Metadata-Driven Technology for Implementing CDISC SDTM

Michael Todd
President
Nth Analytics
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
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
Commercial ETL Product

• Commercial products
  – Obvious advantages:
    • Already built!
    • Validation is simplified
    • Vendor training available
  – Downside
    • Expensive!
  – Products
    • SAS Data Integration (DI) Studio
    • Oracle Warehouse Builder (OWB)
In-House ETL Product

• While in-house products have obvious downsides, they represent the only feasible solution for smaller companies, due to cost of commercial products

• Advantages:
  – Incremental development
  – Cost
  – Achieves same result as commercial products

• Disadvantages
  – Requires SDLC validation
  – Usual in-house development problems
ETL Transformation 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
  - Metadata for the SDTM files
  - Basis for the FDA-mandated “DEFINE.XML” documentation
Process Without Automation

Project 1

Data

Project Programs

Reports

Project 2

Data

Project Programs

Reports

... etc.
ETL Process

- SDTM variable = source variable
- Direct assignment for information not in the database
- Algorithm, join from multiple sources, other complex transform
Why it Works

• Role of standards
  – Standards drive the process. Target has standard structure so it makes standardizing tools economically worthwhile.
  – While source variables differ, there are certain commonalities that can be exploited

• Knowledge required
  – CDISC Standards
  – Understanding of raw data issues
  – Study design
  – Limited derivation
Elements

- Data
- People
- Dataflow
- Software
- Validation
# The Data

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADVEVNT Lab Normals</td>
<td>DEFINE.XML</td>
</tr>
<tr>
<td>DEMOG Protocol-specific dictionary</td>
<td>Trial Design Datasets TE, TA, TV, TI, TS</td>
</tr>
<tr>
<td>MEDHIST Abnormality Criteria</td>
<td>Subject Datasets SV, SE</td>
</tr>
<tr>
<td>PHYSEXAM Protocol</td>
<td>AE</td>
</tr>
<tr>
<td>INCLEXCL Annotated CRF</td>
<td>SUPPAE</td>
</tr>
<tr>
<td>PROTDEV Statistical Analysis Plan</td>
<td>CM</td>
</tr>
<tr>
<td>LABS</td>
<td>SUPPCM</td>
</tr>
<tr>
<td>VITSIGN</td>
<td>DM</td>
</tr>
<tr>
<td>CMED</td>
<td>DS</td>
</tr>
<tr>
<td>STUDYMED ... etc.</td>
<td>SUPPEX ... etc.</td>
</tr>
</tbody>
</table>

---

**Source**
- ADVEVNT: Lab Normals
- DEMOG: Protocol-specific dictionary
- MEDHIST: Abnormality Criteria
- PHYSEXAM: Protocol
- INCLEXCL: Annotated CRF
- PROTDEV: Statistical Analysis Plan
- LABS: CMED
- VITSIGN: SUPPAE
- CMED: SUPPCM
- STUDYMED: SUPPEX

**Target**
- DEFINE.XML
- Trial Design Datasets TE, TA, TV, TI, TS
- Subject Datasets SV, SE
- AE
- SUPPAE
- CM
- SUPPCM
- DM
- DS
- EX
- SUPPEX
# People: Tasks and Job Roles

<table>
<thead>
<tr>
<th>Task</th>
<th>Job Role</th>
<th>Requires</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development of the Annotated CRF</td>
<td>Mapping Specialist</td>
<td>Knowledge of SDTM V3.1.1 Implementation Guide</td>
</tr>
<tr>
<td>Development of mapping specifications</td>
<td>Statistician</td>
<td>Ability to translate abstract concepts into datasets</td>
</tr>
<tr>
<td>Create trial design datasets</td>
<td>Data Integration Specialist</td>
<td>SAS programming</td>
</tr>
<tr>
<td>Development of conversion jobs in ETL Environment</td>
<td></td>
<td>Knowledge of ETL tool</td>
</tr>
<tr>
<td>QC of the SDTM files</td>
<td>QC Specialist</td>
<td>All of the above</td>
</tr>
</tbody>
</table>
### Sample SDTM Metadata

