Octagon Research Solutions, Inc.

Using SAS ETL Studio to Create a CDISC-Compliant Clinical Data Warehouse

Barry Cohen

PhilaSUG Spring Meeting
June 15, 2005

Background

Octagon Research Solutions

(1) E-publication of NDA documents

(2) CDISC/SDTM expertise ➔ Clinical Data Strategies dept:

- **Migrate data to SDTM**
- **Clinical Data Warehouse & strategic uses**

- **Earlier w/o ETL Studio, now with it**
- **Building and refining processes**
- **Topical, hence present some thoughts**

Overview

- CDISC/SDTM background
- Migration scenarios
- Migration process
- Two approaches to migration
- Why use ETL Studio?

Octagon Research Solutions, Inc.

CDISC / SDTM

Octagon Research Solutions, Inc.

CDISC SDTM (1)

- CDISC – Clinical Data Interchange Standards Consortium
- SDTM – Study Data Tabulation Model (the std)
- FDA issued guidance for SDTM use
- FDA desires SDTM data. Future: require SDTM
- FDA developing tools to process SDTM data, and build warehouse on this standard

CDISC SDTM (2)

- No tutorial about SDTM today
- Today: Process of migrating data to SDTM, and use of ETL Studio in the process
- SDTM info and training resources available through CDISC…
- Published info on web site
- CDISC-sponsored training by certified orgs
- Octagon Research is certified

Octagon Research Solutions, Inc.

Migration Scenarios

Migration Scenarios (1)

From what and where, to what (SDTM) and when

- **Many scenarios, based on:**
  - **Data mgmt history**
  - **Data to be used for reporting & analysis**

<u>Multiple in-house structures</u>

- **No in-house standard – all dif. strucs**
- **In-house standard new – mix std and dif. strucs**
- **Work done by CROs using own standards**
- **“Many-to-one” migration**

Migration Scenarios (2)

Multiple in-house structures

- **Migrate to SDTM early:**
  - **SDTM for reporting and analysis, per study**
  - **SDTM to integrate data for ISS/ISE**
  - **SDTM for submission**
- **Migrate to SDTM late:**
  - **Various struc for reporting & analysis, per study**
  - **Ad hoc integrate various struc for ISS/ISE**
  - **(or SDTM for ISS/ISE)**
  - **SDTM for submission**

Migration Scenarios (3)

Single in-house structure

- **Stable standard used in-house**
- **CROs deliver in standard**
- **“One-to-one” migration**
- **May have some “multiple-to-one” cases, too**

Migration Scenarios (4)

- **Migrate to SDTM early:**
  - **OperDB for collect/clean/manage**
  - **SDTM for rpting and analysis, per CSR**
  - **SDTM to integrate data for ISS/ISE rpting and analysis**
  - **SDTM for submission**
- **Migrate to SDTM late:**
  - **OperDB for collect/clean/manage**
  - **OperDB for rpting and analysis, per CSR**
  - **OperDB to integrate data for ISS/ISE rpting and analysis**
  - **SDTM for submission**

Migration Scenarios (5)

“Three-Levels” Scenario

- Legacy data in various structures
- Building in-house std now (prospective use)
- Migrate legacy data to in-house std
  - For current subm. (CSR, ISS)
  - For future Clinical Data Warehouse
- Analysis and reporting from in-house std
- Migrate from in-house std to SDTM for subm.
- “Many-to-one-to-one” migration

Migration Process

Migration Process (1)

Process: From source to target

Source

- **Clinical data in DB (Oracle, SAS)**
- **Multiple tables/datasets – one per “domain”**
- **Data sets today**

Target

- **Same**
- **Individual files, or integrated in warehouse**

Migration Process (2)

Top-level steps

- Build and maintain metadata for the target (std)
- Build metadata for each source
- Map source to target - specs
- Develop programs as per specs
- Execute programs
- Integrate std data in warehouse

Migration Process (3)

Build and maintain metadata for the target (std)

- Tables and structure
- Items and attributes (name, label, type, length)
- Item controlled content (terminology)
- Items display format
- Study CRFs annotated to SDTM

Migration Process (4)

Build metadata for *each* source

- **Same pieces as for target (except controlled terms)**
- **Study CRFs annotated to source**

Migration Process (5)

Map source to target - specs

- **“what goes to what and how (rules)” – not pgms**

Transformations:

- **One-to-one, no other transform**
- **To single table, single item, w transform**
- **To single table, multiple items**
- **To multiple tables, multiple items**

Migration Process (6)

Develop programs as per specs

- Programming task
- Programs may be written, else auto generated by ETL tool

Execute Programs/Job

- Test and Production
- Error check / resolve
- Manage production jobs (logs, audits)

Migration Process (7)

Integrate std data in warehouse

- **Another (final) step in the migration process. The Load step.**
- **Immediate: For ISS/ISE reporting, if reporting from SDTM data ("Submission-level")**
- **Long-term: Multiple studies, subm's, and compounds stored together for analysis and reporting ("Company-level")**

