An Introduction to the Data Documentation Initiative (DDI)

ICPSR OR Meeting 2001

Wendy L. Thomas
Data Access Core Director

William C. Block
Information Technology Core Directory

Minnesota Population Center
26 October 2001
What is the DDI ...

- DDI = Data Documentation Initiative
- XML = eXtensible Markup Language
- DTD = Document Type Definition
- Archive quality machine readable metadata designed to be human AND computer understandable and processable
- ... and so much more
...and why is it important to you?

- Increases the depth of access to your collection
- Allows sharing of discovery tools
- Allows functional sharing of all metadata materials
- Encourages cooperative metadata collection development
- Encourages FULL documentation of data
Jakob Nielsen,  
Distinguished Engineer at Sun Microsystems

XML is one of the greatest advances in the Web in a long time. Whereas most other Web innovations since 1993 have focused on glitz and on making superficially glamorous but useless fancy layouts, XML attacks the usefulness of the Web by adding structure and meaning to its vast seas of information."
Stewart Brand,
Founder of the Whole Earth Catalog

Perpetually obsolescing and thus losing all data and programs every 10 years (the current pattern) is no way to run an information economy or a civilization."
Brian Behlendorf,
President, Apache Software Foundation

"XML has become increasingly crucial throughout the software industry, as well as the Open Source community, as a non-proprietary method of storing and exchanging complex data."
James Clark,
interview with Dr. Dobbs Journal

"[What’s the next step for XML?] That’s a difficult question...it’s like asking me, "What’s the next application for ASCII text?"
The Session:

- XML & where you might encounter DDI
- The ‘Bill’ Experience: helping the hapless
- Using and exploiting DDI compliant files
- Managing large scale coding projects
- Tools of the trade
- Questions
XML basics

• XML is to a document’s intellectual content what HTML is to the physical structure of that document
• Elements  <element></element>
• Attribute  <element attribute="xxx">  
• Attribute types (imposing controls)
• Hierarchies and nesting
<docDscr>
  <citation>
    <titlStmt>
      <titl>World Population Table</titl>
      <subTitl>Example of Final Proposed Aggregate Tagging Model</subTitl>
    </titlStmt>
    <rspStmt>
      <AuthEnty>Wendy L. Thomas</AuthEnty>
    </rspStmt>
    <prodStmt>
      <prodDate date="2001-06-13">13. June 2001</prodDate>
    </prodStmt>
  </citation>
</docDscr>
<var ID="AGE" additivity="Y">
  <labl level="var">Age</labl>
  <catgry>
    <catValu>1</catValu>
    <labl level="catgry">0-14</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl level="catgry">15-64</labl>
  </catgry>
  <catgry>
    <catValu>3</catValu>
    <labl level="catgry">65+</labl>
  </catgry>
</var>
<nCube ID="Cube1" dmnsQnty="3" dmnsQnty="3" cellQnty="42">
  <location locMap="LM"/>
  <labl level="nCube">Population by Gender, Continent, and Year</labl>
  <universe>Persons</universe>
  <timeDmns rank="3" varRef="YEAR"/>
  <dmns rank="1" varRef="GENDER"/>
  <dmns rank="2" varRef="GEOG"/>
  <measure aggrMeth="count"
    measUnit="Persons" scale="x1000"
    additivity="Y"/>
</nCube>
Is XML DDI?

- The DDI is often used to refer to the specific XML document type definition file(s) created to describe social science data files.
- Understanding the basics of XML will help you understand the ‘DDI’.
Where you might encounter DDI

- DDI compliant documents distributed with data
- Creating DDI codebooks for your own collection
- Assisting researchers with creating DDI codebooks for their own research projects
The ‘Bill’ Experience:  
helping the hapless :-)  

- What I was doing  
- Why I documented using DDI  
- Issues raised in this experience:  
  - Broad to specific or specific to broad?  
  - The glories of the ID attribute  
  - OR’s support role
Specific to Broad Learning:
Learning every element at once is NOT recommended
<codeBook xml:lang="en">
  <docDscr>
    <citation>
      <titlStmt>
        <titl></titl>
        <subTitl></subTitl>
        <altTitl></altTitl>
        <parTitl></parTitl>
        <IDNo></IDNo>
      </titlStmt>
      <rspStmt>
        <AuthEnty></AuthEnty>
        <othId></othId>
      </rspStmt>
    </citation>
  </docDscr>
</codeBook>

