DDI as a Common Format for Export and Import from Statistical Packages

Larry Hoyle
Institute for Policy & Social Research, University of Kansas
&
Joachim Wackerow
GESIS - Leibniz Institute for the Social Sciences

DDI - Moving Data across Space and Time

- Across space – one organization to another
- Across time – via an archive
- Across software
  - Different organizations use different software
  - Software, and preferences for software evolve over time

- Optimize for clarity and completeness, not necessarily for speed/efficiency

DDI 3.1 as Common Format

- Dagstuhl 2009 paper (Hoyle, Wackerow & Hopt)
  - Metadata elements in software packages and DDI

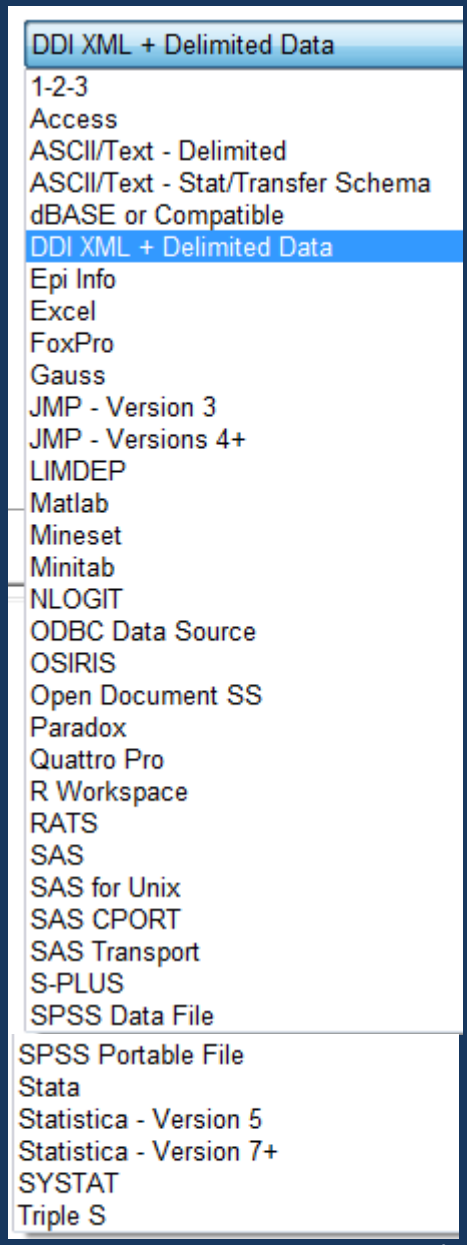
Metadata Element	CSV File	Excel	R	SPSS	Stata	SAS	JMP	MS Access	Triple-S	Stat Transfer	DDI Parent Element	DDI Element
Dataset											ddi:DDIInstance	s:StudyUnit
name	-	x	x	x	x	x	x	x	x	x	l:LogicalProduct	l:LogicalProductName
label	-	-	-	x	x	x	x	x	x		l:LogicalProduct	r:Label
user defined attributes	-	-	-	x	-	-	-	-	-		l:LogicalProduct	r:Description

Oops – Stata command can make a characteristic on a dataset:
"Define characteristic one attached to the data
. char _dta[one] this is char named one of _dta"

Stat/Transfer (http://www.stattransfer.com/)

- Data conversion software
  - Added DDI 3.1 with version 11
  - DDI plus 35 other file formats
  - Metadata aware

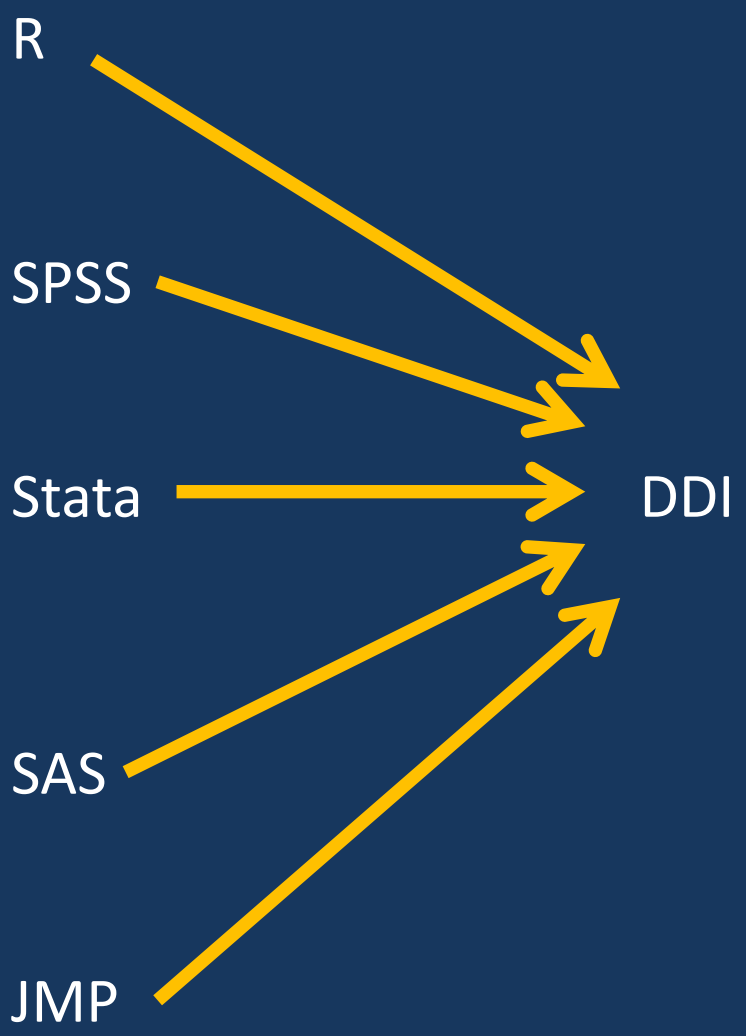
- This paper not intended as a critique of Stat/Transfer. Any suggestions for changes are offered with the intent of improving a very useful tool.



