Map Metadata – Going Beyond the Obvious/Connecting the Dots

Gregory Steffens
Associate Director, Technology Innovations
Novartis Pharmaceuticals

Praveen Garg
Director, SAS Programming & Global Strategic Resourcing
ICON

Agenda

- Data Standard Journey
- Objectives of CDISC Standards
- Missed Opportunities
- Implementation Challenges
- Current State
- Next Steps
- RECAP

Period1: Inconsistency

- Where is the industry in evolution?

  Periods contain epochs contain ages.
- Epoch 1 of no data standards made it difficult to...
  - Pool data and do meta-analyses – the prime directive of CDISC and the FDA
  - Share program code across studies to get high quality with low effort
  - Be transparent about how data is collected, transformed, derived and reported
- Epoch 2 of Data Standards resident in documents and even early metadata (often in inadequate products like excel) did not adequately solve the problems. Compliance to standards and study specifications was difficult to implement and monitor. Lots of reentry of document-resident information into SAS programs. Sharing code was copy and tweak.

Period2: Corporate Consistenc*ies*

- Epoch 1 of robust and rigorously standardized "**database metadata**" (DBMD). Realization that standardizing metadata is just as important as standardizing data because it supports metaprogramming.
- Epoch 2 of "**database metaprogramming**" (DBMP) started showing the real value and started to solve the problems. Moved beyond copy/tweak to macro libraries shared across all studies.
  - Specs in metadata does not do any more than Spec in documents without metaprogramming
  - Metadata: a standard list of database attributes put in a standardized database structure to support metaprogramming. Metaprogramming blurs the line between data and code – populating metadata does what writing programming code used to do.
  - But this was done in each company without adequate standards for the industry or even across TAs within a company. Standards? Yes, we have many!

