Creating DDI Compliant Codebooks

Wendy L. Thomas
William C. Block
Robert P. Wozniak
Joshua J. Buysse

A workshop presented at IASSIST 2001
Amsterdam NL -15 May 2001

Structure for the Workshop

- 9:15 – 10:15 DDI compliant codebooks
  - Contents
  - Points and perspectives to keep in mind
  - Best practices
- 10:15 – 11:30 MADDIE (break in here somewhere)
  - Walk-through of functions
  - Practice entries
- 11:30 – 12:15 Playtime and questions

Workshop Materials

- CD-ROM
  - Copy of Maddie
  - Quick Reference Guide
- Tag Library
  - Awe-inspiring reference tool for every element and attribute in the DDI
- Codebook
  - Stripped down model of an ICPSR codebook to use as a source for the workshop
  - Do NOT try to use this with the data set described under this study number (its been edited beyond recognition)

Overall DDI Structure

- Document Description (1.0)
  - Describes the XML document itself and the source materials
- Study Description (2.0)
  - Describes the overall study
- Data File Descripton (3.0)
  - Describes the physical data files
- Variables Description (4.0)
  - Describes the variables themselves
- Other Materials (5.0)

Basic Concepts to Remember:

- It is NOT just your basic codebook
- Machine readable vs. Machine processable
- Human understandable vs. Machine understandable
- Information needs to be entered in discrete bits

Principles to follow:

- Use attributes
- Use ID attribute so you can use IDRefs
- Make implicit information explicit
  - source, XML:Lang, level
- Follow ISO standards where available
- Inheritance

ID Attribute

- Provides a unique name for each specific element
- Must start with an alpha character and contain no spaces
- Must be unique within the XML document
- Create your own scheme for easy application and reference

Example of an ID scheme:

```
<docDscr ID="doc0">
        <citation ID="doc1"></citation>
        <docSrc ID="doc4"> </docSrc>
</docDscr>
<stdyDscr ID="s0">
        <stdyInfo ID="s2">
                <sumDscr ID="s2_3">
                        <universe ID="s2_3u1">Persons living on farms</univese>
                        <universe ID="s2_3u2">Farms over 100 acres</univese>
                </sumDscr>
        </stdyInfo>
</stdyDscr>
```

Using ID references

```
<stdyDscr ID="s0">
        <stdyInfo ID="s2">
                <sumDscr ID="s2_3">
                        <universe ID="s2_3u1">Persons living on farms</univese>
                        <universe ID="s2_3u2">Farms over 100 acres</univese>
                </sumDscr>
        </stdyInfo>
</stdyDscr>
<dataDscr ID="d0">
        <var ID="v01" sdatrefs="s2_3u1"> </var>
        <var ID="v01" sdatrefs="s2_3u2"> </var>
        <var ID="v01" sdatrefs="s2_3u1"> </var>
</dataDscr>
```

Best Practices: Multi-country data sets

- Example: EuroBarometer
- Questions vary by country
- Response category value varies by country
- Identify countries under<nation> and use sdatref attribute to identify variants

```
<stdyDscr>
 <stdyInfo>
  <sumDscr>
    <nation ID=`NL´>The
    Netherlands</nation>
    <nation ID=`FR´>France
    </nation>
  </sumDscr>
 </stdyInfo>
</stdyDscr>
```

Use of sdatRefs, methRefs and pubRefs:

- Under verison 1.01 these attributes have been made broadly available
- Their use varies only in the sections of the dtd to which they refer
- Each can contain references to one or more element IDs

Examples of use:

- When two or more universe statements are used these can be stated in the study description and then variables can be associated to the correct universe by sdatRefs
- Changes in response category labels by country. The appropriate label is linked to the country by sdatRefs

sdatRefs

Summary data description references that record the ID values of all elements within the summary data description section of the Study Description that might apply.

These elements include: *time period covered, date of collection, nation or country, geographic coverage, geographic unit, unit of analysis, universe, and kind of data*.

methRefs

methodology and processing references which record the ID values of all elements within the study methodology and processing section of the Study Description which might apply.

