

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

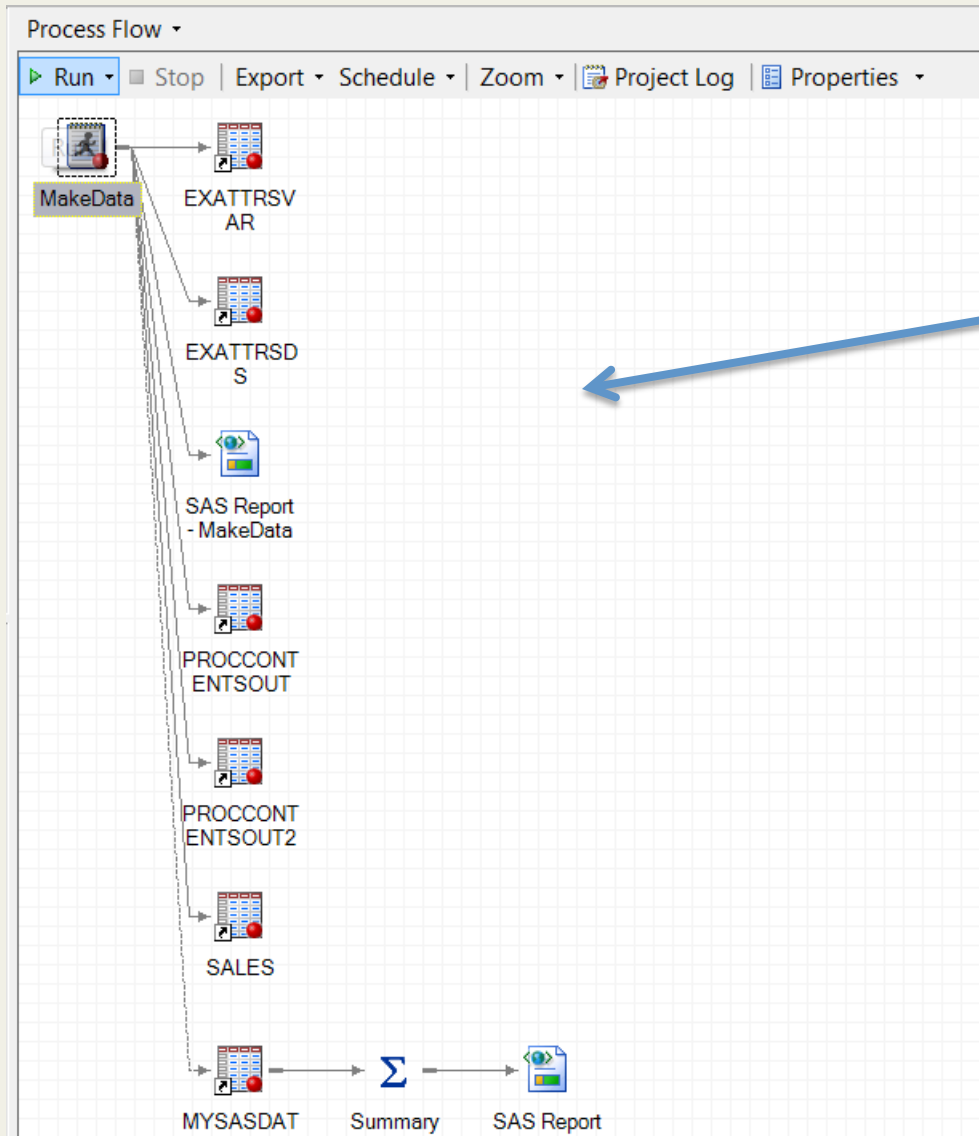
- SAS, Stata, SPSS, Excel, Relational Databases, R
 - All have some metadata which can be extracted
 - Variable names, labels, data types, formats
 - Missing value representation
 - Codelists

 - Integrity constraints (e.g. $0 \leq \text{age} \leq 130$)
 - 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

