PhUSE 2012 DH07

Using Metadata to Integrate and Standardise Clinical Trial Data: The ICON Study Data Mapper
Existing situation

- Manual processes: High level of effort to specify, program and deliver data
- ETL custom written from scratch for each study
- Limited reuse of data deliveries within a study
- Limited reuse of programming across studies
- No single source of clinical data

Oracle® Life Sciences Data Hub already in place

- Implemented in ICON Data Management in 2010
Project Overview

Project goals

– Build on the strengths of LSH (global repository of clinical data, program and data traceability)
– Replace multiple data flows with single, streamlined flow using standard data models
– Reengineer process for data specification and programming
– Facilitate reuse of code in other ICON departments (e.g., DM)

Three areas of focus

– Standard data flow
– Standard data models
– Standard tools
New Dataflow 1: Centralise

- Diverse sources with little standardisation
- Permanent staging area with full audit history of “raw” data
New Dataflow 2: Standardise

- Standardised access point for all study clinical data
- Standard clinical data model is an SDTM-like structure
New Dataflow 3: Deliver

- Distribute conformed data to standard delivery models
- Re-structured targeted copies of data for data consumers
Study Sources to Clinical Standard

- Involves a high level of study-specific ETL
- Limited use of standard code
Clinical Standard to Delivery Standards

- Majority of ETL is standard library code
- Little or no study-specific ETL
Challenges Arising from This Approach

Data standardisation work is done “up front”
- Most of the standardisation happens between source and the standard clinical model

Data standardisation must happen earlier in the life of study
- Users are reliant on the data being standardised to access new standard reports, etc.

Manual ETL development is costly and time consuming
Transform specifications are already required for traceability

- Describe related structures (source and target)
- Specify rules for transforming data between sources and targets

Use this asset as a source for metadata

Prerequisites for ETL derivation

- Complete, detailed specification of data structures
- Predictable approach to defining transformation rules
- Machine readable spec
Study Data Mapper is a tool to
  – Describe the metadata describing the tables, variables and value lists for sources and targets
  – Specify how those source tables will be transformed into the target table structures
  – Generate SAS code based on the recorded metadata

In-house software application
  – Developed in partnership with BioPharm Systems
  – Phase 1 delivered in March 2012, Phase 2 currently rolling out
  – Development ongoing...

Number of alternative tools available
Components

- A web based user interface for managing mapping projects
- An integrated Excel spreadsheet for specifying metadata structures and transformation maps
- An Oracle database for securely storing the metadata repository
- An ETL generator that produces SAS PROC SQL code
Project Setup

1. Create metadata for sources from LSH
2. Create mapping project and import sources metadata
3. Copy chosen target standard from repository
4. Save project to repository and generate starting specification
Study and mapping project metadata

---Mapping Project
- Project Name: A001_HUB
- Project Version: 3
- Description: Map study A001 to Clinical standard model (HUB)

---Study Details
- Sponsor: ACME
- Program: ABC
- Study code: A001
- Protocol: PRO_A
- EDC Study Location: ACME>Oncology>A001.DATA>HUB

---Document Details (Replace with version number/date information)
- Authored by: 
- Approved by: 
- Current version: 
- Date of completion: 
- Date of approval: 

Screenshot from specification template
### Table Metadata

#### Table set metadata

<table>
<thead>
<tr>
<th>Tableset Name</th>
<th>Author</th>
<th>Source System</th>
<th>Source Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDC</td>
<td>Paul Crean</td>
<td>RAVE</td>
<td>ACME&gt;Oncology&gt;A001&gt;DATA&gt;EDC Source</td>
</tr>
<tr>
<td>EXT</td>
<td>Paul Crean</td>
<td>LSH</td>
<td>ACME&gt;Oncology&gt;A001&gt;DATA&gt;External Source</td>
</tr>
</tbody>
</table>

