2010	SRCED
Research assistantship		Research assistantship	
index: 14 start: 40 end: 42 width: 3		index: 14 start: 40 end: 42 width: 3	
n/a		n/a	
Type: Numeric		Which of the following were sources of financial support during graduate school?	
		Type: Numeric	
		1 Yes	1 Yes
		2 No	2 No
		Missing	Missing

Done

# Next Steps

- Production and maintenance mode
  - Enhance user manual
  - Apply tool to SED 2012
  - Apply minor fixes and enhancements
- Metadata driven environment
  - Currently data driven
  - Metadata must be considered earlier in process
  - Establish variable/classification/question banks
- Integrate in NCSES Data Repository
- Extend to other NCSES surveys
  - NSF SED Manager based on existing tool
  - Based on common framework, DDI with extensions
  - Open source

# Thank you!

Contact information:

Kimberly Noonan  
[knoonan@nsf.gov](mailto:knoonan@nsf.gov)

Pascal Heus  
[pascal.heus@metadatatechnology.com](mailto:pascal.heus@metadatatechnology.com)

Tim Mulcahy  
[Mulcahy-Tim@norc.org](mailto:Mulcahy-Tim@norc.org)