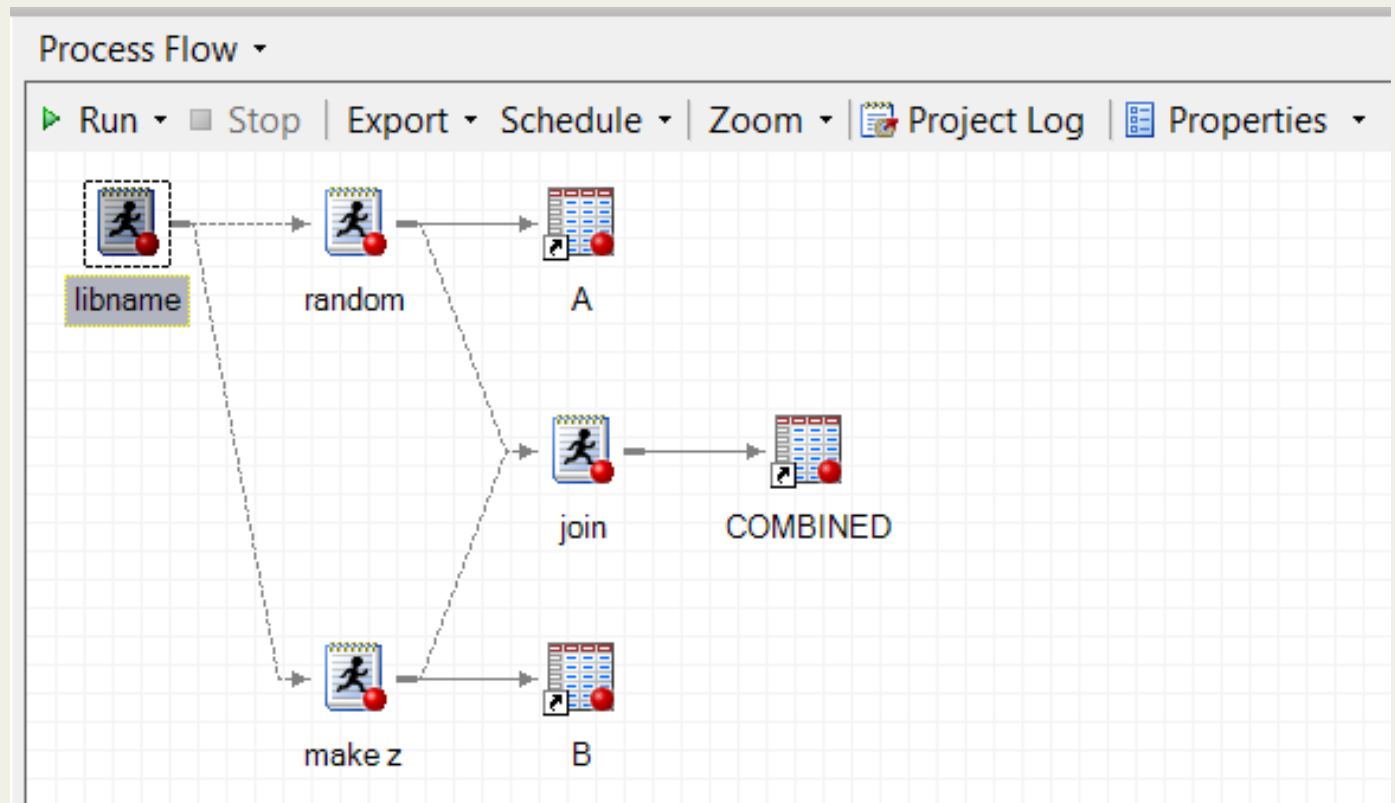
A Prototype Tool



Runs in SAS
Enterprise Guide

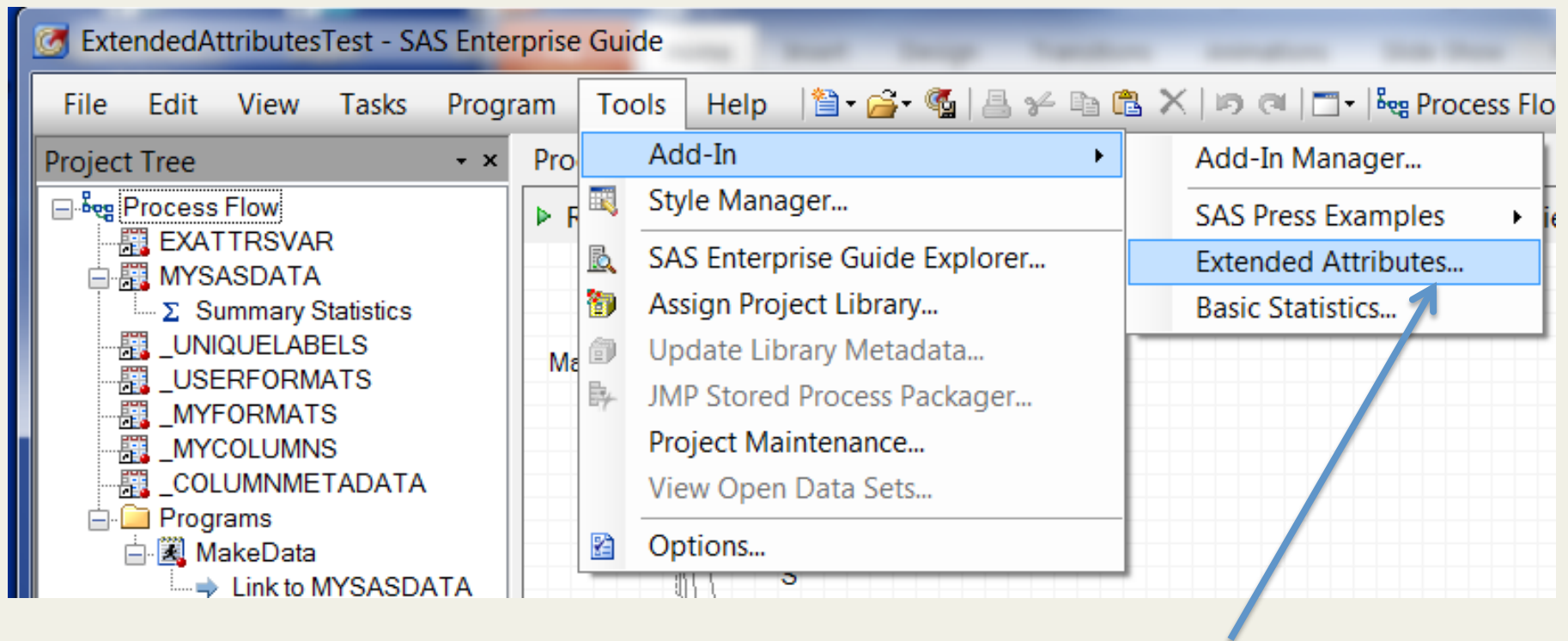
Part of
Documented
Process Flow

Why Enterprise Guide?



- Document sequence of programs
- Rerun whole sequence
- REPLICATION !

