Evolution Towards a Metadata Driven Clinical Data Repository -

The Sanofi-Aventis Approach
In the beginning...

- When CDISC civilization first started in our company

- We created SDTM domains,

- and analysis domains
Incompleteness of SDTM Domains

SDTM

- Mostly CRF variables
- Mostly character variables
- Date variables are ISO datetime
- Any derived variables not in the model must go in suppqual
- Mostly vertical data structures

Not useful for analyses
Incompleteness of Our Original Analysis Datasets

Analysis Datasets

- Mostly derived variables
- CRF variables mostly processed (imputed)
- Lack of standardization

- Hard to trace derived variables back to source
- Lack of standardization made developing macros and programming conventions difficult
Redundancy

- “Variables in SDTM are redundant to the analysis datasets!”
- “Variables in the analysis datasets are redundant to SDTM!”
Working on a Solution

- To resolve redundancy issue and stand-alone incompleteness of the SDTM and analysis datasets

  - SDTM and analysis datasets were integrated
Integrated Clinical Metadata

❖ **SDTM component**
  ➢ All CRF variables
  ➢ SDTM assumptions, definitions, naming conventions, attributes, and controlled terminology
  ➢ SDTM standard derived variables

❖ **Analysis dataset model component**
  ➢ Standard derived analysis variables
  ➢ Standard assumptions, definitions, naming conventions, attributes, and controlled terminology
    • Decode and code (ex. SEX = F SEXN = 1)
    • Detailed algorithms and programming conventions
  ➢ Common variables

❖ **Both components created from raw data in same step**
Integrated Standard Domain Components

- Raw ClinTrial Data
- RAW SAS datasets
- SDTM Variables
- Analysis variables
- Domains
- Supp Qual Domains
- CSR
- FDA
- **Used as our analysis datasets (ADS)**
  - Used to program TLGs
  - Submitted to FDA as analysis datasets

- **Used to create the SDTM dataset (SDS)**
  - The SDS dataset is a subset of the variables and records of the ADS
  - Supplemental qualifier variables, which remain horizontal in the ADS, are split out to the vertical suppqual dataset
Shortcomings

- Some non-standard domains are very time-consuming to map to SDTM, so if a decision is made not to submit the study, the added riggers may not have been necessary.

- The vertical design of the SDTM does not lend itself to some analyses like a horizontal structure would.

- Lack of automation made this time-consuming and risked being these datasets being inconsistently created.
The Process Flow

Note: The color pink stands for ADS and are used for reporting. The dotted line means when needed.
The Process Flow Detailed

- Standard datasets
  - Data that fits into SDTMIG

- Non-standard datasets
  - Not in SDTMIG
  - May retain structure of raw data if makes easier to program off
  - To be remapped to SDTM before submission
  - Follow as much of SDTMIG conventions as possible

- Additional analysis datasets
  - Summary level (1 proc away) for analysis purposes only
  - No part is used to create SDTM domains

- SDTM datasets
  - Created only for submission
  - Only collected plus required and expected variables and a few derived
  - Created from ADS/SDS (remapping and structural change, if necessary, for non-standard)
Standards

- Standard integrated ADS/SDS metadata
- Standard controlled terminology
- Standard statistical programming conventions
- Standard CRFs design
- Standard data management conventions
- Standard statistical output TLGs

Tools

- Metadata Management and Publishing Tool
- Submission & Statistical Analysis Dataset Creation Macro Library
- Global Statistical Analysis and Reporting System Macro Library
- Template programs using reporting system to produce TLGs
Metadata Management and Publishing Tool (MDMAP)

- Repository for ADS/SDS metadata dictionary

- Used to manage the metadata for a study
  - Maintains domain, variable, and value level attributes and controlled terminology
  - Imports metadata directly from domain SAS datasets
  - Reconciles metadata with the domain SAS dataset data to ensure the metadata matches the data exactly

- Annotates CRFs
  - Creates Define.xml and/or Define.pdf
  - Creates hyperlinks to annotated CRF and supplemental data definitions

- Has audit trail
Submission and Statistical Analysis Dataset Creation Macro Library (SuSADS)

- Used to map raw data and derive/format variables according to the ADS/SDS metadata and statistical programming conventions

- User modifies data specification templates (Excel files) which define source, target, and attributes for all variables in each domain

- Macro calls:
  - Map the data based on these specification files
  - Use look-up tables to apply derivation algorithms
  - Derive/format variables
  - Perform checks on the data
  - Can “break off” SDTM datasets (with associated supplemental qualifier datasets) from ADS/SDS datasets
  - Has audit trail
Produces standard and non-standard statistical output TLGs

Is not dependant on standards – can work with any data

Has audit trail
Template Programs using Reporting System to Produce TLGs

- Produces standard statistical output TLGs using the Global Statistical Analysis and Reporting System Macro calls
- Is dependant on standard variable names and attributes
Putting It All Together

- Model
- Process Flow
- Standards
- Tools
How It All Fits Together
Metadata dictionary repository (within MDMAP) does not yet store the information on the raw data source variables and, if data specification templates of the submission dataset creation process (SuSADS) which do have this information are imported into MDMAP, the metadata must be maintained in two places from that point on.

There is not yet any centralized oversight, so non-standard SDTM domains and derived variables are not uniformly created across the company.

The process does not work well for legacy data where the source data is not standardized.

The process is not completely automated since the raw data, although standardized, varies greatly for each study, so the programmer must tailor the process to the uniqueness of the study at each step.
Changing Submission Standards Challenges

- Evolving SDTM Challenges
  - Changing variable names and attributes
  - Changing assumptions
  - New controlled terminology
  - Added domains

- This requires changes to our clinical metadata definitions and standards which, in turn, requires trickle down changes in all other standards and tools

- It takes considerable time to implement these changes
Changing Submission Standards Challenges

❖ The ADaM vision challenges

➢ Analysis datasets should be created from the SDTM datasets

➢ The same named variables could exist in both ADaM and SDTM and have different definitions (example: as collected in SDTM and imputed in ADaM)

➢ The same variables could exist in both ADaM and SDTM and have different names

❖ This would require a major redesign of our model and tools