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

Larry Hoyle

Institute for Policy & Social Research

University of Kansas

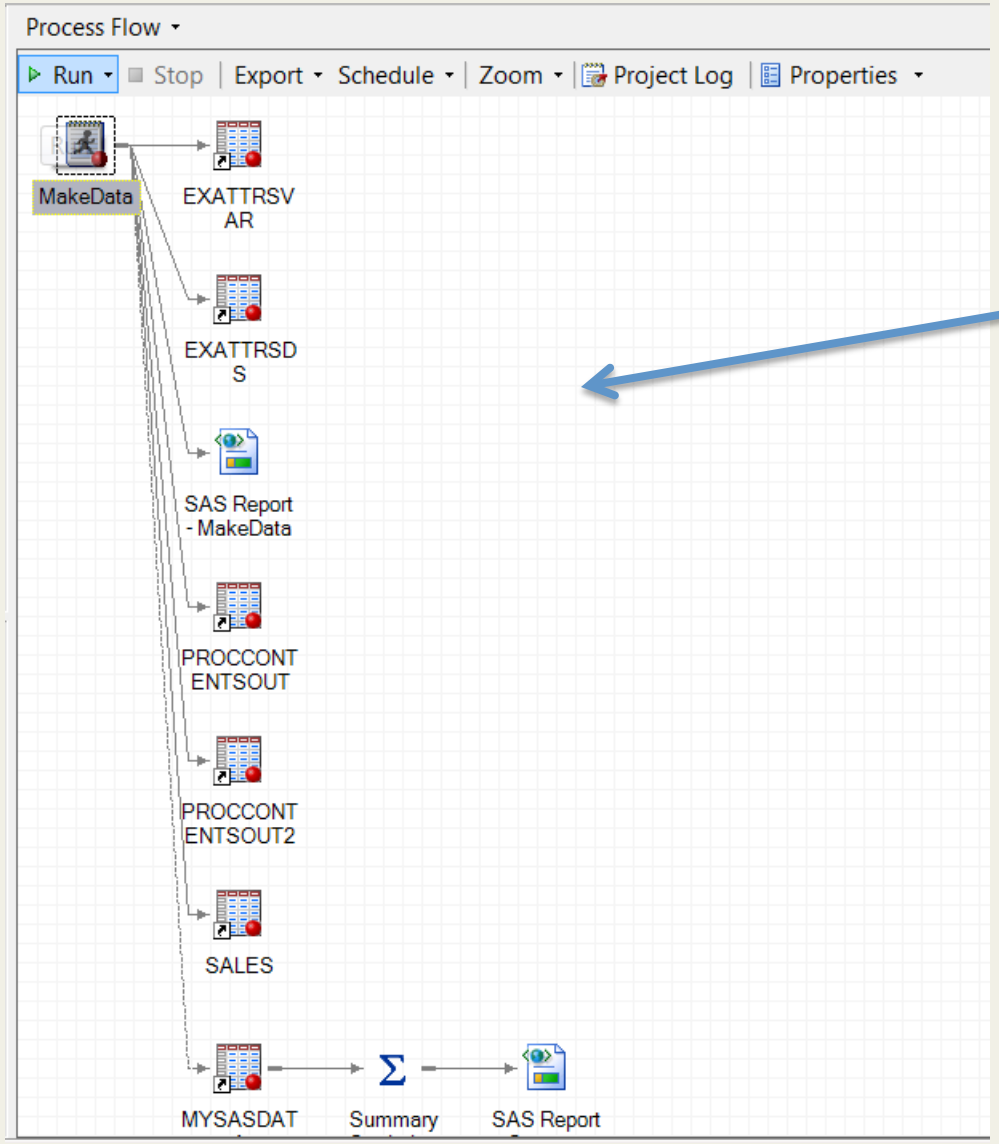
Metadata Embedded in Proprietary Data Files