SAS – Addin (C# .Net Program) Tools... Add-in...”Extended Attributes”



See: Hemedinger, Chris. 2012. *Custom Tasks for SAS® Enterprise Guide® Using Microsoft .NET*. Cary, NC: SAS Institute Inc.

Just needs the appropriate .dll file copied to the system

Editing Dataset Attributes

Dataset

Attribute

Value

Explore or Set Extended Attributes

Edit Test SAS Properties DDI32 Codebook Glossary

Server
Local

Library (Libname)
SASDATA

Input Dataset (Memname)
MYSASDATA

Select a Variable
{DATASET_Attribute}

Select or Enter an Extended Variable Attribute Name
Abstract

Enter an Attribute Value for the Variable/Attribute Combination
This is an abstract for the dataset.....

Agency
exampleagency.subagency

Default Version
1.0

Output Dataset Name
MYSASDATA_andXATTRs

DDI2.5

DDI3.2

DDI RDF

CDISC ???

EnterAttribute Delete Attribute

OK-Finish Cancel

Output Dataset

Editing Variable Attributes

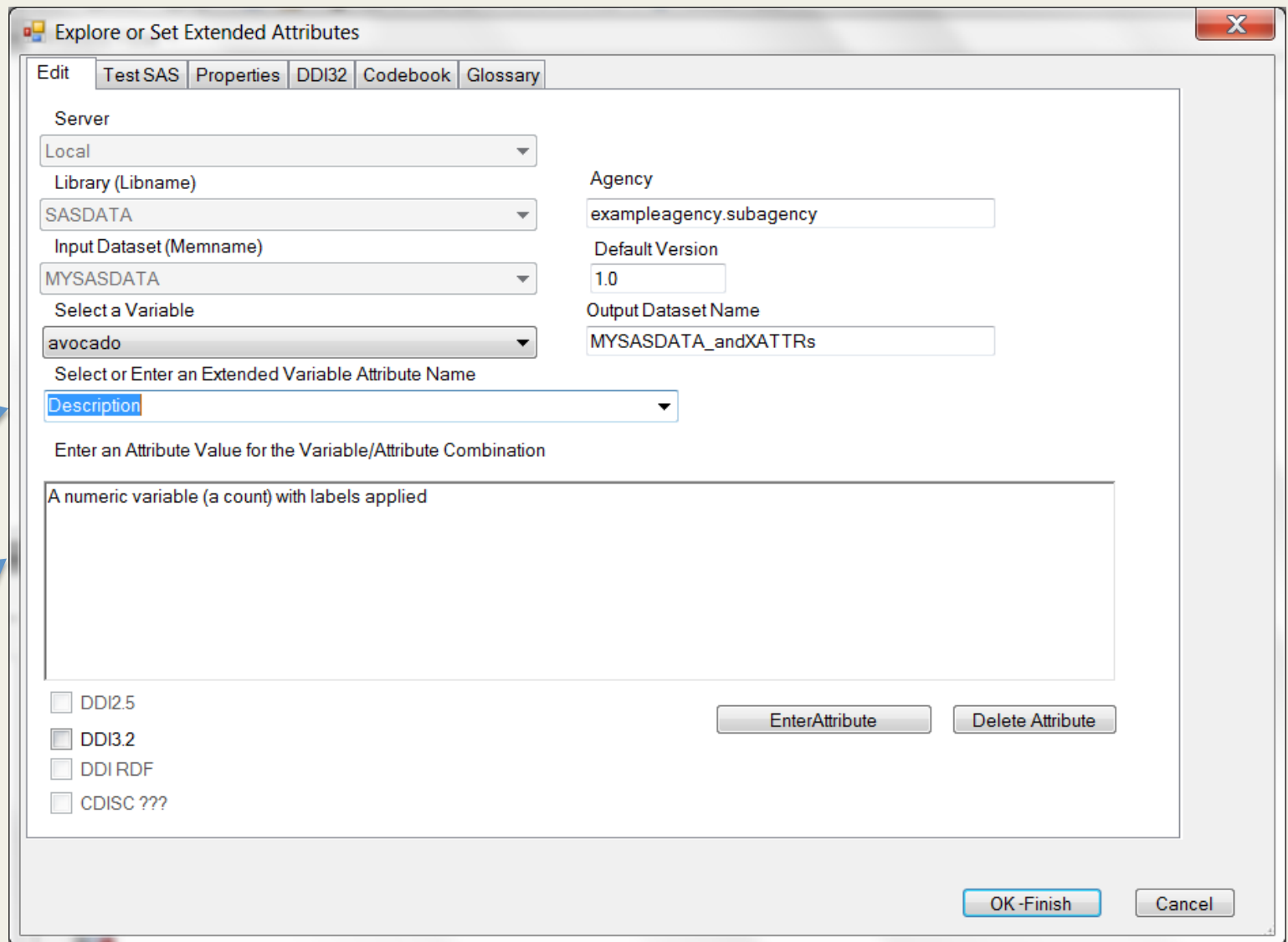
Variable



Attribute



Value



Explore or Set Extended Attributes

Edit Test SAS Properties DDI32 Codebook Glossary

Server: Local

Library (Libname): SASDATA

Input Dataset (Memname): MYSASDATA

Select a Variable: avocado

Select or Enter an Extended Variable Attribute Name: Description

Agency: exampleagency.subagency

Default Version: 1.0

Output Dataset Name: MYSASDATA_andXATTRs

Enter an Attribute Value for the Variable/Attribute Combination

A numeric variable (a count) with labels applied

DDI2.5 DDI3.2 DDI RDF CDISC ???