These elements include: *information on data collection and data appraisal (e.g., sampling, sources, weighting, data cleaning, response rates, and sampling error estimates).*

pubRefs

Provides a link to publication/citation references and records by listing the ID values of all citations elements within Section 2.5 or Section 5.0 that pertain to the element.

source, XML:lang, level

- Source attribute provides the source of the information in the element
  - Remember that not all elements may be passed to another person/system and it is always good to know who to blame ☺
- XML:lang provides language identifier
  - The `default´ language to you may not be the `default´ language of the user
- Level indicates nesting patterns
  - Some elements such as <labl> and <txt> occur in many locations in the dtd. This lets you identify the level of label (var, file, etc)

Using ISO standards

<prodDate> 1.1.3.3 (Generic element A.6.3.3)

Description: Date the marked-up document was produced (not distributed or archived). The ISO standard for dates (YYYY-MM-DD) is recommended for use with the date attribute. Equivalent to Dublin Core Date.

Example:

<prodDate date='1999-01-25'>January 25, 1999</prodDate>

Inheritance

- Lower levels in hierarchies inherit information from higher levels
- If a piece of information is true for the entire subset of elements, move it up to the next level
- This means consciously looking for common pieces of information and entering them appropriately

Referencing standard catagory lists

<stdCatgry> 4.2.16

Description: Standard category group used in a variable, like industry codes, employment codes, or social class codes. The attribute of "date" is provided to indicate the version of the code in place at the time of the study. The attribute of "URI" is provided to indicate a URN or URL that can be used to obtain the electronic form of the category group.

Example:

<var><stdCatgry date='1981' source='producer' >Census of Population, Classified Index of Industries and Occupations </stdCatgry></var>

Attributes: ID, xml:lang, source, date, URI

Recording or creating variable groups

- Variable groups can contain both variables and other variable groups.
- Variable groups are created this way in order to permit variables to belong to multiple groups.
- Variables that are linked by use of the same question need not be identified by a Variable Group element because they are linked by a common unique question identifier in the Variable element.
- All Variable Groups must be marked up before the Variable element is opened.

Types of Variable Groups:

- Section: Questions from the same section of the questionnaire, e.g., all variables located in Section C.
- Multiple response: respondent can select more than one answer from a variety of choices, e.g., what newspapers have you read in the past month.
- Grid: Sub-questions of an introductory or main question but which do not constitute a multiple response group, e.g., I‘m going to read a list of candidates and I would like you to tell me whether you have heard of them.

Type of groups *continued*

- Display: Questions which appear on the same interview screen (CAI) together or are presented to the interviewer or respondent as a group.
- Repetition: The same variable (or group of variables) which are repeated for different groups of respondents or for the same respondent at a different time.
- Subject: Questions which address a common topic or subject, e.g., income, poverty, children.

Type of groups *continued*

- Version: Variables, often appearing in pairs, which represent different aspects of the same question, e.g., pairs of variables (or groups) which are adjusted/unadjusted for inflation or season or whatever, pairs of variables with/without missing data imputed, and versions of the same basic question.
- Iteration: Questions that appear in different sections of the data file measuring a common subject in different ways, e.g., a set of variables which report the progression of respondent income over the life course.

Type of groups *continued*

- Analysis: Variables combined into the same index, e.g., the components of a calculation, such as the numerator and the denominator of an economic statistic.
- Pragmatic: A variable group without shared properties.
- Record: Variables from a single record in a hierarchical file.
- File: Variables from a single file in a multifile study.

Type of groups *continued*

- Randomized: Variables generated by CAI surveys produced by one or more random number variables together with a response variable, e.g, random variable X which could equal 1 or 2 (at random) which in turn would control whether Q.23 is worded "men" or "women", e.g., would you favor helping [men/women] laid off from a factory obtain training for a new job?

Type of groups *continued*

And finally….

- Other: Variables which do not fit easily into any of the categories listed above, e.g., a group of variables whose documentation is in another language.