- 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

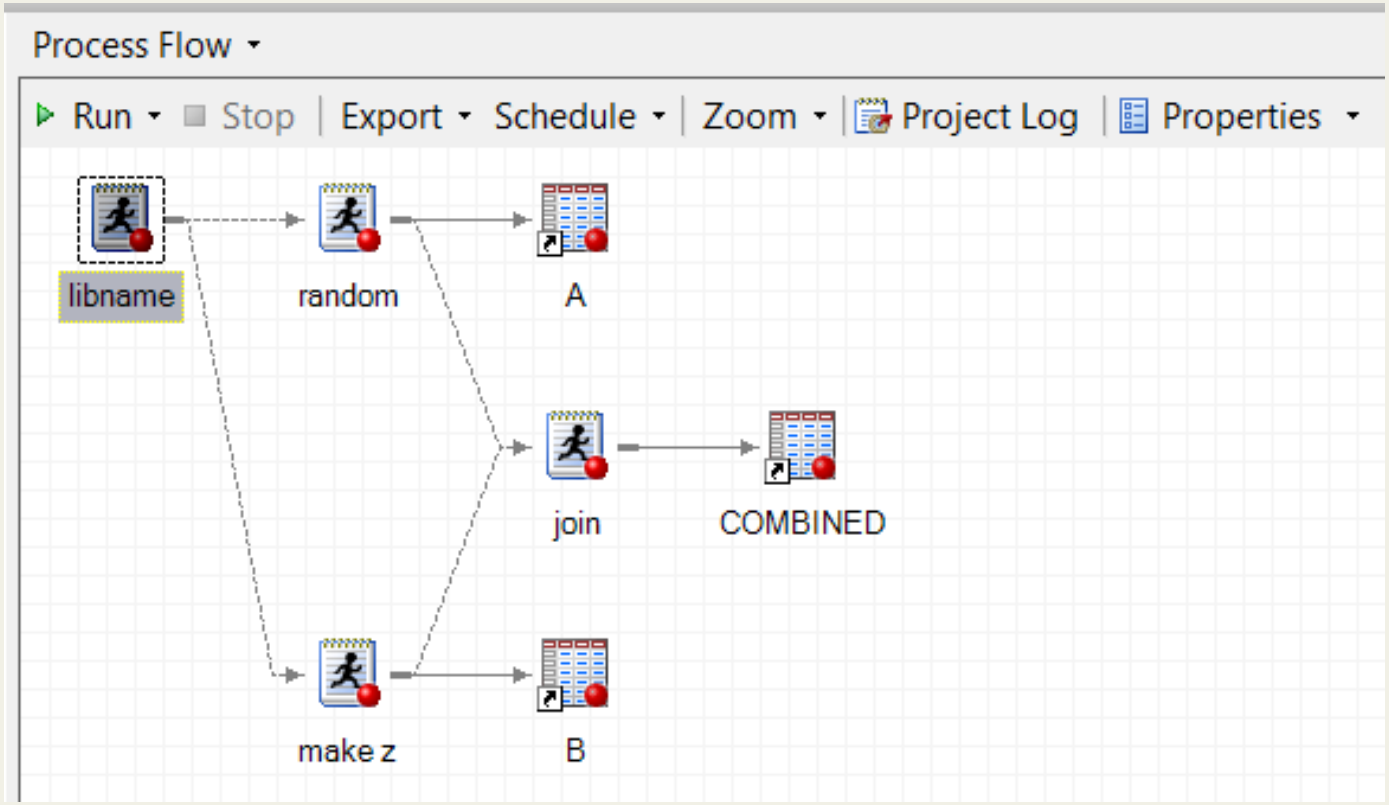
A Prototype Tool



Runs in SAS **Enterprise Guide**

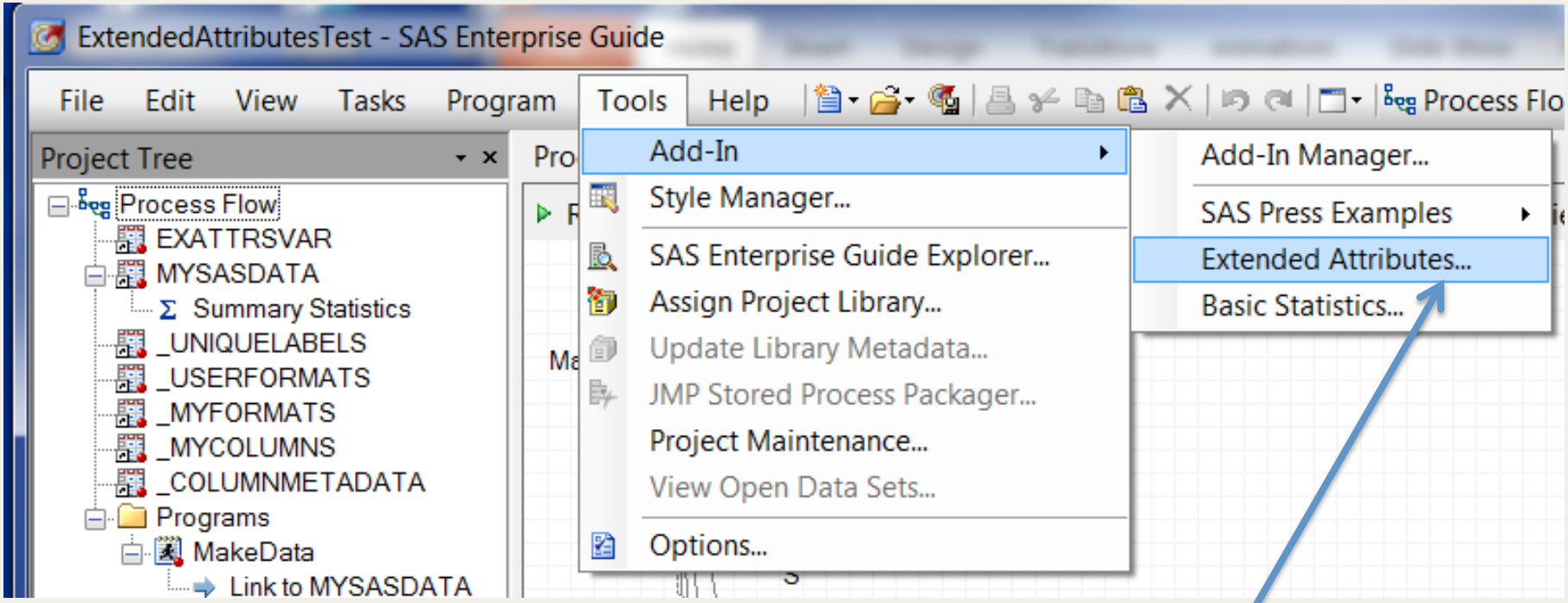
Part of Documented Process Flow

Why Enterprise Guide?



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

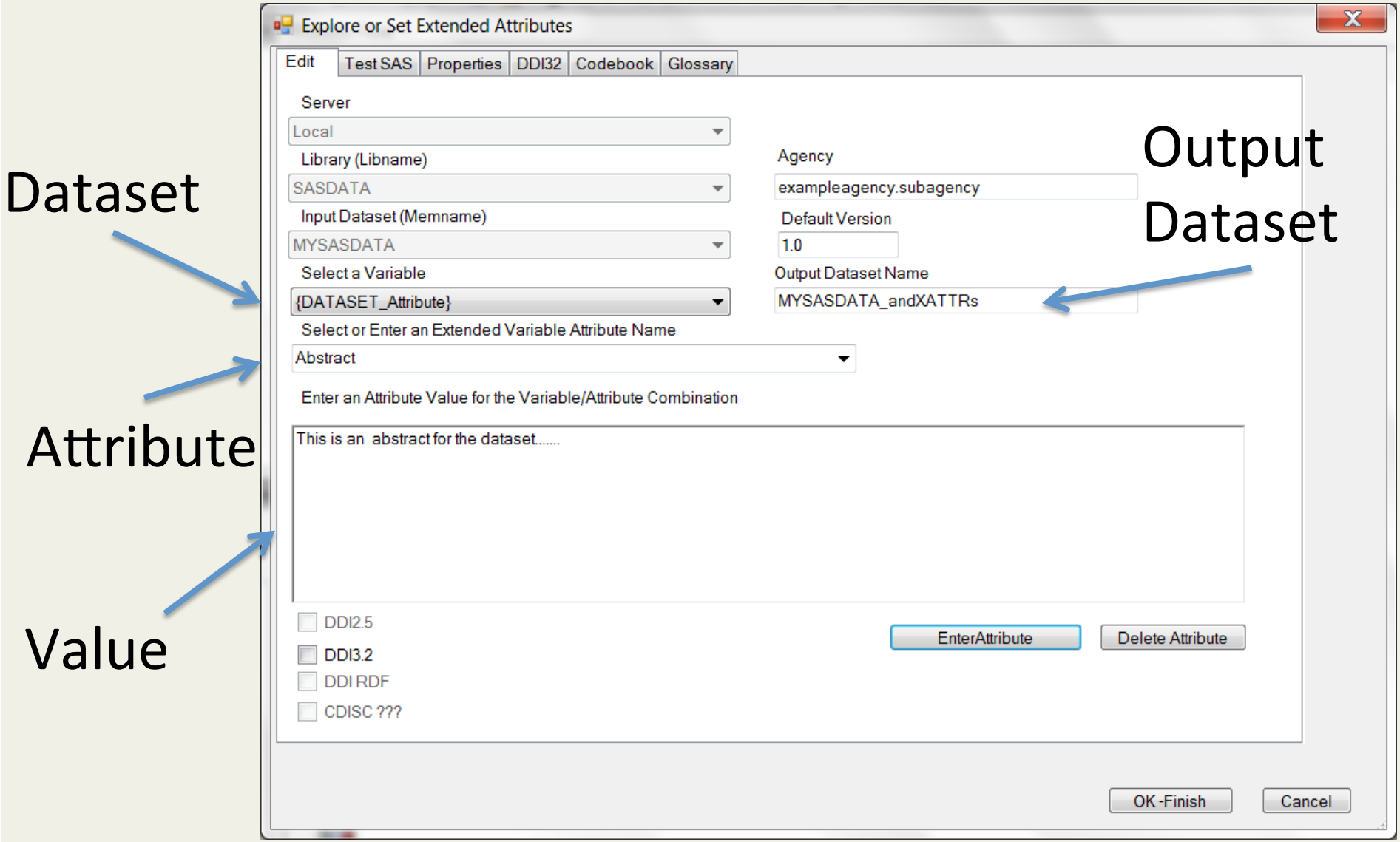
SAS – Addin (C# .Net Program)
Tools… Add-in…"Extended Attributes"