EnterAttribute Delete Attribute

OK-Finish Cancel

Pick from Drop-down List or Enter Value

Explore or Set Extended Attributes

Edit Test SAS Properties DDI32 Codebook Glossary

Server
Local

Library (Libname)
SASDATA

Input Dataset (Memname)
MYSASDATA

Select a Variable
avocado

Agency
exampleagency.subagency

Default Version
1.0

Output Dataset Name
MYSASDATA_andXATTRs

Select or Enter an Extended Variable Attribute Name

Description

AccessRights
Additivity
AggregationMethod
AnalysisUnit
CategoryStandard
Concept
Description
Embargo
VariableIdentifier
ImputationDescription
LevelOfMeasurement
MeasurementUnits
Notes
ProcessingDescription
Question
RelevantFormats
Role
Scale
SourceUnit
Universe
WeightVariable
Derivation
Minimum

EnterAttribute Delete Attribute

OK-Finish Cancel

Attribute

Pick from Drop-down List or Enter Value

Populated with [DDI](#) based terms plus any other extended attributes in the dataset

The screenshot shows a dialog box titled "Explore or Set Extended Attributes" with a menu bar containing "Edit", "Test SAS", "Properties", "DDI32", "Codebook", and "Glossary". The main area contains several input fields and a dropdown menu:

- Server: Local
- Library (Libname): SASDATA
- Input Dataset (Memname): MYSASDATA
- Select a Variable: avocado
- Select or Enter an Extended Variable Attribute Name: Description (dropdown menu is open, showing a list of attributes including AccessRights, Additivity, AggregationMethod, AnalysisUnit, CategoryStandard, Concept, Description, Embargo, VariableIdentifier, ImputationDescription, LevelOfMeasurement, MeasurementUnits, Notes, ProcessingDescription, Question, RelevantFormats, Role, Scale, SourceUnit, Universe, WeightVariable, Derivation, and Minimum)
- Agency: exampleagency.subagency
- Default Version: 1.0
- Output Dataset Name: MYSASDATA_andXATTRs

Buttons at the bottom include "EnterAttribute", "Delete Attribute", "OK-Finish", and "Cancel".

Glossary of Terms

Explore or Set Extended Attributes

Edit Test SAS Properties DDI32 Codebook **Glossary**

Study and Dataset

- Abstract An abstract describing the study and dataset
- AccessRights Describes access conditions and terms of use for the data
- AlternativeTitle Any alternative title for the study or dataset
- Study_AnalysisUnit A description of the type of object studied e.g. persons
- bibliographicCitation Recommended citation for the dataset
- Study_CleaningOperation A text description of the cleaning done on the data
- Study_CollectionMethodology The methodology and processing involved in a data collection.
- Contributor Contributor to the creation of the dataset or study

Variables

- AccessRights Describes access conditions and terms of use for the variable
- Additivity e.g. ("stock" | "flow" | "non-additive" | "other")
- AggregationMethod e.g. ("sum" | "average" | "count" | "mode" | "median" | "maximum" | "minimum" | "percent" | "other")
- AnalysisUnit information regarding whom or what the variable describes
- CategoryStandard Standard category codes used in the variable, like industry codes, employment codes, or social class codes

OK-Finish Cancel

View of Internal Representation of Selections

Explore or Set Extended Attributes

Edit Test SAS Properties DDI32 Codebook Glossary

PropertiesEntered

```
<VariableAttributeElement>Minimum</VariableAttributeElement>
</VariableAttributeElements>
<XATTRs>
  <XATTR>
    <XVariable>percentTime</XVariable>
    <XName>Role</XName>
    <XValue>reject</XValue>
  </XATTR>
  <XATTR>
    <XVariable>{DATASET_Attribute}</XVariable>
    <XName>TemporalCoverage</XName>
    <XValue>dataset time</XValue>
  </XATTR>
  <XATTR>
```

subject

predicate

object

These should also map to a Semantic Web (RDF) style representation like DDI-Discovery

Codebook Generation (HTML)

Codebook for SAS Dataset: MYSASDATA - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Generic Statistical Information Model... Codebook for SAS Dataset: MYSASD...

file:///D:/junk/Codebook4.html GSIM

Latest Headlines Google News Weather KU ipsr SAS IMLSgrant Data

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

_____ extended attributes _____

Abstract: dataset abstract

Dataset, Variables, Codelists

II Variable: fee

Label: Fee in Euros
SASFormat: EUROX10.2
FormatDescription: Writes numeric values with a leading euro symbol (E), a period that separates every three digits, and a comma that separates the decimal fraction
Type: Numeric, internal bytes : 8
Transcode: yes
SortedBy: 0
Represents: Currency-euros

Codelists (Formats, Value Labels)

There were 8 formats defined in the SAS session which generated this documentation. Note that not all of these formats were necessarily in use by a variable.

List Name: AVOCADONUMBER (Numeric Ranges AVOCADONUMBER.N)

| Low | (exclusive) | High | (exclusive) | Label |
|---------------|-------------|---------------|-------------|------------------|
| 1 | N | 1 | N | lonley avocado |
| 1 | Y | 6.02214149E23 | N | too few avocados |
| 6.02214149E23 | Y | 6.02214209E23 | N | guaca mole |
| 6.02214209E23 | Y | HIGH | N | a party |
| LOW | N | 0 | Y | avocados owed |

List Name: SEX (Character Codelist SEX.C)

| Value | Label |
|-------|--------|
| 'f' | female |
| 'm' | male |

This is another reason to use a machine actionable representation for metadata – refactoring into representations like [codebooks](#), or discovery web pages

SAS Informats

Codebook for SAS Dataset: MYSASDATA - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Generic Statistical Information Model... Codebook for SAS Dataset: MYSASD...

file:///D:/junk/Codebook4.html

Latest Headlines Google News Weather KU ipsr SAS IMLSgrant Data software

SAS variables may also have an associated 'informat' which describes how the variable is to be read from a text representation.