#### Target Tablesets

<table>
<thead>
<tr>
<th>Tableset Name</th>
<th>Author</th>
<th>Source System</th>
<th>Source Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGET</td>
<td>Standards</td>
<td>Study Data Mapper</td>
<td>Standards Library</td>
</tr>
<tr>
<td>WORK</td>
<td>Mapper</td>
<td>Temp</td>
<td>Temp</td>
</tr>
</tbody>
</table>

### Table metadata

#### SOURCE Tables

<table>
<thead>
<tr>
<th>Order</th>
<th>TablesetName</th>
<th>Name</th>
<th>ShortName</th>
<th>IsRepeating</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EDC</td>
<td>AE</td>
<td>AE</td>
<td>N</td>
<td>Adverse Events</td>
<td>Adverse Events</td>
</tr>
<tr>
<td>2</td>
<td>EXT</td>
<td>ADVEC</td>
<td>ADVEC</td>
<td>N</td>
<td>Coded Terms AE</td>
<td>Adverse Event Coded Terms</td>
</tr>
</tbody>
</table>

#### Target Tables

<table>
<thead>
<tr>
<th>Order</th>
<th>TablesetName</th>
<th>Name</th>
<th>ShortName</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TARGET</td>
<td>STD_AE</td>
<td>AE</td>
<td>STD_AE</td>
<td>One record per adverse event per subject.</td>
</tr>
<tr>
<td>2</td>
<td>WORK</td>
<td>RFSTDTC</td>
<td>RFSTDTC</td>
<td>RFSTDTC</td>
<td>One record per subject</td>
</tr>
</tbody>
</table>
Variable Metadata

Variable metadata

---Source Variables---

<table>
<thead>
<tr>
<th>Order</th>
<th>TableName</th>
<th>VariableName</th>
<th>ShortName</th>
<th>Pkey</th>
<th>Label</th>
<th>DataType</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EDC.AE</td>
<td>PROJECTID</td>
<td>projec00</td>
<td></td>
<td>projectid</td>
<td>Num</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>EDC.AE</td>
<td>PROJECT</td>
<td>project</td>
<td></td>
<td>project</td>
<td>Char</td>
<td>255</td>
</tr>
<tr>
<td>3</td>
<td>EDC.AE</td>
<td>STUDYID</td>
<td>studyid</td>
<td></td>
<td>Internal id for the study</td>
<td>Num</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>EDC.AE</td>
<td>ENVIRONMENTNAME</td>
<td>enviro00</td>
<td></td>
<td>Environment</td>
<td>Char</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>EDC.AE</td>
<td>SUBJECTID</td>
<td>subjco00</td>
<td></td>
<td>Internal id for the subject</td>
<td>Num</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>EDC.AE</td>
<td>STUDYSITEID</td>
<td>StudyS00</td>
<td></td>
<td>Internal id for study site</td>
<td>Num</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>EDC.AE</td>
<td>SUBJECT</td>
<td>Subject</td>
<td></td>
<td>Subject name or identifier</td>
<td>Char</td>
<td>50</td>
</tr>
</tbody>
</table>

---Target Variables---

<table>
<thead>
<tr>
<th>Order</th>
<th>TableName</th>
<th>VariableName</th>
<th>ShortName</th>
<th>Label</th>
<th>DataType</th>
<th>Length</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TARGET.STD_AE</td>
<td>STUDYID</td>
<td>STUDYID</td>
<td>Study Identifier</td>
<td>Char</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TARGET.STD_AE</td>
<td>DOMAIN</td>
<td>DOMAIN</td>
<td>Domain Abbreviation</td>
<td>Char</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>TARGET.STD_AE</td>
<td>USUBJID</td>
<td>USUBJID</td>
<td>Unique Subject Identifier</td>
<td>Char</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>TARGET.STD_AE</td>
<td>AESEQ</td>
<td>AESEQ</td>
<td>Sequence Number</td>
<td>Num</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>TARGET.STD_AE</td>
<td>AEGRPID</td>
<td>AEGRPID</td>
<td>Group ID</td>
<td>Char</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>TARGET.STD_AE</td>
<td>AEREFID</td>
<td>AEREFID</td>
<td>Reference ID</td>
<td>Char</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>TARGET.STD_AE</td>
<td>AESPID</td>
<td>AESPID</td>
<td>Sponsor-Defined Identifier</td>
<td>Char</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>TARGET.STD_AE</td>
<td>AETERM</td>
<td>AETERM</td>
<td>Reported Term for the Adverse Event</td>
<td>Char</td>
<td>200</td>
<td></td>
</tr>
</tbody>
</table>