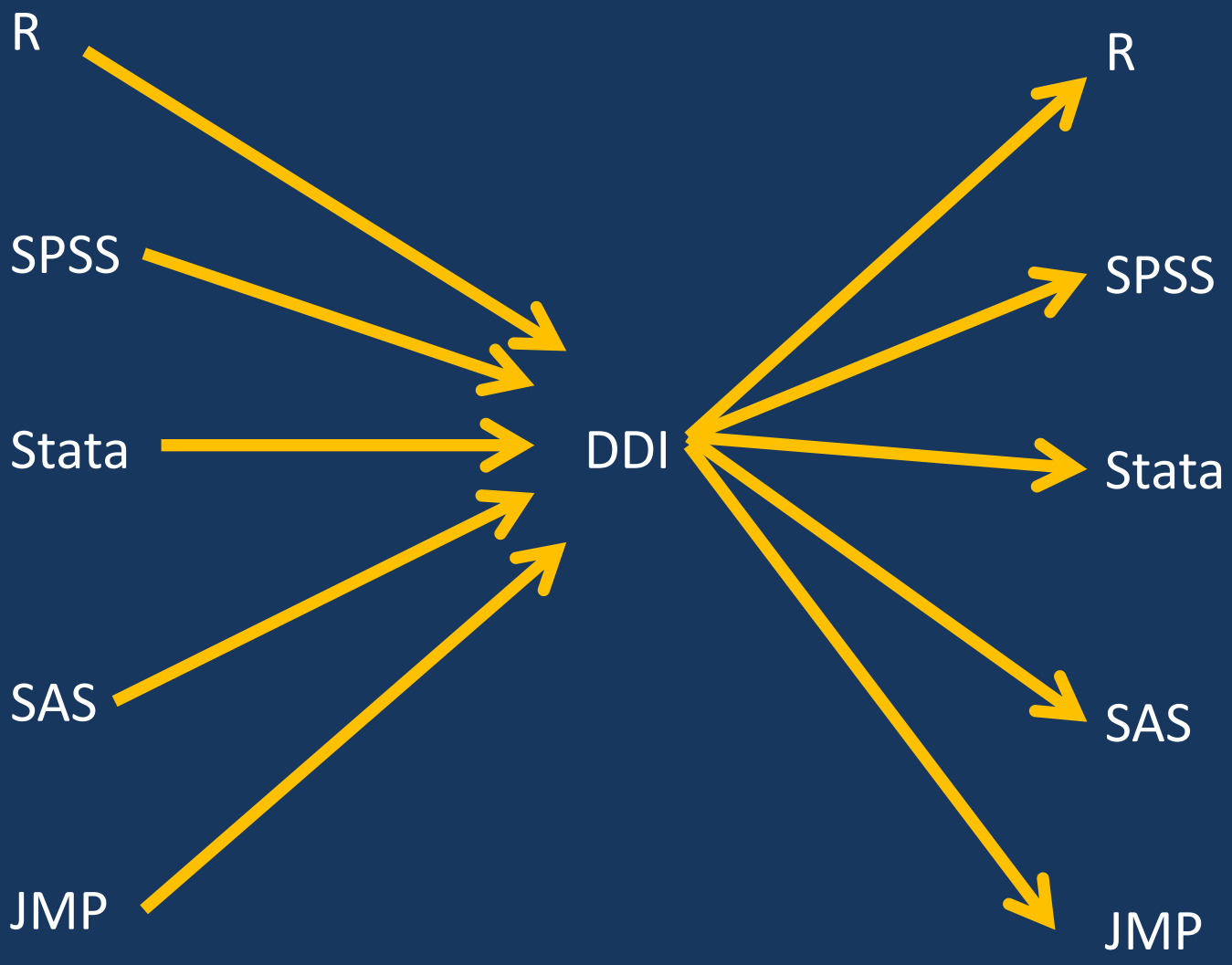
Our Experiment With Stat/Transfer (S/T) Questions

- What is currently **automatic** in moving data and metadata among software packages through DDI?
  - (no scripts needed)
- What else does DDI support that S/T does not ?
- What more could DDI support?

Our Experiment With Stat/Transfer



Our Experiment With Stat/Transfer



Our Experiment With Stat/Transfer

- Create master DDI 3.1 file and associated data file
  - Export to R, SPSS, Stata, SAS, JMP
  - What metadata features carry over?
- Create a dataset in each package, convert to DDI3.1
  - Include all identified metadata features for that package
    - Example: a characteristic named "Universe" if supported
  - Check which metadata features are included
- Build matrices showing which metadata would survive transition from one package to the other with DDI as an intermediary

The Dataset

Without labels

ID	Weight	DOB	DTOB	GenderChar	Gender	Group	Measure	MeasureMissing	Fee	bmi	Comment	IDFormatted
1	1	07-Jun-1944	7-Jun-1944 02:14:27.00	m	1	1	€197.500	0	11.11	15.0	This is a co...	1.00
2	1	05-Apr-1949	5-Apr-1950 15:23:45.00	f	2	1	€188.600	0	12.12	17.0	This is a co...	2.00
3	1	29-Feb-1948	29-Feb-1948 23:59:59.00	m	1	1	€201.400	0	13.13	22.0	This is a co...	3.00
4	1	13-Jan-1948	13-Jan-1948 01:02:03.00	m	1	0	€222.200	0	14.14	27.0	This is a co...	4.00
5	1	09-May-1949	9-May-1950 16:20:30.00	f	2	0	€196.200	0	15.15	33.0	This is a co...	5.00
6	1	22-Aug-1944	22-Aug-1944 07:30:00.00	m	1	0	.	1	16.16	37.0	This is a co...	6.00
7	2	14-Feb-1943	14-Feb-1943 14:14:14.00	f	2	1	.	2	17.17	44.0	This is a co...	7.00
8	2	22-Nov-1944	22-Nov-1944 09:09:00.00	m	1	0	€170.700	0	18.18	23.8	This is a co...	8.00

With labels

	ID	Weight	DOB	DTOB	GenderChar	Gender	Group	Measure	MeasureMissing	Fee	bmi	Comment	IDFormatted
1	1	1	07-Jun-1944	7-Jun-1944 02:14:27.00	male	male	Treatment	€197.500	0	11.11	15.0	This is a co...	Treatment
2	2	1	05-Apr-1949	5-Apr-1950 15:23:45.00	female	female	Treatment	€188.600	0	12.12	17.0	This is a co...	Treatment
3	3	1	29-Feb-1948	29-Feb-1948 23:59:59.00	male	male	Treatment	€201.400	0	13.13	22.0	This is a co...	Treatment
4	4	1	13-Jan-1948	13-Jan-1948 01:02:03.00	male	male	Control	€222.200	0	14.14	27.0	This is a co...	Control
5	5	1	09-May-1949	9-May-1950 16:20:30.00	female	female	Control	€196.200	0	15.15	33.0	This is a co...	Control
6	6	1	22-Aug-1944	22-Aug-1944 07:30:00.00	male	male	Control	.	Don't Know	16.16	37.0	This is a co...	Control
7	7	2	14-Feb-1943	14-Feb-1943 14:14:14.00	female	female	Treatment	.	Refused	17.17	44.0	This is a co...	Treatment
8	8	2	22-Nov-1944	22-Nov-1944 09:09:00.00	male	male	Control	€170.700	0	18.18	23.8	This is a co...	Control

