Standards-based Metadata Management for Data Collection: an introduction
Agenda

- Barriers to sharing data
- DDI: the metadata standard for survey data
- GSIM: information model for official statistics
- DDI use cases
Barriers to Data Sharing
Barriers to Sharing Data: #1

- Data are meaningless without metadata
- Data require good documentation for understanding
Metadata are like punctuation ...

it was the best of times it was the worst of times it was the age of wisdom it was the age of foolishness it was the epoch of belief it was the epoch of incredulity it was the season of light it was the season of darkness it was the spring of hope it was the winter of despair we had everything before us we had nothing before us we were all going direct to heaven we were all going direct the other way in short the period was so far like the present period that some of its noisiest authorities insisted on its being received for good or for evil in the superlative degree of comparison only
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—
in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>119</td>
<td>18</td>
<td>70</td>
<td>90</td>
<td>2124</td>
<td>14</td>
<td>73</td>
</tr>
<tr>
<td>120</td>
<td>19</td>
<td>122</td>
<td>85</td>
<td>2310</td>
<td>19</td>
<td>73</td>
</tr>
<tr>
<td>121</td>
<td>21</td>
<td>155</td>
<td>107</td>
<td>2472</td>
<td>14</td>
<td>73</td>
</tr>
<tr>
<td>122</td>
<td>26</td>
<td>98</td>
<td>90</td>
<td>2265</td>
<td>16</td>
<td>73</td>
</tr>
<tr>
<td>123</td>
<td>15</td>
<td>350</td>
<td>145</td>
<td>4082</td>
<td>13</td>
<td>73</td>
</tr>
<tr>
<td>124</td>
<td>16</td>
<td>400</td>
<td>230</td>
<td>4278</td>
<td>10</td>
<td>73</td>
</tr>
<tr>
<td>125</td>
<td>29</td>
<td>68</td>
<td>49</td>
<td>1867</td>
<td>20</td>
<td>73</td>
</tr>
<tr>
<td>126</td>
<td>24</td>
<td>116</td>
<td>75</td>
<td>2158</td>
<td>16</td>
<td>73</td>
</tr>
<tr>
<td>127</td>
<td>20</td>
<td>114</td>
<td>91</td>
<td>2582</td>
<td>14</td>
<td>73</td>
</tr>
<tr>
<td>128</td>
<td>19</td>
<td>121</td>
<td>112</td>
<td>2868</td>
<td>16</td>
<td>73</td>
</tr>
<tr>
<td>129</td>
<td>15</td>
<td>318</td>
<td>150</td>
<td>3399</td>
<td>11</td>
<td>73</td>
</tr>
<tr>
<td>130</td>
<td>24</td>
<td>121</td>
<td>110</td>
<td>2660</td>
<td>14</td>
<td>73</td>
</tr>
<tr>
<td>131</td>
<td>20</td>
<td>156</td>
<td>122</td>
<td>2807</td>
<td>14</td>
<td>73</td>
</tr>
<tr>
<td>132</td>
<td>11</td>
<td>350</td>
<td>180</td>
<td>3664</td>
<td>11</td>
<td>73</td>
</tr>
<tr>
<td>133</td>
<td>20</td>
<td>198</td>
<td>95</td>
<td>3102</td>
<td>17</td>
<td>74</td>
</tr>
</tbody>
</table>
How satisfied with life now

Variable: AIPD1
Label: How satisfied with life now

Question:
Text: At present, how satisfied are you with your LIFE? Would you say A LOT, SOMewhat, A LITTLE, or NOT at ALL?

Pre-Question:
Text: And now a few questions about you.

Dataset: MIDUS 1

<table>
<thead>
<tr>
<th>Value</th>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A LOT</td>
<td>4383</td>
</tr>
<tr>
<td>2</td>
<td>SOMewhat</td>
<td>2166</td>
</tr>
<tr>
<td>3</td>
<td>A LITTLE</td>
<td>380</td>
</tr>
<tr>
<td>4</td>
<td>NOT AT ALL</td>
<td>140</td>
</tr>
<tr>
<td>7</td>
<td>DON'T KNOW</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>REFUSED/MISSING</td>
<td>1</td>
</tr>
</tbody>
</table>

Valid: 7069
Invalid: 39
Min: 1
Max: 4
Barriers to Sharing Data: #2

- Different agencies have different systems
  - Taking over a survey from another agency often requires re-inputting everything
  - Questionnaire specification quality and format differences
  - This makes re-use and comparability difficult
Barriers to Sharing Data: #3

- Barriers are also internal within organisations
  - Different disciplines have different attitudes to what is most important
  - Different departments speak different languages
  - Communication is always an issue
DDI: the Metadata Standard for Survey Data
DDI: a Shared Vocabulary

- Survey design and specification
- Data documentation
- Data lifecycle documentation
- Foundational metadata
DDI is an international, open standard for describing survey data

- XML standard
- Since 1995
Questions with many response types
Conditional logic and flow control
Dynamic text fills
Reusable questions and blocks of questions
Custom computations
Link collected data to source questions
Benefits of DDI

- Rich, machine-actionable metadata
- Common, interoperable vocabulary to describe surveys and data
- Question Banks
- Classification Management

Queries like
- What are all the datasets that have information from this question?
- What are all the versions of this classification used within my institution?
Metadata-Driven Processes

- DRY: Don’t Repeat Yourself
  - Define things once and create multiple outputs from that canonical information
- Generate documentation as a byproduct of the process
- Populate CAI systems
- Track changes over time
- Generate multiple reports from the same information
One Specification, Many Outputs

DDI Survey Instrument

- PDF documentation
- Web survey
- Blaise survey
- Paper forms
- Generic Statistical Information Model
- Common language to describe the whole statistical production process
- UN High-Level Group for the Modernisation of Official Statistics

- GSIM is a conceptual model
- DDI is an implementation model

- DDI implements GSIM
Case Studies
DDI Adopters

- National Statistical Organizations
- University Research Groups
- Data Archives
- Other Data Producers and Publishers

- Used in over 80 countries

- Collaborative community: talk to your colleagues
INSEE

- Specify questionnaires in DDI 3.2
- Build a central metadata repository to enable reuse
- Active in the DDI community
Statistics Denmark

- Statistical register documentation
- Statistical product descriptions
- Classification management
- Eurostat quality statements
- Central metadata repository
Statistics New Zealand

- Concept and classification management
- Statistical product documentation
- Variable-level documentation
- Central metadata repository
Recap

- Barriers to sharing data exist
- Using a metadata standard for survey data can help overcome these
- Metadata – including survey specifications – should be treated as first class information objects
Document, Discover and Interoperate

The Data Documentation Initiative (DDI) is an international standard for describing statistical and social science data. Documenting data with DDI facilitates interpretation and understanding -- both by humans and computers. The freely available international DDI standard describes data that result from observational methods in the social, behavioral, economic, and health sciences. Use DDI to Document, Discover, and Interoperate!

Why Use DDI?
Thank you