

# Exporting SAS Datasets to DDI 3 XML files: Data, Metadata, and More Metadata

**Larry Hoyle**, Institute for Policy and Social Research,  
University of Kansas

**Joachim Wackerow**, GESIS-ZUMA  
(Centre for Survey Research and Methodology,  
German Social Science Infrastructure Services)

# What would you do with these data if this were all you were given?

```
1M1                .11100029feb5211:49Joe <In the know> Schmo
2M6.021e23         .21000001jan7214:01Bill Hill
3F6.02214139e23.30900008jun8505:25Donna O'Fauna
4M6.02214149e23.40800025dec6401:23Rob "Bob" Cobb
5m6.02214159e23.50700015mar7515:15Tom Applebaum
6f6.02214179e23.60600005jun0708:09Louise Mac&Cheese
7m6.02214209e23.70500011nov1111:11Jack Black
8F6.02214219e23.80400001jan7214:04Jill Hill
9m-3               .90300001apr9903:03Gno Avocado
```

# Would this be better?

1	M	1	.1	11000	29feb52	11:49	Joe <In the know>	Schmo
2	M	6.021e23	.2	10000	01jan72	14:01	Bill Hill	
3	F	6.02214139e23	.3	9000	08jun85	05:25	Donna O'Fauna	
4	M	6.02214149e23	.4	8000	25dec64	01:23	Rob "Bob" Cobb	
5	m	6.02214159e23	.5	7000	15mar75	15:15	Tom Applebaum	
6	f	6.02214179e23	.6	6000	05jun07	08:09	Louise Mac&Cheese	
7	m	6.02214209e23	.7	5000	11nov11	11:11	Jack Black	
8	F	6.02214219e23	.8	4000	01jan72	14:04	Jill Hill	
9	m	-3	.9	3000	01apr99	03:03	Gno Avocado	

# What about this?

MySASData.xls [Compatibility Mode] - Microsoft Excel

Home Insert Page Layout Formulas Data Review View

Clipboard Font Alignment Number Styles Cells Editing

C17

	A	B	C	D	E	F	G	H	I
1	ID	name	avocado	sex	percentTime	DOB	TOB	fee	
2	1	Joe <In the know> Schmo	1	1	0.1	2/29/1952	11:49:00 AM	11000	
3	2	Bill Hill	6.021E+23	1	0.2	1/1/1972	2:01:00 PM	10000	
4	3	Donna O'Fauna	6.02214E+23	2	0.3	6/8/1985	5:25:00 AM	9000	
5	4	Rob "Bob" Cobb	6.02214E+23	1	0.4	12/25/1964	1:23:00 AM	8000	
6	5	Tom Applebaum	6.02214E+23	1	0.5	3/15/1975	3:15:00 PM	7000	
7	6	Louise Mac&Cheese	6.02214E+23	2	0.6	6/5/2007	8:09:00 AM	6000	
8	7	Jack Black	6.02214E+23	1	0.7	11/11/2011	11:11:00 AM	5000	
9	8	Jill Hill	6.02214E+23	2	0.8	1/1/1972	2:04:00 PM	4000	
10	9	Gno Avocado	-3	1	0.9	4/1/1999	3:03:00 AM	3000	

MySASData

Ready 100%

# Data in these forms leave us with questions

	A	B	C	D	E	F	G	H	
1	ID	name	avocado	sex	percentTime	DOB	TOB	fee	
2	1	Joe <In the know> Schmo	1	1	0.1	2/29/1952	11:49:00 AM	11000	
3	2	Bill Hill	6.021E+23	1	0.2	1/1/1972	2:01:00 PM	10000	
4	3	Donna O'Fauna	6.02214E+23	2	0.3	6/8/1985	5:25:00 AM	9000	
5	4	Rob "Bob" Cobb	6.02214E+23	1	0.4	12/25/1964	1:23:00 AM	8000	
6	5	Tom Applebaum	6.02214E+23	1	0.5	3/15/1975	3:15:00 PM	7000	
7	6	Louise Mac&Cheese	6.02214E+23	2	0.6	6/5/2007	8:09:00 AM	6000	
8	7	Jack Black	6.02214E+23	1	0.7	11/11/2011	11:11:00 AM	5000	
9	8	Jill Hill	6.02214E+23	2	0.8	1/1/1972	2:04:00 PM	4000	
10	9	Gno Avocado	-3	1	0.9	4/1/1999	3:03:00 AM	3000	