Custom Attributes (e.g. in SPSS)

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role	[Concept]	[Note]	[Universe]
ID	Numeric	1	0	Identifier	None	None	3	Right	Nominal	None		ID must not ...	
Weight	Numeric	8	0		None	None	5	Right	Scale	Input			
DOB	Date	11	0	Date of Birth	None	None	8	Right	Nominal	Input			
DTOB	Date	23	2	Date-Time of ...	None	None	16	Right	Nominal	Input			
GenderChar	String	1	0	GenderMF	{f, femal...	None	7	Left	Nominal	Input			
Gender	Numeric	1	0	Gender	{1, male}...	None	8	Right	Nominal	Input	Self-identifie...		
Group	Numeric	1	0	Treatment Gr...	{0, Contr...	None	8	Right	Nominal	Input			
Measure	Custom	9	3	Dependent M...	None	None	8	Right	Scale	Target			Persons bor...
MeasureMis...	Numeric	8	0		{1, Don't ...	None	8	Right	Unknown	Input			
Fee	Numeric	8	2	Fee in Euros	None	None	4	Right	Scale	Input			
bmi	Numeric	8	1	Body Mass I...	None	None	8	Right	Scale	Input			
Comment	String	100	0	Unstructured ...	None	None	9	Left	Nominal	Input			
IDFormatted	Numeric	8	2		{1.00, Tr...	None	8	Right	Ordinal	Input			

ResourcePackage vs StudyUnit?

- Stat/Transfer uses ResourcePackage

```
31  <g:ResourcePackage id="759e6a68
32  <r:Citation>
37  </r:Citation>
38  <g:Purpose id="62fb391d-3e78-4
40  </g:Purpose>
41  <r:UniverseReference>
50  </r:UniverseReference>
51
52  <g:DataCollection>
102 </g:DataCollection>
103 <g:LogicalProduct>
251 </g:LogicalProduct>
252 <g:PhysicalDataProduct>
434 </g:PhysicalDataProduct>
435 <pi:PhysicalInstance id="4e4c7
451 </pi:PhysicalInstance>
452
453 <c:ConceptScheme id="35a8fd51-
459 </c:ConceptScheme>
460 <c:UniverseScheme id="11059e5c
467 </c:UniverseScheme>
```

```
469 <l:CategoryScheme id="88f56365-
479 </l:CategoryScheme>
480 <l:CategoryScheme id="832035cb-
494 </l:CategoryScheme>
495 <l:CategoryScheme id="c609b4ed-
505 </l:CategoryScheme>
```

```
532  <l:CategoryScheme id="e6f5c6b0
550  </l:CategoryScheme>
551  <l:CategoryScheme id="b8c79e8d
586  </l:CategoryScheme>
587  <l:CodeScheme id="5c706c37-d19
605  </l:CodeScheme>
606  <l:CodeScheme id="88fb9048-23d
624  </l:CodeScheme>
625  <l:CodeScheme id="5524bead-328
643  </l:CodeScheme>
644  <l:CodeScheme id="048006cf-0f5
670  </l:CodeScheme>
671  <l:CodeScheme id="c95a20cf-1eb
737  </l:CodeScheme>
738  <l:VariableScheme id="b2c91c69
1046 </l:VariableScheme>
1047 </g:ResourcePackage>
1048 </ddi:DDIInstance>
```

StudyUnit

```
35  <s:StudyUnit id="87b6c137-590
36    <r:Citation>
49    </r:Citation>
50    <s:Abstract id="177b3ee6-
52    </s:Abstract>
53    <r:UniverseReference>
62    </r:UniverseReference>
63    <s:Purpose id="c464000a-5
65    </s:Purpose>
66    <c:ConceptualComponent id
84    </c:ConceptualComponent>
85
86    <d:DataCollection xmlns="
134   </d:DataCollection>
```

