



NCSES



NORC
at the UNIVERSITY of CHICAGO

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

Kimberly Noonan (NSF NCSES)

Pascal Heus (MTNA)

Tim Mulcahy (NORC)

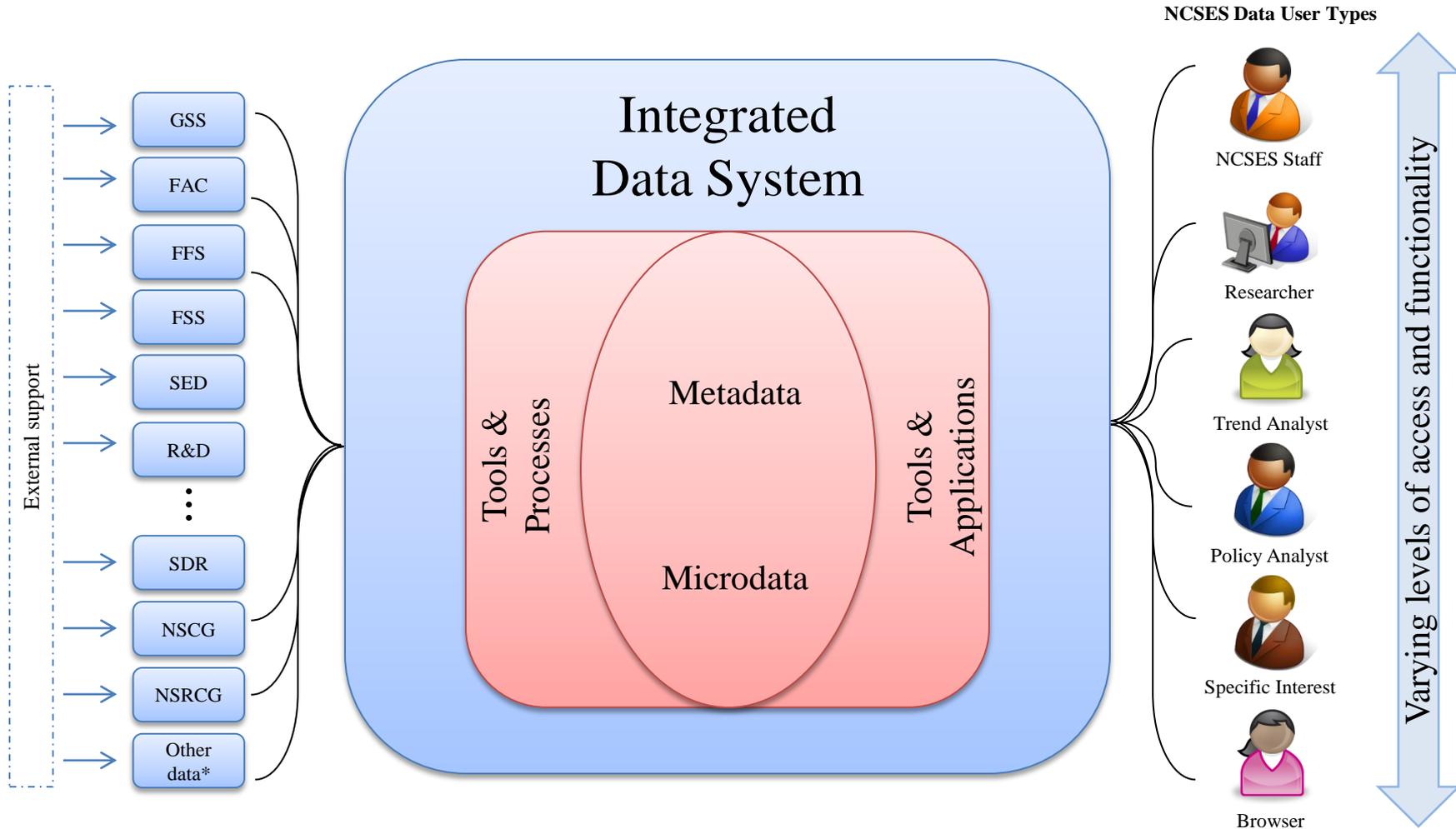
November 5, 2013