This goes on for 6 pages in 10 point type
Broad to Specific Learning:
Learn one section at a time

- **Document Description:** *Items describing the marked-up document itself as well as its source documents*
- **Study Description:** *Items describing the overall data collection (title, citation, methodology, study scope, data access, etc)*
- **Data Files Description:** *Items relating to the format, size, and structure of the data files (physical descriptions)*
- **Variables Description:** *Items relating to variables in the data collection (logical descriptions)*
- **Other Study-Related Materials:** *Other study-related material not included in the other sections (bibliography, separate questionnaire file, etc.)*
Lowering the Learning Curve: Creating customized views and subsets

<table>
<thead>
<tr>
<th>Field</th>
<th>Marked Up</th>
<th>Source</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>1.1.1.1</td>
<td>1.4.1.1</td>
<td>2.1.1.1</td>
</tr>
<tr>
<td>Document Collection</td>
<td>1.1.1.2</td>
<td>1.4.1.2</td>
<td>2.1.1.2</td>
</tr>
<tr>
<td>&lt;titl&gt;</td>
<td>1.1.1.3</td>
<td>1.4.1.3</td>
<td>2.1.1.3</td>
</tr>
<tr>
<td>&lt;altTitl&gt;</td>
<td>1.1.1.5</td>
<td>1.4.1.5</td>
<td>2.1.1.5</td>
</tr>
<tr>
<td>&lt;IDNo&gt;</td>
<td>1.1.2.1</td>
<td>1.4.2.1</td>
<td>2.1.2.1</td>
</tr>
<tr>
<td>&lt;AuthEny&gt;</td>
<td>1.1.2.2</td>
<td>1.4.2.2</td>
<td>2.1.2.2</td>
</tr>
<tr>
<td>&lt;othID&gt;</td>
<td>1.1.3.1</td>
<td>1.4.3.1</td>
<td>2.1.3.1</td>
</tr>
<tr>
<td>&lt;copyright&gt;</td>
<td>1.1.3.2</td>
<td>1.4.3.2</td>
<td>2.1.3.2</td>
</tr>
<tr>
<td>&lt;prodDate&gt;</td>
<td>1.1.3.3</td>
<td>1.4.3.3</td>
<td>2.1.3.3</td>
</tr>
<tr>
<td>&lt;prodPlac&gt;</td>
<td>1.1.3.4</td>
<td>1.4.3.4</td>
<td>2.1.3.4</td>
</tr>
<tr>
<td>&lt;software&gt;</td>
<td>1.1.3.5</td>
<td>1.4.3.5</td>
<td>2.1.3.5</td>
</tr>
<tr>
<td>&lt;fundAg&gt;</td>
<td>1.1.3.6</td>
<td>1.4.3.6</td>
<td>2.1.3.6</td>
</tr>
<tr>
<td>&lt;grantNo&gt;</td>
<td>1.1.3.7</td>
<td>1.4.3.7</td>
<td>2.1.3.7</td>
</tr>
<tr>
<td>&lt;distrbtr&gt;</td>
<td>1.1.4.1</td>
<td>1.4.4.1</td>
<td>2.1.4.1</td>
</tr>
<tr>
<td>&lt;contact&gt;</td>
<td>1.1.4.2</td>
<td>1.4.4.2</td>
<td>2.1.4.2</td>
</tr>
<tr>
<td>&lt;depositr&gt;</td>
<td>1.1.4.3</td>
<td>1.4.4.3</td>
<td>2.1.4.3</td>
</tr>
<tr>
<td>&lt;depDate&gt;</td>
<td>1.1.4.4</td>
<td>1.4.4.4</td>
<td>2.1.4.4</td>
</tr>
</tbody>
</table>

Automatically filled in by, must be site-specific
Automatically filled in by MADDIE, same info for all locations
Entered by collaborators
The resident rural population of the United States on June 1, 1880 living in sampled states and counties.