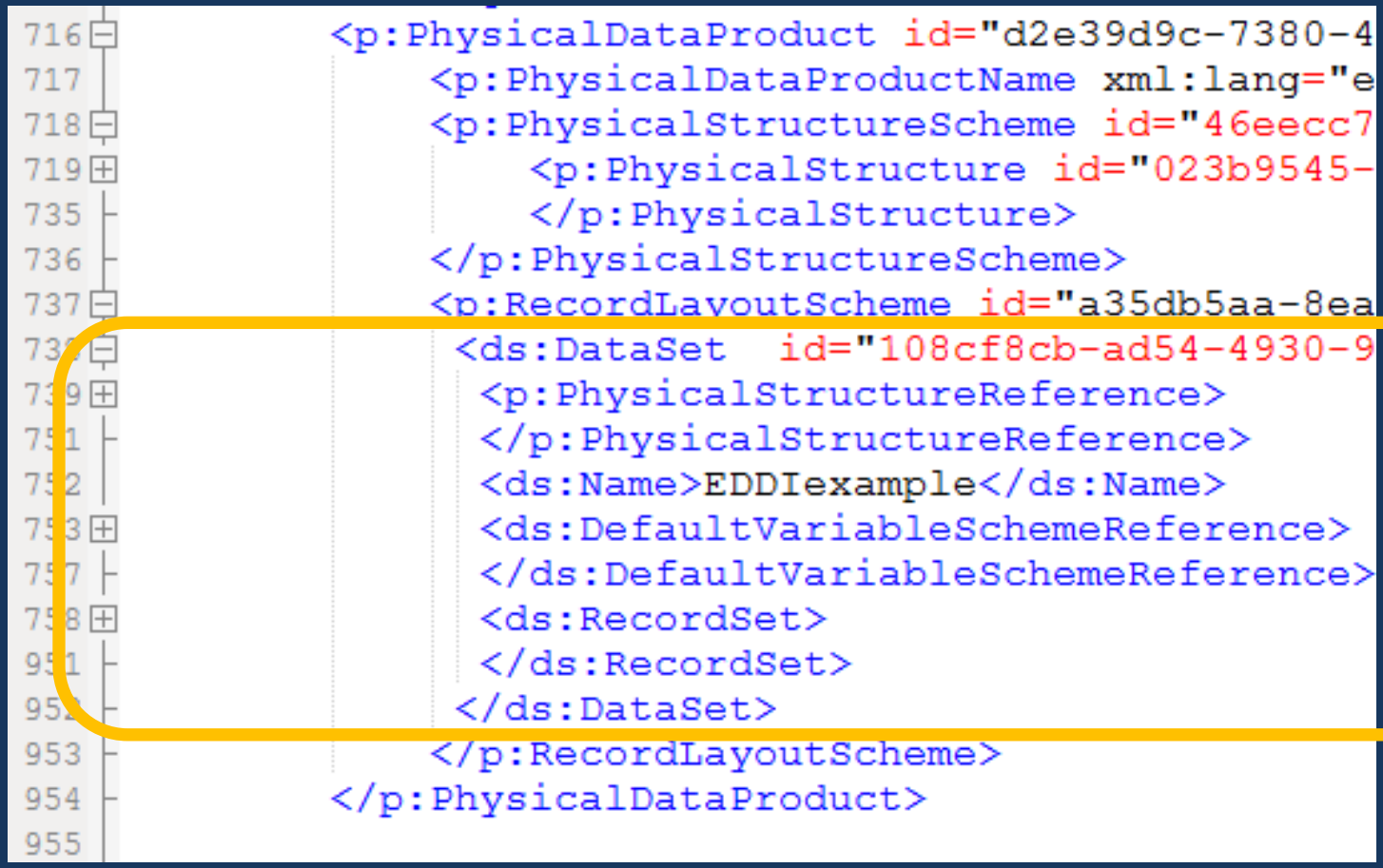
```
136   <l:LogicalProduct id="c15
137     <l:CategoryScheme id="8
147     </l:CategoryScheme>
148     <l:CategoryScheme id="8
162     </l:CategoryScheme>
163     <l:CategoryScheme id="c
173     </l:CategoryScheme>
174     <l:CategoryScheme id="b
184     </l:CategoryScheme>
185     <l:CategoryScheme id="9
```

```
255     <l:CodeScheme id="5c706
273     </l:CodeScheme>
274     <l:CodeScheme id="88fb9
292     </l:CodeScheme>
293     <l:CodeScheme id="5524b
311     </l:CodeScheme>
312     <l:CodeScheme id="04800
```

```
406       <l:VariableScheme id="b2c91c69-f10f-4ba7
714       </l:VariableScheme>
715     </l:LogicalProduct>
716     <p:PhysicalDataProduct id="d2e39d9c-7380-4
717       <p:PhysicalDataProductName xml:lang="e
718       <p:PhysicalStructureScheme id="46eecc7
719         <p:PhysicalStructure id="023b9545-
735         </p:PhysicalStructure>
736       </p:PhysicalStructureScheme>
737       <p:RecordLayoutScheme id="a35db5aa-8ea
738         <ds:DataSet  id="108cf8cb-ad54-4930-9
739           <p:PhysicalStructureReference>
751           </p:PhysicalStructureReference>
752           <ds:Name>EDDIexample</ds:Name>
753           <ds:DefaultVariableSchemeReference>
757           </ds:DefaultVariableSchemeReference>
758           <ds:RecordSet>
951           </ds:RecordSet>
952         </ds:DataSet>
953       </p:RecordLayoutScheme>
954     </p:PhysicalDataProduct>
955
956
957   </s:StudyUnit>
958 </ddi:DDIInstance>
```

Currently does not appear to work with Stat/Transfer

Embedded Data



Currently No Option with Stat/Transfer
Missed opportunity?

One Output: A Grid of Success/Failure

	Location In DDI (StatTransfer style)	Via StatTransfer From DDI3.1 to:					Via StatTransfer To DDI3.1					
metadata element	Parent / element or @attrubute	R data.frame	SPSS	Stata	SAS	JMP	R	SPSS	Stata	SAS	JMP	Notes
Dataset	ddi:DDIInstance / g:ResourcePackage											
name	l:LogicalProduct / l:LogicalProductName	+	+	+	+	+	-	~	~	+	+	~ Dataset name is in l:DataRelationshipName In R this is just "R", not the data.frame name
label	l:LogicalProduct / r:Label		-	-	~	~		+	+	-	-	~ indicates label generated by StatTransfer
date created	l:LogicalProduct / r:Description				+					-		Date created isfor the initial creation of the dataset, before any changes
date modified	l:LogicalProduct / @versionDate		+	~	+			+	+	+		the last time the metadata and data were modified ~ indicates available from operating system date on file
Notes on dataset	l:LogicalProduct / r:Note	-	-	-		~	-	-	-	-	-	~ indicates note generated by StatTransfer
user defined attrubutes	l:LogicalProduct / r:Description	-	-	-			-	-	-			
script stored with data	l:LogicalProduct / r:Description					-					-	

Indicates feature not supported in package

+ Feature translated successfully

- Feature didn't translate

~ Partial success e.g. there, but in an unexpected element

What Generally Worked (Classic Codebook)

- Dataset Name
- Variable Names, Labels, Order
- Data type (e.g. Dates and DateTimes ok)
- Missing or not
- Data

i.e. Elements supported by all of the packages

What Mostly Worked

- Dataset Labels
- Date Modified
- Value Labels for Numerics <–> Categories and Codes
  - R is different (factors)
  - SAS formats should work soon

Problems, Dataset

- Notes
- User defined attributes
- Scripts

Problems - Variables

- Weight (pretty important)
- Display formats (no standards across packages)

- Measurement units (important , most don't support)
- Measurement level

- Number of decimal positions
- Scale (where supported)
- Role
- User defined attributes (could be useful)
- Notes

Problems - Values

- Multiple distinct missing
- Ranges labeled
- Multiple sets of labels for a variable
- Range restrictions

- Labeling text values (will be fixed)
- Colors (only for JMP)