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



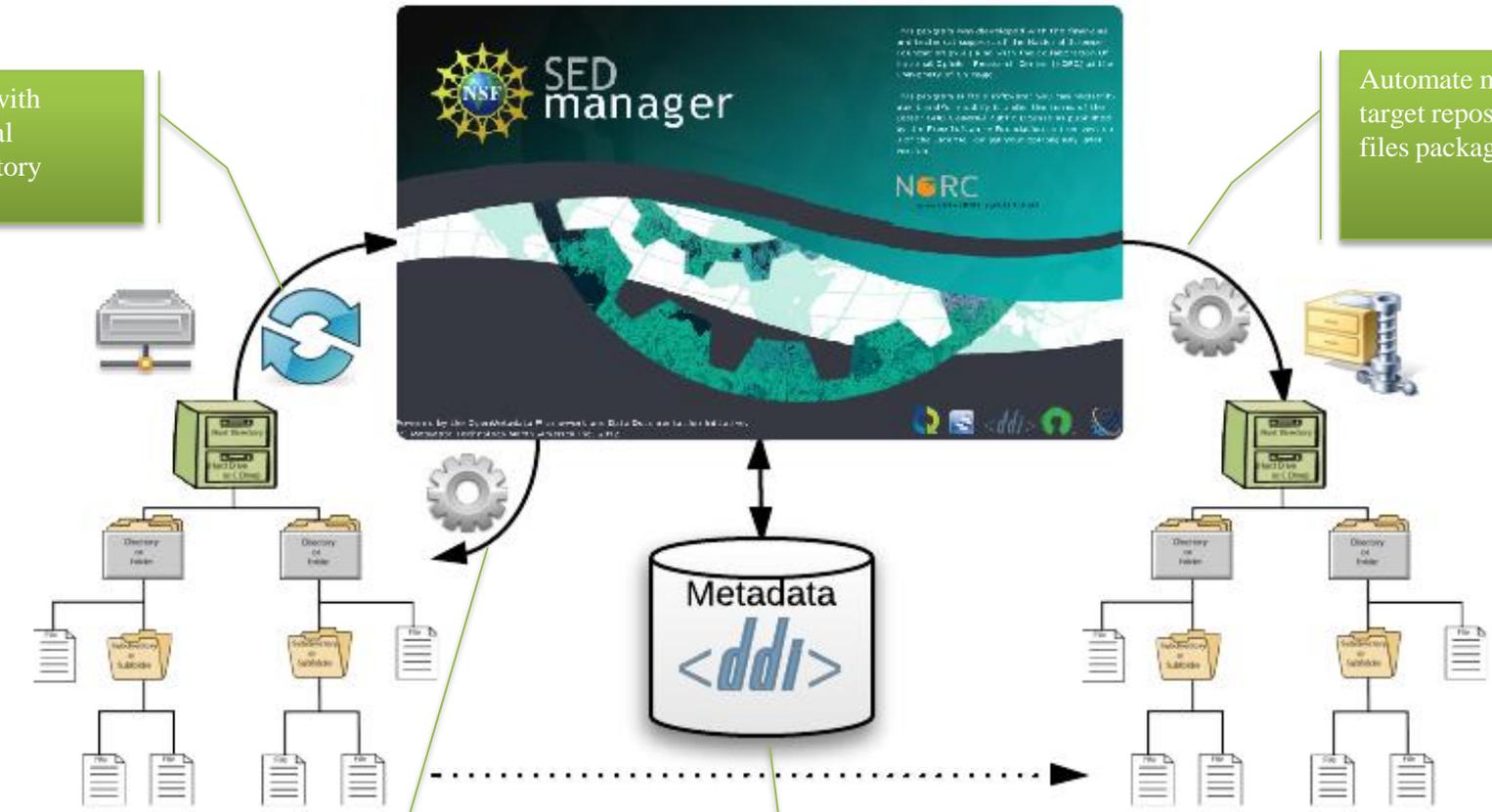
SED Metadata Project Plan

- **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 NCSSES 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

