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

```
<?xml version="VC"?>
<codebook ID=”wpop.xml”>
  <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>
```

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 questionaire file, etc.)*

Lowering the Learning Curve:
Creating customized views and subsets

Automatically filled in by, must be site-specific			
Automatically filled in by MADDIE, same info for all locations			
Entered by collaborators			
	Marked Up Document	Source Document	Data Collection
<titl>	1.1.1.1	1.4.1.1	2.1.1.1
<subTitl>	1.1.1.2	1.4.1.2	2.1.1.2
<altTitl>	1.1.1.3	1.4.1.3	2.1.1.3
<IDNo>	1.1.1.5	1.4.1.5	2.1.1.5
<AuthEnty>	1.1.2.1	1.4.2.1	2.1.2.1
<othID>	1.1.2.2	1.4.2.2	2.1.2.2
<producer>	1.1.3.1	1.4.3.1	2.1.3.1
<copyright>	1.1.3.2	1.4.3.2	2.1.3.2
<prodDate>	1.1.3.3	1.4.3.3	2.1.3.3
<prodPlac>	1.1.3.4	1.4.3.4	2.1.3.4
<software>	1.1.3.5	1.4.3.5	2.1.3.5
<fundAg>	1.1.3.6	1.4.3.6	2.1.3.6
<grantNo>	1.1.3.7	1.4.3.7	2.1.3.7
<distrbtr>	1.1.4.1	1.4.4.1	2.1.4.1
<contact>	1.1.4.2	1.4.4.2	2.1.4.2
<depositr>	1.1.4.3	1.4.4.3	2.1.4.3
<depDate>	1.1.4.4	1.4.4.4	2.1.4.4

```
<sumDscr>
  <timePrd event="start" date="1879-01-01">January 1,
1879</timePrd>
  <timePrd event="end" date="1880-06-01">June 1,
1880</timePrd>
  <collDate ID="PCS" event="start" date="1989-11-
01">November 1, 1989</collDate>
  <collDate ID="PCE" event="end" date="1993-07-21">July
21, 1993</collDate>
  <collDate ID="ACS" event="start" date="1990-08-
01">August 1, 1990</collDate>
  <collDate ID="ACE" event="end" date="1998-07-21">July
21, 1998</collDate>
  <universe ID="PU" clusion="I">The resident rural
population of the United States on June 1, 1880 living
in sampled states and counties.</universe>
  <universe ID="AU" clusion="I">agline > 0.  Owners,
Tenants, or Managers of farms greater than 3 acres in
size or producing and selling at least $500 in product
during the year.</universe>
  <dataKind>census/enumeration data</dataKind>
</sumDscr>
```

```
<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>
```

```
<var ID="A20" name="farmval" format="numeric"
Dcml="0" sdatref="ACS ACE AU">
  <location StartPos="60" EndPos="65"
width="6"></location>
  <labl>Value of farm, including land, fences and
building.</labl>
  <security>public</security>
  <respUnit>Respondent</respUnit>
  <anlysUnit>Farm</anlysUnit>
  <qstn>
    <qstnLit>Farm Values.  Of farm, including land,
fences and buildings.</qstnLit>
  </qstn>
  <valrng>
    <range min="0" max="36400"></range>
    <key>Dollars</key>
  </valrng>
  <TotlResp>2006</TotlResp>
</var>
```