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

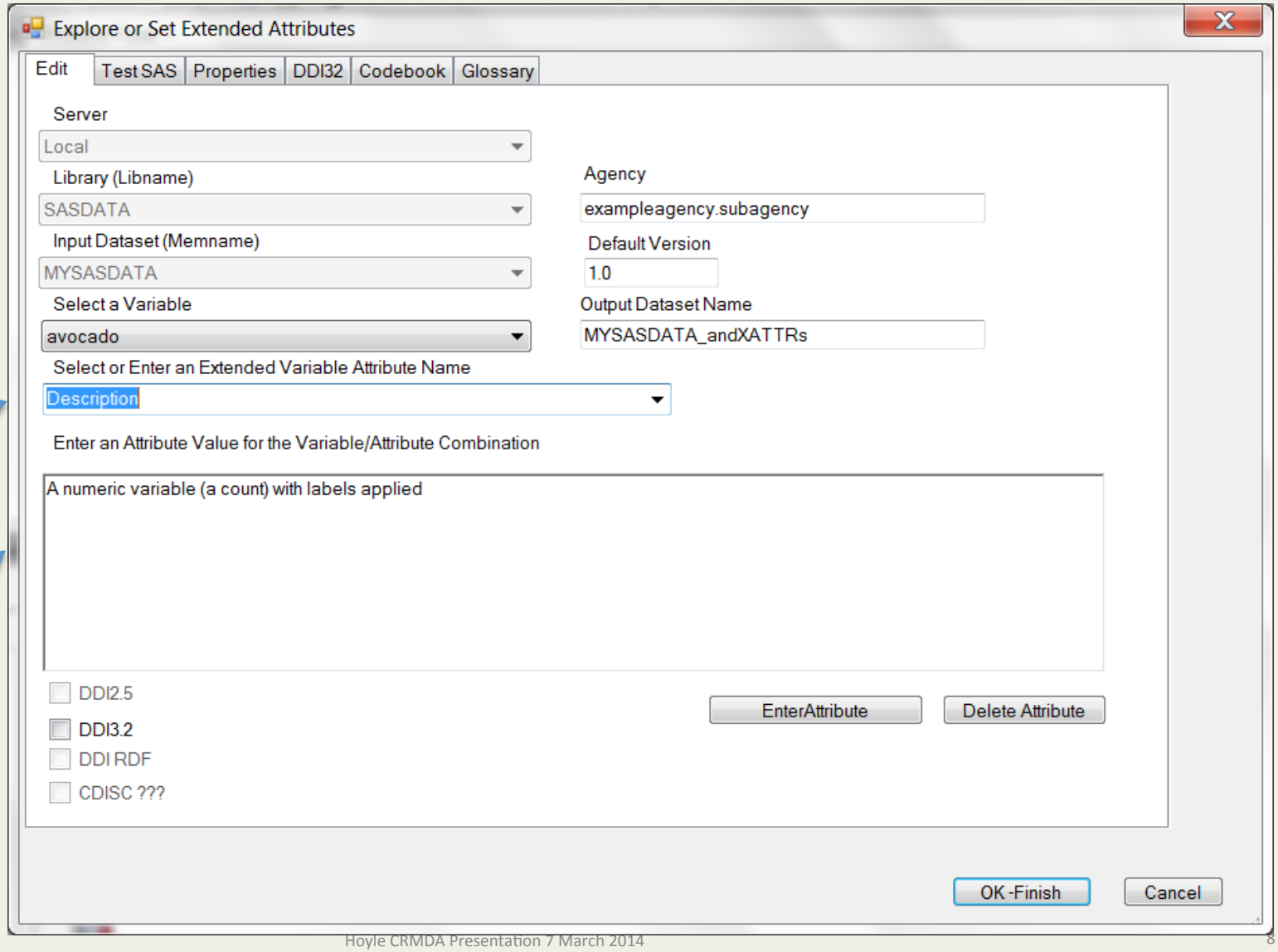


Editing Variable Attributes

Variable

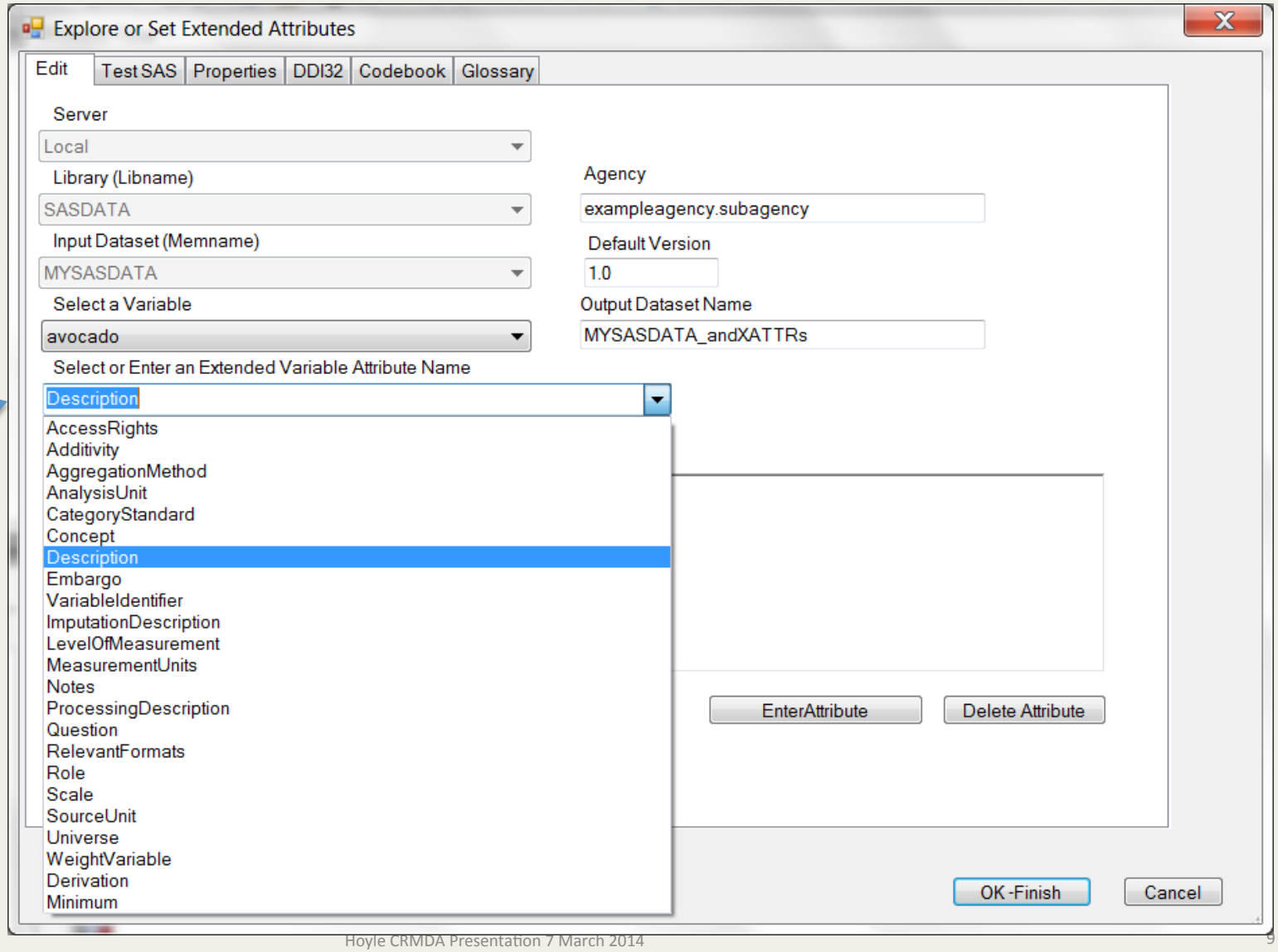
Attribute

Value



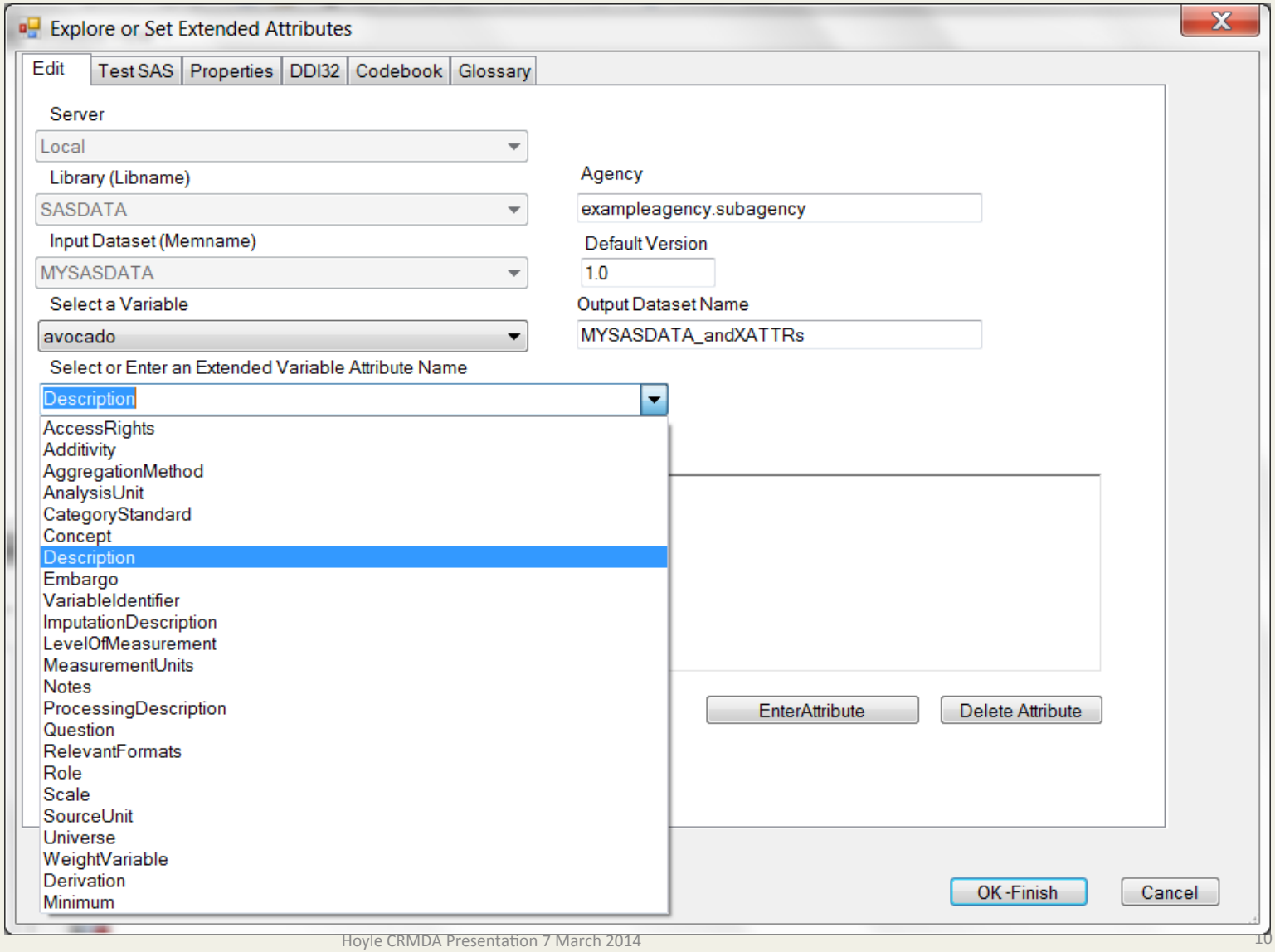