Sync with internal repository

Automate mapping into target repository and files packaging



Metadata driven production of documentation

Capture standard metadata around surveys

NSF SED Manager

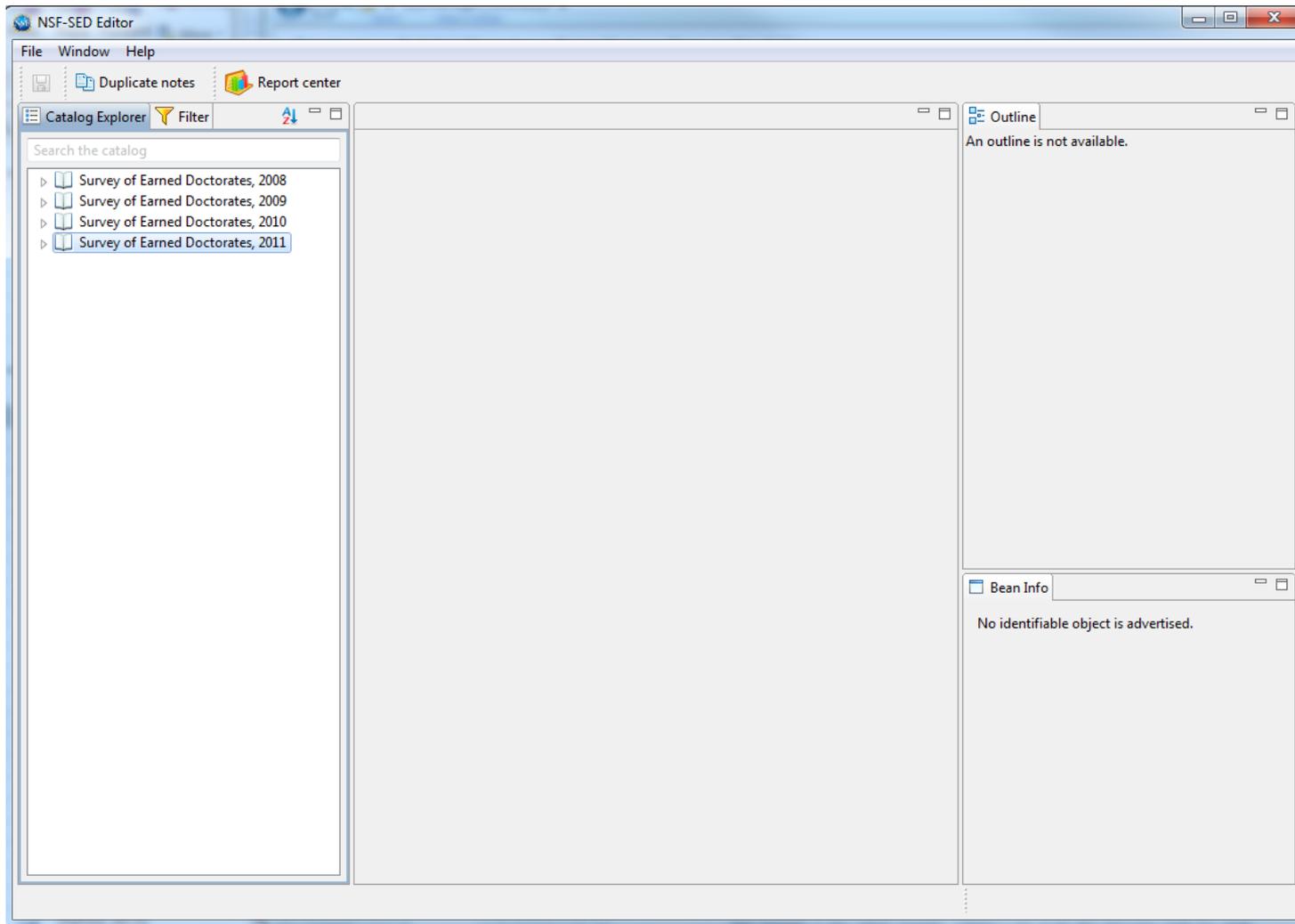
Features:

- **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

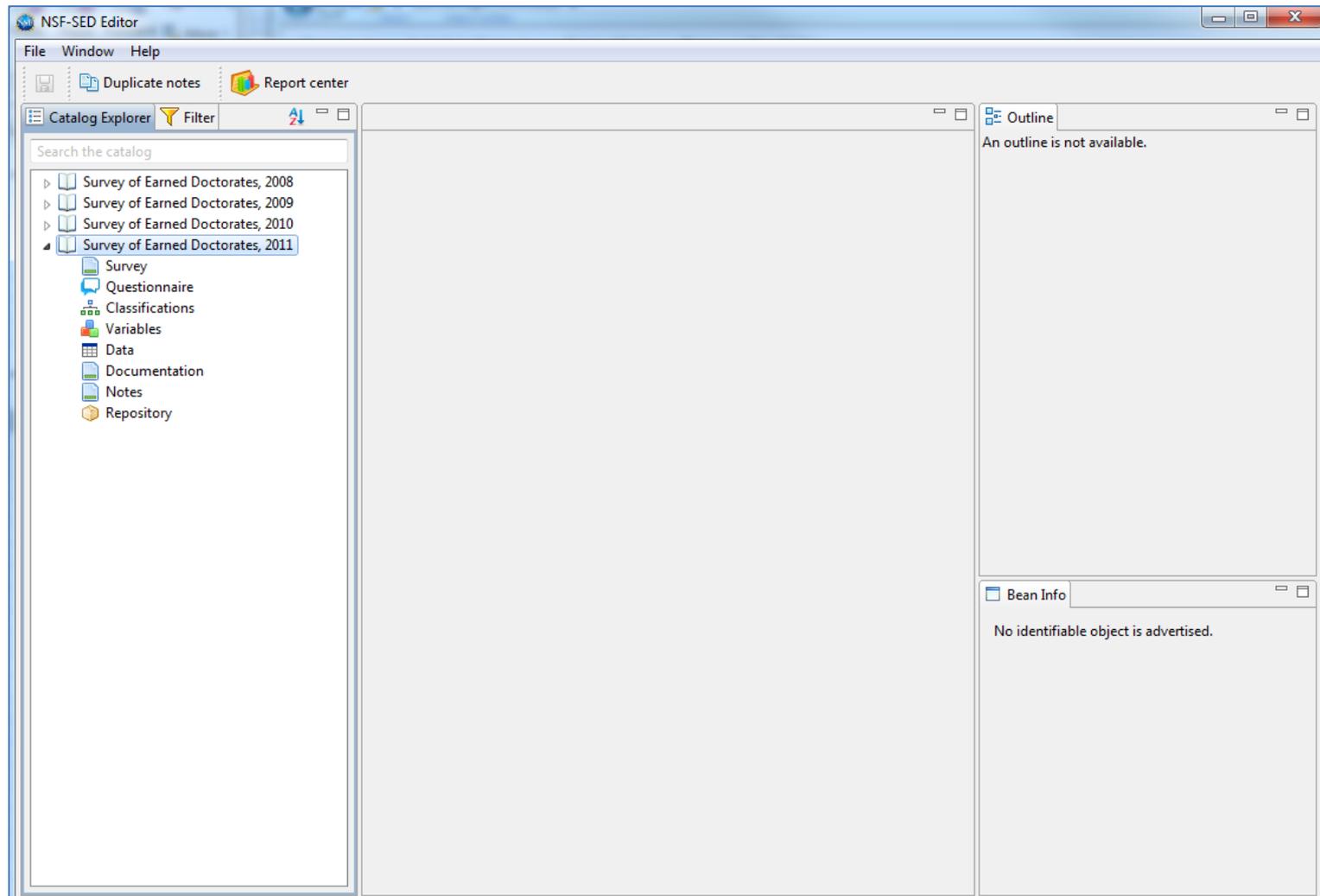
Benefits:

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

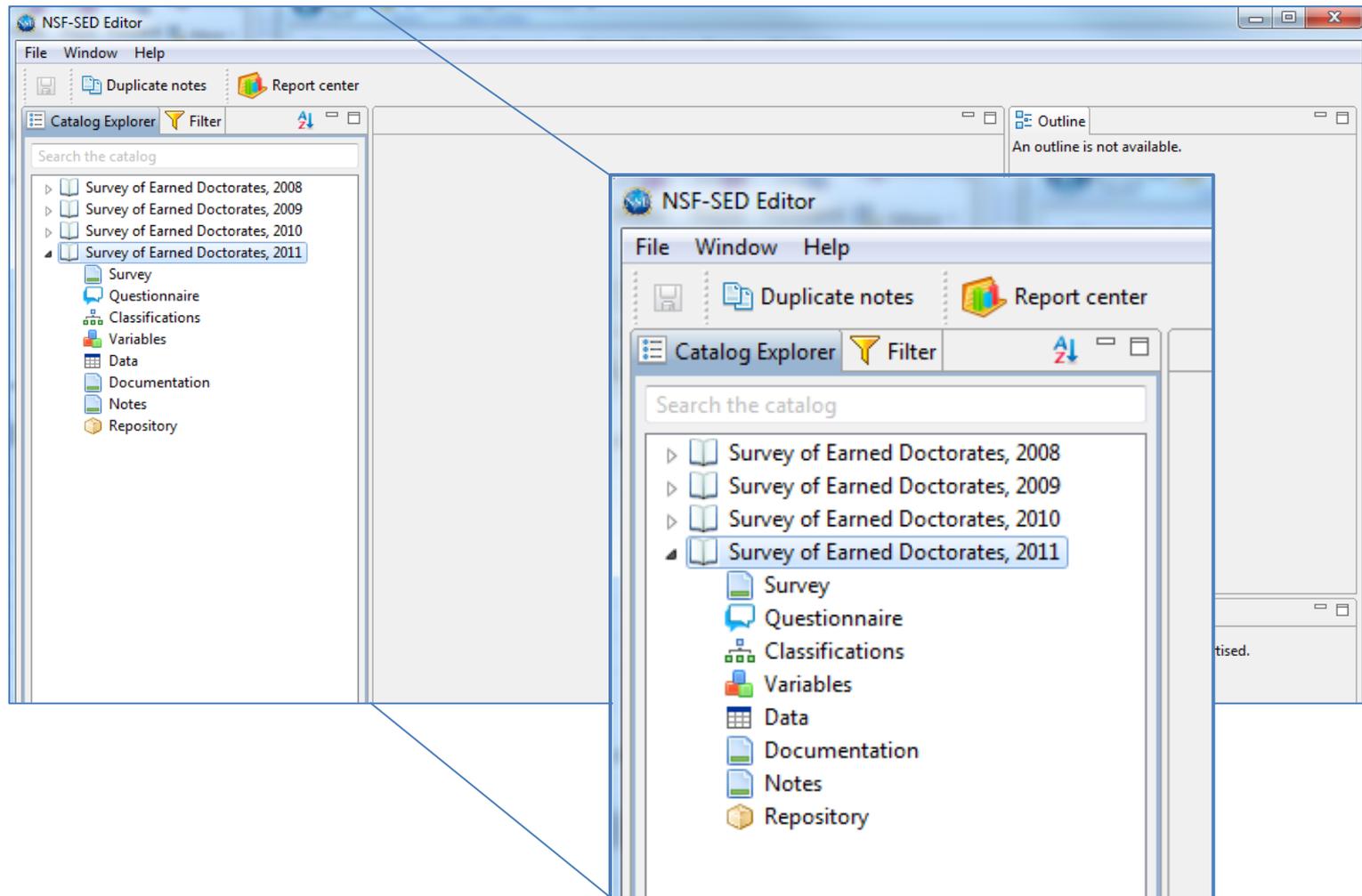
Catalog Explorer



Metadata Editors



Metadata Editors



Survey Editor

The screenshot displays the NSF-SED Editor application window. The title bar reads "NSF-SED Editor". The menu bar includes "File", "Window", and "Help". Below the menu bar are buttons for "Duplicate notes" and "Report center". The main interface is divided into three panes:

- Catalog Explorer:** A tree view on the left showing a search bar and a list of survey records. The selected record is "Survey of Earned Doctorates, 2011", which is expanded to show sub-items: "Survey", "Questionnaire", "Classifications", "Variables", "Data", "Documentation", "Notes", and "Repository".
- Main Editor:** A central pane titled "Study - NSF-SED 2011" containing a form for editing the "Survey of Earned Doctorates, 2011". The form is organized into sections:
 - Identification:** Includes fields for "Study Year" (2011) and "URN" (urn:ddi:us.norc:StudyUnit.NSF_SED_2011.1.0.0).
 - Citation:** Includes fields for "English Title" (Survey of Earned Doctorates, 2011), "Abbreviation" (NSF-SED 2011), "Date" (2011), "Creators", "Contributors", "Copyright", and "Description".
 - Abstract:** A large text area for entering the abstract.
- Outline:** A pane on the right showing a hierarchical list of sections: "Identification", "Citation", "Abstract", and "Purpose".

At the bottom right, there is a "Bean Info" pane with the message: "No identifiable object is advertised."

Variable Editor

NSF-SED Editor

File Window Help

Duplicate notes Report center

Catalog E... Filter Variables - NSF-SED 2011

Search the catalog

- Survey of Earned Docto
- Survey
- Questionnaire
- Classifications
- Variables
- Data
- Documentation
- Notes
- Repository

Search: Columns LIST New variable

Showing 1 to 100 of 134 entries

Name	Label	Data Type	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 (demographics) (19 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

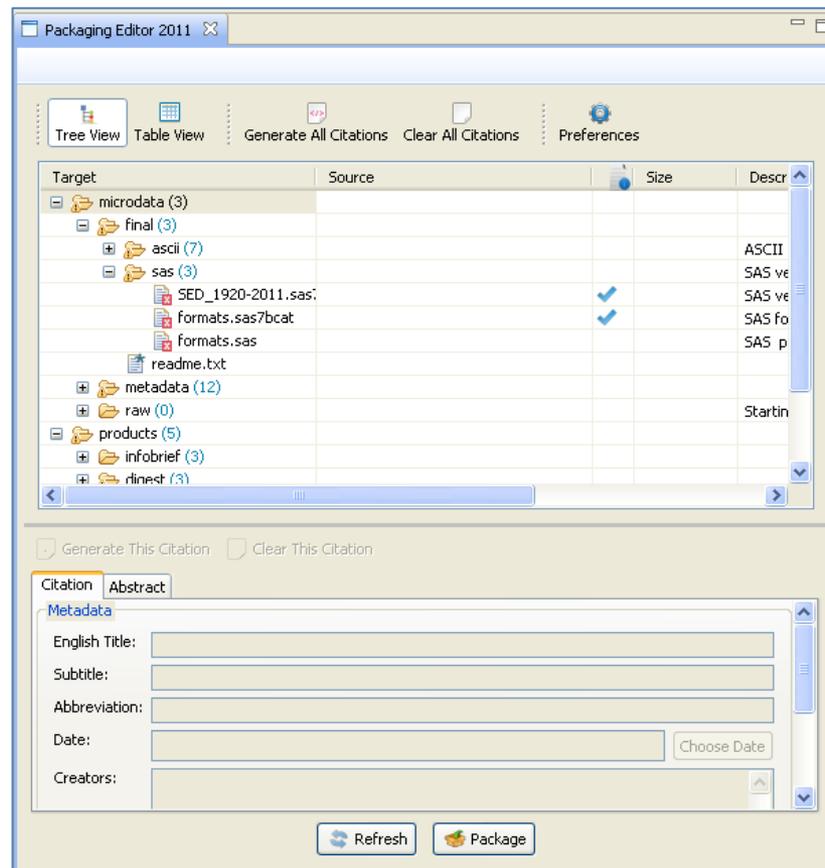
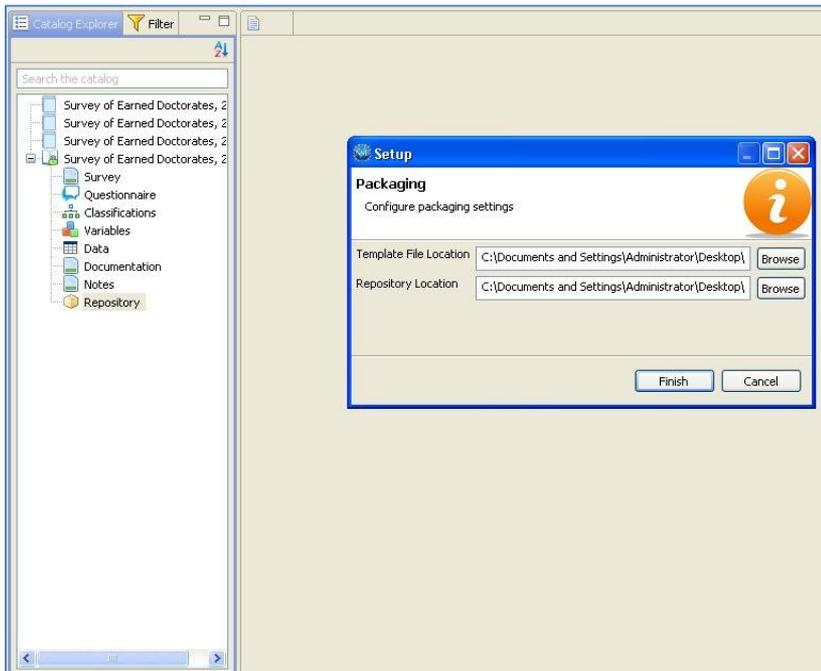
The screenshot displays the NSF-SED Editor interface, which is used for editing questionnaires. The main window shows the 'Survey of Earned Doctorates, 2011' questionnaire. The interface is divided into several panes:

