

Tracing Data Elements through a Standard Data Flow

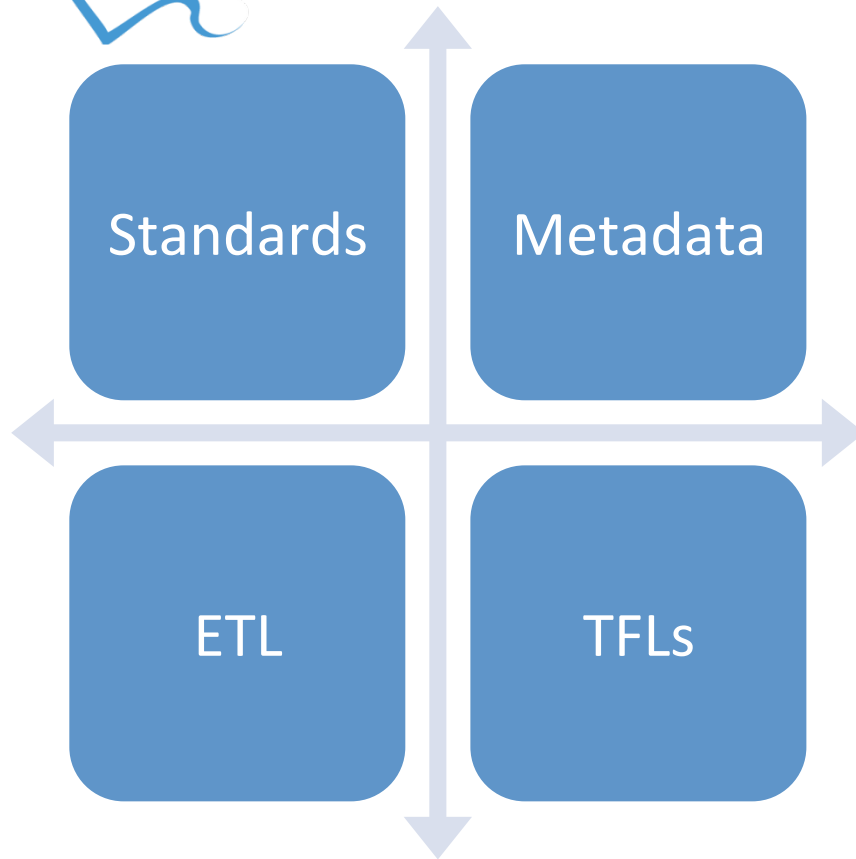
Alistair Dootson

Dootsonic Ltd, Manchester, UK

The missing piece of the puzzle

www.dootsonic.com





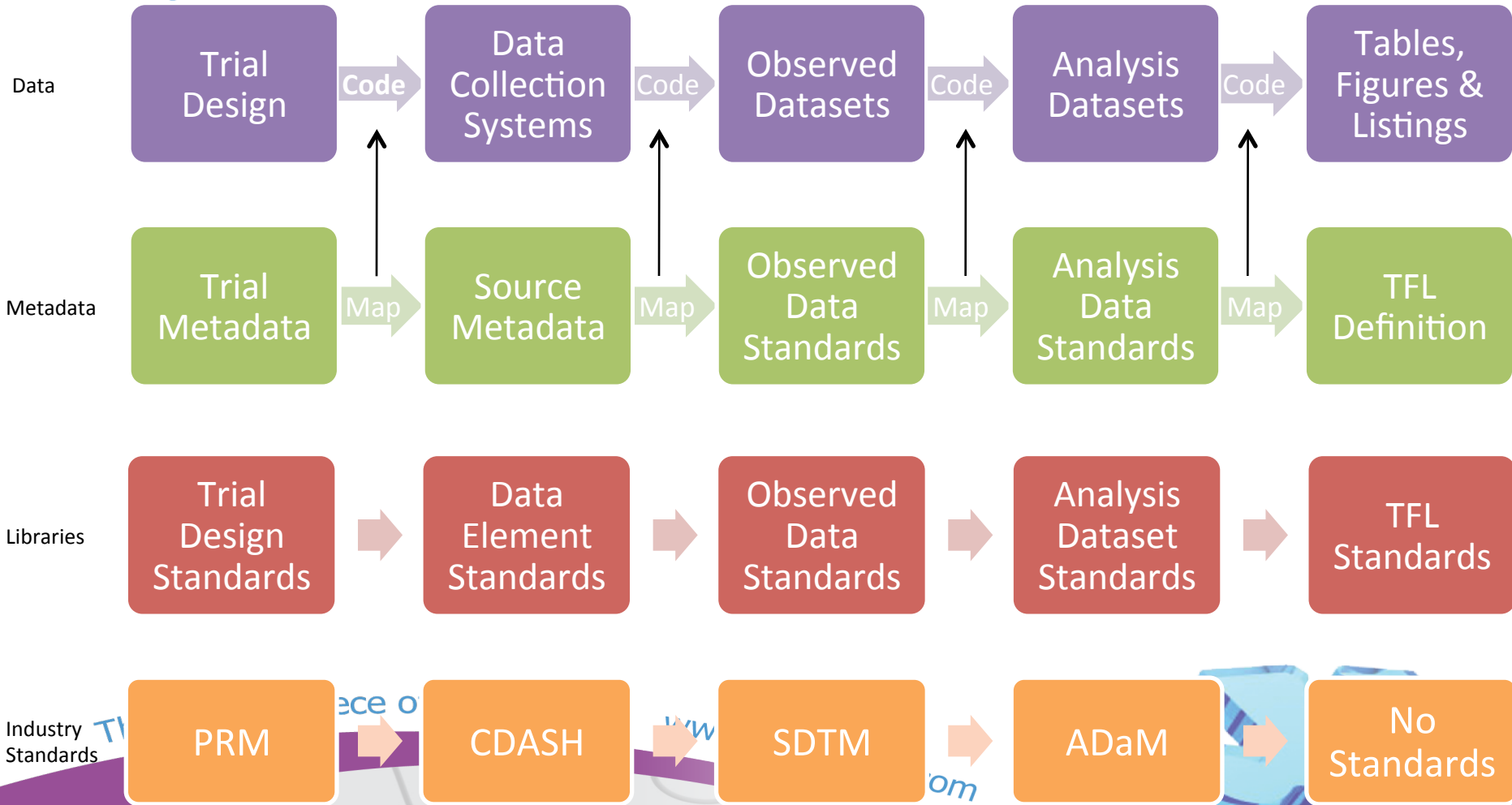
Traceability

The missing piece of the puzzle

www.dootsonic.com



“Floverview”



- Industry Standards
 - PRM
 - CDASH
 - SDTM
 - ADaM
- Internal Standards
- CDISC-SHARE[®]

The missing piece of the puzzle

www.dootsonic.com



- What is Metadata?
- How do we “get” metadata?
- Proc Contents?

The missing piece of the puzzle

www.dootsonic.com



- Proc SCAPROC

```
proc scaproc;  
    record "C:\phuse\record.txt" attr;  
run;
```

... code ...

```
proc scaproc;  
    write;
```

```
run;
```

The missing piece of the puzzle

www.dootsonic.com



```
/* JOBSPLIT: DATASET INPUT SEQ WORK.DEMO.DATA */  
/* JOBSPLIT: LIBNAME WORK V9 'C:\DOCUME~1\LOCALS~1\Temp\SAS  
Temporary Files\_TD5868' */  
/* JOBSPLIT: DATASET OUTPUT SEQ WORK.DEMO1.DATA */  
/* JOBSPLIT: LIBNAME WORK V9 'C:\DOCUME~1\LOCALS~1\Temp\SAS  
Temporary Files\_TD5868' */  
/* JOBSPLIT: ATTR WORK.DEMO.DATA INPUT VARIABLE:subjid  
TYPE:NUMERIC LENGTH:8 LABEL:Subject ID FORMAT:BEST12.  
INFORMAT:F12. */  
/* JOBSPLIT: ATTR WORK.DEMO.DATA INPUT VARIABLE:siteid  
TYPE:CHARACTER LENGTH:2 LABEL:Center or Site ID FORMAT:  
$F2. INFORMAT:$F2. */
```

The missing piece of the puzzle

www.dootsonic.com



DM Properties

General Columns Indexes Keys Parameters Physical Storage Options Notes Extended Attributes Authorization Clinical Domain







