

Legacy to SDTM Conversion Workshop: Tools and Techniques

Mike Todd
President
Nth Analytics

A thick, dark blue horizontal bar with rounded ends, positioned below the speaker's name.

Legacy Data

- Old studies never die ...
- Legacy studies are often required for submissions or pharmacovigilance.
- Often there are multiple Legacy systems, disparate from each other
- Problem: design an efficient method for converting legacy data to SDTM

Legacy CDISC Implementation Goals

- Design a strategy such that:
 - No knowledge needed of system that originally produced the legacy data
 - Applicable to files from any system
 - Implementation is flexible enough to adapt to different study designs
 - Minimal programming support required for maintenance
 - Reasonable cost

Study Scenarios

- Well-documented
 - Raw data available
 - Analysis data reliable
 - Study report, SAP, CRF available
 - Someone familiar with the study available to answer questions
- Less well-documented
 - Analysis data either not available, or not reliable
 - No SAP
 - Study report missing appendices
 - No one remembers the study
 - Requires a lot of “pre-work” before automated methods can be applied

Well-Documented Studies

- This presentation focuses on the best case
- These studies lend themselves to automated, non-expert driven solutions

The Other Ones, Briefly ...

- Becomes a data management problem
- Problematic data may be excluded or imputed, so long as a reasonable, well-documented process can be defined
 - Replace unreliable lab normal ranges with published ranges
- Requires expert data manager to identify problems and propose solutions

Well-Documented Studies

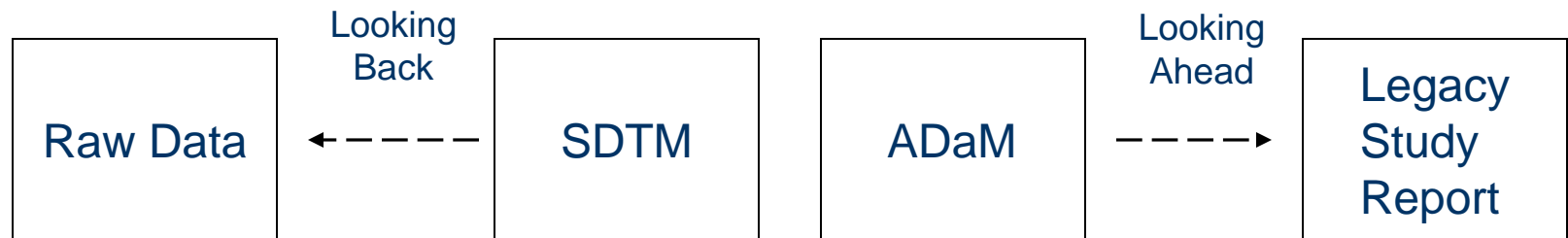
ETL Process

A thick, dark blue horizontal bar with rounded ends, positioned below the "ETL Process" text.

Our approach

- Start with the **analysis files**
 - Transform these to SDTM
 - Usually contain most of the data required for SDTM
 - Better ones tend to be self-documenting
- Compare the analysis files to the SDTM domains
 - Ensure that required and expected SDTM variables are available
 - Understand all derivations from SAP or study report

Issues



- Raw data, analysis data, or combination?
- SDTM is designed to represent raw data
 - raw data not included must be documented
- Must match results in legacy CSR
 - or explain why

ETL Process

- Define how raw/analysis data fits into SDTM domains and variables
- Match data to required, permitted and expected SDTM data when possible
- Provide an automated mechanism for specifying the data sources and algorithms
- Basis for the FDA-mandated “DEFINE.PDF” documentation
- Provide the metadata for the SDTM files

Implementing an ETL Process

- Programs read table-driven metadata to translate the analysis data into SDTM formats
 - Tells the SAS code which analysis variables populate the SDTM variables
 - Indicates when specialized code is required
- All code is developed to be generic using the metadata to indicate when variations are required
- New studies only require changes to metadata