- Including SAS format, SDTM flag, supplemental qualifier flag
# Value List Metadata

## Value List metadata

### Source Value List Values

<table>
<thead>
<tr>
<th>SourceValueListName</th>
<th>Start Value</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDC.AESEV</td>
<td>Grade 1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>EDC.AESEV</td>
<td>Grade 2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>EDC.AESEV</td>
<td>Grade 3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>EDC.AESEV</td>
<td>Grade 4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>EDC.AESEV</td>
<td>Grade 5</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

### Target Value List Values

<table>
<thead>
<tr>
<th>TargetValueListName</th>
<th>Start Value</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGET.AESEV</td>
<td>1</td>
<td>MILD</td>
<td>Grade 1; 1</td>
</tr>
<tr>
<td>TARGET.AESEV</td>
<td>2</td>
<td>MODERATE</td>
<td>Grade 2; 2</td>
</tr>
<tr>
<td>TARGET.AESEV</td>
<td>3</td>
<td>SEVERE</td>
<td>Grade 3; 3</td>
</tr>
</tbody>
</table>
Mapping Process Phase 2: Project Mapping

**Project Mapping**

1. Update mapping specification
2. Parse and save specification to repository
3. Generate and export code
4. Execute and extend code with custom mapping operations
Table Mapping

- Supports horizontal and vertical mappings
- Temporary target tables and variables supported
- SQL JOIN completed by mapping author

Source to Target Mapping

<table>
<thead>
<tr>
<th>Target Table</th>
<th>Source Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGET.STD_AE</td>
<td>EDC.AE</td>
</tr>
<tr>
<td>TARGET.STD_AE</td>
<td>TARGET.STD_DM</td>
</tr>
<tr>
<td>TARGET.STD_AE</td>
<td>EXT.ADVEC</td>
</tr>
</tbody>
</table>

Conditions

Join:
FROM EDC.AE AE
LEFT JOIN EXT.ADVEC ADVEC ON PUT(AE.RECORDID,8,) = ADVEC.EXTERNAL_ID
LEFT JOIN TARGET.STD_DM STD_DM ON AE.SUBJECT = STD_DM.SRC_SUBJID

Filter:
Variable Mapping

- **1.** Simple assignments
- **2.** SAS functions
- **3.** Global mapping macros
- **4.** Target variables used as a source within the same table
- **5.** Study specific mapping macros
PROC SQL
- Generated code creates and populates the target structures
- Each table map is generated separately and called in the order that takes care of dependencies

Portable code
- No dependency on Study Data Mapper, metadata or LSH
- Can be distributed to sponsors or regulatory authorities with a compiled macro catalog of mapping functions
- Clear, readable code: only basic PROC SQL knowledge required

Migration to LSH (including targets and ETL)
- Currently a manual process that takes about 20 minutes
/*
SAS Code Generated By Study Data Mapper

Mapping Project: A001_HUB
Map Set: A001_HUB version 1
Generated By: ICON-EU\creanp
Database: SDMDEV
Date: 2012-10-10 17:48:28 +01:00
Generation ID: 254294
*/