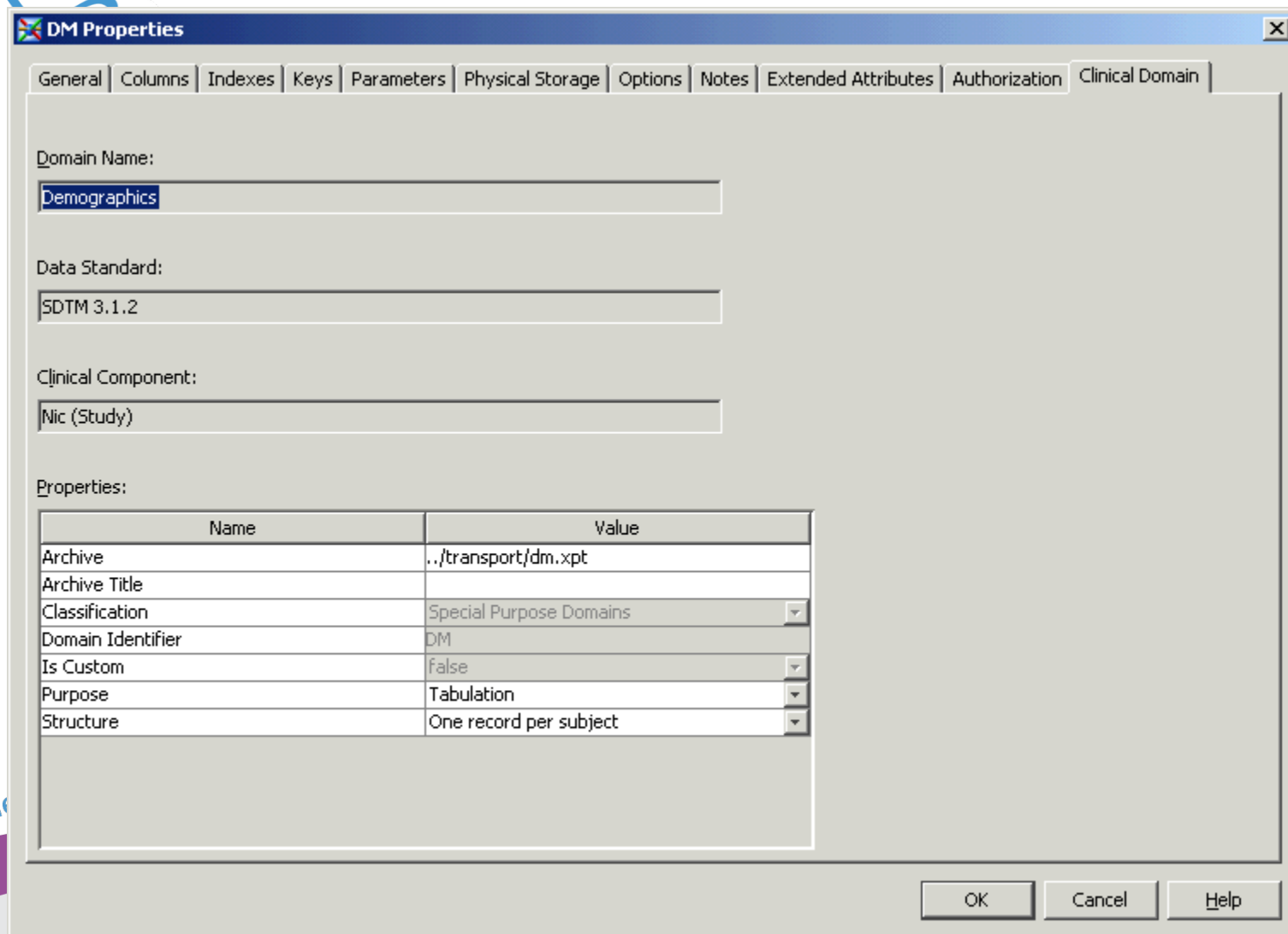






#	Name	Description	Type	Length	Informat	Format	Is Nullable	Summary
1	 STUDYID	Study Identifier	Character	40	(None)	(None)	Yes	(None)
2	 DOMAIN	Domain Abbreviation	Character	8	(None)	(None)	Yes	(None)
3	 USUBJID	Unique Subject Ide...	Character	40	(None)	(None)	Yes	(None)
4	 SUBJID	Subject Identifier f...	Character	40	(None)	(None)	Yes	(None)
5	 RFSTDTC	Subject Reference ...	Character	64	(None)	(None)	Yes	(None)
6	 RFENDTC	Subject Reference ...	Character	64	(None)	(None)	Yes	(None)
7	 SITEID	Study Site Identifier	Character	40	(None)	(None)	Yes	(None)
8	 INVID	Investigator Identifier	Character	40	(None)	(None)	Yes	(None)
9	 INVNAM	Investigator Name	Character	40	(None)	(None)	Yes	(None)
10	 BRTHDTC	Date/Time of Birth	Character	64	(None)	(None)	Yes	(None)
11	 AGE	Age	Numeric	8	(None)	(None)	Yes	(None)
12	 AGEU	Age Units	Character	10	(None)	(None)	Yes	(None)
13	 SEX	Sex	Character	2	(None)	(None)	Yes	(None)
14	 RACE	Race	Character	40	(None)	(None)	Yes	(None)
15	 ETHNIC	Ethnicity	Character	40	(None)	(None)	Yes	(None)
16	 ARMCD	Planned Arm Code	Character	20	(None)	(None)	Yes	(None)
17	 ARM	Description of Plann...	Character	40	(None)	(None)	Yes	(None)
18	 COUNTRY	Country	Character	3	(None)	(None)	Yes	(None)
19	 DMDTC	Date/Time of Collec...	Character	64	(None)	(None)	Yes	(None)
20	 DMDY	Study Day of Collec...	Numeric	8	(None)	(None)	Yes	(None)