Sample Mapping Spreadsheet

Microsoft Excel - Book1

File Edit View Insert Format Tools Data S-PLUS Window Help Acrobat

C5 = Subject Identifier for the Study

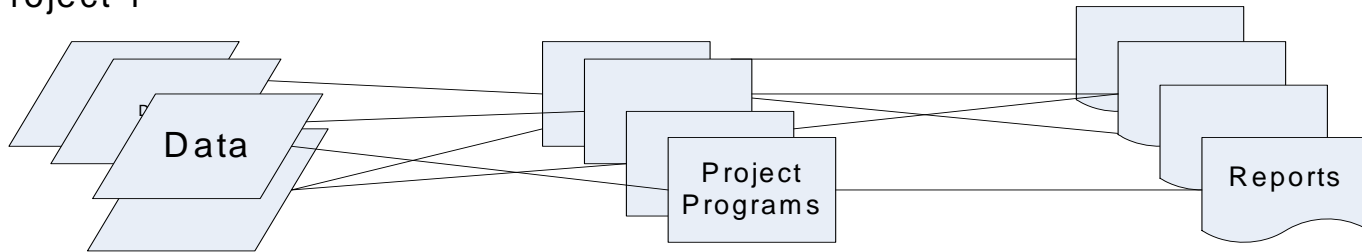
	A	B	C	D	E	F	G	H
1	Domain	VariableName	VariableLabel	Type	Origin	Role	Comments	Core
2	DM	STUDYID	Study Identifier	Char	CRF	Identifier	[default]	Req
3	DM	DOMAIN	Domain Abbreviation	Char	Derived	Identifier	[default]	Req
4	DM	USUBJID	Unique Subject Identifier	Char	Sponsor Defined	Identifier	[default]	Req
5	DM	SUBJID	Subject Identifier for the Study	Char	CRF	Topic	%USUBJID(VARNAME=SUBJID)	Req
6	DM	RFSTDTC	Subject Reference Start Date/Time	Char	Sponsor Defined	Timing	%RFSTDTC	Exp
7	DM	RFENDTC	Subject Reference End Date/Time	Char	Sponsor Defined	Timing	%RFENDTC	Exp
8	DM	SITEID	Study Site Identifier	Char	Derived	Record Qualifier	SUBSTR(DEMOG.INVSITE,5,3)	Req
9	DM	INVID	Investigator Identifier	Char	Derived	Record Qualifier	DEMOG.INV	Perm
10	DM	INVNAM	Investigator Name	Char	Derived	Synonym Qualifier	%INVNAM	Perm
11	DM	BIRTHDTC	Date/Time of Birth	Char	CRF	Result Qualifier	%ISO_DATETIME(DATE=DEMOG.DMDOB DT, TIME=0)	Perm
12	DM	AGE	Age in AGEU at RFSTDTC	Num	Derived	Result Qualifier	%AGE	Exp

