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

NCSES Data User Types
- NCSES Staff
- Researcher
- Trend Analyst
- Policy Analyst
- Specific Interest
- Browser

Integrated Data System
- Metadata
- Microdata
- Tools & Processes
- Tools & Applications

External support
- GSS
- FAC
- FFS
- FSS
- SED
- R&D
- SDR
- NSCG
- NSRCG
- Other data*

* 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

  - 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
Variable Editor

<table>
<thead>
<tr>
<th>Variable</th>
<th>Representation</th>
<th>Notes</th>
<th>Concept/Universe</th>
<th>Generation Instruction</th>
<th>Files</th>
<th>Summary Statistics</th>
<th>Used By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name:</td>
<td>PHDDISS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label:</td>
<td>Dissertation field</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is Time</td>
<td></td>
<td>Is Geographic</td>
<td>Is Weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description:</td>
<td>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.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response Unit:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Questionnaire Editor
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
--- | ---
Survey of Earned Doctorates, 2011 | 
Survey of Earned Doctorates, 2009 | 
Survey of Earned Doctorates, 2003 | 

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

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

< Back | Next > | Finish | Cancel
Example Reports

Codebook

Comparison report

Variable Comparison

<table>
<thead>
<tr>
<th>Year</th>
<th>DRF_ID</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Number</td>
<td>D Number</td>
<td>index:1 start:1 and 7 width:7</td>
<td>index:1 start:1 and 7 width:7</td>
</tr>
<tr>
<td>n/a</td>
<td>Type: Character</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHDFY</th>
<th>Fiscal year of Doctorate</th>
<th>index:2 start:8 and 11 width:4</th>
<th>n/a</th>
<th>Type: Numeric</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>FORMIND</th>
<th>Form type indicator</th>
<th>index:1 start:12 and 13 width:2</th>
<th>n/a</th>
<th>Type: Character</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>SRCED</th>
<th>Research assistantship</th>
<th>index:14 start:46 and 42 width:3</th>
<th>n/a</th>
<th>Type: Numeric</th>
</tr>
</thead>
</table>
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