A DDI3.2 Style for Data and Metadata Extracted from SAS

Larry Hoyle
Institute for Policy & Social Research
University of Kansas
Metadata Embedded in Proprietary Data Files

- SAS, Stata, SPSS, Excel, Relational Databases, R
  
  - All have some metadata which can be extracted
  - Variable names, labels, data types, formats
  - Missing value representation
  - Codelists

  - Integrity constraints (e.g. 0 <= age <= 130)
  - Implied information format (Euro.)
  - indexes
Tools

• Growing list of tools can harvest this information. Here are some:
  – Colectica
  – Dataset Documentation Manager
  – OpenDataForge Toolkit
  – DExT
  – Nesstar
  – RSpssConversion
  – SPSSOMS2DDI
  – StatTransfer
A Prototype Tool

Runs in SAS Enterprise Guide

Part of Documented Process Flow
Why Enterprise Guide?

- Document sequence of programs
- Rerun whole sequence
- REPLICATION!
SAS – Addin (C# .Net Program)
Tools... Add-in...”Extended Attributes”

Just needs the appropriate .dll file copied to the system

Editing Dataset Attributes

Dataset
Attribute
Value

Output Dataset

This is an abstract for the dataset:.....

DDI 5
DDI 3.2
DDI RDF
CDISC ???
Editing Variable Attributes

Variable
Attribute
Value
Pick from Drop-down List or Enter Value

Attribute
Pick from Drop-down List or Enter Value

Populated with DDI based terms plus any other extended attributes in the dataset
These should also map to a Semantic Web (RDF) style representation like DDI-Discovery
Codebook Generation (HTML)

Codebook for SAS Dataset: MYSASDATA

Dataset

Dataset Identifier: dataset identifier
Dataset Label: Test Data for SAS to DDI 3.2 program
Date Created: 2014-02-18T11:35:39.7
Date Last Modified: 2014-02-18T11:35:39.7
Number of Observations: 10
Number of Variables: 11
Encoding: wlatin1 Western (Windows)
Engine: V9

Abstract: dataset abstract
This is another reason to use a machine actionable representation for metadata – refactoring into representations like codebooks, or discovery web pages.
SAS INFORMATS

SAS variables may also have an associated informat which describes how the variable is to be read from a text representation. There were 4 informats defined in the SAS session which generated this documentation. Note that not all of these informats were necessarily in use by a variable.

**List Name: RANGEIN (String Range to Numeric Value RANGEIN.I)**

<table>
<thead>
<tr>
<th>Low</th>
<th>(exclusive)</th>
<th>High</th>
<th>(exclusive)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>' '</td>
<td>N</td>
<td>'5'</td>
<td>N</td>
<td>3</td>
</tr>
<tr>
<td>'0'</td>
<td>N</td>
<td>'9'</td>
<td>N</td>
<td>7.5</td>
</tr>
</tbody>
</table>

**List Name: RANGEIN (String Range to Character Value RANGEIN.J)**

<table>
<thead>
<tr>
<th>Low</th>
<th>(exclusive)</th>
<th>High</th>
<th>(exclusive)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>' '</td>
<td>N</td>
<td>'9'</td>
<td>N</td>
<td>'3'</td>
</tr>
<tr>
<td>'0'</td>
<td>N</td>
<td>'9'</td>
<td>N</td>
<td>'7.5'</td>
</tr>
</tbody>
</table>

**List Name: SEXIN (String to Character Value SEXIN.J)**

<table>
<thead>
<tr>
<th>Text</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>'F'</td>
<td>Female</td>
</tr>
<tr>
<td>'M'</td>
<td>Male</td>
</tr>
<tr>
<td>'f'</td>
<td>Female</td>
</tr>
<tr>
<td>'m'</td>
<td>Male</td>
</tr>
</tbody>
</table>

**List Name: SEXIN (String to Numeric Value SEXIN.I)**

<table>
<thead>
<tr>
<th>Text</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>'F'</td>
<td>2</td>
</tr>
<tr>
<td>'FEMALE'</td>
<td>2</td>
</tr>
<tr>
<td>'M'</td>
<td>1</td>
</tr>
<tr>
<td>'MALE'</td>
<td>1</td>
</tr>
</tbody>
</table>
Metadata Structured as DDI 3.2 (XML)
Review and Run the SAS Code

SAS Program:

data SASDATA.MYSASDATA_andXATTRS; set SASDATA.MYSASDATA; run;
title 'Contents of the Revised Dataset'; proc datasets lib=SASDATA nolist;
modify MYSASDATA_andXATTRS;
   XATTR SET VAR percentTime (Role='reject');
   XATTR SET DS TemporalCoverage='dataset time';
   XATTR SET VAR avocado (VariableIdentifier='identifier');
   XATTR SET VAR sex (LevelOfMeasurement='nominal');
   XATTR SET VAR avocado (Minimum='min');
   XATTR SET VAR avocado (Embargo='embargo');
   XATTR SET DS Study_DataCollectionDescription='study collected info';
   XATTR SET DS Abstract='dataset abstract';
   XATTR SET DS CopyRight='dataset copyright notice';
   XATTR SET DS BibliographicCitation='biblio citation';
   XATTR SET DS License='dataset license agreement';