Pick from Drop-down List or Enter Value

Attribute

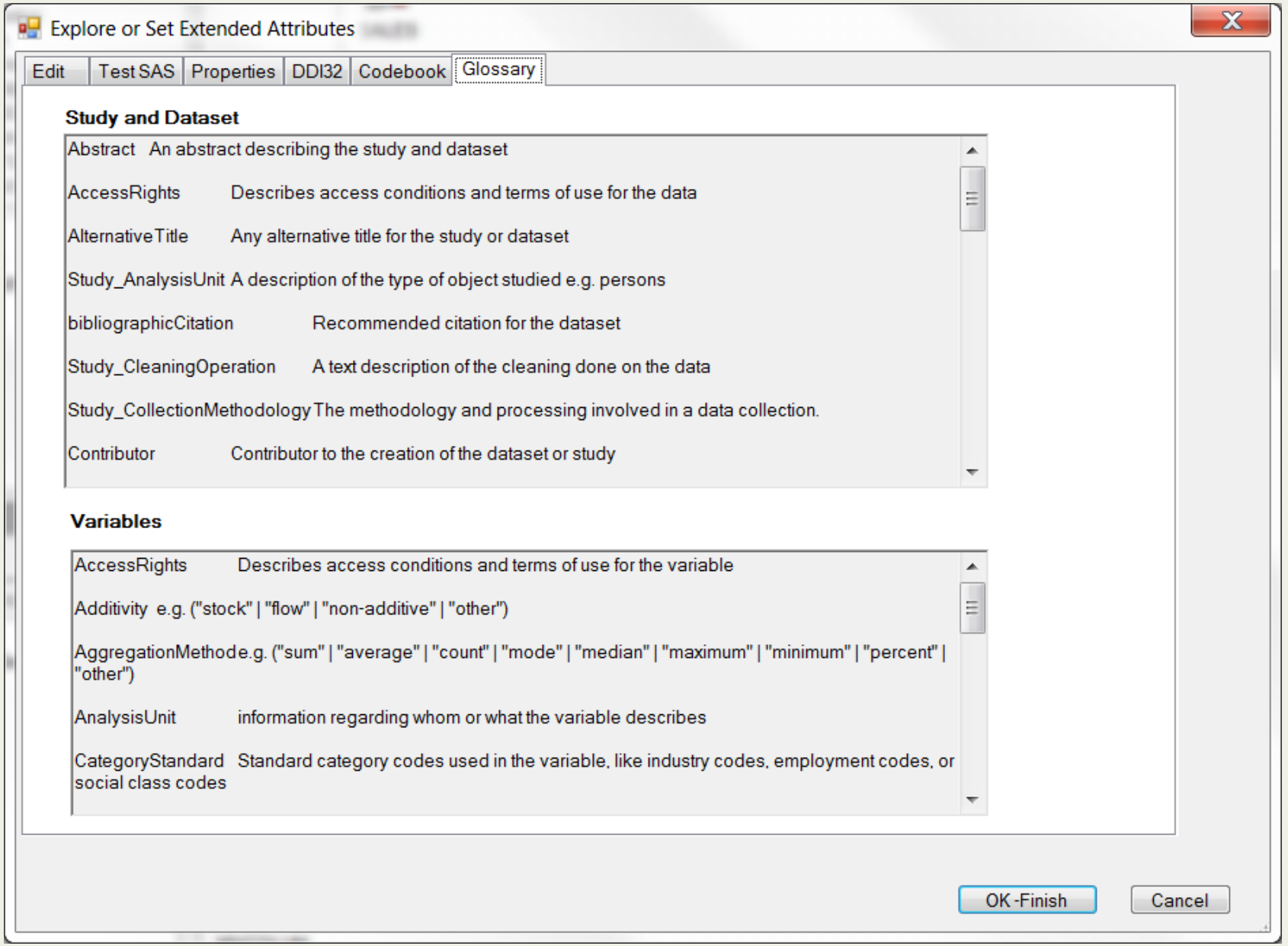


Pick from Drop-down List or Enter Value

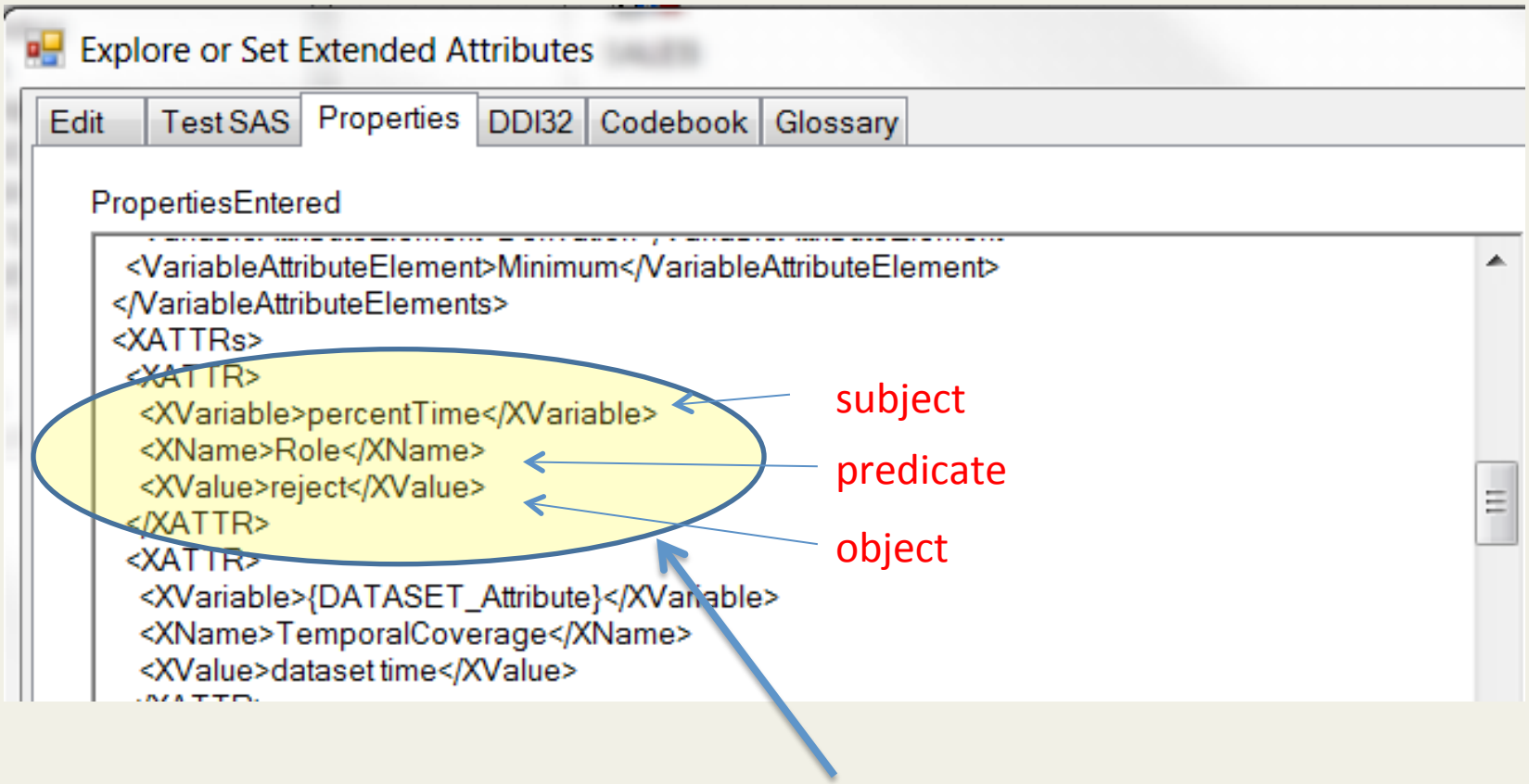
Populated with DDI based terms plus any other extended attributes in the dataset