Owners, Tenants, or Managers of farms greater than 3 acres in size or producing and selling at least $500 in product during the year.
<var ID="P13" name="hhsize" format="numeric" Dcml="0" sdatref="PCS PCE PU">
  <location StartPos="643" EndPos="646" width="4"></location>
  <labl>Number of persons in household.</labl>
  <security>public</security>
  <respUnit>Respondent</respUnit>
  <anlysUnit>Person</anlysUnit>
  <qstn>
    <qstnLit></qstnLit>
  </qstn>
  <valrng>
    <range min="0" max="1515"></range>
    <key>9999 missing</key>
  </valrng>
  <TotlResp>23806</TotlResp>
</var>
Value of farm, including land, fences and building.

Farm Values. Of farm, including land, fences and buildings.

Dollars

2006
The BIGGEST Lesson

The importance of the TAG LIBRARY!!

“If you could only take one thing to a deserted island to do DDI...make it the Tag Library.”
Using/Exploiting DDI compliant files

- The key lies in uniformity and consistency within an XML instance or within a series
- Never forget that a computer as well as a human being will be reading this
  - Element contents are for people
  - Attribute contents are for machines
The Concept of Inheritance

The idea that lower elements within an intellectual tree ‘inherit’ the attributes of the higher levels *unless a new value is provided*

Inheritance allows you to:
- Increase uniformity
- Reduce entry time
- Speed up processing
Looking for inheritance options

• Within a single xml instance
  – Within an element type
  – Within a section
  – Within the ‘codebook’

• Within a series of xml instances
  – External references
  – Cut and paste
The power of the ID attribute

- Every element should have an ID
- Developing a schema for ID’s
- IDRef and IDRefs:
  - sdatRef
  - methRef
  - pubRef
  - Others (var, nCube, varGrp, locMap...)
Managing large scale coding projects

- The order of things: *complete a document vs. completing all like parts*
- Specialization: *everyone learn everything vs. creating section experts*
- Notification: *automatic notification of step completion*
- Training: *mid-process training*
- Contact: *established “chain of command”*
- Models: *creating a “Model Book”*
The World According to the Unfortunates

- Is MADDIE the tool we want to use?
- Will there be models to guide our work?
- What’s the difference between universe and measurement unit?
- How uniform do the lettered/numbered variables need to be?
- Are there standard names for geography levels?
- When do I use category and when cohort?
- At what level do we describe units of measurement?
Tools of the Trade

- Free Resources
- Commercial Resources
- Plug-ins to Word
- DDI specific editors
  - NESSTAR
  - MADDIE
Free Resources

1. XED  
   www.ltg.ed.ac.uk/~ht/xed.html
2. MERLOT  www.merlotxml.org
3. SIXPACK  
   www.trafficstudio.com/sixpack
4. Others worth checking out:
   - LOGILAB’s XML Editor  
     www.logilab.org/xmltools/xmleditor.html
   - VISUAL XML  
     www.pierlou.com/visxml/

1. Best for small to medium sized XML documents; does *not* validate
2. Runs on any Java 2 virtual machine; extensible via custom editor interface
3. Works on Macintosh
Commercial Resources

1. AuthorIT  
   www.author-it.com
2. X-Ray XML Editor  
   www.xmlspy.com/products.html
3. Xmetal  
   www.softquad.com/top_frame.sq
4. XMLwriter  
   www.xmlwriter.com/
5. Morphon XML-Editor  
   www.morphon.com/xmleditor/index.shtml
6. XML Spy 4.0 Document Editor  
   www.xmlspy.com/products_doc.html/

1. Ideal for large multi-user documentation projects
2. Diagnoses XML errors in real time
3. “open and scriptable” development environment
4. Customizable interface
5. Multi-platform
6. For non-tech types
Plug-ins to Word