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

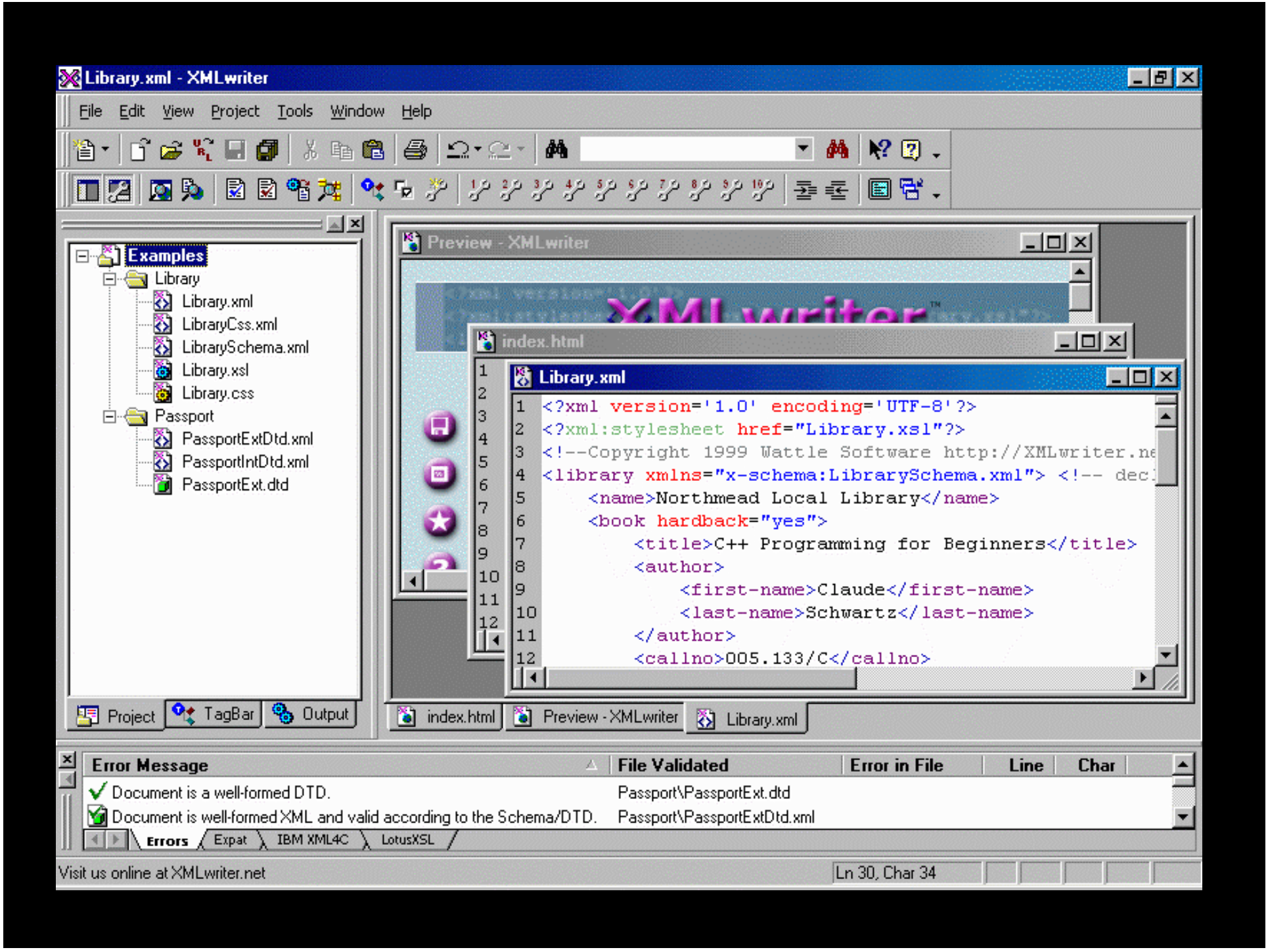
1. AuthorIT
   www.author-it.com
2. X-Ray XML Editor
   www.xmlspy.com/products.html
3. Xmetal
   www.softquad.com/top_frame.sq
4. XMLwriter
   www.xmlwriter.com/
5. Morphon XML-Editor
   www.morphon.com/xmleditor/index.shtml
6. XML Spy 4.0 Document Editor www.xmlspy.com/products_doc.html/

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=inforamtion/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"


Library.xml - XMLwriter
File Edit View Project Tools Window Help
Examples
Library
Library.xml
LibraryCss.xml
LibrarySchema.xml
Library.xsl
Library.css
Passport
PassportExtDtd.xml
PassportIntDtd.xml
PassportExt.dtd
Project TagBar Output
Preview - XMLwriter
XMLwriter™
index.html
Library.xml
<?xml version='1.0' encoding='UTF-8'?>
<?xml:stylesheet href="Library.xsl"?>
<!--Copyright 1999 Wattle Software http://XMLwriter.ne
<library xmlns="x-schema:LibrarySchema.xml"> <!-- dec
<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>
index.html Preview - XMLwriter Library.xml
Error Message
File Validated
Error in File
Line
Char
Document is a well-formed DTD.
Passport\PassportExt.dtd
Document is well-formed XML and valid according to the Schema/DTD.
Passport\PassportExtDtd.xml
Errors Expat IBM XML4C LotusXSL
Visit us online at XMLwriter.net
Ln 30, Char 34