OK Cancel Help



DM Properties

General | Columns | Indexes | Keys | Parameters | Physical Storage | Options | Notes | Extended Attributes | Authorization | Clinical Domain

Domain Name:
Demographics

Data Standard:
SDTM 3.1.2

Clinical Component:
Nic (Study)

Properties:

Name	Value
Archive	../transport/dm.xpt
Archive Title	
Classification	Special Purpose Domains
Domain Identifier	DM
Is Custom	false
Purpose	Tabulation
Structure	One record per subject

OK Cancel Help

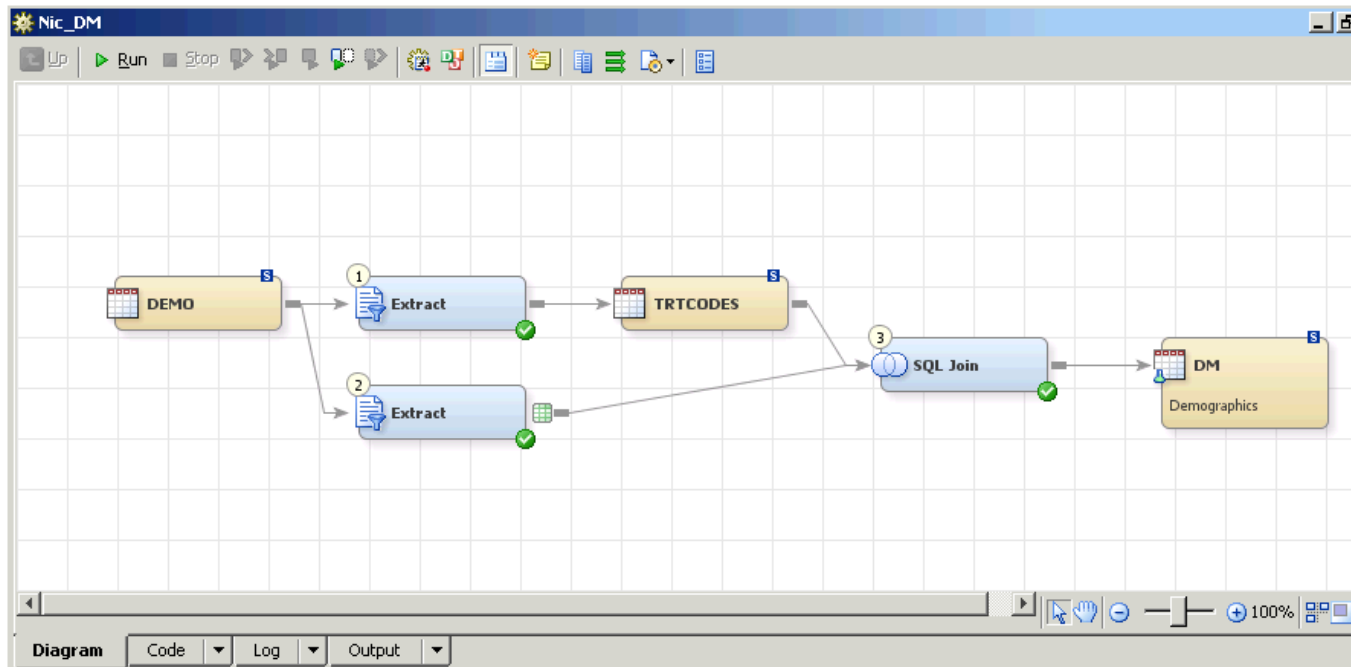
- Clinical DI Studio [®]
- Use Metadata to Produce Jobs
- Trace that Metadata