- Left Pane:** A navigation tree showing the questionnaire structure, including sections like 'Survey of Earned Doctorates', 'Survey', 'Questionnaire', 'Classifications', 'Variables', 'Data', 'Documentation', 'Notes', and 'Repository'.
- Top Pane:** A search bar and a list of variables (First Name, Middle Name, Last Name, Suffix, etc.) that can be used in the questionnaire.
- Center Pane:** The main editing area, currently showing a list of questions under the 'Groups' tab. The selected question is 'Which of the following were sources of financial support during graduate school?'. Below this, a table lists the number of referers for each question.
- Right Pane:** A 'Bean Info' section showing the URN (Uniform Resource Name) for the questionnaire: `urn:ddius.norc:ControlConstructScheme.UGL`.

The 'Number of referers: 15' table is as follows:

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

Repository Editor



Report Center

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

Primary Study

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

< Back 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

< Back 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 institution code reported or no baccalaureate degree was received (see BACCNE to distinguish)

Variable	Column	Column	Column
Length	Start	End	End
1	3	236	238
2	3	239	241
3	3	242	244
4	3	245	247
5	3	248	250
6	2	251	252
7	3	253	256
8	3	256	258
9	3	259	261
10	3	262	264
11	3	265	267
12	2	268	269
13	2	270	271
14	3	272	274
15	3	275	277
16	2	278	279
17	3	280	282
18	3	283	285
19	2	286	287
20	6	288	293
21	2	294	296
22	3	296	298
23	8	299	308
24	3	307	309
25	3	310	312
26	3	313	316
27	3	316	318
28	3	319	321
29	3	322	324
30	3	325	327
31	2	328	329
32	2	330	331
33	3	332	334
34	3	335	337
35	3	338	340
36	3	341	343
37	3	344	346

Comparison report

Variable Comparison - Windows Internet Explorer

C:\Documents and Settings\Administrator\Desktop\SED_Repository\Report_Center\Variable