SAS Log:

   The SAS System               09:37 Thursday, February 27, 2014
   OPTIONS PAGENO=MIN;
   data SASDATA.MYSASDATA_andXATTRS; set SASDATA.MYSASDATA; run;
   n
   NOTE: There were 10 observations read from the data set SASDATA.MYSASDATA.
   NOTE: The data set SASDATA.MYSASDATA_ANDXATTRS has 10 observations and 11 variables.
   n
   real time          0.03 seconds
   cpu time           0.01 seconds
Upon Finishing Runs Proc Dataset to Create New Dataset

### Name (attribute), value pairs

<table>
<thead>
<tr>
<th>Extended Attribute</th>
<th>Numeric Value</th>
<th>Character Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td></td>
<td>dataset abstract</td>
</tr>
<tr>
<td>AccessRights</td>
<td></td>
<td>dataset access rights</td>
</tr>
<tr>
<td>AlternativeTitle</td>
<td></td>
<td>dataset alternativetitle</td>
</tr>
<tr>
<td>BibliographicCitation</td>
<td></td>
<td>biblio citation</td>
</tr>
<tr>
<td>Concept</td>
<td></td>
<td>software test</td>
</tr>
<tr>
<td>Contributor</td>
<td></td>
<td>dataset contributor</td>
</tr>
<tr>
<td>Copyright</td>
<td></td>
<td>dataset copyright notice</td>
</tr>
<tr>
<td>Creator</td>
<td></td>
<td>dataset creator</td>
</tr>
<tr>
<td>DatasetIdentifier</td>
<td></td>
<td>dataset identifier</td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td>Test dataset for Extended Attributes Addin, extracting metadata</td>
</tr>
<tr>
<td>Embargo</td>
<td></td>
<td>dataset embargo</td>
</tr>
<tr>
<td>InstrumentDescription</td>
<td></td>
<td>dataset instrument</td>
</tr>
<tr>
<td>License</td>
<td></td>
<td>dataset license agreement</td>
</tr>
<tr>
<td>ProcessingStatus</td>
<td></td>
<td>processingstatus</td>
</tr>
<tr>
<td>Publisher</td>
<td></td>
<td>dataset publisher</td>
</tr>
<tr>
<td>SpatialCoverage</td>
<td></td>
<td>dataset place</td>
</tr>
<tr>
<td>Study_AnalysisUnit</td>
<td></td>
<td>study analysis unit</td>
</tr>
<tr>
<td>Study_CleaningOperation</td>
<td></td>
<td>study cleaning operation</td>
</tr>
</tbody>
</table>
Becomes Part of Process Flow

Documented data
What Structure to use?
DDI-L Style

• Valid DDI-L can be written in several ways for the same set of metadata
Software Incompatibilities

• DDI-L Style:
• Even though it shouldn’t matter, software importing DDI-L may handle only one style.
Alternative Styles

- Embed in StudyUnit
- ResourcePackage
- Fragments
StudyUnit

• “A primary packaging and publication module within DDI representing the purpose, background, development, data capture, and data products related to a study”

• There may be no study level metadata in a data file.
ResourcePackage

• “The Resource Package is a specialized structure which is intended to hold reusable metadata outside of the structures of a single StudyUnit or Group”

• DDI 3.1 alternative when no study level information.
Fragment (DDI 3.2)

- “A Fragment is a means of transporting a maintainable or versionable object plus any associated notes and other material.”
Custom Attributes

• SAS, Stata, SPSS, Excel, R all can now have custom, or extended attributes as (name, value) pairs on variables or the dataset.

• Opens up the possibility of embedded lifecycle information, including at the study level
Example Custom Attribute

(Embargo, one year)

name

value

Applies to

the dataset
Example Custom Attribute

(Embargo, one year)

name

the dataset

value

Applies to

You could also think of this as a sentence (a triple)

“The dataset has an embargo “one year”

subject, predicate, object
Representing Custom Attributes

(Embargo, one year)

As if attribute has no meaning within DDI

```xml
<r:UserAttributePair>
  <r:AttributeKey>Embargo</r:AttributeKey>
  <r:AttributeValue>one year</r:AttributeValue>
</r:UserAttributePair>
```

This also explicitly documents that this pair was in the dataset
Alternative – Assign Selected Terms Meaning

(Embargo, one year)

<l:Variable>
...

