Alternatives for Representing Coding of Qualitative Data in DDI

Larry Hoyle
Institute for Policy & Social Research
University of Kansas
Qualitative Data

- Digital or real-world object (analog?)
- Purpose collected, gathered, or referenced
  - Text
  - XML
  - Web pages
  - Video
  - Audio
  - Images
Three Scenarios

1. Documentation at the object level
2. Segments of objects need documentation
   E.g. CAQDAS – codes associated with defined segments
3. Segments have documentation and quantitative data have been generated from the qualitative segments
   E.g. text mining

DDI Qualitative Data Model Working Group http://www.ddialliance.org/alliance/working-groups#qdewg
Segments Examples

Segment defined by a rectangle

Segment defined by a polygon

Segments can also be marked in text. Like in this text example.

Text segment

Overlapping text segment
Overall model
(many current DDI elements assumed to be applicable and not shown)

Methods and Instruments

Segment Definition
Segments can have “Codes”, “Categories” and “Memos”

Modeled after terms from qualitative data analysis packages like Atlas/ ti or NVIVO
Codes and Memos

“Dinner site”

Memo: This is marked here for an example in a PowerPoint Presentation

“Presentations”

Segments can also be marked in text. Like in this text example.

“Marking”

“Text reference”
Dataset

Off by itself is this Dataset

This might be produced by text mining, or might be quantitative data associated with an open-ended question.
This is a data record which can be described by existing DDI elements.

It can contain codes, categories, memos, and any quantitative data associated with the segment.
This dataset can be included in the DDI representation.
### Example Records

#### Dataset

<table>
<thead>
<tr>
<th>SegmentID</th>
<th>SegmentName</th>
<th>CampusBuilding</th>
<th>EveningVenue</th>
<th>YearBuilt</th>
<th>Address</th>
<th>Cluster1</th>
<th>MiningVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlumniCenter</td>
<td></td>
<td>0</td>
<td>1</td>
<td>19831266 Oread Ave.</td>
<td>0</td>
<td>1.27</td>
</tr>
<tr>
<td>2</td>
<td>StudentUnion</td>
<td></td>
<td>1</td>
<td>0</td>
<td>19261301 JayHawk Blvd.</td>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>
### Example Record - Codes

**Codes are handled like “Select all that apply” questions.**

<table>
<thead>
<tr>
<th>SegmentID</th>
<th>SegmentName</th>
<th>CampusBuilding</th>
<th>EveningVenue</th>
<th>YearBuilt</th>
<th>Address</th>
<th>Cluster1</th>
<th>MiningVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlumniCenter</td>
<td>0</td>
<td>1</td>
<td>1983</td>
<td>19831266 Oread Ave.</td>
<td>0</td>
<td>1.27</td>
</tr>
<tr>
<td>2</td>
<td>StudentUnion</td>
<td>1</td>
<td>0</td>
<td>1926</td>
<td>19261301 Jayhawk Blvd.</td>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

**Codes**

1 = has code
0 = does not

**Could have Categories and Memo here as well**
# Example Record – Other Variables

## Codes

<table>
<thead>
<tr>
<th>SegmentID</th>
<th>SegmentName</th>
<th>CampusBuilding</th>
<th>EveningVenue</th>
<th>YearBuilt</th>
<th>Address</th>
<th>Cluster1</th>
<th>MiningVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlumniCenter</td>
<td></td>
<td>0</td>
<td>1</td>
<td>19831266 Oread Ave.</td>
<td>0</td>
<td>1.27</td>
</tr>
<tr>
<td>2</td>
<td>StudentUnion</td>
<td></td>
<td>1</td>
<td>0</td>
<td>19261301 Jayhawk Blvd.</td>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Search for Evening Venue=1

<table>
<thead>
<tr>
<th>SegmentID</th>
<th>SegmentName</th>
<th>CampusBuilding</th>
<th>EveningVenue</th>
<th>YearBuilt</th>
<th>Address</th>
<th>Cluster1</th>
<th>MiningVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlumniCenter</td>
<td>0</td>
<td>1</td>
<td>1983</td>
<td>1266 Oread Ave.</td>
<td>0</td>
<td>1.27</td>
</tr>
<tr>
<td>2</td>
<td>StudentUnion</td>
<td>1</td>
<td>0</td>
<td>1926</td>
<td>1301 Jayhawk Blvd.</td>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>
### Search for YearBuilt<1950