- Technical
  - How is sex coded? How is fee scaled? Is percent a proportion?
- “Business”
  - In what currency is fee? What does avocado mean?
- Discovery
  - Where do the data live? Who created? When? Where? Why?

# These questions are addressed by **metadata**

- Data about data
  - Sometimes categorized as “Technical” and “Business”
  - Paper vs electronic
  - Structured vs unstructured

# Machine Actionable **Metadata**

- Not just in machine readable format
- In a well defined structure
  - Could be XML or could be properties of objects
  - A program can use this information
  - Metadata can be used in various combinations
    - Presented as a codebook
    - Offered as a Web service

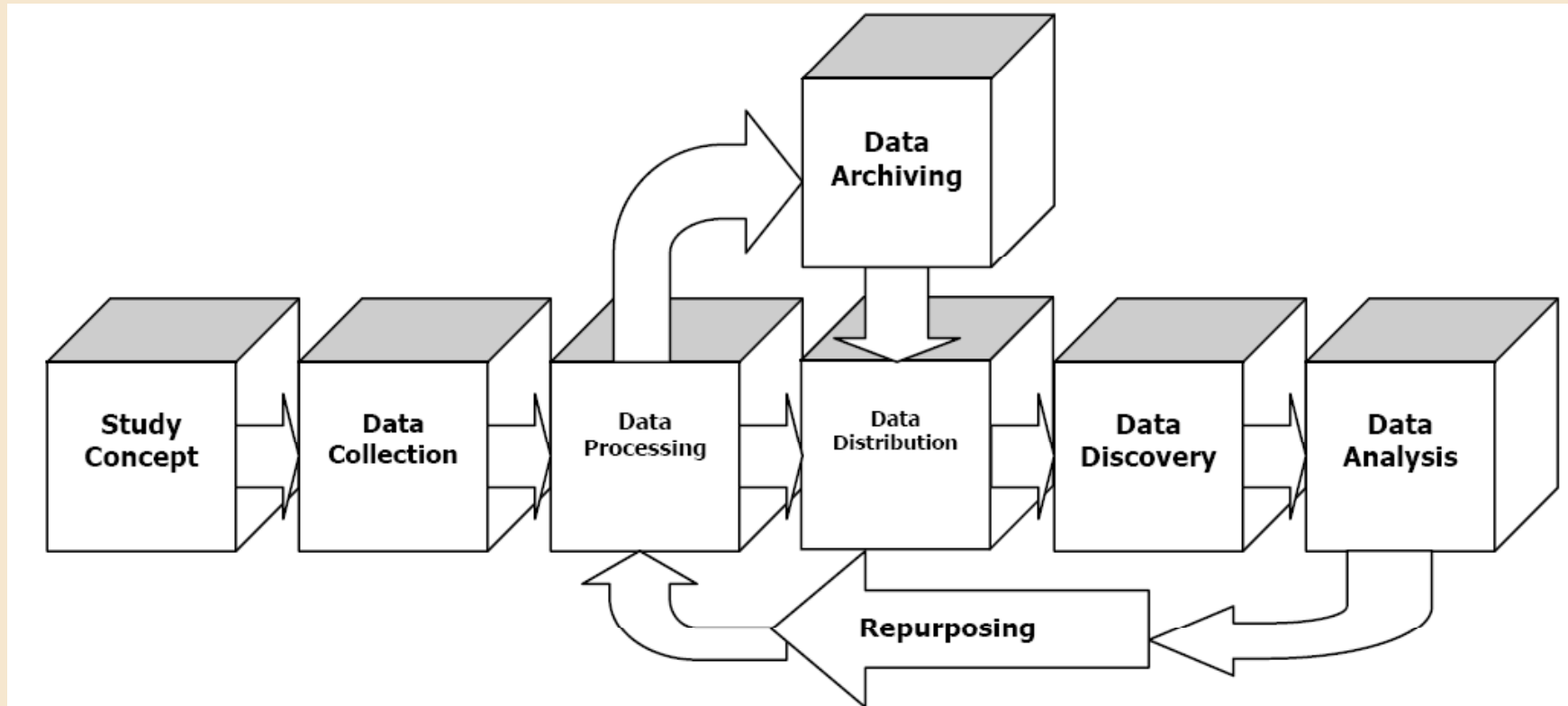
# Data Documentation Initiative (DDI)