The missing piece of the puzzle

www.dootsonic.com



- A Basic Job

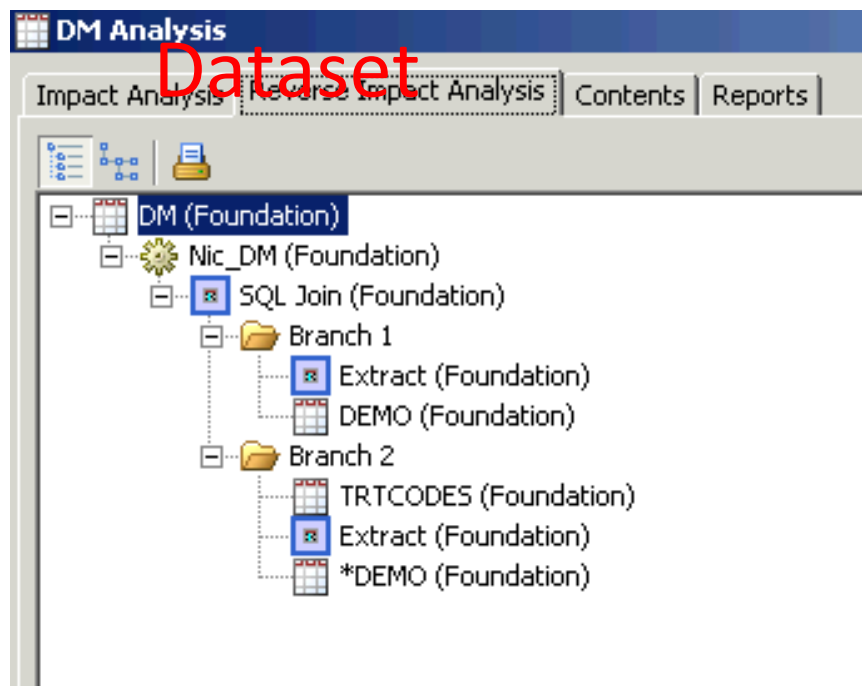


The missing piece of the puzzle

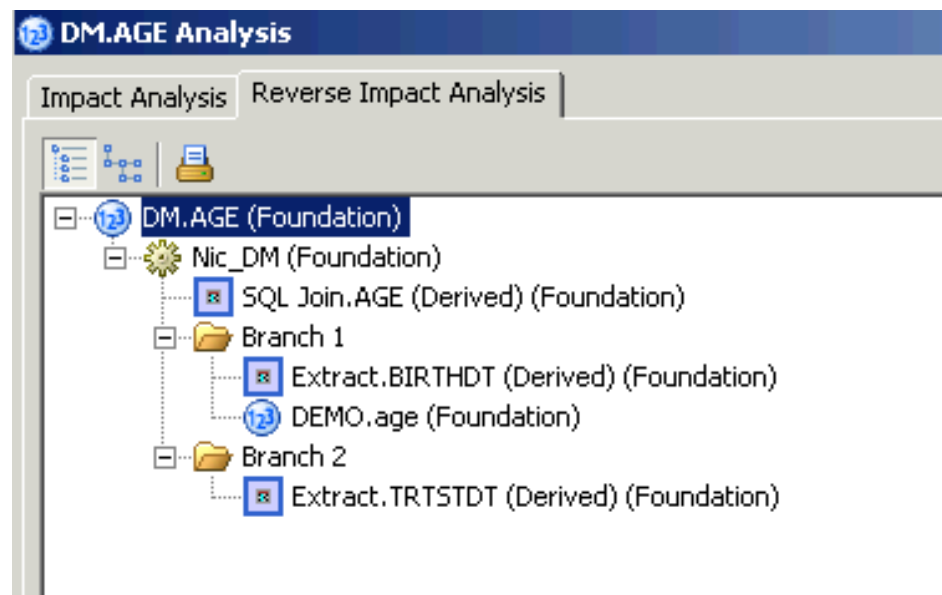
www.dootsonic.com

- Use Analyze Features

Dataset



Column

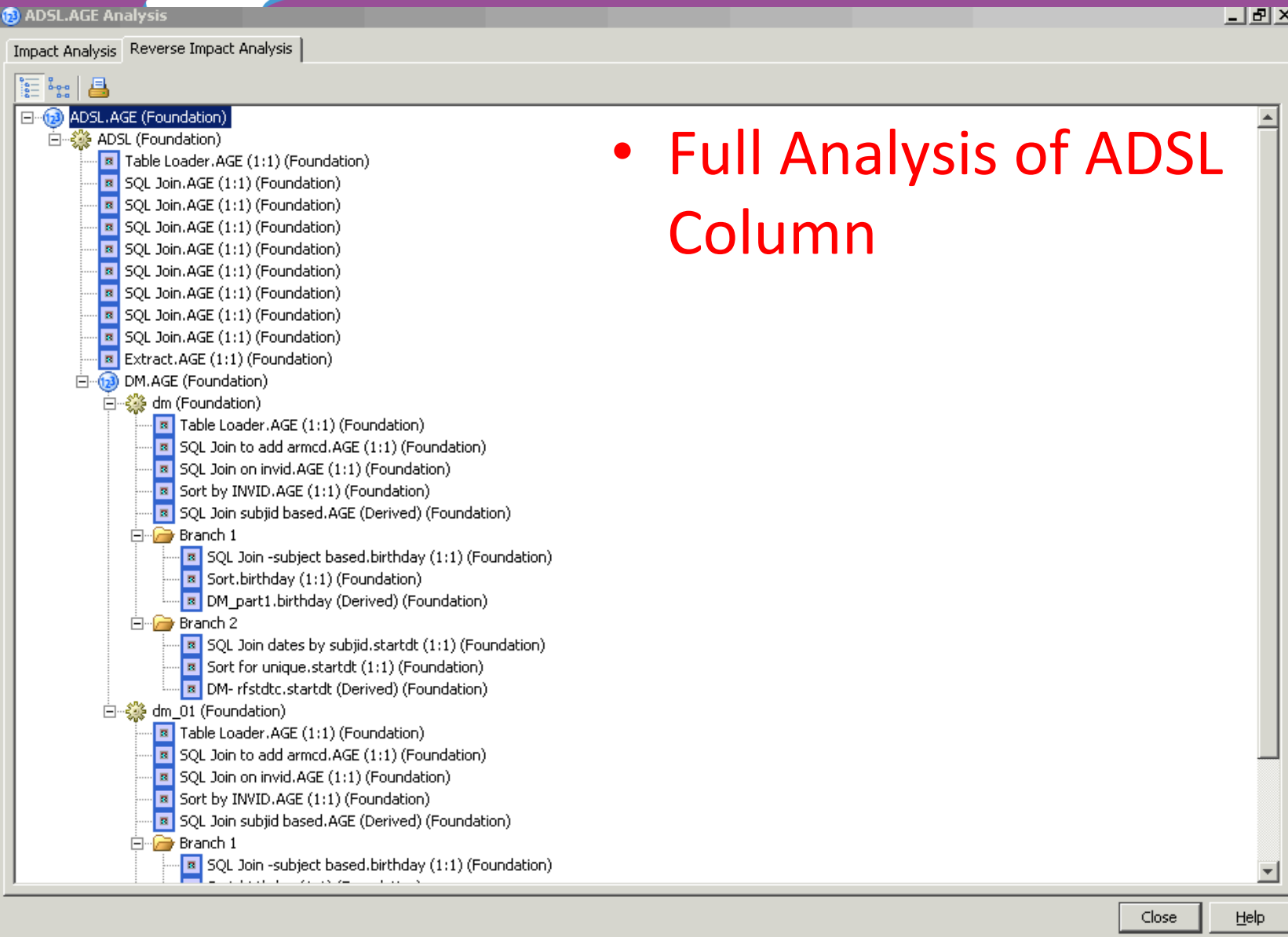


The missing piece of the puzzle

www.dootsonic.com



- Full Analysis of ADSL Column



The screenshot shows the 'ADSL.AGE Analysis' window with two tabs: 'Impact Analysis' and 'Reverse Impact Analysis'. The 'Impact Analysis' tab is active, displaying a hierarchical tree of data processes. The tree is organized as follows:

- ADSL.AGE (Foundation)
 - ADSL (Foundation)
 - Table Loader.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - SQL Join.AGE (1:1) (Foundation)
 - Extract.AGE (1:1) (Foundation)
 - DM.AGE (Foundation)
 - dm (Foundation)
 - Table Loader.AGE (1:1) (Foundation)
 - SQL Join to add armcd.AGE (1:1) (Foundation)
 - SQL Join on invid.AGE (1:1) (Foundation)
 - Sort by INVID.AGE (1:1) (Foundation)
 - SQL Join subjid based.AGE (Derived) (Foundation)
 - Branch 1
 - SQL Join -subject based.birthday (1:1) (Foundation)
 - Sort.birthday (1:1) (Foundation)
 - DM_part1.birthday (Derived) (Foundation)
 - Branch 2
 - SQL Join dates by subjid.startdt (1:1) (Foundation)
 - Sort for unique.startdt (1:1) (Foundation)
 - DM- rfstdtc.startdt (Derived) (Foundation)
 - dm_01 (Foundation)
 - Table Loader.AGE (1:1) (Foundation)
 - SQL Join to add armcd.AGE (1:1) (Foundation)
 - SQL Join on invid.AGE (1:1) (Foundation)
 - Sort by INVID.AGE (1:1) (Foundation)
 - SQL Join subjid based.AGE (Derived) (Foundation)
 - Branch 1
 - SQL Join -subject based.birthday (1:1) (Foundation)

The interface includes a toolbar with icons for file operations and a status bar at the bottom with 'Close' and 'Help' buttons.

Subject Demographics

Page x of y
hh:mm ddmmmyyyy
status

ITT Population

STUDY #

Parameter	Arm A (pem+cis) (N=xxx)		Arm B (placebo+cis) (N=xxx)		Total (N=xxx)	
	n	(%)	n	(%)	n	(%)
Gender						
Female	xx	(xx.x)	xx	(xx.x)	xx	(xx.x)
Male	xx	(xx.x)	xx	(xx.x)	xx	(xx.x)
Race						
Caucasian	xx	(xx.x)	xx	(xx.x)	xx	(xx.x)
Hispanic	xx	(xx.x)	xx	(xx.x)	xx	(xx.x)
Asian	xx	(xx.x)	xx	(xx.x)	xx	(xx.x)
<u>Age (years)</u>						
Number of Subjects	xx		xx		xx	
Mean	xx.x		xx.x		xx.x	
SD	xx.x		xx.x		xx.x	
Median	xx.x		xx.x		xx.x	
Minimum	xx.x		xx.x		xx.x	
Maximum	xx.x		xx.x		xx.x	

The missing piece of the puzzle

www.dootsonic.com



Raw Data

Raw
→
SDTM

```
( floor(( intck( 'month',DEMO.BIRTHDT,datepart( DEMO.TRTSTDY )) -  
( day( datepart( DEMO.TRTSTDY )) < day( DEMO.BIRTHDT ))) / 12 ));
```

SDTM
→
ADaM

DM.AGE

ADaM
→
Display

```
proc MEANS data=DM noprint;  
var AGE;  
output out=MEANS median=MedianAge max=MaxAge min=MinAge std=StdAge mean=MeanAge;  
run;
```

The missing piece of the puzzle

www.dootsonic.com

- How is it calculated for display in the table?
- What are the inputs for that calculation in the ADaM data?
- What are the observed data elements (SDTM) that feed into the ADaM derivation?
- What are the original data elements in CDASH?

The missing piece of the puzzle

www.dootsonic.com



Variable Name	Variable Label	Variable Type	Core	Display Format	Codelist / Controlled Terms	Source / Derivation
STUDYID	Study Identifier	Char	Req	50		"NIC001"
DOMAIN	Domain Abbreviation	Char	Req	2	**DM	"DM"
USUBJID	Unique Subject Identifier	Char	Req	50		left(put(DEMO.SUBJID1,Z6.));
SUBJID	Subject Identifier for the Study	Char	Req	50		substr(compress(put(DEMO.SUBJID1,Z6.)),4,3);
RFSTDC	Subject Reference Start Date/Time	Char	Exp	50	ISO 8601	put(DEMO.TRTSTDTC, IS8601DT.)
RFENDTC	Subject Reference End Date/Time	Char	Exp	50	ISO 8601	put(DEMO.TRTENDTC, IS8601DT.)
SITEID	Study Site Identifier	Char	Req	8		"01";
INVID	Investigator Identifier	Char	Perm	8	*	" "
INVNAM	Investigator Name	Char	Perm	50		DEMO.INVNAME;
BRTHDTC	Date/Time of Birth	Char	Perm	50	ISO 8601	put(DEMO.BIRTHDTC,IS8601DA.);
AGE	Age in AGEU at RFSTDC	Num	Exp	8		(floor((intck('month',DEMO.BIRTHDTC,datepart(DEMO.TRTSTDTC)) - (day(datepart(DEMO.TRTSTDTC)) < day(DEMO.BIRTHDTC))) / 12));
AGEU	Age Units	Char	Exp	50	** YEARS	"YEARS"
SEX	Sex	Char	Req	1	**M, F	DEMO.SEX
RACE	Race	Char	Exp	100	**	DEMO.RACE