Multiple Missing Values

- Multiple Distinct
  - In-band (SPSS) 998, 999

  vs

  - distinct system missing (SAS, Stata) .D, .R
    - No representation in DDI
    - No way to associate categories & codes

  vs

  - No distinction among missing types (R,JMP)

Multiple Sets of Value Labels for a Variable

- SAS – "formats" and "informats" stored separately in a catalog or "CNTLIN" dataset.

```
proc format cntlout=eddi.sas_Fmts;

value GENDERen
   1="Male" 2="Female";
value GENDERde
   1="Männlich" 2="Weiblich";
value GenderL
   1="Self Identified Male" 2="Self Identified Female";
…
format gender GENDERen.;
```

Multiple Sets of Value Labels for a Variable

- Stata- Script

```
label define GenderE 1 "Male" 2 "Female"
label define GenderG 1 "Männlich" 2 "Weiblich"
label values Gender GenderE
```

- Stata – unassociated value labels not saved to .dta file, but are saved to "dta" xml file.

```
<value_labels>
<vallab name='GenderG'>
<label value='1'>Männlich</label>
<label value='2'>Weiblich</label>
</vallab>
<vallab name='GenderE'>
<label value='1'>Male</label>
<label value='2'>Female</label>
</vallab>
```

DDI – Multiple Labels for a Category

- DDI – xml:lang and type attributes of Label in Category

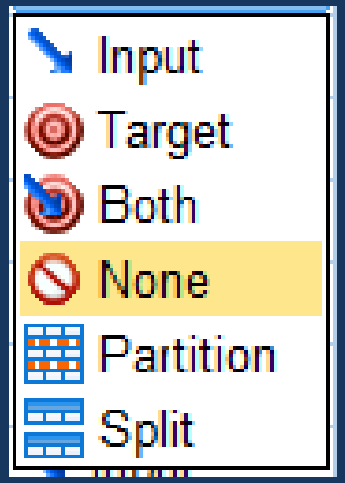
<l:Category id="c1" version="1.0.0" versionDate="2011-10-26T13:33:00" missing="false" >
<r:Label xml:lang="en-US" type="GENDER" >male</r:Label>
<r:Label xml:lang="de" type="GENDER" >männlich</r:Label>
<r:Label xml:lang="en-US" type="GENDERL" >Self Identified Male</r:Label>

Which is the "default?" (first listed?)

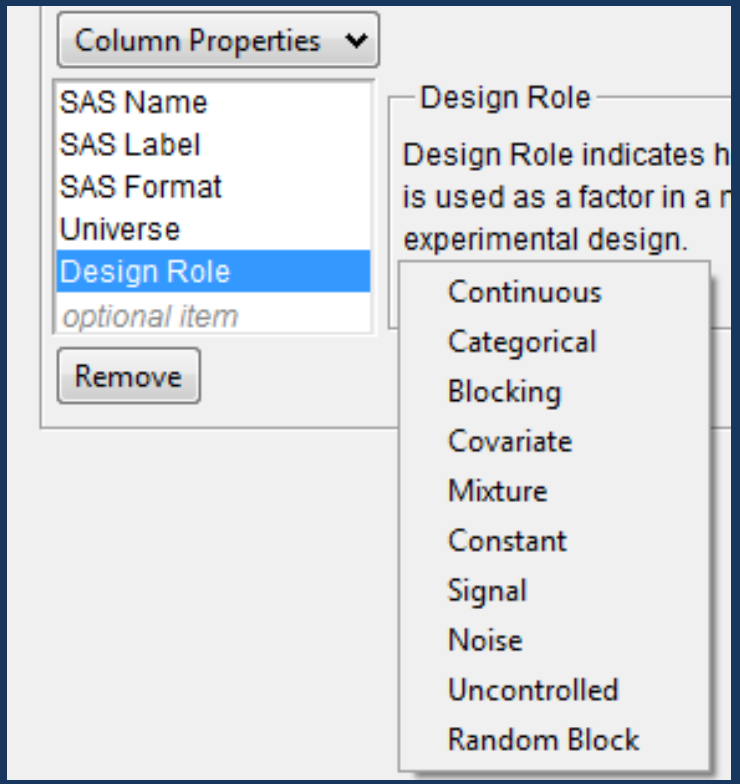
Role

- Several packages have metadata for "role"
- No standards

SPSS



JMP



Custom/User Variable Attributes

- R – **attributes**
  - function - attr() (column and data.frame)
- Stata – "**Characteristics**"
  - function – char (variable or table)
- SPSS
  - VARIABLE **ATTRIBUTE** VARIABLES=Age Gender Region ATTRIBUTE=DemographicVars ('1').
- Jmp
  - Column **Properties** … Other

Labeled Ranges in SAS and JMP

- Can be used dynamically in analyses, output.
- Probably not the best practice for a preservation dataset

Built-in Display Formats

- Currency symbols
- Thousands separators, decimal separator
- Date/Time formats

- Some of these (like currency symbols) convey units of measurement)