<http://www.ddialliance.org/>

- A standard for the compilation, presentation, and exchange of documentation for datasets in the social and behavioral sciences
- XML based
- Begun 1995, first public release 2000
- Version 3 planned for June
- Life-cycle of data - from conception to re-use
- Metadata and data can be included in the same file



# Life Cycle of Data



# DDI Features

- Metadata capture from planning and production to dissemination and analysis
- An underlying data model that permits the expression of the model in alternative technologies
- Coverage of more of the data life cycle, with an emphasis on data collection
- Modular design
- Enhanced support for multiple languages
- Support for variable comparison and harmonization
- Structured mechanisms for identification and versioning that enable the creation of registries like question banks
- Core HTML for formatting of unstructured text

# More DDI Features

- Elimination of redundancies through a new grouping model and an extensive set of reusable elements
- Grouping of study series for longitudinal and comparative research
- Capturing comparative information for the creation of harmonized data
- ISO/IEC 11179 compliant data registries such as question, variable, and concept banks
- Capability to create "DDI profiles" for specific uses
- Mechanism to carry data inline
- Alignment with other metadata standards, including Dublin Core (cross-domain information resource description), SDMX (time-series data), ISO/IEC 11179 (metadata registry), and FGDC and ISO 19115 (geographic standards)
- Extensibility

# Central DDI Modules

## Study Unit

- Identification
  - Coverage
    - Topical
    - Temporal
    - Spatial
  - Conceptual Components
    - Universe
    - Concept
    - Representation (optional replication)
  - Purpose, Abstract, Proposal, Funding
- Identification is mapped to Dublin Core and basic Dublin Core is included as an option
  - Geographic coverage mapped to FGDC
    - bounding box
    - spatial object
    - polygon description of levels and identifiers
  - Universe Scheme, Concept Scheme
    - link of concept, universe, representation through Variable also allows storage as a ISO/IEC 11179 compliant registry

## Data Collection

- Methodology
- Question Scheme
  - Question
  - Response domain
- Instrument
- Coding Instructions
  - question to raw data
  - raw data to public file
- Question and Response Domain designed to support Question Banks
  - Question Scheme is a maintainable object
- Organization and flow of questions into Instrument
  - Used to drive CAI systems
- Coding Instructions
  - Reuse by Questions, Variables, and comparison

## Physical Storage

- Physical Data Structure
  - Links to Data Relationships
  - Links to Variable or NCube Coordinate
  - **Description of physical storage structure**
    - fixed, delimited or proprietary
- Physical Instance
  - One-to-one relationship with a data file
  - Coverage constraints
  - **Variable and category statistics**
- Data set
  - **Inline data items**

## Logical Product

- **Category Schemes**
- **Coding Schemes**
- **Variables**
- NCubes
- Variable and NCube Groups
- Data Relationships
- Used as both question response domains and variable representations
- Used as both question response domains and variable representations
- Link representations to concepts and universes through references
- Built from variables (dimensions and attributes)
  - Map directly to SDMX structures
  - More generalized to accommodate legacy data