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)

| Low | (exclusive) | High | (exclusive) | Value |
|-----|-------------|------|-------------|-------|
| '1' | N | '5' | N | 3 |
| '6' | N | '9' | N | 7.5 |

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

| Low | (exclusive) | High | (exclusive) | Value |
|-----|-------------|------|-------------|-------|
| 'a' | N | 'e' | N | '3' |
| 'v' | N | 'z' | N | '7.5' |

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

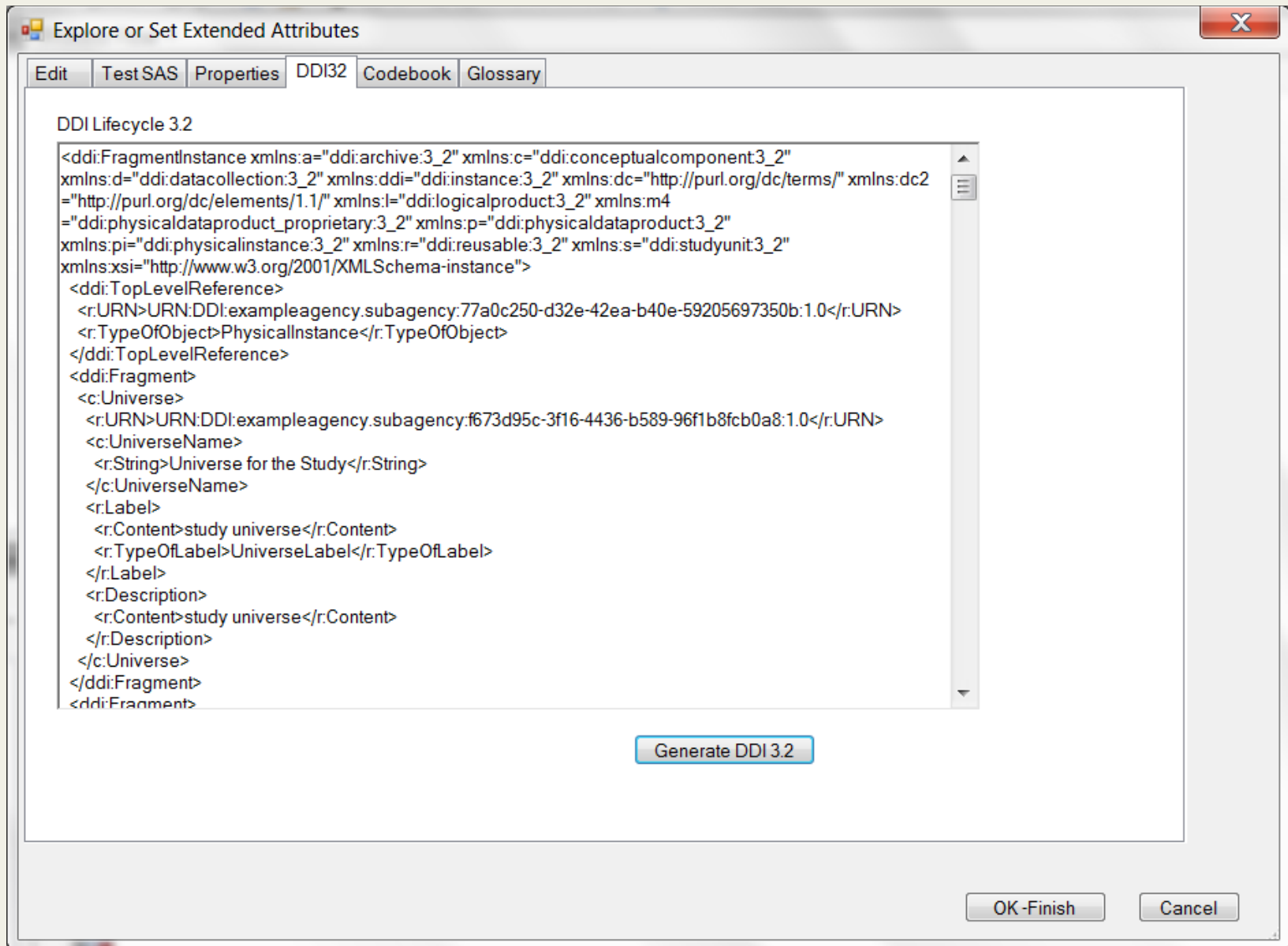
| Text | Value |
|------|----------|
| 'F' | 'Female' |
| 'M' | 'Male' |
| 'f' | 'Female' |
| 'm' | 'Male' |

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

| Text | Value |
|----------|-------|
| 'F' | 2 |
| 'FEMALE' | 2 |
| 'M' | 1 |
| 'MALE' | 1 |

codebook generated at 2/22/2014 12:42:17 PM (2/22/2014 06:42:17 PM UTC)

Metadata Structured as DDI 3.2 (XML)