- Again – not standardized

R DateTime and UTC conversion

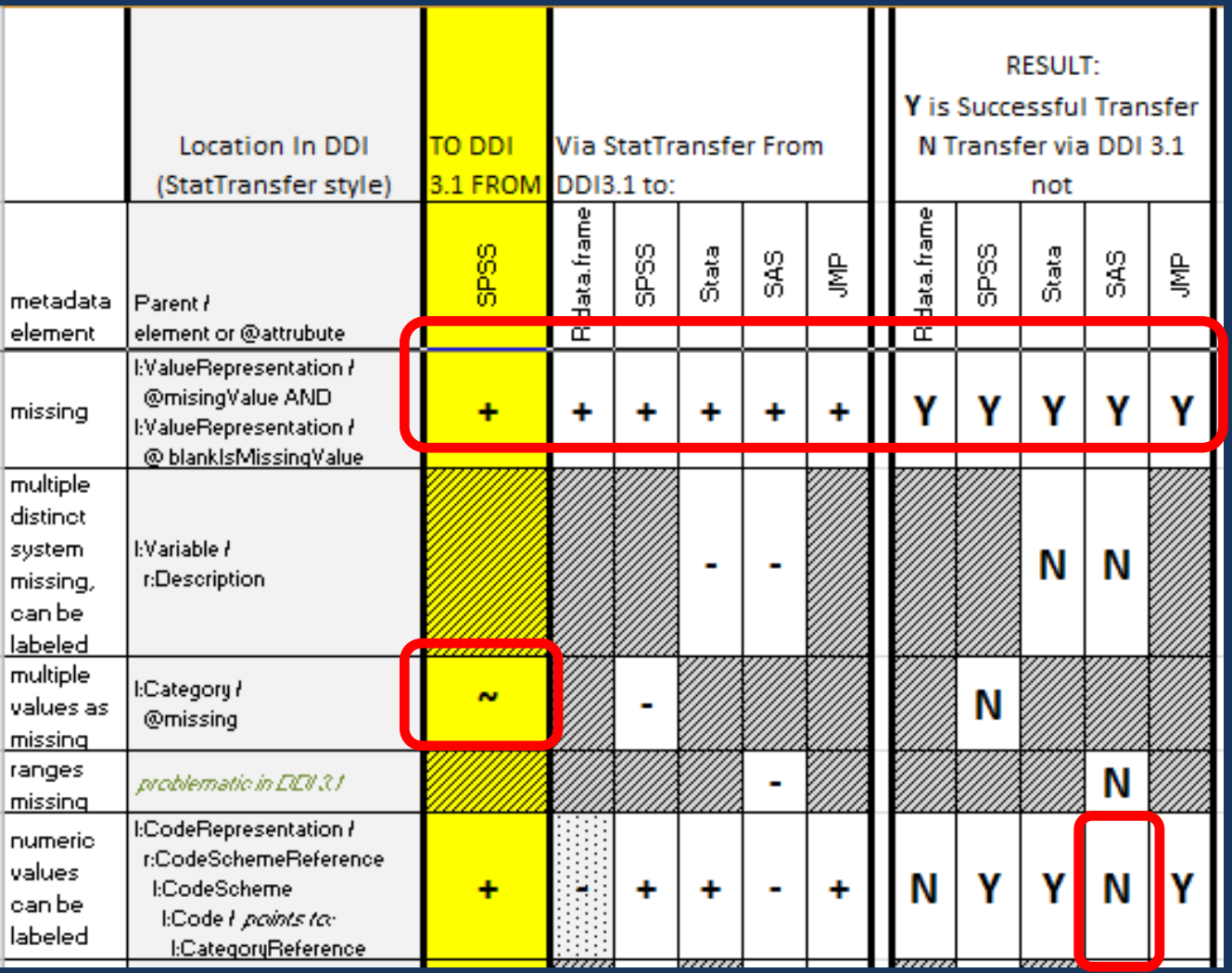
- Conversion may alter DateTime values if assumptions differ about **local**

  vs

  Coordinated Universal Time (**UTC**).

USING THE GRID FROM HERE TO THERE

Here to There (and Back Again?)



	Location In DDI (StatTransfer style)	TO DDI 3.1 FROM	Via StatTransfer From DDI3.1 to:					RESULT: Y is Successful Transfer N Transfer via DDI 3.1 not				
metadata element	Parent / element or @attrubute	SPSS	Rdata.frame	SPSS	Stata	SAS	JMP	Rdata.frame	SPSS	Stata	SAS	JMP
missing	l:ValueRepresentation / @misingValue AND l:ValueRepresentation / @ blankIsMissingValue	+	+	+	+	+	+	Y	Y	Y	Y	Y
multiple distinct system missing, can be labeled	l:Variable / r:Description				-	-				N	N	
multiple values as missing	l:Category / @missing	~		-					N			
ranges missing	*problematic in DDI 3.1*					-					N	
numeric values can be labeled	l:CodeRepresentation / r:CodeSchemeReference l:CodeScheme l:Code / *points to:* l:CategoryReference	+	-	+	+	-	+	N	Y	Y	N	Y

- "Missing" transfers to all packages.
- R does not support value labels in the same way as other packages
- Fix for importing formats to SAS is pending

R to the Others

1		Step 1: TO DDI 3.1 FROM	Step 2: From DDI3.1 to:					RESULT: Y is Successful Transfer N Transfer via DDI 3.1 not				
2	metadata element	R	R data.frame	SPSS	Stata	SAS	JMP	R data.frame	SPSS	Stata	SAS	JMP
3	**Dataset**											
4	name	-	+	+	+	+	+	n	n	n	n	n
5	Notes on dataset	-	-	-	-		~	n	n	n		n
6	user defined attrubutes	-	-	-	-			n	n	n		
7	**VARIABLES**											
8	name	+	+	+	+	+	+	Y	Y	Y	Y	Y
9	basic datatype	+	+	+	+	+	+	Y	Y	Y	Y	Y
10	display format	-	-	-	-	-	-	n	n	n	n	n
11	position	+	+	+	+	+	+	Y	Y	Y	Y	Y
12	label	+	+	+	+	+	~	Y	Y	Y	Y	n
13	specific type	+	+	+	+	+	+	Y	Y	Y	Y	Y
14	measurement level	-	-	-			-	n	n			n
15	user defined attributes	-	-	-	-		-	n	n	n		n
16	notes	-	-	-	-		~	n	n	n		n
17	**VALUES**											
18	missing	+	+	+	+	+	+	Y	Y	Y	Y	Y
19	numeric values can be labeled	+	+	+	+	-	+	n	Y	Y	n	Y
20	**Scriping Language**											
21	structured comments possible	-	-	-	-	-	-	n	n	n	n	n

Looking at Just what R contains:

The basics are preserved

SPSS to the Others

		TO DDI 3.1 FROM	Via StatTransfer From DDI3.1 to:						RESULT: Y is Successful Transfer N Transfer via DDI 3.1 not				
1													
2	metadata element	SPSS	R data.frame	SPSS	Stata	SAS	JMP		R data.frame	SPSS	Stata	SAS	JMP
3	**Dataset**												
4	name	~	+	+	+	+	+		n	n	n	n	n
5	label	+		-	-	~	~			n	n	n	n
6	date modified	+		+	~	+				Y	n	Y	
7	Notes on dataset	-	-	-	-		~		n	n	n		n
8	user defined attrubutes	-	-	-	-				n	n	n		
9	**VARIABLES**												
10	name	+	+	+	+	+	+		Y	Y	Y	Y	Y
11	basic datatype	+	+	+	+	+	+		Y	Y	Y	Y	Y
12	display format	-	-	-	-	-	-		n	n	n	n	n
13	position	+	+	+	+	+	+		Y	Y	Y	Y	Y
14	label	+	+	+	+	+	~		Y	Y	Y	Y	n
15	decimal positions	-		+	-	-	-			n	n	n	n
16	specific type	+	+	+	+	+	+		Y	Y	Y	Y	Y
17	measurement level	-	-	-			-		n	n			n
18	weight, variable is a	-		-			-			n			n
19	role	-		-						n			n
20	user defined attributes	-	-	-	-		-		n	n	n		n
21	notes	-	-	-	-		~		n	n	n		n
22	**VALUES**												
23	missing	+	+	+	+	+	+		Y	Y	Y	Y	Y
24	multiple values as missing	~		-						n			
25	numeric values can be labeled	+	-	+	+	-	+		n	Y	Y	n	Y
26	text values can be labeled	+		-		-	-			n		n	n