Glossary of Terms

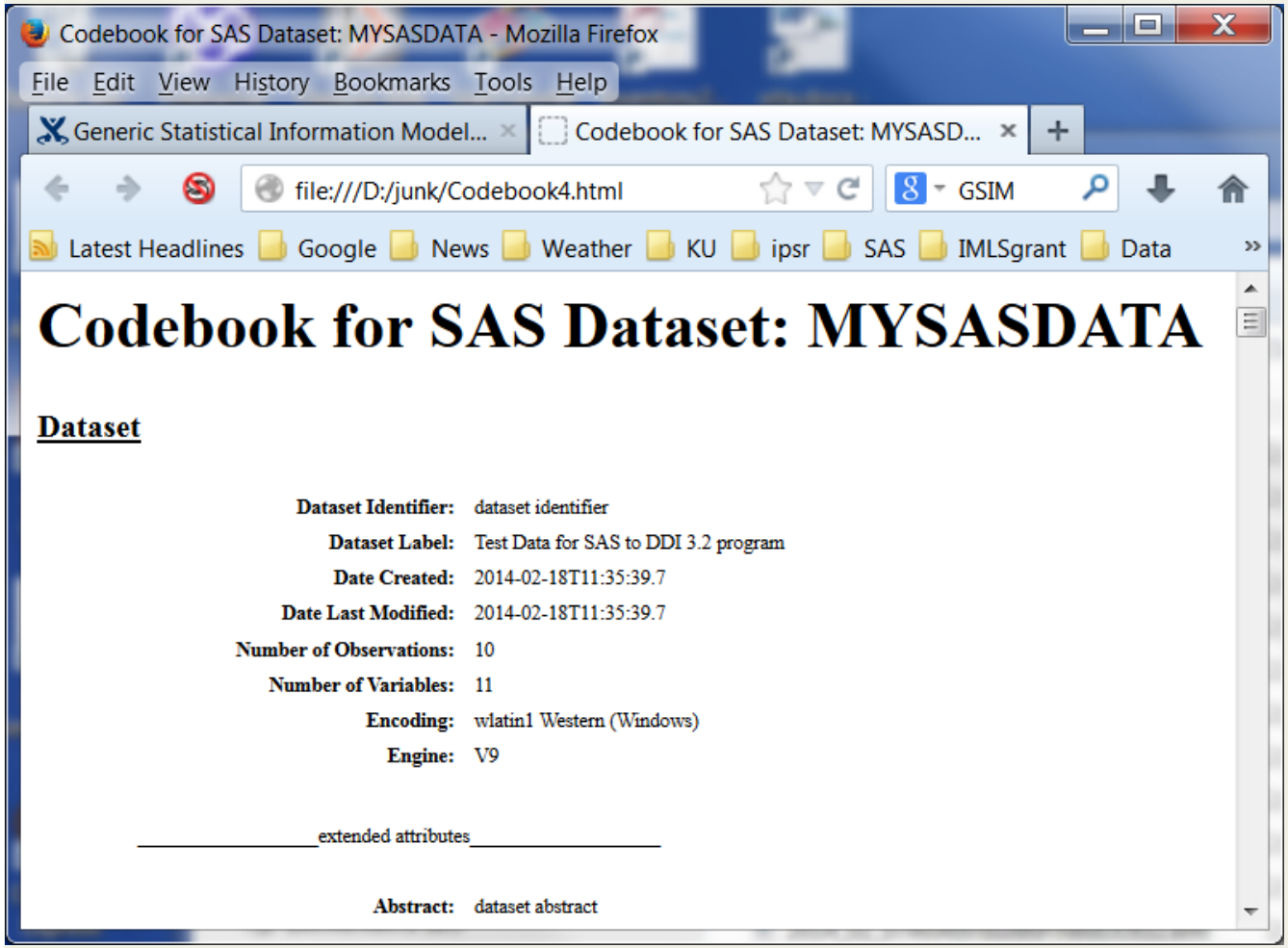


View of Internal Representation of Selections

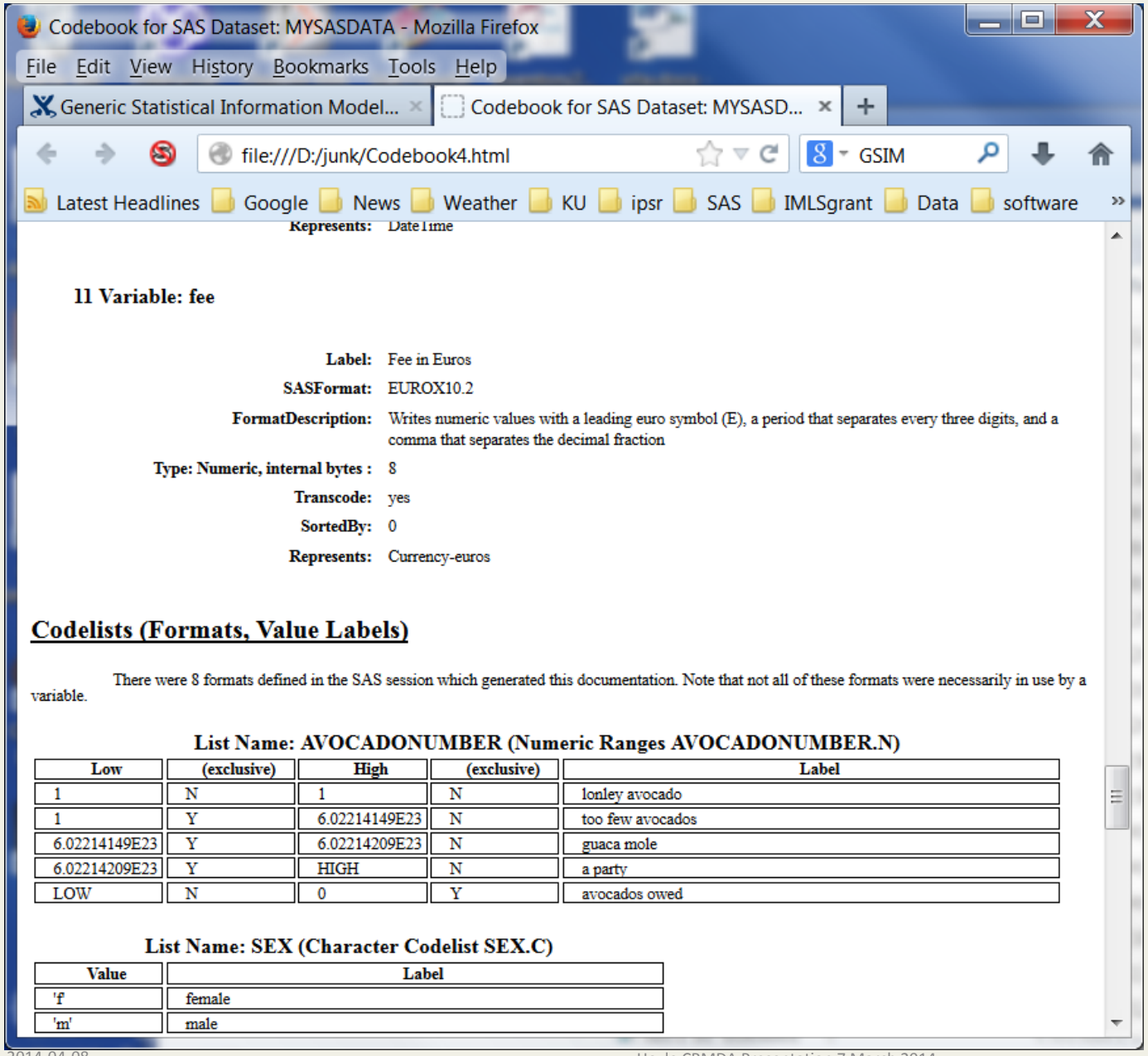


These should also map to a Semantic Web (RDF) style representation like DDI-Discovery

Codebook Generation (HTML)

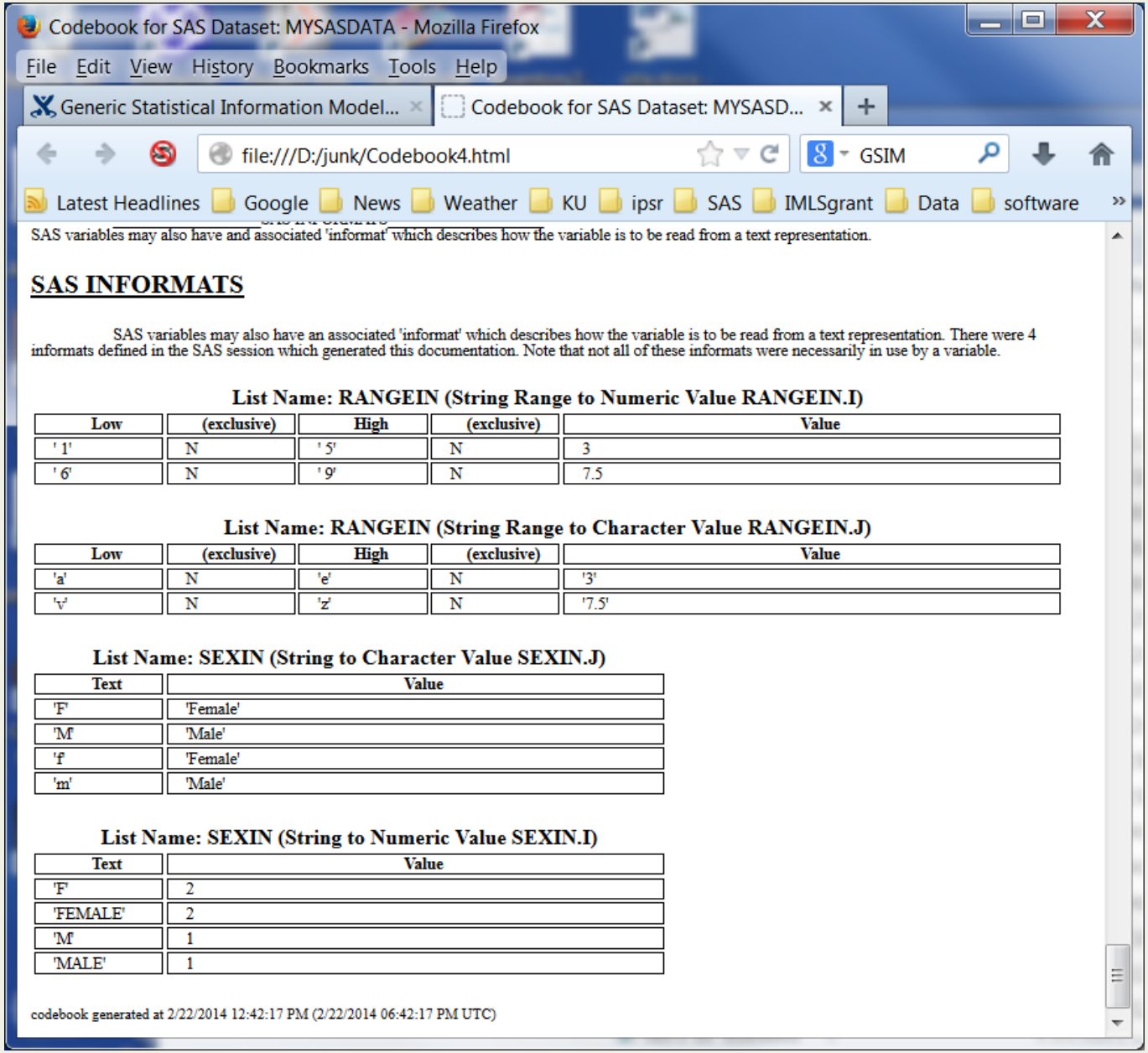


Dataset, Variables, Codelists



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 and 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)

