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

Kimberly Noonan (NSF NCSES)
Pascal Heus (MTNA)
Tim Mulcahy (NORC)

May 22, 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

Integrated Data System

Metadata

Tools & Processes

Microdata

Tools & Applications

External support

GSS
FAC
FFS
FSS
SED
R&D

SDF
NSCG
NSRCG
Other data*

NCSES Data User Types

NCSES Staff
Researcher
Trend Analyst
Policy Analyst
Specific Interest
Browser

Varying levels of access and functionality

* Other data used regularly in NCSES publications
NCSES Data Delivery

Specify data delivery requirements
  • Microdata
  • Metadata

• Ensure comprehensive documentation

• Standardize delivery formats

• Automate data processing

• Adopt metadata standards
  • Data Documentation Initiative (DDI)
  • Globally recommended practices
  • Industry standards and technologies
NSF Metadata Project: SED Case Study

Objectives

• Capture comprehensive survey metadata, preferably 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)

Project Team

• NSF NCSES
  • Building next generation data system and management framework
• NORC SED and Data Enclave team
  • Public interest research years of survey and data/metadata specific knowledge
• Metadata Technology North America (MTNA)
  • Domain and technology experts in statistical data management
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 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
- Prepare metadata for SED 2011
  - Develop software tool, extend existing MTNA open source application
  - Extend to additional survey cycles
- Recommend maintenance and future steps
  - Findings
  - 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

Metadata

| Metadata technology | NORC at the UNIVERSITY of CHICAGO | NSF | NCSSES | North America |

[Diagram showing the integration of metadata, repository synchronization, and documentation production processes]
Features

- **Catalog Explorer**
  - Import/create survey

- **Metadata Editors**
  - Survey, Questionnaire, Classification, Variables, Data, Documentation, Notes

- **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
Survey Editor
Questionnaire Editor
Variable 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)

<table>
<thead>
<tr>
<th>Study</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey of Earned Doctorates, 2011</td>
<td></td>
</tr>
<tr>
<td>Survey of Earned Doctorates, 2009</td>
<td></td>
</tr>
<tr>
<td>Survey of Earned Doctorates, 2003</td>
<td></td>
</tr>
</tbody>
</table>

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
Report Center

Codebook

Comparison report

Variable Comparison

<table>
<thead>
<tr>
<th>2011</th>
<th>DRF_ID</th>
<th>2012</th>
<th>DRF_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Number</td>
<td>ID Number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index: 1 start:1 and 7 width:7</td>
<td>Index: 1 start:1 and 7 width:7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type: Character</td>
<td>Type: Character</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2011</th>
<th>PHDFY</th>
<th>2012</th>
<th>PHDFY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiscal year of Doctorate</td>
<td>Fiscal year of Doctorate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>index: 2 start:8 and 11 width:4</td>
<td>index: 2 start:8 and 11 width:4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type: Numeric</td>
<td>Type: Numeric</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2011</th>
<th>FORMIND</th>
<th>2012</th>
<th>FORMIND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form type indicator</td>
<td>Form type indicator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>index: 3 start:12 and 13 width:2</td>
<td>index: 3 start:12 and 13 width:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type: Character</td>
<td>Type: Character</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. AVRIN/Bush (2011)
3. Reid/Bush (2009)
5. Frame/Bush (2007)
Next Steps

• Transition to production & maintenance mode
  – Training, document SED 2012, tweak reports / metadata / templates, minor fixes & enhancements

• Transition to metadata driven environment
  – Need further research/enhancements
  – Establish variable/classification/question/concept banks

• Integrate in NCSES Data Repository

• Extend to other NCSES surveys (SDR)
Thank you!

Contact information:

Kimberly Noonan  
knoonan@nsf.gov

Pascal Heus  
pascal.heus@metadatatechnology.com

Tim Mulcahy  
Mulcahy-Tim@norc.org