<table>
<thead>
<tr>
<th>SegmentID</th>
<th>SegmentName</th>
<th>CampusBuilding</th>
<th>EveningVenue</th>
<th>YearBuilt</th>
<th>Address</th>
<th>Cluster1</th>
<th>MiningVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlumniCenter</td>
<td>0</td>
<td>1</td>
<td>1983</td>
<td>1266 Oread Ave.</td>
<td>0</td>
<td>1.27</td>
</tr>
<tr>
<td>2</td>
<td>StudentUnion</td>
<td>1</td>
<td>0</td>
<td>1926</td>
<td>1301 Jayhawk Blvd.</td>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>
### Text Mining Example

<table>
<thead>
<tr>
<th>title</th>
<th>paranum</th>
<th>storynum</th>
<th>paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>763</td>
<td>11</td>
<td>WHAT a funny sight it is to see a brood of ducklings with a hen! --List...</td>
</tr>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>764</td>
<td>11</td>
<td>HER sister-in-law, Mrs. Rebeccah Puddle-duck, was perfectly willing...</td>
</tr>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>765</td>
<td>11</td>
<td>&quot;I wish to hatch my own eggs. I will hatch them all by myself,&quot; quacke...</td>
</tr>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>766</td>
<td>11</td>
<td>SHE tried to hide her eggs; but they were always found and carried o...</td>
</tr>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>767</td>
<td>11</td>
<td>Jemima Puddle-duck became quite desperate. She determined to m...</td>
</tr>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>768</td>
<td>11</td>
<td>SHE set off on a fine spring afternoon along the cart- road that leads...</td>
</tr>
<tr>
<td>THE TALE OF JEMIMA PUDDLE-...</td>
<td>769</td>
<td>11</td>
<td>She was wearing a shawl and a poke bonnet.</td>
</tr>
</tbody>
</table>

What a funny sight it is to see a brood of ducklings with a hen!

—Listen to the story of Jemima Puddle-duck, who was annoyed because the farmer's wife would not let her hatch her own eggs.

Each segment is a paragraph from a Beatrix Potter Story

(downloaded from Project Gutenberg - [http://www.gutenberg.org/](http://www.gutenberg.org/))
A Text Topic Tool Can Build a Set of Topics
Topics are Based on Weighted Combinations of Words

The weights for calculation

Each segment can then be assigned a score for that topic
Future Segments Could be Scored on These

The weights for calculation

<table>
<thead>
<tr>
<th>Term</th>
<th>Role</th>
<th># Docs</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>pigling</td>
<td>Prop</td>
<td>67</td>
<td>90</td>
</tr>
<tr>
<td>bland</td>
<td>Prop</td>
<td>38</td>
<td>46</td>
</tr>
<tr>
<td>alexander</td>
<td>Prop</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>paper</td>
<td>Noun</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>drop</td>
<td>Verb</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>wish</td>
<td>Verb</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Described by a DDI Generation Instruction for a variable shared across studies?

<table>
<thead>
<tr>
<th>SegmentID</th>
<th>paragraph</th>
<th>MiningVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Did you read the paper?</td>
<td>0.305</td>
</tr>
<tr>
<td>1002</td>
<td>I wish spring would come</td>
<td>0.173</td>
</tr>
</tbody>
</table>
Data or Metadata?

 Depends on use, doesn’t it? (c.f. NSA use of phone “metadata”)

 YearBuilt might be metadata when searching for image segments or text descriptions of old buildings

 It might be data if we used the text mining variables to predict building age

 Metadata?  Data?
Similarity Between Models

• A code (its associated category) can refer to a concept
• A variable can refer to a concept
• So both approaches ultimately relate a segment to some concept
Data Record Advantages

• Can handle Mixed Method Research (Quantitative and Qualitative Approaches)
• Works for surveys with open-ended questions
• Allows for sharing of codes and quantitative variables across studies
• Searching for codes is the same as searching for other attributes (e.g. building age from above)
• Flexible
Disadvantages

• More difficult to explain?

• Searching involves more indirect lookup
  – Segment with a record with a value of 1 ("has attribute") on Variable “AnalyticCode”

  Vs
  – Segment with “AnalyticCode” of xxx

• “Codes”, “categories”, and “memos” lose special meaning

• Resistance from qualitative only researchers?
Discussion

• Advantages and disadvantages to data record approach?

• Other comments?
For More About the Qualitative Model

• See:
  • *Toward Qualitative Data in DDI*. Larry Hoyle and Joachim Wackerow
  • Larry Hoyle is a Senior Scientist at the Institute for Policy & Social Research, University of Kansas
  • [LarryHoyle@ku.edu](mailto:LarryHoyle@ku.edu)
  • Joachim Wackerow is a metadata expert at GESIS - Leibniz Institute for the Social Sciences and can be reached at [joachim.wackerow@gesis.org](mailto:joachim.wackerow@gesis.org).

version: November 2013