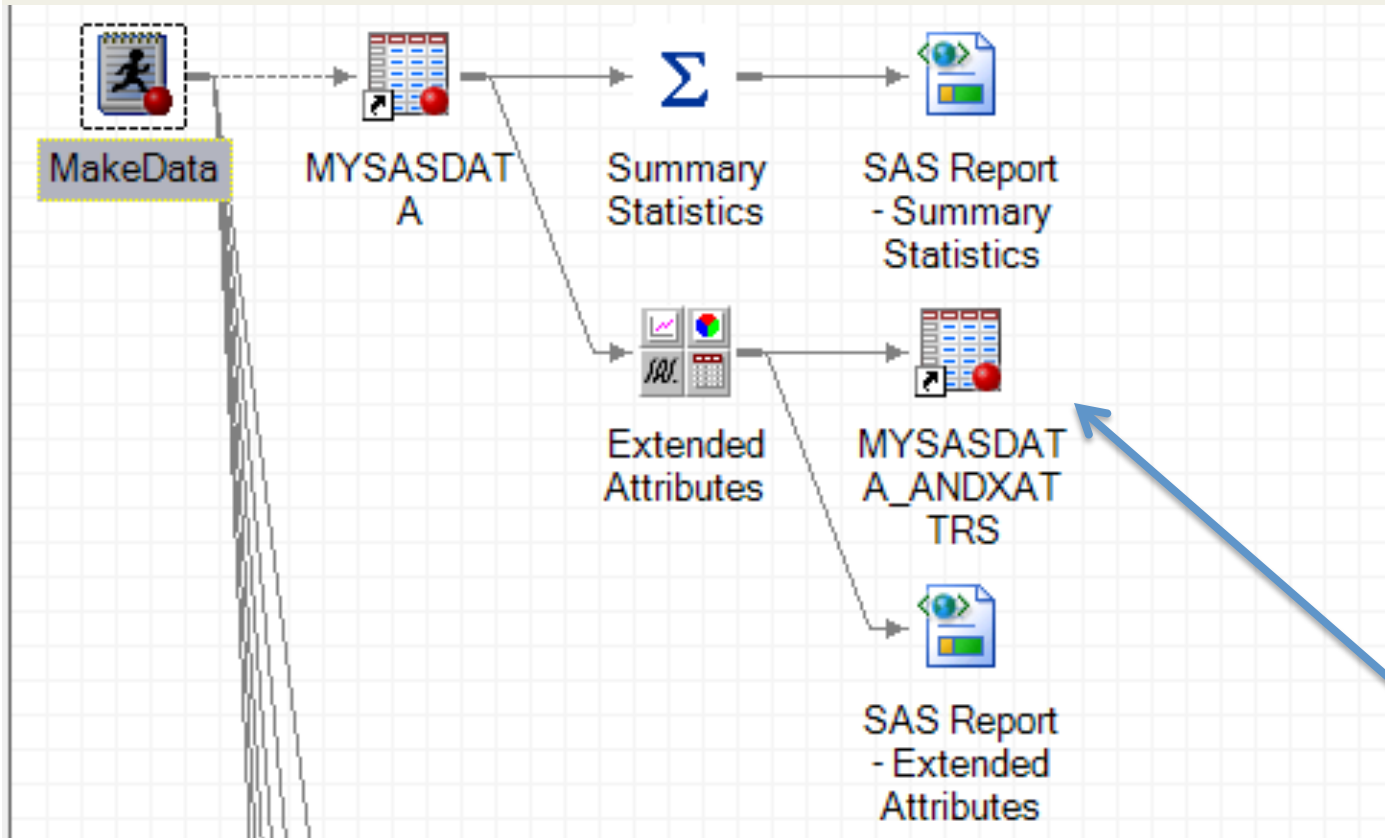
Upon Finishing Runs Proc Dataset to Create New Dataset

| Alphabetic List of Indexes and Attributes | | | | |
|---|---------|---------------|-------------|--------------------|
| # | Index | Unique Option | Owned by IC | # of Unique Values |
| 1 | ID | YES | YES | 10 |
| 2 | avocado | YES | YES | 10 |
| 3 | name | | YES | 10 |