Two Approaches

- **Build process/programs w/o ETL tool**
- **Build process/programs w ETL tool (ETL-Studio)**

Without ETL Tool

Without ETL Tool (1)

Build and maintain metadata for SDTM (target)

- In Excel SS's or SAS data sets or other
- CDISC provides SDTM metadata in Excel SS
- Annotate study CRFs to SDTM

Build metadata for each source

- In Excel SS's or SAS data sets or other
- Annotate study CRFs to source

Without ETL Tool (2)

Map source to target - specs

- **Source-centric SS's - one row per source item**
- **SrcDataSet.Varname to SdtmDataSet.Varname, plus "how" rules**
- **Done by clinical/SDTM-aware person**

Resources:

- **Annotated CRFs (to source, to target)**
- **Metadata SS's**
- **SDTM Implementation Guide**

Without ETL Tool (3)

Develop programs as per specs

- **SAS programmer, using map specs SS**

Levels of sophistication…

- **Fully ad hoc**
- **Table-driven, auto-generated program, for "one-to-one" process, using:**
  - **Map specs SS; Source & target metadata (Excel, SAS)**
- **Macros for common transformations referenced in map specs SS**

Without ETL Tool (4)

Levels of sophistication…

- **Table-driven full job, not just for indv ("one-to-one") processes:**
  - **e.g., transpose, sort, validate, load**

Notes:

- **Plus ad hoc code for remainder of program(s)**
- **May be building robust application**

Without ETL Tool (5)

Execute programs/job

- **Controlled by programmer, unless robust application developed**

Integrate std data in warehouse

- **Last program in the process**
- **Submission-level, company-level**

With ETL Studio

With ETL Studio (1)

Build and maintain metadata for SDTM (target)

- Stored in the target metadata library
- CDISC provides SDTM metadata in Excel SS
- SAS provides a load process
- Annotate study CRFs to SDTM

Build metadata for *each* source

- Stored in the source metadata library
- Build by loading the source SAS dataset
- Annotate study CRFs to source

With ETL Studio (2)

Map source to target - specs

- Same process as w/o ETL Studio
- Map specs Excel SS's
- Done by clinical/SDTM-aware person

Resources:

- Annotated CRFs (to source, to target)
- Metadata SS's
- SDTM Implementation Guide

With ETL Studio (3)

Develop programs as per specs

- **Work shared by ETL developer and SAS pgmr**

ETL Developer

- **Job comprised of processes**
- **Process Design wizard for full job**
- **Process Properties wndw to define ea. process**
- **Process ~ SAS data step(s)**
- **Code for job is auto-generated**
- **Uses map specs SS (CRFs as needed)**

With ETL Studio (4)

Process might have:

- **Map, Transform, Sort, Transpose, Join, Validate**
- **Auto-generated code, or user-written code**

SAS programmer:

- **Develop code and load to ETL-S, to be engaged by ETL developer**
  - **Ad hoc: one-time transformations**
  - **Macros: repeating transformations**

ETL Studio Screen Shots

With ETL Studio (6)

Execute programs/process

- Run by ETL developer
- Using ETL Studio job scheduler
- Define method of error reporting (rpt, table, msg)

Integrate std data in warehouse

- Last ad hoc program in the process
- Submission-level, company-level

Why Use ETL Studio?

Why use ETL-Studio? (1)

Central metadata – uses SAS Metadata Server

- SDTM metadata
- Other metadata (eg, libraries, servers)
- Available for all processes in larger workflow
- Available for all SAS products in larger workflow
  - Enterprise Guide
  - Web Report Studio
- Valuable for Clinical Data Warehouse analysis and reporting

Why use ETL-Studio? (2)

Programmer Perspective

- Code generation, minimize code written
- Metadata management
- Documentation of indv processes and full job

Flexible Staffing

- (Clinical/SDTM-aware staff does map specs SS)
- ETL Developer builds and executes jobs
- SAS programmer generates utility programs
  - Other code by ETL-S; Not managing jobs

Why use ETL-Studio? (3)

Change Management

- For collaborative work
- Valuable: Large projects ➔ Multiple developers ➔ Update the same resources

Project Management

- Organized, structured, managed env. for data, programs, and processes

Why use ETL-Studio? (4)

Impact Analysis

- Easy to assess change
- Valuable: Migrations can be large projects

Deploy Jobs to SAS Drug Development

- Ability to do this now
- Improvement planned

Why use ETL-Studio? (5)

SDTM Metadata loader

- Upon request

Easy Re-use of Work

- Process and job metadata
- Leverage workflow for:
  - New sources
  - New SDTM domains
  - Startpoint for in-house standard

Questions

Contact Information

Barry Cohen

(bcohen@octagonresearch.com)

610.535.6500 x635

585 East Swedesford Road, Suite 200

Wayne, PA 19087

Octagon Research Solutions, Inc.