Stata to the Others

		TO DDI 3.1 FROM	Via StatTransfer From DDI3.1 to:						RESULT: **Y** is Successful Transfer **N** Transfer via DDI 3.1 not				
1													
2	metadata element	Stata	R data.frame	SPSS	Stata	SAS	JMP		R data.frame	SPSS	Stata	SAS	JMP
3	**Dataset**												
4	name	~	+	+	+	+	+		n	n	n	n	n
5	label	+		-	-	~	~			n	n	n	n
6	date modified	+		+	~	+				Y	n	Y	
7	Notes on dataset	-	-	-	-		~		n	n	n		n
8	user defined	-	-	-	-				n	n	n		
9	**VARIABLES**												
10	name	+	+	+	+	+	+		Y	Y	Y	Y	Y
11	basic datatype	+	+	+	+	+	+		Y	Y	Y	Y	Y
12	display format	-	-	-	-	-	-		n	n	n	n	n
13	position	+	+	+	+	+	+		Y	Y	Y	Y	Y
14	label	+	+	+	+	+	~		Y	Y	Y	Y	n
15	decimal positions	-		+	-	-	-			n	n	n	n
16	specific type	+	+	+	+	+	+		Y	Y	Y	Y	Y
17	user defined attributes	-	-	-	-		-		n	n	n		n
18	notes	-	-	-	-		~		n	n	n		n
19	**VALUES**												
20	missing	+	+	+	+	+	+		Y	Y	Y	Y	Y
21	multiple distinct	-			-	-					n	n	
22	numeric values can be	+	-	+	+	-	+		n	Y	Y	n	Y
23	multiple sets of labels				-	-					n	n	

SAS to the Others

1		TO DDI 3.1 FROM	Via StatTransfer From DDI3.1 to:						RESULT: Y is Successful Transfer N Transfer via DDI 3.1 not				
2	metadata element	SAS	R data.frame	SPSS	Stata	SAS	JMP		R data.frame	SPSS	Stata	SAS	JMP
3	**Dataset**												
4	name	+	+	+	+	+	+		Y	Y	Y	Y	Y
5	label	-		-	-	~	~			n	n	n	n
6	date created	-				+						n	
7	date modified	+		+	~	+				Y	n	Y	
8	**VARIABLES**												
9	name	+	+	+	+	+	+		Y	Y	Y	Y	Y
10	basic datatype	+	+	+	+	+	+		Y	Y	Y	Y	Y
11	display format	-	-	-	-	-	-		n	n	n	n	n
12	position	+	+	+	+	+	+		Y	Y	Y	Y	Y
13	label	+	+	+	+	+	~		Y	Y	Y	Y	n
14	scale	-				-						n	
15	decimal positions	-		+	-	-	-			n	n	n	n
16	specific type	+	+	+	+	+	+		Y	Y	Y	Y	Y
17	**VALUES**												
18	missing	+	+	+	+	+	+		Y	Y	Y	Y	Y
19	multiple distinct system missing,	-			-	-					n	n	
20	ranges missing	-				-						n	
21	numeric values can be labeled	+	-	+	+	-	+		n	Y	Y	n	Y
22	text values can be labeled	+		-		-	-			n		n	n
23	ranges can be labeled	-				-	-					n	n
24	multiple sets of labels (formats)	-			-	-					n	n	

JMP to the Others

metadata element	JMP	R data.frame	SPSS	Stata	SAS	JMP		R data.frame	SPSS	Stata	SAS	JMP
Dataset												
name	+	+	+	+	+	+		Y	Y	Y	Y	Y
label	-		-	-	≈	≈			n	n	n	n
Notes on dataset	-	-	-	-		≈		n	n	n		n
script stored with data	-					-						n
VARIABLES												
name	+	+	+	+	+	+		Y	Y	Y	Y	Y
basic datatype	+	+	+	+	+	+		Y	Y	Y	Y	Y
display format	-	-	-	-	-	-		n	n	n	n	n
position	+	+	+	+	+	+		Y	Y	Y	Y	Y
label	≈	+	+	+	+	≈		n	n	n	n	n
decimal positions	-		+	-	-	-			n	n	n	n
specific type	+	+	+	+	+	+		Y	Y	Y	Y	Y
measurement units	-					-					n	
measurement level	-	-	-			-		n	n			n
weight, variable is a	-		-			-			n			n
role	-		-						n			n
user defined attributes	-	-	-	-		-		n	n	n		n
notes	≈	-	-	-		≈		n	n	n		n
VALUES												
missing	+	+	+	+	+	+		Y	Y	Y	Y	Y
numeric values can be labeled	+	-	+	+	-	+		n	Y	Y	n	Y
text values can be labeled	+		-		-	-			n		n	n
ranges can be labeled	-				-	-					n	n
value colors	-					-						n
Scriping Language												
structured comments possible	-	-	-	-	-	-		n	n	n	n	n

Suggestions for DDI – Custom/User Attributes

- Named attributes for variables?

<l:Variable>

<l:VariableAttribute>

<r:Name>

<l:Value>

- Named attributes for dataset?

Suggestions for DDI - Ranges

- Should CodeScheme include a CodeRange element (contains Range and Value, plus CodeRange and Code for hierarchies)?

- Alternatively Code could contain a range – this would not be genericode compliant, not such a good idea

Suggestions for DDI – Multiple Labels

```
<l:Category id="Gm" version="1.0.0" versionDate="2011-10-26T13:33:00" missing="false">

<r:Label xml:lang="sv" type="GENDERshort">kvinna</r:Label>
<r:Label xml:lang="de" type="GENDERshort">weiblich</r:Label>
<r:Label xml:lang="en-US" type="GENDERshort">female</r:Label>

<r:Label xml:lang="en-US" type="GENDERLong">Self Identified
    Female</r:Label>
</l:Category>
```

Which "type" was the primary/default/selected (if any)?