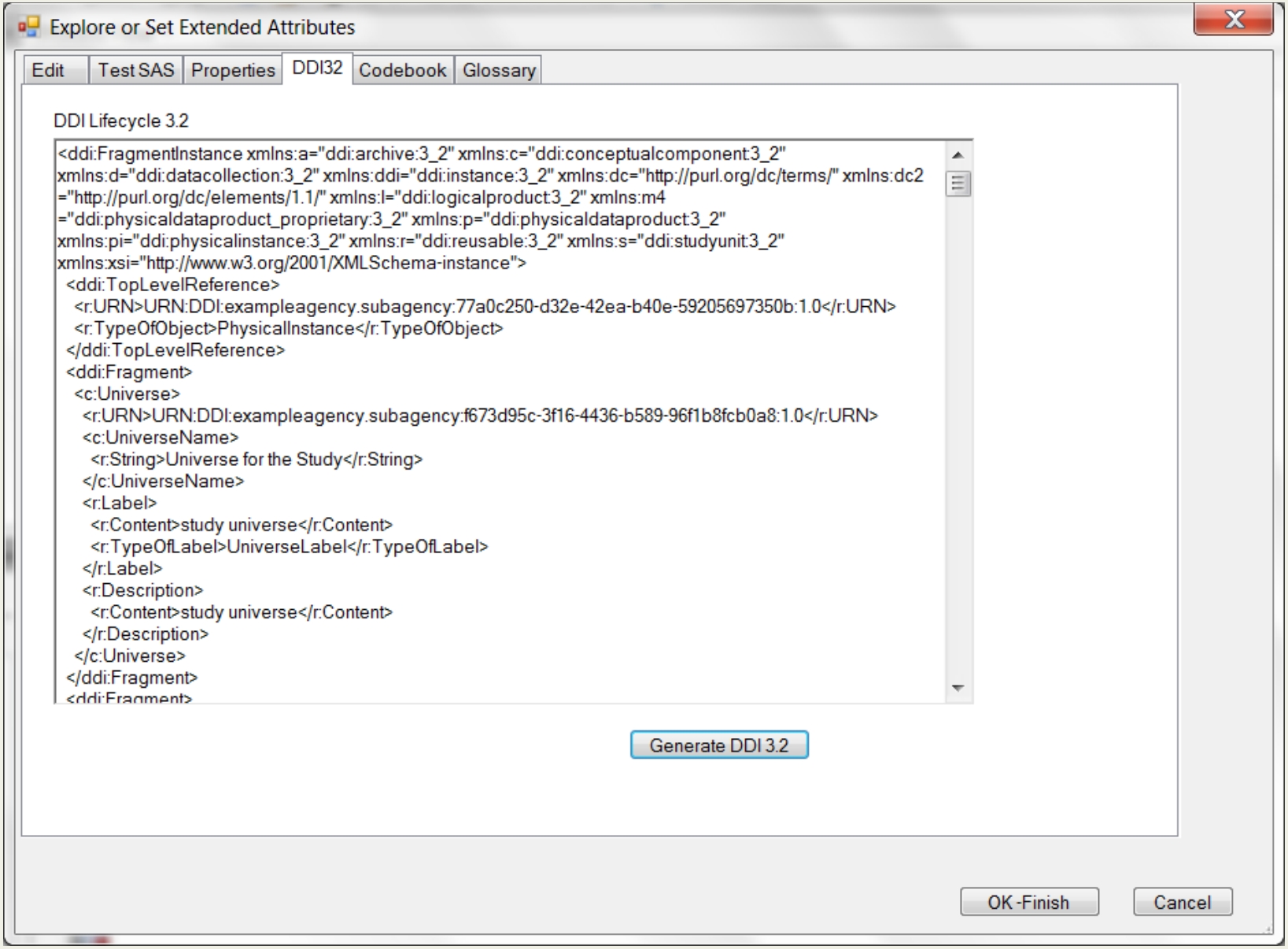
Text	Value
'F'	'Female'
'M'	'Male'
'f'	'Female'
'm'	'Male'

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

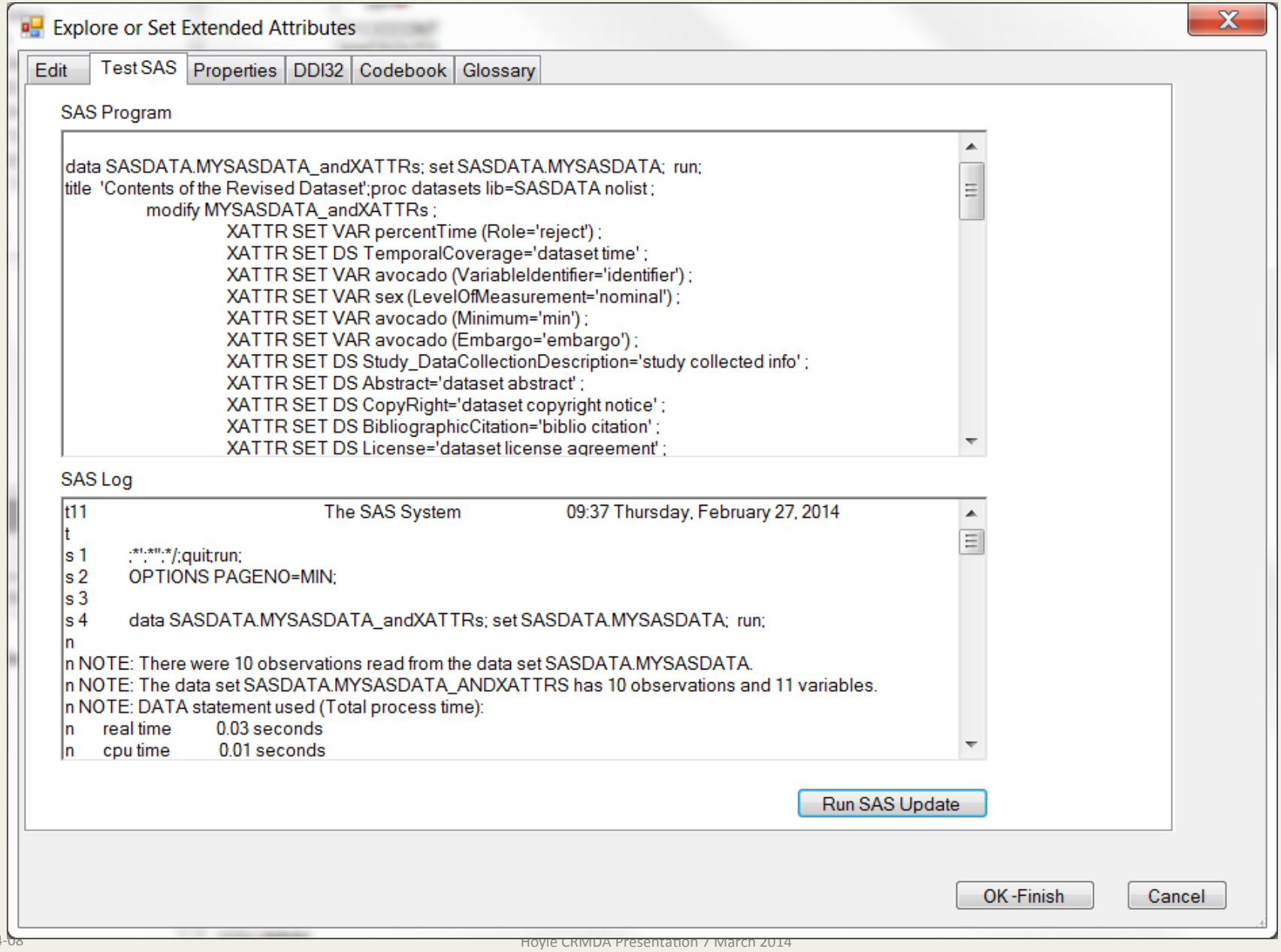
Text	Value
'F'	2
'FEMALE'	2
'M'	1
'MALE'	1