# Study Unit

## Study Unit

- Identification
- Coverage
  - Topical
  - Temporal
  - Spatial
- Conceptual Components
  - Universe
  - Concept
  - Representation (optional replication)
- Purpose, Abstract, Proposal, Funding
- Identification is mapped to Dublin Core and basic Dublin Core is included as an option
- Geographic coverage mapped to FGDC
  - bounding box
  - spatial object
  - polygon description of levels and identifiers
- Universe Scheme, Concept Scheme
  - link of concept, universe, representation through Variable
  - also allows storage as a ISO/IEC 11179 compliant registry

- Physical Instance
  - One-to-one relationship with a data file
  - Coverage constraints
  - Variable and category statistics
- Data set
  - Inline data items

- Category Schemes
- Coding Schemes
- Variables
- NCubes
- Variable and NCube Groups
- Data Relationships
- Used as both question response domains and variable representations
- Used as both question response domains and variable representations
- Link representations to concepts and universes through references
- Built from variables (dimensions and attributes)
  - Map directly to SDMX structures
  - More generalized to accommodate legacy data

# Data Collection Module

## Study Unit

- Identification
- Coverage
- Identification is mapped to Dublin Core and basic Dublin Core is included as an option

## Data Collection

- Methodology
- Question Scheme
  - Question
  - Response domain
- Instrument
- Coding Instructions
  - question to raw data
  - raw data to public file
- Question and Response Domain designed to support Question Banks
  - Question Scheme is a maintainable object
- Organization and flow of questions into Instrument
  - Used to drive CAI systems
- Coding Instructions
  - Reuse by Questions, Variables, and comparison

- Data Relationships

## Study Unit

- Identification
- Coverage
  - 
  - 
  -
- Conceptual
  - 
  - 
  -
- Purpose, Ab

## Physical Structure

- Physical Data
  - 
  - 
  -
- Physical Information
  - 
  - 
  -
- Data set
  -

## Logical Product

- **Category Schemes**
  - Used as both question response domains and variable representations
- **Coding Schemes**
  - Used as both question response domains and variable representations
- **Variables**
  - Link representations to concepts and universes
- **NCubes**
  - Built from variables (dimensions and attributes)
    - Map directly to SDMX structures
    - More generalized to accommodate legacy data
- **Variable and NCube Groups**
- **Data Relationships**

## Physical Storage

- Physical Data Structure
  - Links to Data Relationships
  - Links to Variable or NCube Coordinate
  - **Description of physical storage structure**
    - fixed, delimited or proprietary
- Physical Instance
  - One-to-one relationship with a data file
  - Coverage constraints
  - **Variable and category statistics**
- Data set
  - **Inline data items**

esponse Domain  
port Question Banks  
stion Scheme is a  
tainable object  
id flow of questions

d to drive CAI systems  
ons  
ie by Questions,  
ables, and comparison

able representations  
able representations  
rough references

gacy data



# From SAS Dataset to DDI File

## Two Approaches

- Two approaches
  - DATA steps and PROCs wrapped in macros
  - Tagset for ODS (Output Delivery System)
    - User written
    - ODS with default tagset plus XSLT transformation
- Both need to gather metadata

# From SAS Dataset to DDI File

## Gathering the Metadata

- A SAS dataset contains data (of course)
- It also contains a mix of technical and business metadata
  - Labels: dataset and variables
  - Formats: **Links** to native and user formats
  - Integrity constraints

# Join information from:

- **DICTIONARY.COLUMNS**
  - Name, length, type, Fmtname, informat, precision, scale, sortedby, idxusage, notnull
- **PROC FORMAT, CNTLOUT dataset**
  - Information from formats – ranges and labels
- **Format documentation dataset (coded in program)**
  - Represents e.g. Currency-euros
  - FormatDocumentation e.g. Writes numeric values with a leading euro symbol (E), a comma that separates every three digits, and a period that separates the decimal fraction
- **Proc Contents**
  - Integrity constraints – type, variables, whereClause, ForeignReference, OnDelete, OnUpdate
- **DICTIONARY.CONSTRAINT\_COLUMN\_USAGE**
  - ColumnName
- **DICTIONARY.TABLES**
  - Memlabel, crdate, modate, nobs, nvar,
- **The dataset**
  - data