Suggestions for DDI – Multiple Labels

```
<r:Label xml:lang="en-US" type="GENDERshort">female</r:Label>
</l:Category>
```

Could l:Representation for l:Variable contain "PrimaryLabelType"?

```
<l:Representation>
  <l:CodeRepresentation blankIsMissingValue="true" classificationLevel="Nominal">
    <r:RecommendedDataType>string</r:RecommendedDataType>
    <l: PrimaryLabelType > GENDERshort </l: PrimaryLabelType >
    <r:CodeSchemeReference>
      <r:ID>5c706c37-d19b-4b8e-ac6d-40094024421f</r:ID>
      <r:IdentifyingAgency>example.org</r:IdentifyingAgency>
      <r:Version>1</r:Version>
    </r:CodeSchemeReference>
  </l:CodeRepresentation>
</l:Representation>
```

Could be More Machine Actionable Than Using r:Description

metadata element	Parent / element or @attrubute
Dataset	
date created	l:LogicalProduct / r:Description
user defined attributes	l:LogicalProduct / r:Description
script stored with data	l:LogicalProduct / r:Description
VARIABLES	
user defined attributes	l:Variable / r:Description
VALUES	
multiple distinct system missing, can be labeled	l:Variable / r:Description
ranges missing	*problematic in DDI 3.1*
ranges can be labeled	*problematic in DDI 3.1* *l:NumberRange has no associated label*
multiple sets of labels (formats)	multiple relevant l:CategoryScheme elements possible but only one l:ValueRepresentation allowed l:Category can have multiple l:Label elements, distinguished by @type and @xml:lang ??Need something like *l:AlternateValueRepresentation*
value colors	l:Variable / r:Description

Could be More Machine Actionable

INTEGRITY CONSTRAINTS	
list restrictions	l:CodeScheme / l:Code No way to indicate ONLY CodeScheme Codes are valid?
restriction by expression (integrity constraints)	l:Variable / r:Description
foreign key	l:LogicalProduct / l:DataRelationship
Scriping Language	
structured comments possible	

Suggestions for Archival Datasets

- Use auxiliary variables to indicate reason for missing
  - These variables could be shared in a ResourcePackage

Measure	MeasureMissing
197.500	0
188.600	0
201.400	0
222.200	0
196.200	0
.	1
.	2
170.700	0

Codes

Measure	MeasureMissing
197.500	0
188.600	0
201.400	0
222.200	0
196.200	0
.	Don't Know
.	Refused
170.700	0

Categories

Suggestions for Archival Datasets

Auxiliary Variable for Missing: Pairing Indicated With Variable Group

```
<l:VariableGroup id="6d23978-f7fd-4328-804b-cabe0c5c3896" version="1.0.0" versionDat
<l:GroupType>Content variable and associated Indicator of reasons for missing values
<l:VariableGroupName>Measure Group</l:VariableGroupName>
<r:Label>Set of variables for Measure</r:Label>
<r:Description>MeasureMissing is linked to Measure. MeasureMissing indicates distinc
<l:VariableReference>
 <r:Scheme>
 </r:Scheme>
 <r:ID>9fed70df-8a4a-4f18-a5a0-fe8eadc1ff54</r:ID>
 <r:IdentifyingAgency>example.org</r:IdentifyingAgency>
 <r:Version>1.0.0</r:Version>
</l:VariableReference>
<l:VariableReference>
 <r:Scheme>
 </r:Scheme>
 <r:ID>034b2ac2-3c5e-49c5-ae57-bc5b445987c2</r:ID>
 <r:IdentifyingAgency>example.org</r:IdentifyingAgency>
 <r:Version>1.0.0</r:Version>
</l:VariableReference>
```

Suggestions for Archival Datasets

- Create additional variables for alternative formats?
  - Long labels
  - Languages
- An alternative would be to put multiple labels in keyed relational tables, but having multiple tables creates its own set of complications

Suggestions for Archival Datasets

- Create additional variables for coded ranges
  - Information is lost – depends on what values are present.

Suggestions for Archival Datasets

- Use controlled vocabulary for user attributes (characteristics, properties)
  - DDI based?
  - Useful for Semantic Data form of DDI element names ?

Conclusions

- Adoption of DDI by tools like Stat/Transfer is encouraging.
- The current state still means that some important metadata that might be contained in proprietary format data files still must be either
  - hand entered into DDI or
  - harvested and entered by user-written or other code.

Conclusions

- Basic metadata is transferrable among all 5 packages through DDI
- No one package has a superset of the others, several desirable metadata elements are not universally supported
- DDI almost supports a superset of the packages considered – a worthwhile goal
- Representation as a ResourcePackage vs a StudyUnit can require intermediate transformation
  - Need best practice recommentation?

References

- Hoyle, Larry and Joachim Wackerow with Oliver Hopt *DDI 3: Extracting Metadata from the Data Analysis Workflow.* DDI Working Paper Series, Schloss Dagstuhl, Germany, 2010. http://dx.doi.org/10.3886/DDIUseCases04
- R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
- Wright, Philip A. *Eliminating Redundant Custom Formats* SAS Global Forum 2011 paper 217-2011 http://support.sas.com/resources/papers/proceedings11/217-2011.pdf

- JMP - http://www.jmp.com/
- R - http://www.r-project.org/
- SAS - http://www.sas.com/
- SPSS - http://www-01.ibm.com/software/analytics/spss/
- Stata - http://www.stata.com/
- Stat/Transfer - http://www.stattransfer.com/

DISCUSSION?

Metadata

- Shoe reference
- http://www.shoecomics.com/archives/shoe_daily/shoe_daily100211.jpg

Contact

Larry Hoyle

University of Kansas,

Institute for Policy & Social Research

LarryHoyle@ku.edu

For files from this presentation see:

http://www.ipsr.ku.edu/ksdata/DDI/

Acknowledgements

The authors view the inclusion of DDI into Stat/Transfer as an important development and look forward to its development into a very useful tool for the DDI community.

Dmitry Basko and Steven Dubnoff at Circle Systems have been very responsive in improving import and export between DDI and Stat/Transfer as suggestions have been made during the development of this paper.