/* Processing table TARGET_STD_AE. */
%put NOTE: Processing TARGET_STD_AE. ;
%setLibname(tsnme=EDC, sponsor=ACME, study=A001);
%setLibname(tsnme=EXT, sponsor=ACME, study=A001);
%setLibname(tsnme=TARGET, sponsor=ACME, study=A001);
PROC SQL;
CREATE TABLE TARGET_STD_AE(label="STD_AE"
  ,STUDYID VARCHAR(15) LABEL="Study Identifier"
  ,USUBJID VARCHAR(25) LABEL="Unique Subject Identifier"
  ,AEBODSYS VARCHAR(67) LABEL="Body System or Organ Class"
  ,AESTDTC VARCHAR(19) LABEL="Start Date/Time of Adverse Event"
  ,AEENDTC VARCHAR(19) LABEL="End Date/Time of Adverse Event"
  ,AESTDY NUMERIC(8,0) LABEL="Study Day of Start of Adverse Event"
  ,AEENDY NUMERIC(8,0) LABEL="Study Day of End of Adverse Event"
  ,AEDUR VARCHAR(12) LABEL="Duration of Adverse Event"
);
INSERT INTO TARGET.STD_AE

{ STUDYID , USUBJID , AEBODSYS , AESTDTC , AEENDTC , AESTDY , AEENDY , AEDUR }

SELECT AE.PROJECT AS STUDYID , CATX('-', AE.PROJECT, AE.SITENUMBER, AE.SUBJECT) AS USUBJID , %SDMUPCSTR(ADVEC.LEVEL_1_TEXT) AS AEBODSYS , %SDMCREATEISODATE_RAVE(AE.AESTDAT) AS AESTDTC , %SDMCREATEISODATE_RAVE(AE.AEENDAT) AS AEENDTC , %SDMDERIVESTUDYD(%SDMCREATEISODATE_RAVE(AE.AESTDAT), STD_DM.RFSTDTC) AS AESTDY , %SDMDERIVESTUDYD(%SDMCREATEISODATE_RAVE(AE.AEENDAT), STD_DM.RFSTDTC) AS AEENDY , %DERIVEDURAO001(%SDMCREATEISODATE_RAVE(AE.AEENDAT), %SDMCREATEISODATE_RAVE(AE.AESTDAT)) AS AEDUR FROM EDC.AE AE
LEFT JOIN EXT.ADVEC ADVEC ON PUT(AE.RECORDID,8) = ADVEC.EXTERNAL_ID
LEFT JOIN TARGET.STD_DM STD_DM ON AE.SUBJECT = STD_DM.SRC_SUBJID
;
ICON Study Data Mapper: Repository

Mapping Projects
- Study A001 RAW to Clinical Std
- Study A002 RAW to SDTM plus
- Pool A001 & A002 to SDTM plus

ICON Standard Models
- Standard Clinical Model
- Visualisation Models
- SDTM plus (3.1.2)

Sponsor Standard Models
- Sponsor A Delivery Model
- Sponsor B SDTM plus

Study Data Mapper Library

Project Types
- Data model projects (ICON standard, sponsor standard...)
- Mapping projects (standard-to-standard, raw-to-standard...)
- All projects are version controlled
Conclusions

Metadata is not a silver bullet

- The use of metadata has reduced programming effort on transforming studies by 90%
- BUT, it hasn’t reduced the time it takes to transform a study

Why?

- We’re standardising and delivering more data to a wider audience
- Programmers take 10-12 weeks to gain expertise in the new tools and processes
- A higher quality specification process takes longer
- Not enough reuse of metadata

More standards reuse required

- Metadata reuse depends on standards in the source systems
Contact Details

Paul Crean
Clinical Data Warehouse Mgr, ICONIK
ICON Clinical Research
Email: paul.crean@iconplc.com
Web: www.iconplc.com

Jennie McGuirk
Manager Clinical Data Delivery, ICONIK
Email: jennie.mcguirk@iconplc.com