Legacy to SDTM Conversion Workshop: Tools and Techniques

Mike Todd
President
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
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
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

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Domain</td>
<td>Variable Name</td>
<td>Variable Label</td>
<td>Type</td>
<td>Origin</td>
<td>Role</td>
<td>Comments</td>
</tr>
<tr>
<td>2</td>
<td>DM</td>
<td>STUDYID</td>
<td>Study Identifier</td>
<td>Char</td>
<td>CRF</td>
<td>Identifier</td>
<td>[default]</td>
</tr>
<tr>
<td>3</td>
<td>DM</td>
<td>DOMAIN</td>
<td>Domain Abbreviation</td>
<td>Char</td>
<td>Derived</td>
<td>Identifier</td>
<td>[default]</td>
</tr>
<tr>
<td>4</td>
<td>DM</td>
<td>USUBJID</td>
<td>Unique Subject Identifier</td>
<td>Char</td>
<td>Sponsor Defined</td>
<td>Identifier</td>
<td>[default]</td>
</tr>
<tr>
<td>5</td>
<td>DM</td>
<td>SUBJID</td>
<td>Subject Identifier for the Study</td>
<td>Char</td>
<td>CRF</td>
<td>Topic</td>
<td>%USUBJID(VARNAME=SUBJID)</td>
</tr>
<tr>
<td>6</td>
<td>DM</td>
<td>RFSTDTC</td>
<td>Subject Reference Start Date/Time</td>
<td>Char</td>
<td>Sponsor Defined</td>
<td>Timing</td>
<td>%RFSTDTC</td>
</tr>
<tr>
<td>7</td>
<td>DM</td>
<td>RFENDTC</td>
<td>Subject Reference End Date/Time</td>
<td>Char</td>
<td>Sponsor Defined</td>
<td>Timing</td>
<td>%RFENDTC</td>
</tr>
<tr>
<td>8</td>
<td>DM</td>
<td>SITEID</td>
<td>Study Site Identifier</td>
<td>Char</td>
<td>Derived</td>
<td>Record Qualifier</td>
<td>SUBSTR(DEMOG.INVSITE,5,3)</td>
</tr>
<tr>
<td>9</td>
<td>DM</td>
<td>INVID</td>
<td>Investigator Identifier</td>
<td>Char</td>
<td>Derived</td>
<td>Record Qualifier</td>
<td>DEMOG.INV</td>
</tr>
<tr>
<td>10</td>
<td>DM</td>
<td>INVNAM</td>
<td>Investigator Name</td>
<td>Char</td>
<td>Derived</td>
<td>Synonym Qualifier</td>
<td>%INVNAM</td>
</tr>
<tr>
<td>11</td>
<td>DM</td>
<td>BRTHDTC</td>
<td>Date/Time of Birth</td>
<td>Char</td>
<td>CRF</td>
<td>Result Qualifier</td>
<td>%ISO_DATETIME(DATE=DEMOG.DMDOB DT, TIME=0)</td>
</tr>
<tr>
<td>12</td>
<td>DM</td>
<td>AGF</td>
<td>Age in AGEU at RFSTDTC</td>
<td>Num</td>
<td>Derived</td>
<td>Result Qualifier</td>
<td>%AGF</td>
</tr>
</tbody>
</table>
Process Without Automation

Project 1

Project 2

... etc.
ETL Process

Project 1
- Metadata
- DBMS Extract
- Data

Project 2
- Metadata
- DBMS Extract
- Data

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 ...

Etc ...
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
Data Flow Process

Data Manager
- Annotated CRF Raw Data
- SDTM Implementation Guide

Statistician
- Legacy Study Report
- Labor-Intensive

Raw Data
- Annotated CRF SDTM Data
- Metadata

Legacy Analysis Data
- ETL System
- DEFINE.XML
- SAS Macros
- Standalone SDTM Programs

Programmer
- SDTM Data
- SDTM Data
- DM SUPPDM
- CO SUPPAE
- AE Etc...

Validation Reports
- Automated

Data Flow Process Diagram
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.