# Labels

- Dataset
  - mySASdata(label='Test Data for SAS to DDI 3 program')
- Variable
  - label avocado = 'Number of avacados';
  - label sex = 'Respondant's Gender';
  - label percentTime = 'Percent of time counting Avacados';
  - label fee = 'Fee in Euros';

# Formats – native SAS format

- format percentTime percent8.1;
  - **A proportion to be displayed as a percent**
- format fee EUROX10.2;
  - **Also tells us fee is in Euros**
- format DOB IS8601DA.;
  - **Number of days since January 1, 1960**
- format TOB IS8601TM.;
  - **Seconds since midnight of the current day**

# Formats - user

- format avocado avocadoNumber;

value avocadoNumber

low-<0 = 'avocados owed'

1 = 'lonley avocado'

1<-6.02214149e23 = 'too few avocados'

6.02214149e23-6.02214209e23 = **'guaca mole'**

6.02214209e23<-high = 'a party';

Labels ranges of data

(tells us someone likes a bad pun)

# Formats – there might also be extraneous user formats

value sexAll

0 = 'Young Male'  
1 = 'Adult Male'  
2 = 'Young Female'  
3 = 'Adult Female'  
;

value sex

0 = 'Male'  
1 = 'Male'  
2 = 'Female'  
3 = 'Female'  
;

/\* format BMI is not used and is here to be ignored later \*/

value BMI

low-<18.5 = "Underweight"  
18.5-24.9 = "Normal weight"  
25-29.9 = "Overweight"  
30-high = "Obesity";

# Integrity Constraints

add constraint prim\_key **Primary key**(id)

add constraint DOB\_present **Not Null**(DOB)

add constraint id\_GT\_0 check(**id GT 0**)

**gives us information about valid range**

add constraint sex\_MF check(**sex in (.,1,2)**)

add constraint avocado\_unique **Unique**(avocado)

add constraint name\_fkey foreign key(name)

**references work.RealPeople**

**valid values come from another table**



# Data Step Approach

- XML is just text
- Data step can write text to a file
- Static XML structure, known schema
- Content from SAS variables

# “Non Technical” Metadata

```
<?xml version="1.0" encoding="UTF-8"?>
<ns1:DDIInstance xmlns:r="ddi:reusable:3_0_CR" .....>
  <r:MaintainableID>
    <r:ID>testDDIFromSAS</r:ID>
  </r:MaintainableID>
  <r:Citation>
    <r:Title>DDI file from SAS dataset</r:Title>
    <dce:DCElements>
      <dc:title>DDI file from SAS dataset</dc:title>
    </dce:DCElements>
  </r:Citation>

  <s:StudyUnit>
    <r:MaintainableID><r:ID>StudyUnit_001</r:ID>
    <r:Version>1.0</r:Version>
    <r:VersionResponsibility>IPSR - The University of Kansas</r:VersionResponsibility>
  </r:MaintainableID>
```

**Can be lots more, including extensive explanatory text**

# Technical Metadata

<l:Variable>

<r:IdentifiableID><r:ID>**ID**</r:ID><r:Name>**ID**</r:Name></r:IdentifiableID>

<r:Label>**Identification Number**</r:Label>

<l:VariableDefinition> **SAS varnum: 1**

**SAS idxusage SIMPLE**

**SAS transcode: yes**

**SAS Integrity Constraints: Check(Where ID>0)**

**Primary Key(Variables ID)** </l:VariableDefinition>

<l:ConceptReference><r:Reference><r:ID>**ID**</r:ID></r:Reference>

</l:ConceptReference>

<l:Representation >

<l:NumericRepresentation type="**Double**"></l:NumericRepresentation>

</l:Representation>

</l:Variable>

**This identifier complexity  
allows for generating a URN  
which can be referenced in a  
global metadata structure**

# Metadata – implied from format

<l:Variable>

<r:IdentifiableID><r:ID>**fee**</r:ID><r:Name>**fee**</r:Name></r:IdentifiableID>