Name (attribute), value pairs

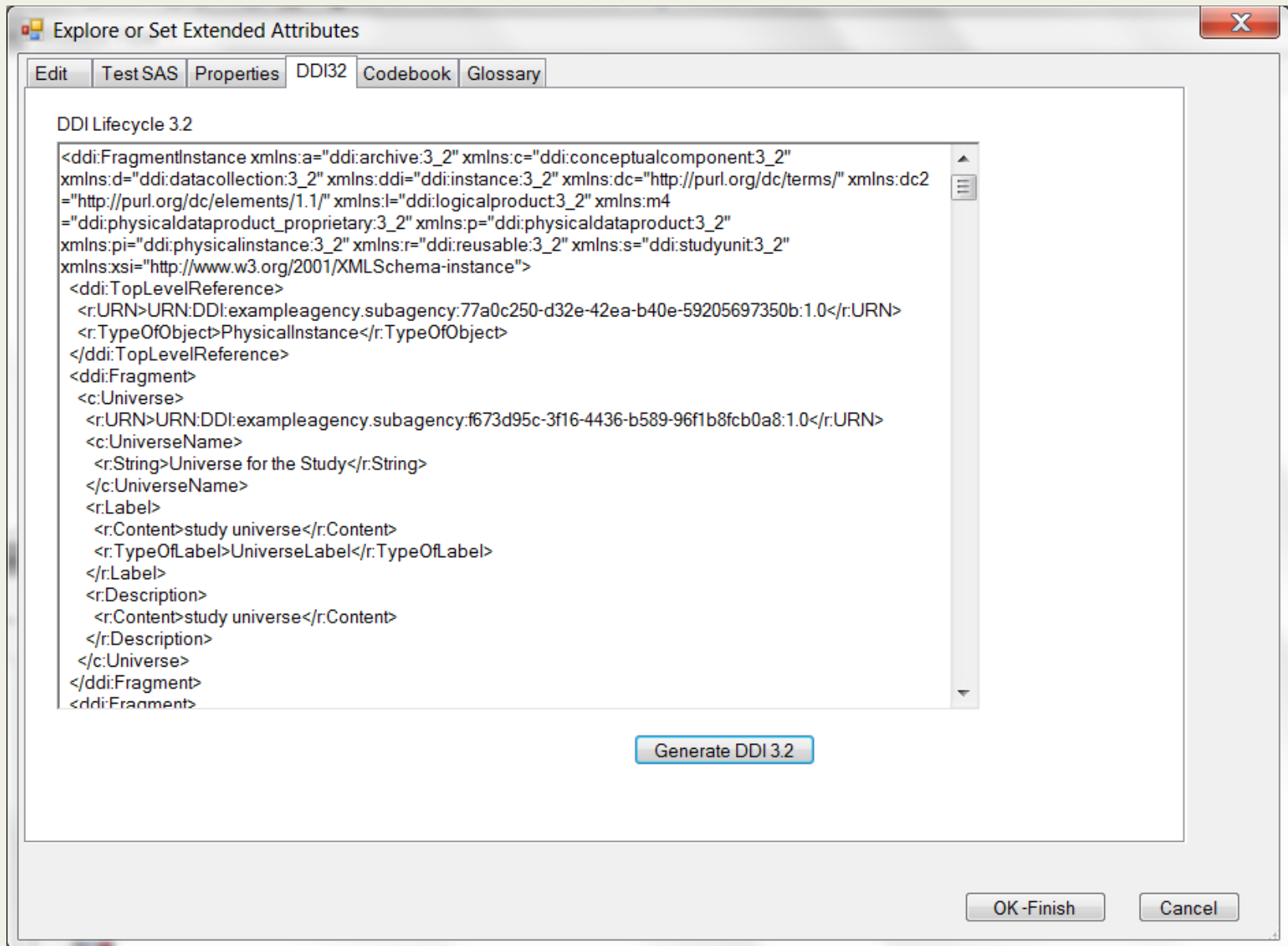
| Alphabetic List of Data Set Extended Attributes | | |
|---|---------------|---|
| Extended Attribute | Numeric Value | Character Value |
| Abstract | . | dataset abstract |
| AccessRights | . | dataset access rights |
| AlternativeTitle | . | dataset alternativetitle |
| BibliographicCitation | . | biblio citation |
| Concept | . | software test |
| Contributor | . | dataset contributor |
| CopyRight | . | dataset copyright notice |
| Creator | . | dataset creator |
| DatasetIdentifier | . | dataset identifier |
| Description | . | Test dataset for Extended Attributes Addin, extracting metadatata |
| Embargo | . | dataset embargo |
| InstrumentDescription | . | dataset instrument |
| License | . | dataset license agreement |
| ProcessingStatus | . | processingstatus |
| Publisher | . | dataset publisher |
| SpatialCoverage | . | dataset place |
| Study_AnalysisUnit | . | study analysis unit |
| Study_CleaningOperation | . | study cleaning operation |

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.

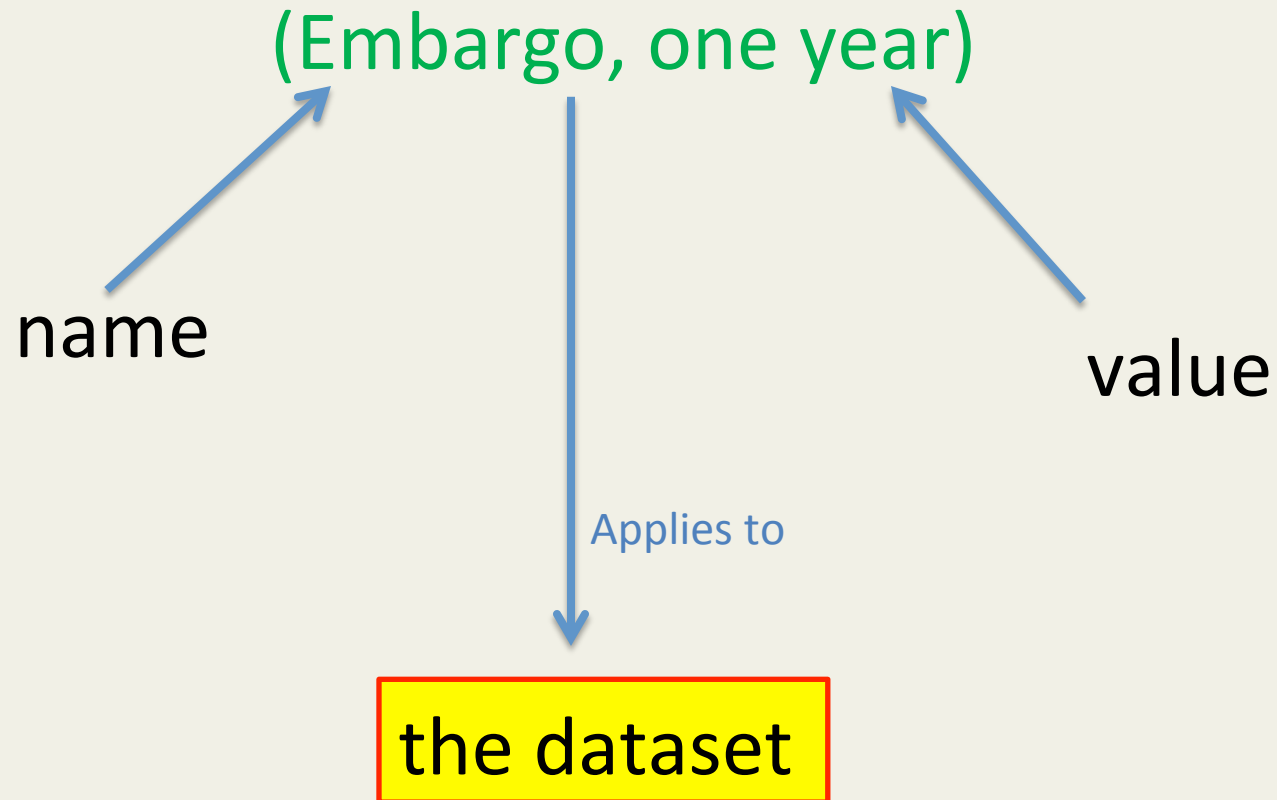
Fragment (DDI 3.2)