File Edit View Favorites Tools Help

Variable Comparison

Variable Comparison

2011	DRF_ID	2010	DRF_ID
ID Number	ID Number	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	Grant	Grant																							
index:2 start:8 end:11 width:4	index:12 start:34 end:36 width:3	index:12 start:34 end:36 width:3	index:12 start:34 end:36 width:3																							
n/a	Which of the following were sources of financial support during graduate school?	n/a	Which of the following were sources of financial support during graduate school?																							
Type: Numeric	Type: Numeric	Type: Numeric	Type: Numeric																							
	<table border="1"> <tr><td>1</td><td>Yes</td><td>1</td><td>Yes</td></tr> <tr><td>2</td><td>No</td><td>2</td><td>No</td></tr> <tr><td></td><td>Missing</td><td></td><td>Missing</td></tr> </table>	1	Yes	1	Yes	2	No	2	No		Missing		Missing	<table border="1"> <tr><td>1</td><td>Yes</td><td>1</td><td>Yes</td></tr> <tr><td>2</td><td>No</td><td>2</td><td>No</td></tr> <tr><td></td><td>Missing</td><td></td><td>Missing</td></tr> </table>	1	Yes	1	Yes	2	No	2	No		Missing		Missing
1	Yes	1	Yes																							
2	No	2	No																							
	Missing		Missing																							
1	Yes	1	Yes																							
2	No	2	No																							
	Missing		Missing																							

2011	FORMIND	2010	SRCEC																																	
Form type indicator	Teaching assistantship	Teaching assistantship	Teaching assistantship																																	
index:3 start:12 end:13 width:2	index:13 start:37 end:39 width:3	index:13 start:37 end:39 width:3	index:13 start:37 end:39 width:3																																	
n/a	Which of the following were sources of financial support during graduate school?	n/a	Which of the following were sources of financial support during graduate school?																																	
Type: Character	Type: Numeric	Type: Numeric	Type: Numeric																																	
<table border="1"> <tr><td>11</td><td>Purple/Black (2011)</td></tr> <tr><td>10</td><td>Green/Black (2010)</td></tr> <tr><td>09</td><td>Red/Black (2009)</td></tr> <tr><td>08</td><td>Indigo/Black (2008)</td></tr> <tr><td>07</td><td>Violet/Black (2007)</td></tr> </table>	11	Purple/Black (2011)	10	Green/Black (2010)	09	Red/Black (2009)	08	Indigo/Black (2008)	07	Violet/Black (2007)	<table border="1"> <tr><td>1</td><td>Yes</td><td>1</td><td>Yes</td></tr> <tr><td>2</td><td>No</td><td>2</td><td>No</td></tr> <tr><td></td><td>Missing</td><td></td><td>Missing</td></tr> </table>	1	Yes	1	Yes	2	No	2	No		Missing		Missing	<table border="1"> <tr><td>1</td><td>Yes</td><td>1</td><td>Yes</td></tr> <tr><td>2</td><td>No</td><td>2</td><td>No</td></tr> <tr><td></td><td>Missing</td><td></td><td>Missing</td></tr> </table>	1	Yes	1	Yes	2	No	2	No		Missing		Missing
11	Purple/Black (2011)																																			
10	Green/Black (2010)																																			
09	Red/Black (2009)																																			
08	Indigo/Black (2008)																																			
07	Violet/Black (2007)																																			
1	Yes	1	Yes																																	
2	No	2	No																																	
	Missing		Missing																																	
1	Yes	1	Yes																																	
2	No	2	No																																	
	Missing		Missing																																	

2011	SRCEC	2010	SRCEC															
Research assistantship	Research assistantship	Research assistantship	Research assistantship															
index:14 start:40 end:42 width:3	index:14 start:40 end:42 width:3	index:14 start:40 end:42 width:3	index:14 start:40 end:42 width:3															
n/a	Which of the following were sources of financial support during graduate school?	n/a	Which of the following were sources of financial support during graduate school?															
Type: Numeric	Type: Numeric	Type: Numeric	Type: Numeric															
	<table border="1"> <tr><td>1</td><td>Yes</td><td>1</td><td>Yes</td></tr> <tr><td></td><td>Missing</td><td></td><td>Missing</td></tr> </table>	1	Yes	1	Yes		Missing		Missing	<table border="1"> <tr><td>1</td><td>Yes</td><td>1</td><td>Yes</td></tr> <tr><td></td><td>Missing</td><td></td><td>Missing</td></tr> </table>	1	Yes	1	Yes		Missing		Missing
1	Yes	1	Yes															
	Missing		Missing															
1	Yes	1	Yes															
	Missing		Missing															



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

Pascal Heus
pascal.heus@metadatatechnology.com

Tim Mulcahy
Mulcahy-Tim@norc.org