Sheet1 / Sheet2 / Sheet3

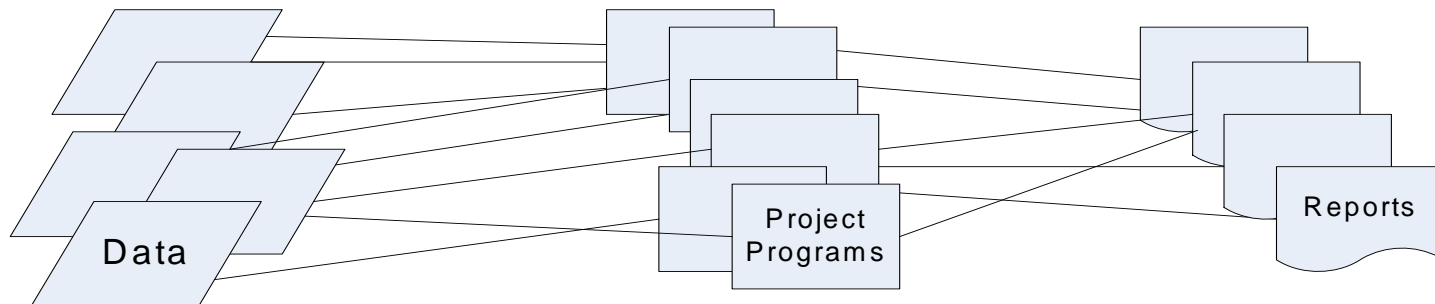
Ready

Process Without Automation

Project 1

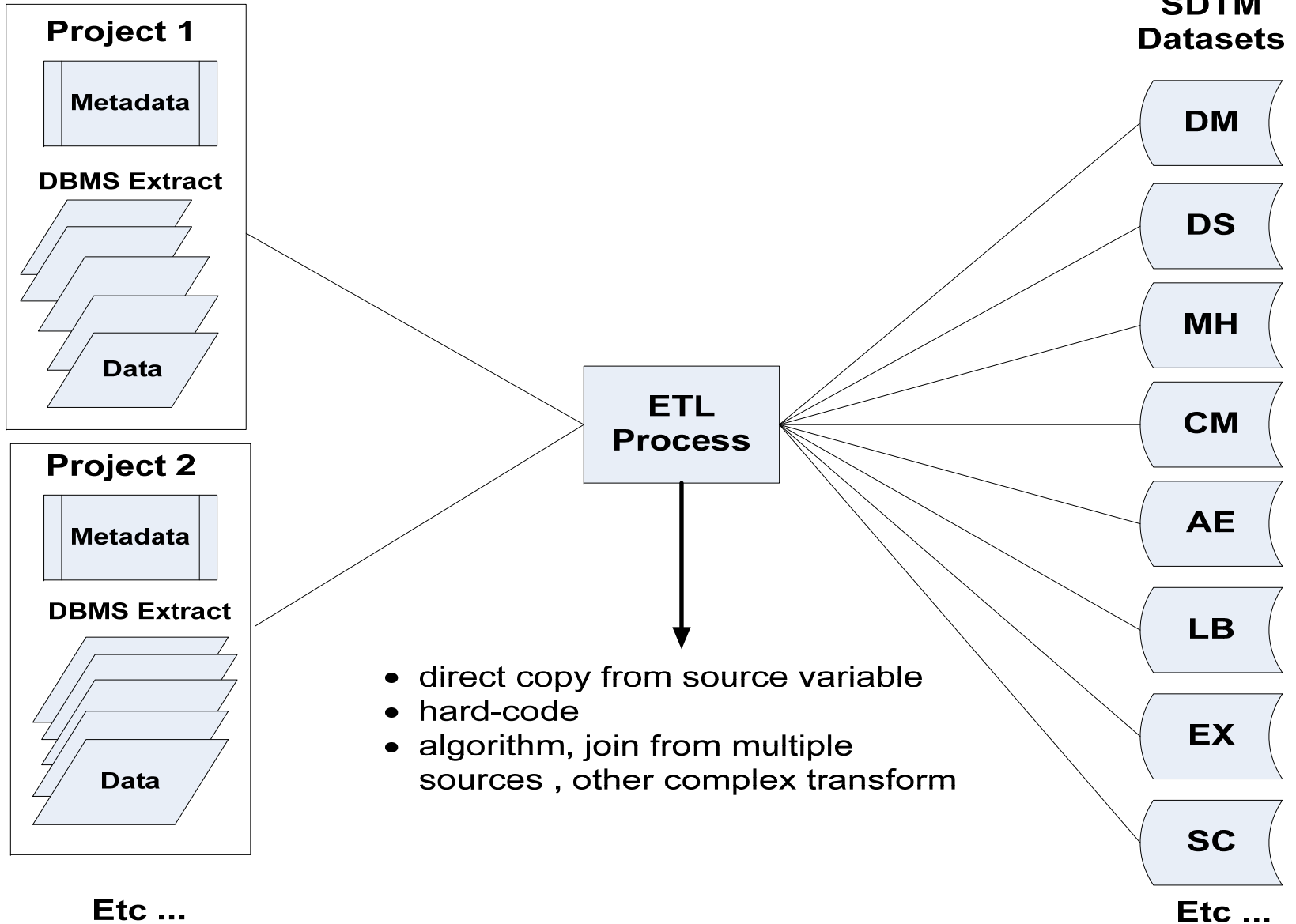


Project 2



... etc.

ETL Process



Software

- Market leader : SAS Data Integration Studio
 - Formerly ETL Studio
 - \$80K per server
- “Off-the-Shelf ETL“
 - Metadata: Excel and Access work very well
 - Converts easily to DEFINE.PDF
 - SAS macros read metadata, generate custom SAS code to create SDTM domains from source data
 - Generates standalone, submission-ready programs

Validation: SDTM Structure

- SDTM compliance checks
 - Conformance with Implementation Guide rules can be automated
 - Variable names, labels, type etc. are correct
 - All required variables have values, etc.

Validation: Source Data Validation

- Verify metadata
 - Data manager/statistician
 - Manual review process
- Possible approaches
 - independent programming
 - for each raw dataset, verify raw to SDTM conversion for a random sample of subjects

Validation: All Raw Data Mapped?

- Leverage metadata
- Use metadata to query raw data structure and determine differences
 - List of raw data variables not mapped
 - Raw datasets vs. domains
 - Raw data variables to SDTM variables

Validation: New SDTM/ADaM Match Legacy Study Report?

- Demonstrate that programs using SDTM (or ADaM) data can reproduce results in the legacy study report
- Nasty – no shortcuts
- Tasks can split among multiple programmers
 - they will be needed!

Advantages

- Does not disrupt existing clinical trial systems
- It works for all legacy data
- Reduces cost since only metadata changes for each new study
- Only new study specific situations cause additional coding
- Changes to CDISC standards are made to the metadata avoiding costly programming revisions

Advantages (continued)

- Self-documenting
- Metadata easily converted to DEFINE.PDF
 - If you follow the system, metadata are guaranteed to provide complete and accurate documentation
- Generates submission-ready SAS programs
 - System macros create standalone code
 - Code looks good because it is machine-written
 - Use PROC COMPARE to verify that standalone programs accuracy create the SDTM datasets

Challenges

- **Incomplete Analysis Data**
 - Have to go back to raw data for domains/variables not covered by analysis data
 - Example: Inclusion/Exclusion
- **Incomplete Raw Data**
 - If annotated CRF is unavailable, it may be difficult to determine what certain raw data actually represent.
 - Data quality issues:
 - was the data cleaned?

Challenges (continued)

- Trial Design and Subject domains
 - Trial Design and Subject domains have to be created manually
- ADaM
 - Ideally, ADaM data are created from SDTM data
 - For derived variables already in analysis datasets, recreate them, then use original variables for validation

Summary

- Messy legacy studies require a lot of data management expertise before automated methods can be used
- Metadata is the key for successful automation
- Off-the-shelf tools can provide a powerful ETL solution