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

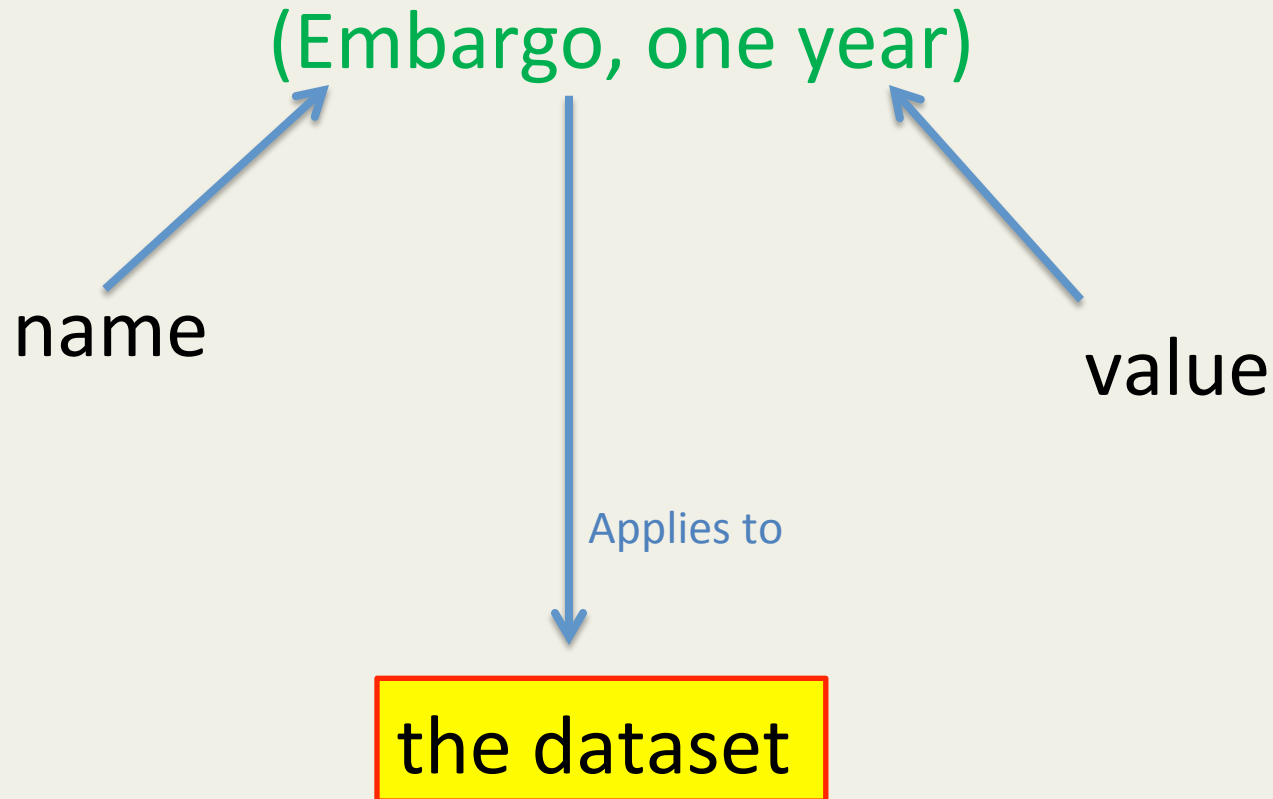
Custom Attributes

- 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



Example Custom Attribute



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

```
<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>
```

```
</l:Variable>
```

```
<r:Embargo>
```

```
  <r:URN>URN:DDI:exampleagency :e1:1.0</r:URN>
```

```
  <r:EmbargoName>
```

```
    <r:String>Embargo for the variable avocado</r:String>
```

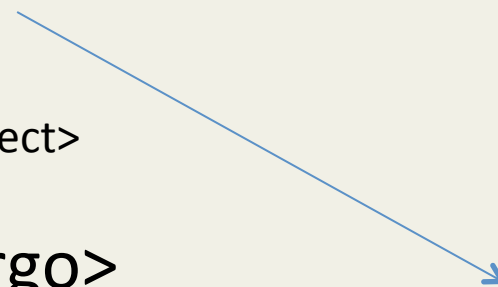
```
  </r:EmbargoName>
```

```
  <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

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>