<r:Label>**Fee in Euros**</r:Label>

<l:VariableDefinition> **SAS varnum: 9**

**SAS format: EUROX10.2**

**SAS transcode: yes** </l:VariableDefinition>

<l:ConceptReference><r:Reference><r:ID>**fee**</r:ID></r:Reference>

</l:ConceptReference>

<l:Representation measurementUnit="**Currency-euros**" >

<l:Role> **SAS format indicates: Writes numeric values with a leading euro symbol (E), a period that separates every three digits, and a comma that separates the decimal fraction**</l:Role>

<l:NumericRepresentation type="**Double**"></l:NumericRepresentation>

</l:Representation>

</l:Variable>

# Data can be written in the XML too

```
<ds:ItemValue>  
  <ds:VariableReference>  
    <r:Reference>  
      <r:ID>fee</r:ID>  
    </r:Reference>  
  </ds:VariableReference>  
  <ds:Value>9000</ds:Value>  
</ds:ItemValue>
```

# User Written Tagset

- Extension of Output Delivery System (ODS) by user-defined tagsets
- Output in ODS as a stream of events
- Capture of events, triggering of specific action
- User-defined tagset defines template definitions for a target format
- User-defined tagsets can be added to existing SAS ODS tagsets

# Tagset – SAS Events can trigger user events

```
define event doc ;  
    start:  
        trigger _parameter ;  
        trigger _head ;  
        trigger _DDIInstance start ;  
    finish:  
        trigger _DDIInstance finish ;  
        putlog '    DDI file written: ' BODY_NAME ;  
end ;
```

# Tagset - User Defined Events

```
define event _DDIInstance ;
```

```
start:
```

```
    put '<ddi:DDIInstance';
```

```
    put ' xsi:schemaLocation="ddi:instance:' $ddi_version ' instance.xsd' ;
```

```
    . . .    '>' ;
```

```
    ndent ;
```

```
    putl '<r:MaintainableID>' ;
```

```
    ndent ;
```

```
    putl '<r:ID>XX</r:ID>' ;
```

```
    xdent ;
```

```
    putl '</r:MaintainableID>' ;
```

```
    trigger _StudyUnit start ;
```

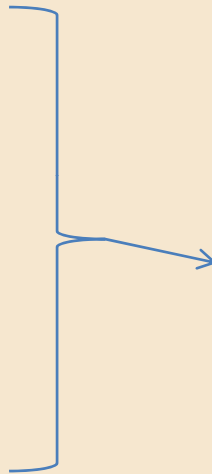
```
finish:
```

```
    trigger _StudyUnit finish ;
```

```
    xdent ;
```

```
    putl '</ddi:DDIInstance>' ;
```

```
end ;
```



```
<ddi:DDIInstance  
xsi:schemaLocation="ddi:instance:3_0_CR  
instance.xsd" xmlns:a=  
    ....  
>  
  <r:MaintainableID>  
    <r:ID>XX</r:ID>  
  </r:MaintainableID>  
  <s:StudyUnit>  
    <r:MaintainableID>  
      <r:ID>XX</r:ID>  
    </r:MaintainableID>
```



# Tagset Streams

The CONTENTS Procedure

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
1	V1	Num	4		ZA Study Number
2	V2	Num	6		Respondent Number
3	V3	Num	3	VA.	Country
4	V4	Num	3	VB.	Good citizen: Always vote in elections
<i>event row</i>					
5	<b>V5</b>	Num	3	VC.	<b>Good citizen: Never try to evade taxes</b>

