Converting Legacy Documentation to DDI: ICPSR’s Approach

Mary Vardigan
Director, Web Resources Development
Variable-level XML markup

- Daunting but necessary
- Provides granularity currently lacking
- Separates content from display
- Creates core document from which others can be generated
- Builds infrastructure for:
  - Data searches
  - Variable-level databases
  - Online analysis systems
First approach: Commercial software

- Marked up a dozen studies using XMetaL
- Tried to select a range of study types
- Tagged them as completely as possible -- best practice
- Recently created stylesheet -- see www.icpsr.umich.edu/DDI/SAMPLES/index.html
CBS News Monthly Poll #2, August 1992 (ICPSR 6084)  
(August National Poll II, Republican National Convention)

| View: | Part 1: Document Description  
| Part 2: Study Description  
| Part 3: Data Files Description  
| Part 4: Variable Description  
| Part 5: Other Study-Related Materials  
| Entire Codebook |

### Document Description

#### Citation

- **Title:** CBS News Monthly Poll #2, August 1992  
- **Alternative Title:** August National Poll II, Republican National Convention  
- **Identification Number:** 6084  
- **Authoring Entity:** CBS News  
- **Producer:** Inter-university Consortium for Political and Social Research  
- **Copyright:** Copyright ICPSR, 2000  
- **Date of Production:** May 10, 2000  
- **Software used in Production:** SoftQuad XMetaL  
- **Funding Agency:** National Science Foundation
Second approach: Macros

- Created set of macros in Unix editor to mark up Current Population Survey, National Survey of Family Growth
- Labor-intensive also
- Must have overall strategy before work starts
Third approach: Conversion

- Identified sets of uniform documentation, like OSIRIS
- Wrote a script to run against 1,000+ OSIRIS codebooks
- Based on typography of codebooks
- About 500 successfully converted
- Only variables portion marked up
Other conversion routines

- Used NESSTAR conversion utility to convert all SPSS portable files -- about 450 files/250 studies
- Working with VDC on other types of files, i.e., data definition statements
Challenges

- Applying aggregate data markup to SF 1 and other Census 2000 data
- Integrating question text
- Applying keywords from thesaurus
- PDF!
- Archiving and distributing marked-up versions
- Getting buy-in from data producers
Payoff: ICPSR of the future

- Search variables of any study (full-text or keywords)
- Compare variables across studies
- Analyze data online
- Download instructional subsets
- Create your own setups from markup
- View codebooks in attractive formats