<l:EmbargoReference>
  <r:URN>URN:DDI:exampleagency:e1:1.0</r:URN>
  <r:TypeOfObject>Embargo</r:TypeOfObject>
</l:EmbargoReference>

<r:Embargo>
  <r:URN>URN:DDI:exampleagency:e1:1.0</r:URN>
  <r:EmbargoName>
    <r:String>Embargo for the variable avocado</r:String>
  </r:EmbargoName>
  <r:Description>
    <r:Content>one year</r:Content>
  </r:Description>
</r:Embargo>
Current Structured Dataset Attributes

- BibliographicCitation
- Abstract
- Title
- Subtitle
- AlternativeTitle
- Creator
- Contributor
- Publisher
- DatasetIdentifier
- Description
- Note
- AccessRights
- CopyRight
- License
- Embargo
- ProcessingStatus

- SpatialCoverage
- TemporalCoverage
- TopicalCoverage
- InstrumentDescription

- Study_AnalysisUnit
- Study_CleaningOperation
- Study_CollectionMethodology
- Study_DataCollectionDescription
- Study_FundingInformation
- Study_KindOfData
- Study_LifecycleEvents
- Study_ProcessingDescription
- Study_Purpose
- Study_SamplingProcedure
- Study_Universe
Require a StudyUnit

All of these require a StudyUnit

- Study_AnalysisUnit
- Study_CleaningOperation
- Study_CollectionMethodology
- Study_DataCollectionDescription
- Study_FundingInformation
- Study_KindOfData
- Study_LifecycleEvents
- Study_ProcessingDescription
- Study_Purpose
- Study_SamplingProcedure
- Study_Universe
Current Structured Dataset Attributes

- BibliographicCitation
- Abstract
- Title
- Subtitle
- AlternativeTitle
- Creator
- Contributor
- Publisher
- DatasetIdentifier
- Description
- Note
- AccessRights
- CopyRight
- License
- Embargo
- ProcessingStatus

- SpatialCoverage
- TemporalCoverage
- TopicalCoverage
- InstrumentDescription

Some of these may apply to the study as well, but they do not require a StudyUnit
DDI Profile Issue?

• Could a recommended DDI profile be enough to specify a recommended style?
  
  – Exclude ResourcePackage?
  
  – Exclude sub-elements in StudyUnit that can be in Fragments?
Profile Possibility?  
In Separate Fragment

ddi:FragmentInstance/ddi:TopLevelReference

ddi:Fragment/c:Concept  
ddi:Fragment/l:CategoryScheme  
ddi:Fragment/l:CodeListScheme  
ddi:Fragment/r:ManagedRepresentationScheme  
ddi:Fragment/r:Note  
ddi:Fragment/d:QuestionItem  
ddi:Fragment/m4:RecordLayout  
ddi:Fragment/c:Universe  
ddi:Fragment/l:VariableScheme
Profile Possibility?
In Archive Elements

ddi:Fragment/a:Archive/r:URN
  ddi:Fragment/a:Archive/
  r:LifecycleInformation
Profile Possibility?
In PhysicalInstance Elements

ddi:Fragment/pi:PhysicalInstance/r:Citation
ddi:Fragment/pi:PhysicalInstance/r:Software
ddi:Fragment/pi:PhysicalInstance/r:URN
ddi:Fragment/pi:PhysicalInstance/r:UserAttributePair
ddi:Fragment/pi:PhysicalInstance/r:UserID
Profile Possibility?
In StudyUnit Elements

ddi:Fragment/s:StudyUnit/r:AnalysisUnit
ddi:Fragment/s:StudyUnit/r:ArchiveReference
ddi:Fragment/s:StudyUnit/r:Embargo
ddi:Fragment/s:StudyUnit/r:FundingInformation
ddi:Fragment/s:StudyUnit/r:KindOfData
ddi:Fragment/s:StudyUnit/r:PhysicalInstanceReference
ddi:Fragment/s:StudyUnit/r:Purpose
ddi:Fragment/s:StudyUnit/r:UniverseReference
ddi:Fragment/s:StudyUnit/r:URN
Profile Possibility?
In StudyUnit/DataCollection Elements

```xml
ddi:Fragment/s:StudyUnit/d:DataCollection/r:Description
ddi:Fragment/s:StudyUnit/
d:DataCollection/d:InstrumentScheme
ddi:Fragment/s:StudyUnit/d:DataCollection/
d:Methodology
ddi:Fragment/s:StudyUnit/d:DataCollection/
d:ProcessingEventScheme
ddi:Fragment/s:StudyUnit/d:DataCollection/
d:ProcessingInstructionScheme
ddi:Fragment/s:StudyUnit/d:DataCollection/r:URN
```
Profile Possibility?
In StudyUnit/LogicalProduct Elements

Fragment/s:StudyUnit/l:LogicalProduct/r:CategorySchemeReference
Fragment/s:StudyUnit/l:LogicalProduct/r:CodeListSchemeReference
Fragment/s:StudyUnit/l:LogicalProduct/r:ManagedRepresentationSchemeReference
Fragment/s:StudyUnit/l:LogicalProduct/r:URN
Fragment/s:StudyUnit/l:LogicalProduct/r:VariableSchemeReference
Contact

Larry Hoyle

Senior Scientist
Institute for Policy & Social Research, University of Kansas
1541 Lilac Lane Suite 607 Blake
Lawrence, KS 66045-3129

LarryHoyle@ku.edu

ExtendedAttributes Addin

• The project has been archived at:
  
  • http://kuscholarworks.ku.edu/dspace/handle/1808/12488