<table>
<thead>
<tr>
<th>Domain</th>
<th>Variable Name</th>
<th>Variable Label</th>
<th>Type</th>
<th>Origin</th>
<th>Role</th>
<th>Comments</th>
<th>Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>STUDYID</td>
<td>Study Identifier</td>
<td>Char</td>
<td>CRF</td>
<td>Identifier</td>
<td>[default]</td>
<td>Req</td>
</tr>
<tr>
<td>DM</td>
<td>DOMAIN</td>
<td>Domain Abbreviation</td>
<td>Char</td>
<td>Derived</td>
<td>Identifier</td>
<td>[default]</td>
<td>Req</td>
</tr>
<tr>
<td>DM</td>
<td>USUBJID</td>
<td>Unique Subject Identifier</td>
<td>Char</td>
<td>Sponsor Defined</td>
<td>Identifier</td>
<td>[default]</td>
<td>Req</td>
</tr>
<tr>
<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>
<td>Req</td>
</tr>
<tr>
<td>DM</td>
<td>RFSTDTC</td>
<td>Subject Reference Start Date/Time</td>
<td>Char</td>
<td>Sponsor</td>
<td>Timing</td>
<td>%RFSTDTC</td>
<td>Exp</td>
</tr>
<tr>
<td>DM</td>
<td>RFENDTC</td>
<td>Subject Reference End Date/Time</td>
<td>Char</td>
<td>Sponsor</td>
<td>Timing</td>
<td>%RFENDTC</td>
<td>Exp</td>
</tr>
<tr>
<td>DM</td>
<td>SITEID</td>
<td>Study Site Identifier</td>
<td>Char</td>
<td>Derived</td>
<td>Qualifier</td>
<td>SUBSTR(DEMOG.INVSITE,5,3)</td>
<td>Req</td>
</tr>
<tr>
<td>DM</td>
<td>INVID</td>
<td>Investigator Identifier</td>
<td>Char</td>
<td>Derived</td>
<td>Qualifier</td>
<td>DEMOG.INV</td>
<td>Perm</td>
</tr>
<tr>
<td>DM</td>
<td>INVNAM</td>
<td>Investigator Name</td>
<td>Char</td>
<td>Derived</td>
<td>Qualifier</td>
<td>%INVNAM</td>
<td>Perm</td>
</tr>
<tr>
<td>DM</td>
<td>BRTHDTC</td>
<td>Date/Time of Birth</td>
<td>Char</td>
<td>CRF</td>
<td>Qualifier</td>
<td>%ISO_DATETIME(DATE=DEMOG.DMDOB DT, TIME=0)</td>
<td>Perm</td>
</tr>
<tr>
<td>DM</td>
<td>AGF</td>
<td>Age in AGEU at REFSTDTC</td>
<td>Num</td>
<td>Derived</td>
<td>Qualifier</td>
<td>%AGF</td>
<td>Exp</td>
</tr>
</tbody>
</table>
Dataflow

SDTM Data Annotated CRF

Mapping Specialist

ETL System

Statistician

Study-Level Metadata

Data Integration Specialist

Study-Level SAS Macros

Project-Level Metadata

Technical Lead

Project-Level SAS Macros

SDTM Datasets and SAS Programs

Trial Design Datasets: TE, TA, TV, TI, TS

SDTM Datasets

XML Stylesheet (from CDISC)

DEFINE.XML

List of raw data variables not mapped

Protocol

Statistical Analysis Plan

Raw Data Annotated CRF

Raw Data

SDTM V3.1.1 Guide

Technical Lead

Study-Level SAS Macros

Project-Level SAS Macros
Software

1. Read project-level metadata
2. Read protocol-level metadata
3. SDTM structure
4. SDTM Content: Mapping Rules
5. Raw Data Keys
6. Iso Date/Time Conversions
7. Compute --SEQ
8. Compute USUBJID
9. Compute ----DY variables
10. Compute --BLFL variables
11. AGE
12. RFSTDTC
13. RFENDTC
14. ARM
15. ARMCD
16. ... etc

Built-in Validation:
- Required variables have values
- Expected variables present
- USUBJID is in DM

- Apply variable order
- Standard sort order

- Generate SAS code to merge raw datasets required for domain
- Apply SAS programming statements in metadata
- Run SAS macros referenced in metadata
- SDTM Dataset
- Standalone SAS Program

SDTM Content:
- Mapping Rules
- Raw Data Keys

SDTM Structure:
- Run SAS macros referenced in metadata
- Apply SAS programming statements in metadata
- Generate SAS code to merge raw datasets required for domain

SDTM Dataset
- Standalone SAS Program

Project-Level Macros

14th Annual Meeting
Sources of Error

Given a validated system, there are still several sources of error in the process:

1. CRF SDTM Annotations
2. Trial design datasets
3. Metadata
4. SAS macros
CRF SDTM Annotations

• **Source of error:**
  – CRF annotations are not automated
  – Dependent on expert knowledge of SDTM Implementation Guide
  – System cannot use metadata to write annotations on the CRF

• **Solution requires:**
  – Better knowledge of XML
Trial Design Datasets

• Source of error:
  – Manual process
  – Depends on detail-oriented statistician with ability to translate study design into abstract concepts
  – System cannot read protocol to generate the trial design datasets

• Solution requires:
  – “Study Designer” tool to enable clinicians or statisticians to generate the trial design datasets
  – XML and machine-readable protocol
Metadata

• **Source of error:**
  – No automatic link between annotated CRF and metadata
  – System cannot read annotations from CRF to write metadata
  – Contains SAS programming statements

• **Solution requires:**
  – XML would enable system to read annotated CRF, but derivations would still be a source of error
SAS Macros

• Source of error due to:
  – Traditional SAS programming approach to handle complex derivations

• Solution requires:
  – Error rate can be minimized, but not eliminated, through good programming practices
  – Use of Stored Processes (black boxes) without access to source code
ETL System: 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.XML
  – 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
Summary

• Metadata is the key for successful automation
• Off-the-shelf tools can provide a powerful ETL solution