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

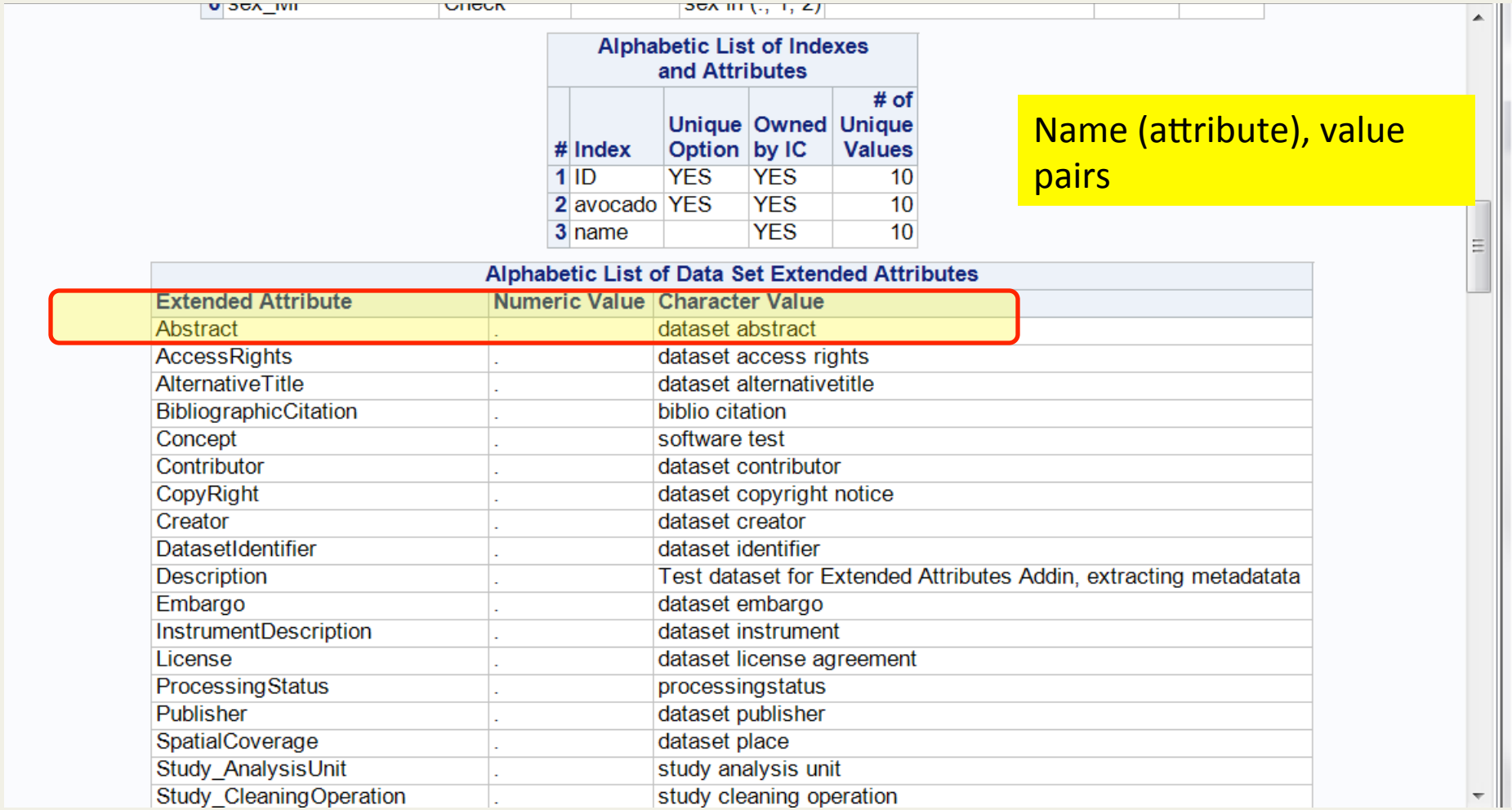
Metadata Structured as DDI 3.2 (XML)



Review and Run the SAS Code



Upon Finishing
Runs Proc Dataset to Create New Dataset

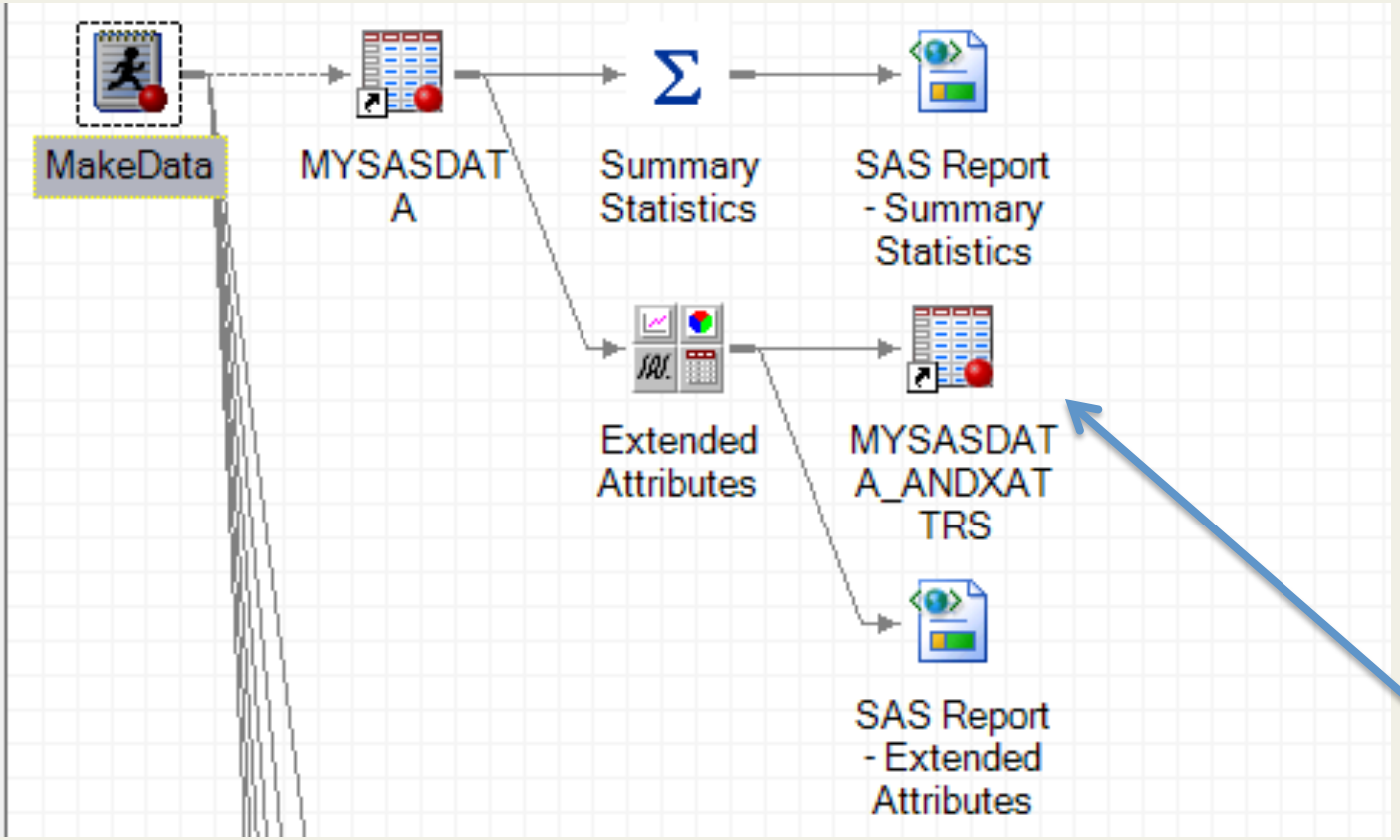


Alphabetic List of Indexes and Attributes				
#	Index	Unique Option	Owned by IC	# of Unique Values
1	ID	YES	YES	10
2	avocado	YES	YES	10
3	name		YES	10