The missing piece of the puzzle

www.dootsonic.com



Variable Name	Variable Label	Variable Type	Core	Display Format	Codelist / Controlled Terms	Source / Derivation	Original Source / Derivation
STUDY IDENTIFIERS: KEY VARIABLES THAT IDENTIFY THE STUDY, SUBJECT, AND SITE							
STUDYID	Study Identifier	Char	Req	9		DM.STUDYID	"NIC001"
USUBJID	Unique Subject Identifier	Char	Req	18		DM.USUBJID	left(put(DEMO.SUBJID1,Z6.));
SUBJID	Subject Identifier for the Study	Char	Req	8		DM.SUBJID	substr(compress(put(DEMO.SUBJID1,Z6.)),4,3);
SITEID	Study Site Identifier	Char	Req	3		DM.SITEID	"01";
COUNTRY	Country	Char	Perm	3		DM.COUNTRY	if upcase(COUNTRY1)="UNITED STATES" then COUNTRY='USA'
DEMOGRAPHICS: INCLUDES SUBJECT DEMOGRAPHICS, GROUPING VARIABLES, AND FLAGS.							
AGE	Age	Num	Req	integer		DM.AGE	(floor((intck('month',DEMO.BIRTHDT,datepart(DEMO.TRTSTDY)) - (day(datepart(DEMO.TRTSTDY)) < day(DEMO.BIRTHDT)) / 12));
AGEU	Age Units	Char	Req	6	(AGEU)	DM.AGEU	"YEARS"
AGECAT1	Age Category 1	Char	Perm	6	AGECAT1	if AGE<50 then '<50' else if AGE>50 then '>=50'	
SEX	Sex	Char	Req	8	(SEX)	DM.SEX	DEMO.SEX
RACE	Race	Char	Req	40	(RACE)	DM.RACE	DEMO.RACE
RACECAT1	Race Category 1	Char	Perm	15	RACECAT1	if RACE='WHITE' then 'Caucasian' else 'Non-Caucasian'	
ARM	Description of Planned Arm	Char	Req	50		DM.ARM	DEMO.TRTGRP
DISCREAS	Reason for Discontinuation	Char	Cond	40		DS.DSDECOD where DSCAT="DISPOSITION EVENT" and EPOCH="XXXXXX"	

The missing piece of the puzzle

www.dootsonic.com



Variable Name	Variable Label	Variable Type	Core	Display Format	TFL Calculation	Source / Derivation
STUDY IDENTIFIERS: KEY VARIABLES THAT IDENTIFY THE STUDY, SUBJECT, AND SITE						
STUDYID	Study Identifier	Char	Req	9		DM.STUDYID
USUBJID	Unique Subject Identifier	Char	Req	18		DM.USUBJID
SUBJID	Subject Identifier for the Study	Char	Req	8		DM.SUBJID
SITEID	Study Site Identifier	Char	Req	3		DM.SITEID
COUNTRY	Country	Char	Perm	3		DM.COUNTRY
DEMOGRAPHICS: INCLUDES SUBJECT DEMOGRAPHICS, GROUPING VARIABLES, AND FLAGS.						
AGE	Age	Num	Req	integer	proc MEANS data=DM noprint; var AGE; output out=MEANS median=MedianAge max=MaxAge min=MinAge std=StdAge mean=MeanAge; run;	DM.AGE

- Note: Some columns excluded

The missing piece of the puzzle

www.dootsonic.com



- Some ETL tools provide the facility to trace data elements through their system.
- The point of traceability is to ensure the quality and integrity of the final variable values.
- When tracing back variables from the tables, would it shed light on how many variables are collected and not used?
- What drives the data flow of a study. Are the specifications and database built first?

The missing piece of the puzzle

www.dootsonic.com



- However, consider defining the tables first and working backwards. Would that make the data capture and flow more efficient?
- There seems to be great benefits available for reusability from ETL tools, but not necessarily for full traceability.
- There's no one system out there that does this traceability completely, but there are ways of manually putting this together to add value

The missing piece of the puzzle

www.dootsonic.com



- Standards
- Be Proactive
- Keep up to date with initiatives
- TFL Development
- No tools can currently trace from TFL back to Raw Data
- Transparency of data is not yet required, but it is becoming more visible

The missing piece of the puzzle

www.dootsonic.com





alistair@dootsonic.com

Questions?