XML Spy - [Datasheet.xml]

File Edit Project XML DTD/Schema Schema design XSL Document Editor Convert Table View Browser Tools Window Help

Assign configuration file...
Hide markup
Show Small markup
Show Large markup
Show Mixed markup
Append row
Insert row
Duplicate row
Move row Up
Move row Down
Delete row

ALTOVA

www.altova.com

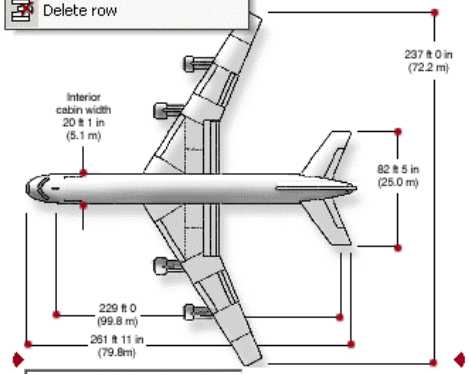
Altova 123xml Data S

The **Altova 123** 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, programmerr-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 Schema based documents, which are some of the most complex content models in the world.

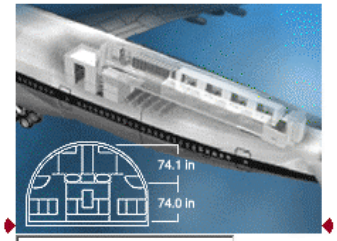
ns



exterior.gif

add description

Interior Design



interior.gif

Moving routers and servers to the crown of the aircraft frees up more room for cubicles, reducing per seat licensing costs by 2 percent.

Principal Characteristics	123	123xml	123xml Stretch
MaxTakeoffWeight	910000	1043000	1043000
MaxLandingWeight	652000	685000	725000
MaxZeroFuelWeight	555000	640000	680000
EngineOfferings	GE, P+W, and R-R	EA and R-R	EA and R-R
FuelCapacity	60305	72853	72853
CruiseMach	0.855	0.86	0.86
Passengers	416	442	522
DesignRange	7500	9175	8000
LowerHoldVolume	4876	5304	6750

