Legacy to SDTM Conversion Workshop: Tools and Techniques

Mike Todd
President
Nth Analytics

Quality to the Nth Degree

Nth Analytics

Legacy Data

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

Quality to the Nth Degree
Nth Analytics

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

Quality to the Nth Degree
Nth Analytics

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

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

Quality to the Nth Degree
Nth Analytics

ETL Process

- Define how 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

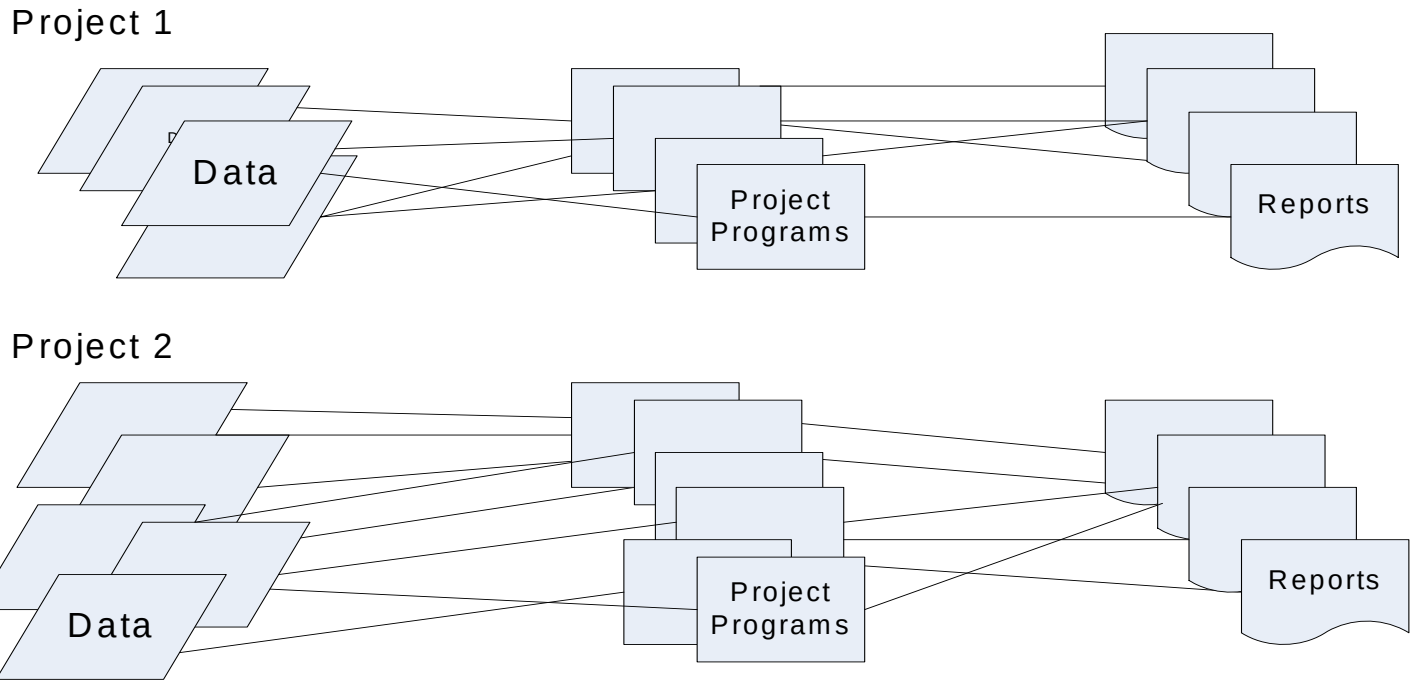
Implementing an ETL Process

- 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

Quality to the Nth Degree

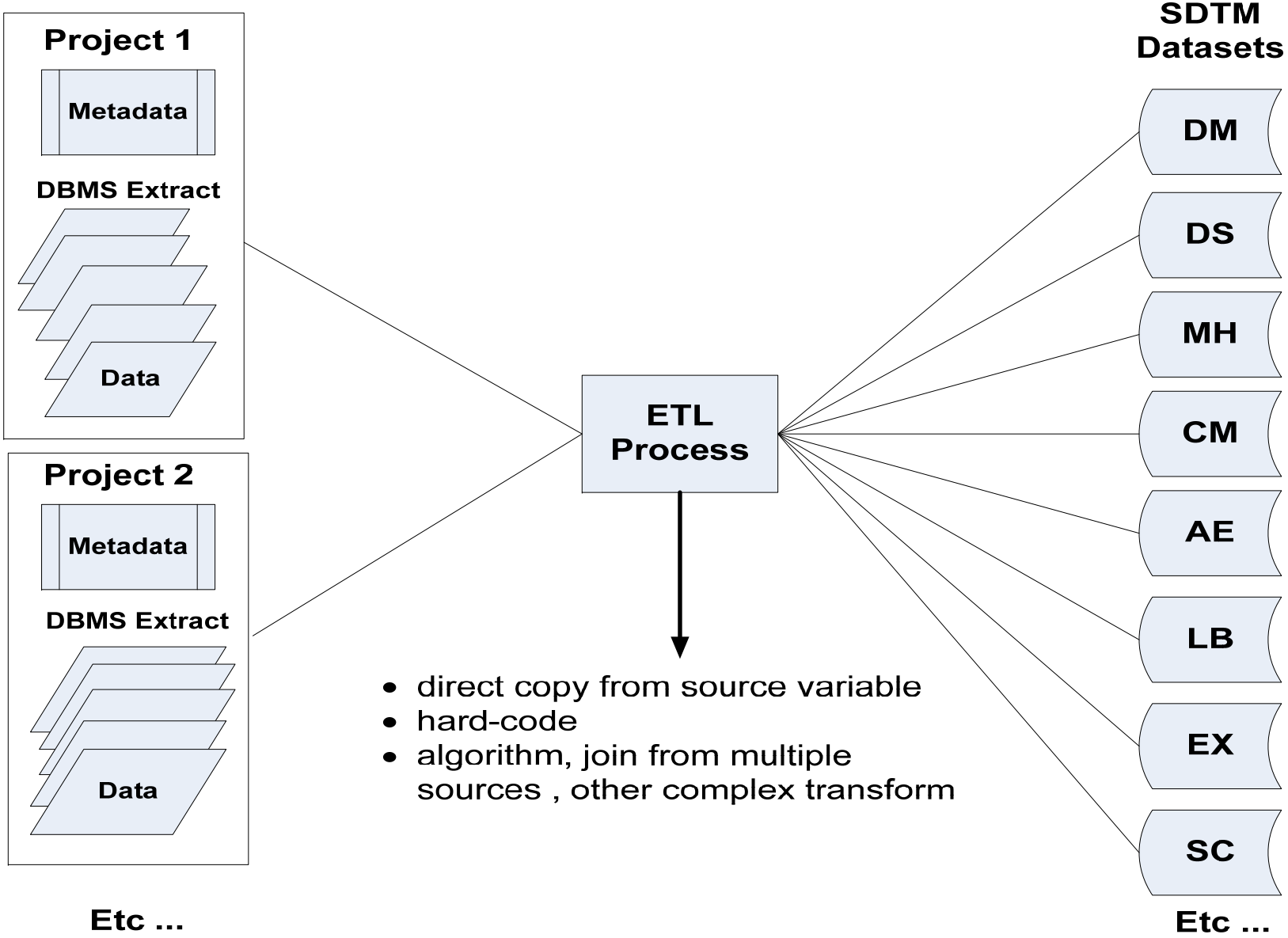
Nth Analytics

Process Without Automation



... etc.

ETL Process


Project 1
Metadata
DBMS Extract
Data
Project 2
Metadata
DBMS Extract
Data
Etc ...
ETL Process
direct copy from source variable
hard-code
algorithm, join from multiple sources , other complex transform
SDTM Datasets
DM
DS
MH
CM
AE
LB
EX
SC
Etc ...

Quality to the Nth Degree
Nth Analytics

Software

- Market leader : SAS Data Integration Studio 3.4
  - Formerly ETL Studio
  - $80K per server
- “Poor Man’s 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

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

Quality to the Nth Degree
Nth Analytics

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
  - In the CM file, variables CMSTRF and CMENRF have controlled terminology of “Before”, “During” and “After”
  - For missing or incomplete dates, need complicated algorithms to map these controlled terms

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

Quality to the Nth Degree

Nth Analytics

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