**Stream VariableScheme**

DDI LogicalProduct - VariableScheme  
Definition of Variable V5 with label

**Stream CategoryScheme**


DDI LogicalProduct - CategoryScheme  
Definition of VC used by variable V5

**Stream PhysicalDataProduct**

DDI PhysicalDataProduct - Width  
Width definition of variable v5: 3

# Row event triggers a cascade of events writing to multiple streams

```
define event row ;
start:
  break / if ^cmp( section, 'body' ) ;
  do /if cmp( $proc, 'Contents' ) ;
    do / if cmp( $leaf, 'Attributes' ) ;
    break ;
    ndent ;
    putl '<OneAttribute>' ;
  else / if cmp( $leaf, 'EngineHost' ) ;
  break ;
  ndent ;
  putl '<OneEngineHost>' ;
  else / if cmp( $leaf, 'Variables' ) ;
    trigger _master_variable start ;
  done ;
else ;
  break ;
done ;
```



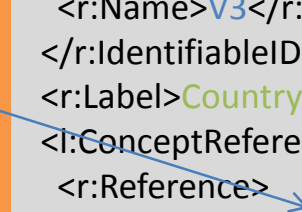
```
define event _master_variable ;
start:
  trigger _Concept start ;
  trigger _Variable start ;
```

# Here the <l:Variable> element is written

```
define event _Variable ;
  start:
    flush ;
    open LogicalProduct ;
    putl '<l:Variable>' ;
    flush ;
    close ;
  finish:
    flush ;
    open LogicalProduct ;
    ndent ;
    putl '<r:IdentifiableID>' ;
    ndent ;
    put '<r:ID>Variable_';
    put '$variable_name';
    putl '</r:ID>' ;
    put '<r:Name>' ;
    put '$variable_name';
    putl '</r:Name>' ;
    xdent ;
```

```
    putl '</r:IdentifiableID>' ;
    put '<r:Label>' ;
    put '$variable_label' ;
    putl '</r:Label>' ;
    putl '<l:ConceptReference>' ;
    ndent ;
    putl '<r:Reference>' ;
    ndent ;
    put '<r:ID>Concept_';
    put '$variable_name';
    putl '</r:ID>' ;
    xdent ;
    putl '</r:Reference>' ;
    xdent ;
    putl '</l:ConceptReference>' ;
    xdent ;
    putl '</l:Variable>' ;
    flush ;
    close ;
end;
```

```
<l:Variable>
  <r:IdentifiableID>
    <r:ID>Variable_V3</r:ID>
    <r:Name>V3</r:Name>
  </r:IdentifiableID>
  <r:Label>Country</r:Label>
  <l:ConceptReference>
    <r:Reference>
      <r:ID>Concept_V3</r:ID>
    </r:Reference>
  </l:ConceptReference>
</l:Variable>
```