Elements
- bold
- italic
- para
- diagram

Attributes
diagram
kind 3d

Datasheet.xml

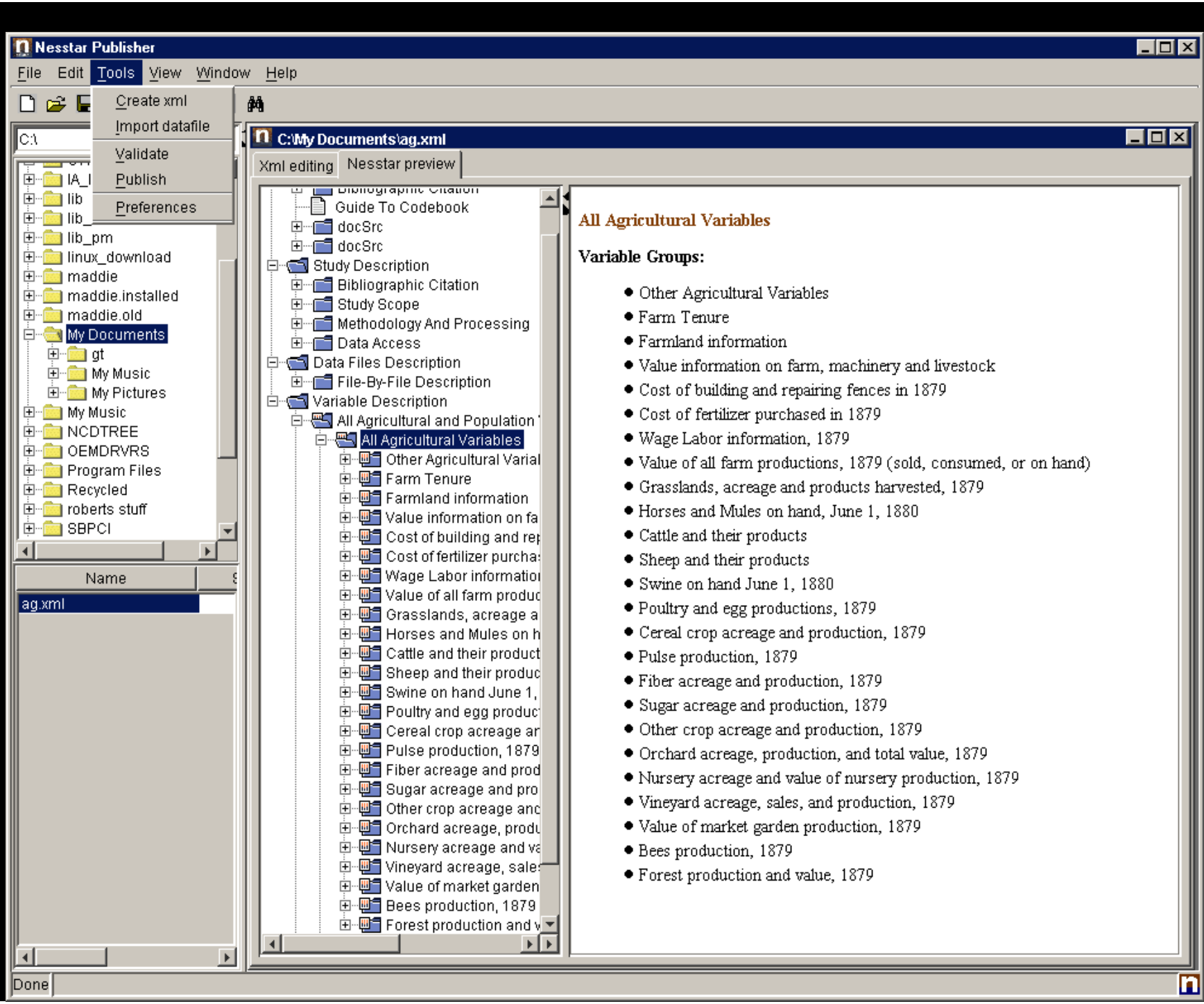
Show small markup

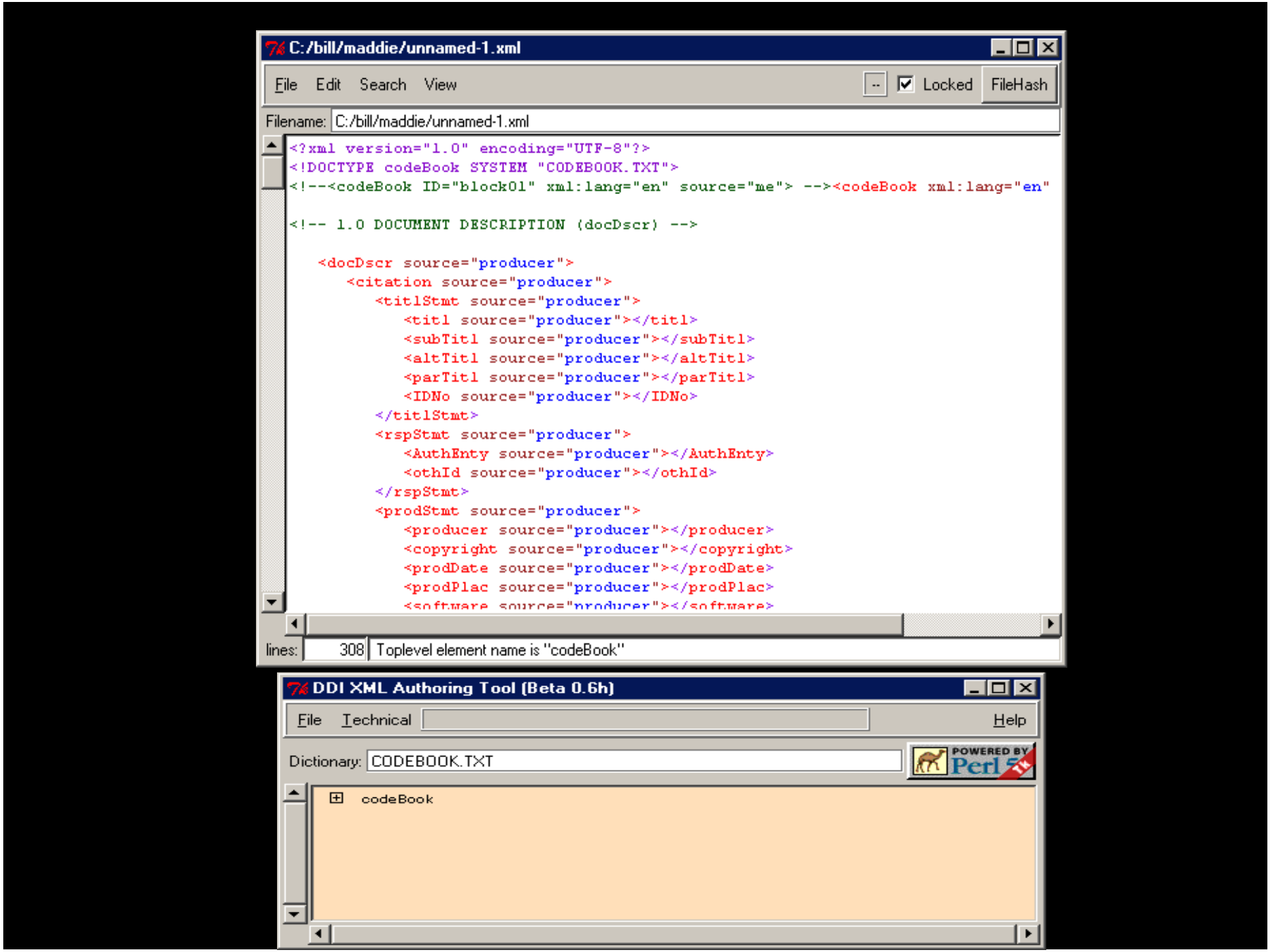
NUM

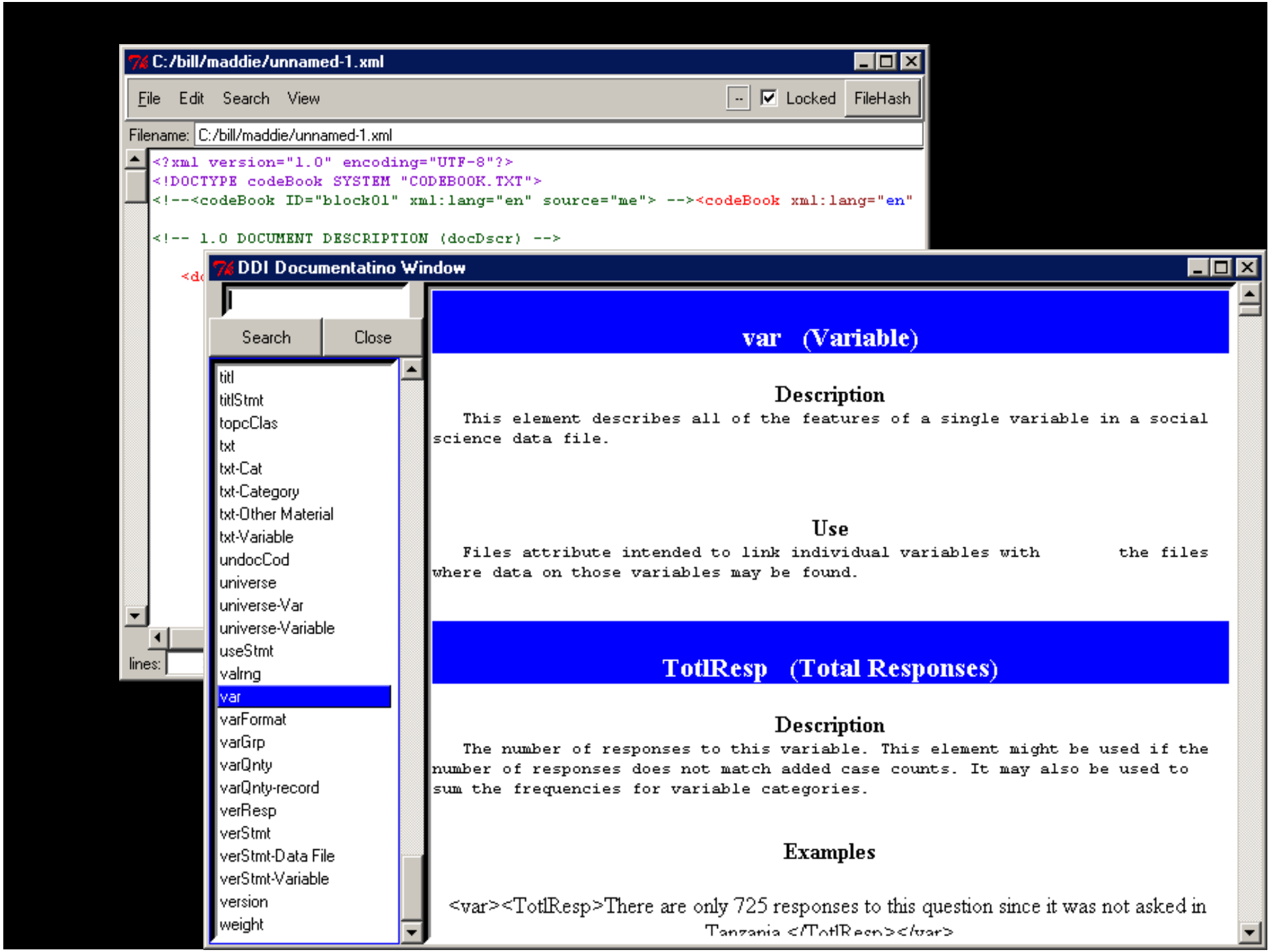
DDI Specific Editors

- NESSTAR Publisher
- MADDIE

Followed by
QUESTIONS


Nesstar Publisher
File Edit Tools View Window Help
Create xml
Import datafile
Validate
Publish
Preferences
C:\
lib
lib_pm
linux_download
maddie
maddie.installed
maddie.old
My Documents
gt
My Music
My Pictures
My Music
NCDTREE
OEMDRVRS
Program Files
Recycled
roberts stuff
SBPCI
Name
ag.xml
C:\My Documents\ag.xml
Xml editing
Nesstar preview
Guide To Codebook
docSrc
docSrc
Study Description
Bibliographic Citation
Study Scope
Methodology And Processing
Data Access
Data Files Description
File-By-File Description
Variable Description
All Agricultural and Population
All Agricultural Variables
Other Agricultural Varial
Farm Tenure
Farmland information
Value information on fa
Cost of building and rep
Cost of fertilizer purcha