1. B-Bop Xfinity Author
   xW
   www.b-bop.com/products_xfinity_author_wX.htm

2. WorX
   www.hvltd.com/default.asp?name=information/xml/worxseOverview.xml&display=information/xsl/default.xsl

1. Unique “Save As” feature allows conversion to any DTD (Industry standard or user-defined)

2. Seybold Reports currently rate WorX as “the most sophisticated tool available for creating structured content in a MS Word environment”
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE library SYSTEM "x-schema:LibrarySchema.xml"
<!-- Copyright 1999 Wattle Software http://XMLwriter.net -->
<library xmlns="x-schema:LibrarySchema.xml">
  <name>Northmead Local Library</name>
  <book hardback="yes">
    <title>C++ Programming for Beginners</title>
    <author>
      <first-name>Claude</first-name>
      <last-name>Schwartz</last-name>
    </author>
    <callno>005.133/C</callno>
</book>
</library>
The 123xml family is the world's premier long-range development airplane. The new 123xml family will have even more capability - more programmers, the best nonstop range, fastest speed, and lowest operating cost in the XML market. New family features include a flight deck with a lead developer type rating surpassing any previous model and a spacious, programmer-pleasing interior. The long-range 123xml and the economical 123xml Stretch also boast a new CPU and higher thrust parser engines.

Interior options include converting overhead space into additional cubicles or server storage - ways to increase space for additional seats or programmer amenities without sacrificing revenue web sites. For example, the overhead space can accommodate up to 11 programmer cubicles, 24 servers, 12 routers, 5 firewalls, and 2 lavatories.

New quiet fuel-efficient parsers on the 123xml family reduce both errors and transformation time. In fact, the parsers will be so much faster that the 123xml family will be able to validate in real-time any XML based documents, which are some of the most complex content models in the world.
DDI Specific Editors

- NESSTAR Publisher
- MADDIE

Followed by
QUESTIONS
All Agricultural Variables

Variable Groups:

- Other Agricultural Variables
- Farm Tenure
- Farmland information
- Value information on farm, machinery and livestock
- Cost of building and repairing fences in 1879
- Cost of fertilizer purchased in 1879
- Wage Labor information, 1879
- Value of all farm productions, 1879 (sold, consumed, or on hand)
- Grasslands, acreage and products harvested, 1879
- Horses and Mules on hand, June 1, 1880
- Cattle and their products
- Sheep and their products
- Swine on hand June 1, 1880
- Poultry and egg productions, 1879
- Cereal crop acreage and production, 1879
- Pulse production, 1879
- Fiber acreage and production, 1879
- Sugar acreage and production, 1879
- Other crop acreage and production, 1879
- Orchard acreage, production, and total value, 1879
- Nursery acreage and value of nursery production, 1879
- Vineyard acreage, sales, and production, 1879
- Value of market garden production, 1879
- Bees production, 1879
- Forest production and value, 1879
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE codeBook SYSTEM "CODEBOOK.TXT">
<!--<codeBook ID="block01" xml:lang="en" source="me" --><codeBook xml:lang="en"

<!-- 1.0 DOCUMENT DESCRIPTION (docDscr) -->

<docDscr source="producer">
  <citation source="producer">
    <titl source="producer">
      <titl source="producer"></titl>
    <subTitl source="producer"></subTitl>
    <altTitl source="producer"></altTitl>
    <partTitl source="producer"></partTitl>
    <IDNo source="producer"></IDNo>
  </titl>
  <rspStat source="producer">
    <AuthEnty source="producer"></AuthEnty>
    <othId source="producer"></othId>
  </rspStat>
  <prodStat source="producer">
    <producer source="producer"></producer>
    <copyright source="producer"></copyright>
    <prodDate source="producer"></prodDate>
    <prodPlac source="producer"></prodPlac>
    <software source="producer"></software>
  </prodStat>
</docDscr>

lines: 308  Toplevel element name is "codeBook:"
**var (Variable)**

**Description**

This element describes all of the features of a single variable in a social science data file.

**Use**

Files attribute intended to link individual variables with the files where data on those variables may be found.

**TotResp (Total Responses)**

**Description**

The number of responses to this variable. This element might be used if the number of responses does not match added case counts. It may also be used to sum the frequencies for variable categories.

**Examples**

```
<var><TotResp>There are only 725 responses to this question since it was not asked in Tanzania</TotResp></var>
```