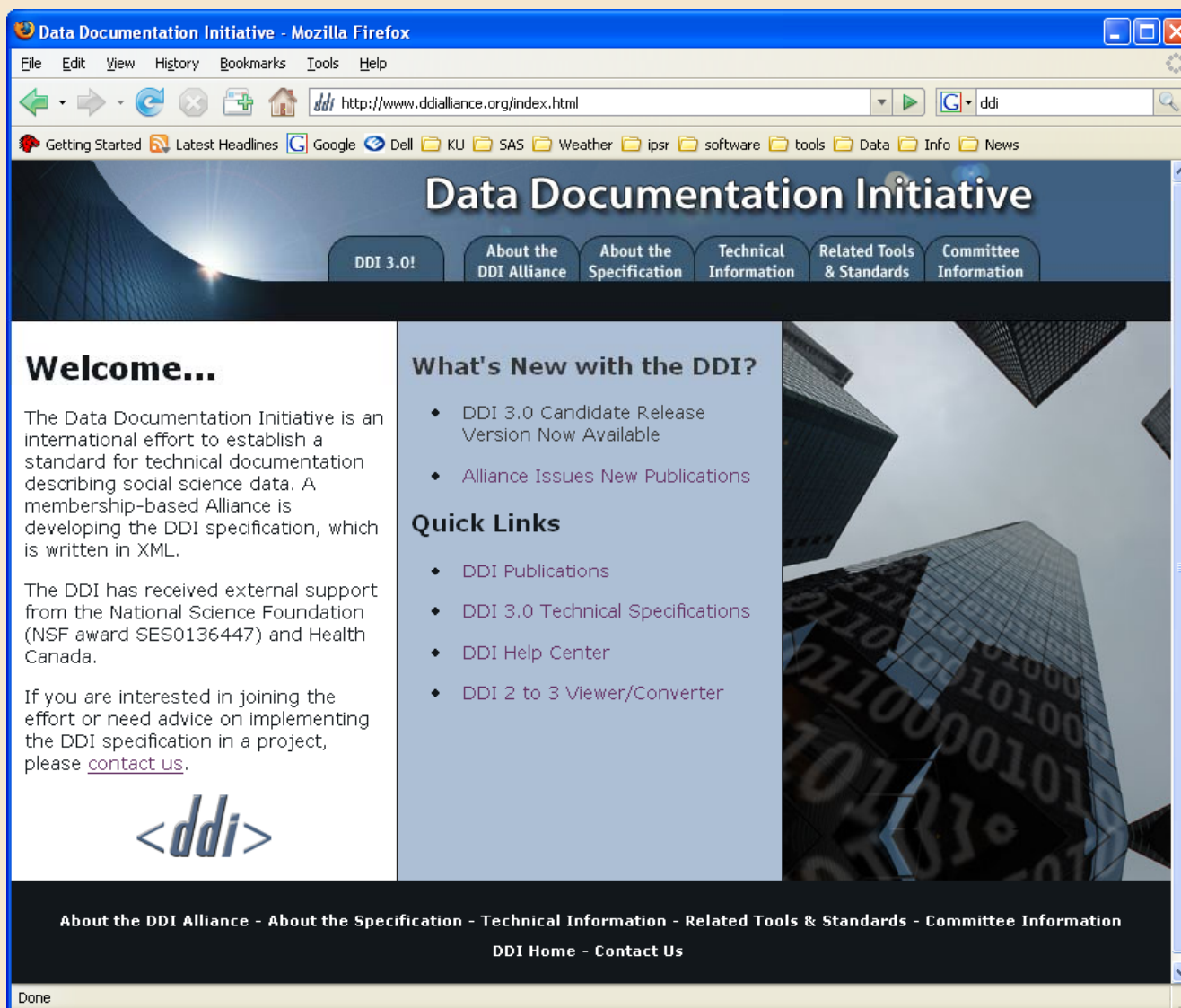
# Outputting the Streams

```
define event _StudyUnit ;  
    ...  
    finish:  
        trigger _ConceptualComponent finish ;  
        putstream ConceptualComponent ;  
        delstream ConceptualComponent ;  
        trigger _LogicalProduct finish ;  
        putstream LogicalProduct ;  
        delstream LogicalProduct ;  
    xdent ;  
    putl '</s:StudyUnit>' ;  
end;
```

# Using the Tagset

```
/* specifying the DDI tagset as ODS destination and opening a file */  
ods tagsets.DDI file='ddi.xml' encoding='utf-8';  
  
proc contents data=library.mySASdata;  
run;  
proc report data=userFormats;  
run;  
proc freq data=library.mySASdata;  
run;  
/* closing the ODS destination for DDI */  
ods tagsets.DDI close;
```

# http://www.ddialliance.org/index.html



# Questions?

- **LARRY HOYLE**

- Institute for Policy and Social Research
- University of Kansas
- 1541 Lilac Road, 607 Blake
- Lawrence, KS 66044-3177
- USA
- +1 785-864-9110
- [LarryHoyle@ku.edu](mailto:LarryHoyle@ku.edu)
- [www.ipsr.ku.edu](http://www.ipsr.ku.edu)

- **JOACHIM WACKEROW**

- GESIS-ZUMA (Centre for Survey Research and Methodology, German Social Science Infrastructure Services)
- B2, 1
- 68159 Mannheim
- Germany
- +49 621 1246 262
- [joachim.wackerow@gesis.org](mailto:joachim.wackerow@gesis.org)
- [www.gesis.org](http://www.gesis.org)