Wage Labor informatio
Value of all farm produc
Grasslands, acreage a
Horses and Mules on h
Cattle and their product
Sheep and their produc
Swine on hand June 1,
Poultry and egg produc
Cereal crop acreage ar
Pulse production, 1879
Fiber acreage and prod
Sugar acreage and pro
Other crop acreage anc
Orchard acreage, produ
Nursery acreage and va
Vineyard acreage, sale
Value of market garden
Bees production, 1879
Forest production and v
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
Done


C:/bill/maddie/unnamed-1.xml
File Edit Search View
Locked
FileHash
Filename: C:/bill/maddie/unnamed-1.xml
<?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">
<titlStmt source="producer">
<titl source="producer"></titl>
<subTitl source="producer"></subTitl>
<altTitl source="producer"></altTitl>
<parTitl source="producer"></parTitl>
<IDNo source="producer"></IDNo>
</titlStmt>
<rspStmt source="producer">
<AuthEnty source="producer"></AuthEnty>
<othId source="producer"></othId>
</rspStmt>
<prodStmt source="producer">
<producer source="producer"></producer>
<copyright source="producer"></copyright>
<prodDate source="producer"></prodDate>
<prodPlac source="producer"></prodPlac>
lines: 308 Toplevel element name is "codeBook"
DDI XML Authoring Tool (Beta 0.6h)
File Technical Help
Dictionary: CODEBOOK.TXT
POWERED BY Perl
codeBook


C:/bill/maddie/unnamed-1.xml
File Edit Search View
Locked
FileHash
Filename: C:/bill/maddie/unnamed-1.xml
<?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) -->
lines:
DDI Documentatino Window
Search
Close
titl
titlStmt
topcClas
txt
txt-Cat
txt-Category
txt-Other Material
txt-Variable
undocCod
universe
universe-Var
universe-Variable
useStmt
valrng
var
varFormat
varGrp
varQnty
varQnty-record
verResp
verStmt
verStmt-Data File
verStmt-Variable
version
weight
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.
TotlResp (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><TotlResp>There are only 725 responses to this question since it was not asked in Tanzania.</TotlResp></var>

Wendy Thomas
wlt@pop.umn.edu

Bill Block
block@pop.umn.edu