Name (attribute), value pairs

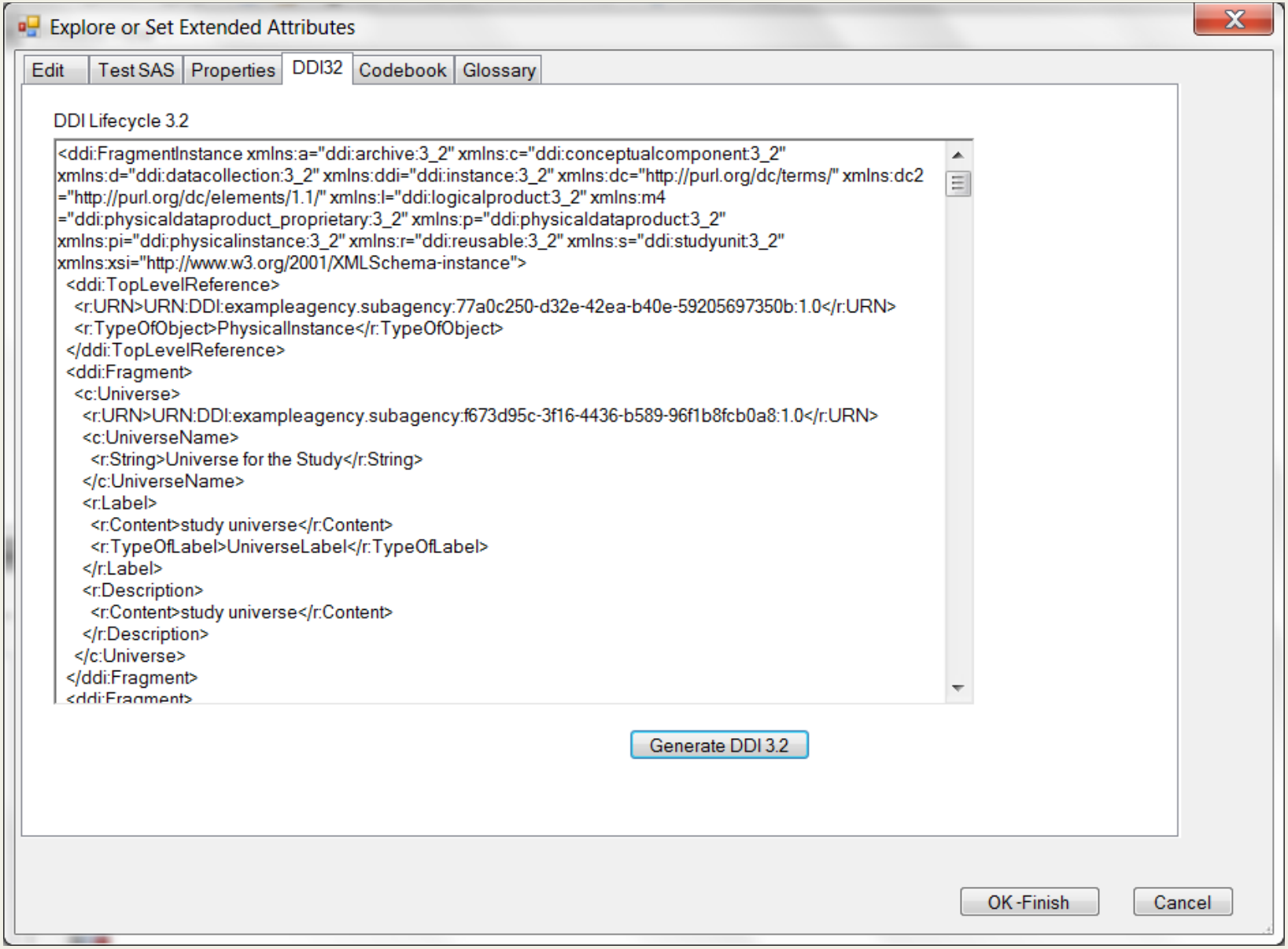
Alphabetic List of Data Set Extended Attributes		
Extended Attribute	Numeric Value	Character Value
Abstract	.	dataset abstract
AccessRights	.	dataset access rights
AlternativeTitle	.	dataset alternativetitle
BibliographicCitation	.	biblio citation
Concept	.	software test
Contributor	.	dataset contributor
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)

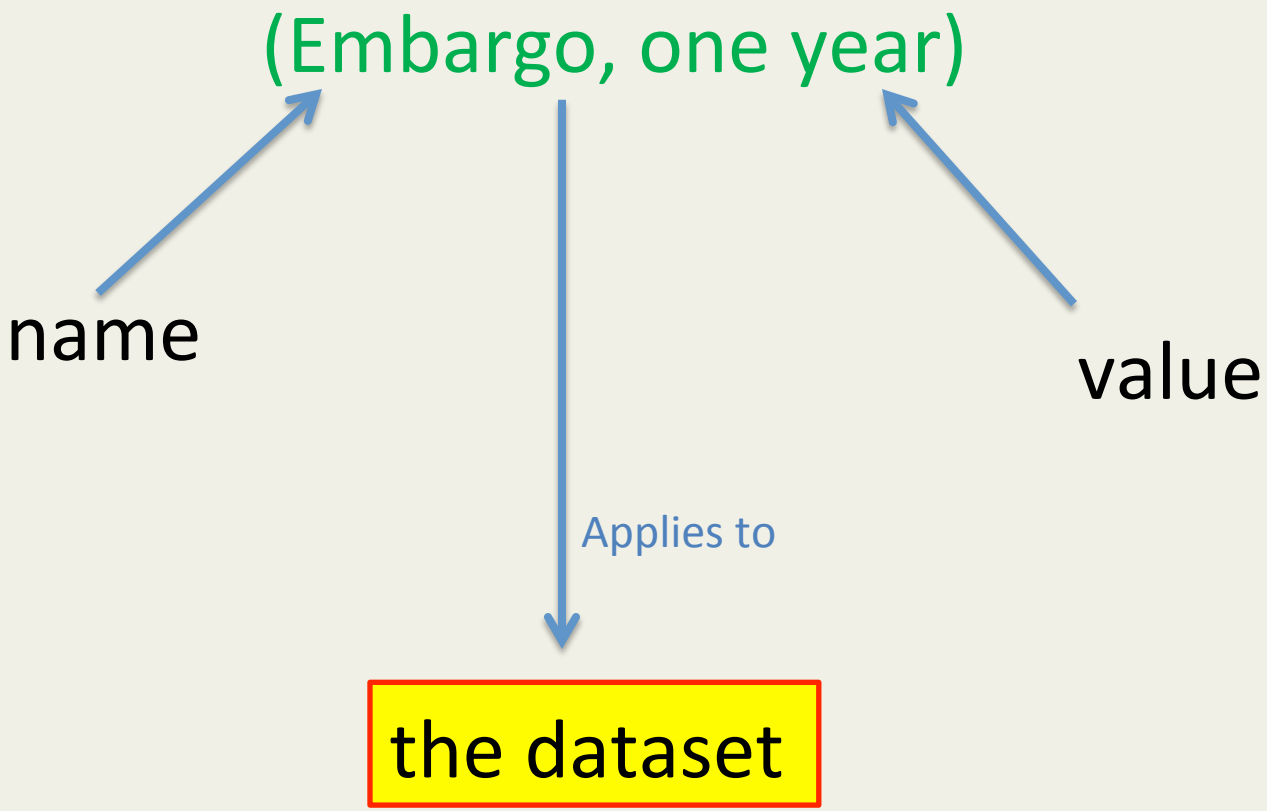
- “A Fragment is a means of transporting a maintainable or versionable object plus any associated notes and other material.”

Custom Attributes

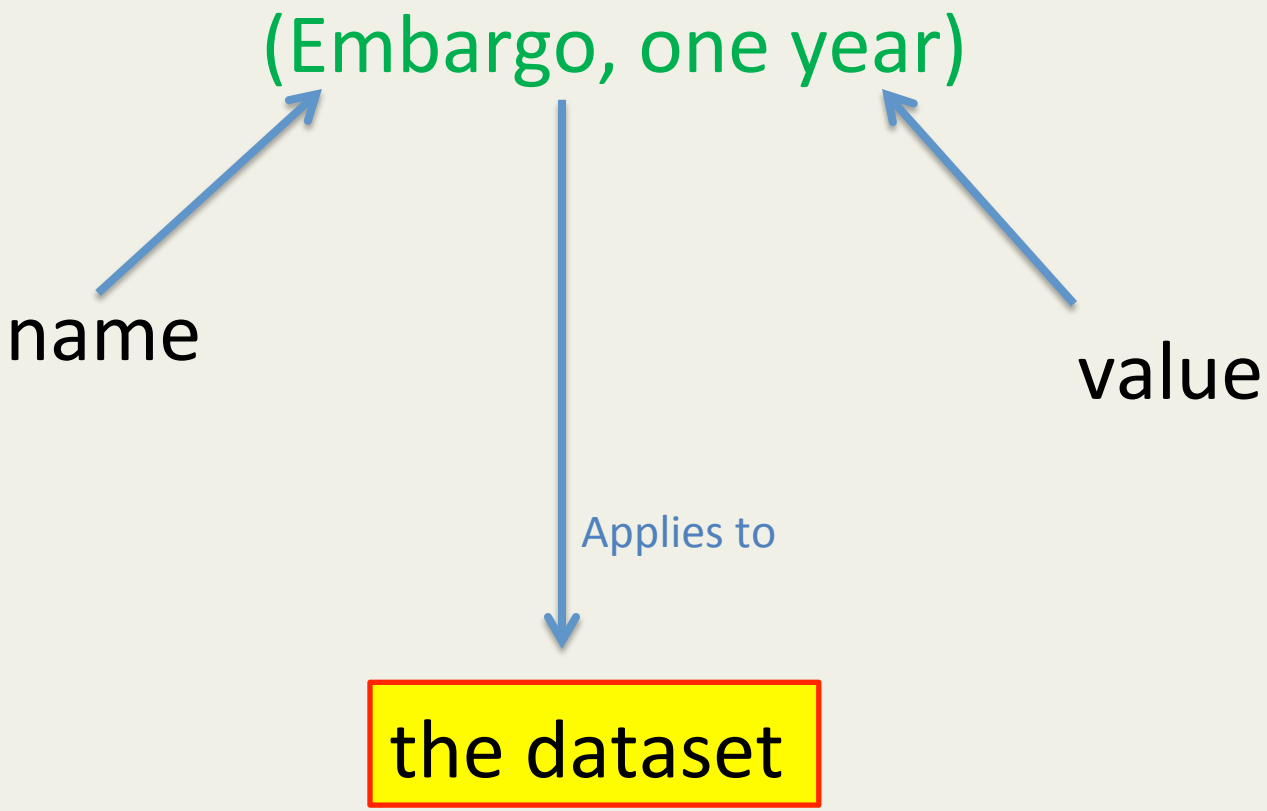
- SAS, Stata, SPSS, Excel, R all can now have custom, or extended attributes as (name, value) pairs on variables or the dataset.

- 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