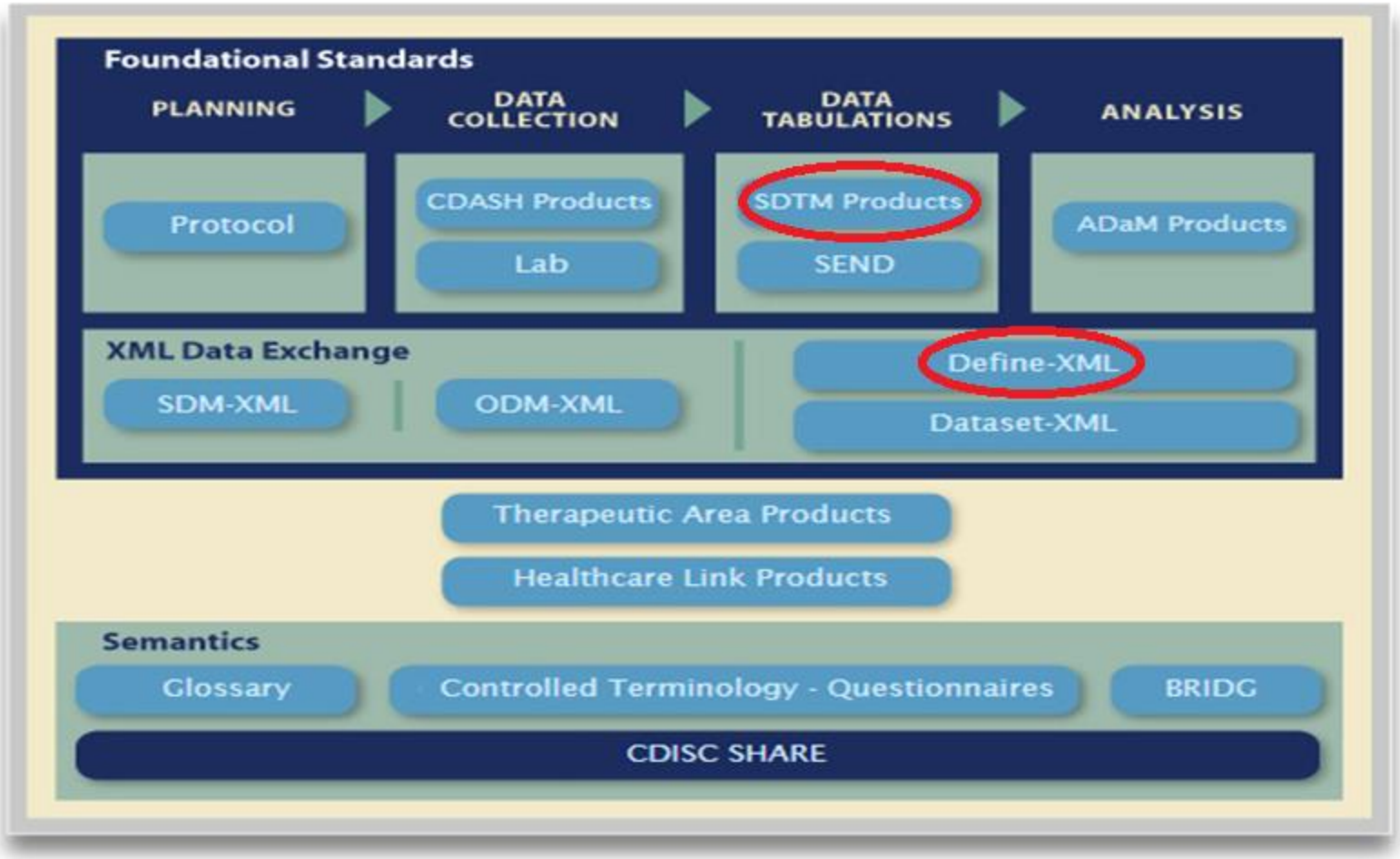
Period3: Industry-Level Consistency

- So every company had standards – this didn't help the FDA do data pooling across companies.
- Epoch 1 of Industry-Level Standardization of data and DBMD
  - CDISC defines data standards in SDTM and ADaM.
  - CDISC defines database metadata standards in the Define.xml to describe the data that doesn't support DBMP.
- Epoch 2 of Industry Level shared software modules and industry DBMP
  - Some work in industry-level database metaprogramming are seen from groups like Open CDISC, but we need more of this!
  - We all use PROC REPORT but why not macros for CREATE_SUPPQUAL, WIDE2THIN, CREATE_DECODES, etc. etc.?

Objective of CDISC standards

- Key Focus Areas
  - Safety signal detection (across companies)
  - Faster integration/pooling of data
  - Faster review by regulatory agencies
- Study Data Consistency
  - Standard data structure
  - Standard Controlled Terms
  - Consistent data handling approach
- Metadata (define.xml)
  - Traceability and Transparency
- May seem like period3, but has traces of each period.
  - The period 1 and 2 components caused implementation challenges

Standards & Implementations



Early Missed Opportunities

- Emphasis on eSubmission
  - Inadequate to support data automation
  - Metadata created retrospectively at the end of the study instead of prescriptively at the start of a study
  - Metadata was (mis)treated as being a document in a document process. Define.xml replaced documents in an unchanged process instead of improving the process.
- Define.xml
  - Data standards definition do not include all attributes required by the standard metadata, define.xml
  - Use of define.xml as document for review
  - Use of XML as industry metadata standard rather than relational metadata
  - Lack of tools to handle XML

Early Implementation Challenges

- Inadequate framework for defining CDISC data standards caused challenges due to
  - Informal metadata lacking consistency in excel “metadata” across CDISC standards
  - Required pharma to create more standardized cross-study metadata than existed in the standards
  - Required pharma to fill in the gaps between define.xml and the data standards
- Missing attributes of elements, such as the primary key seq
- SUPPQUALs as physical study data sets
- Traceability of data points from data state to state not clearly defined. Not all standards adopted as quickly as SDTM.

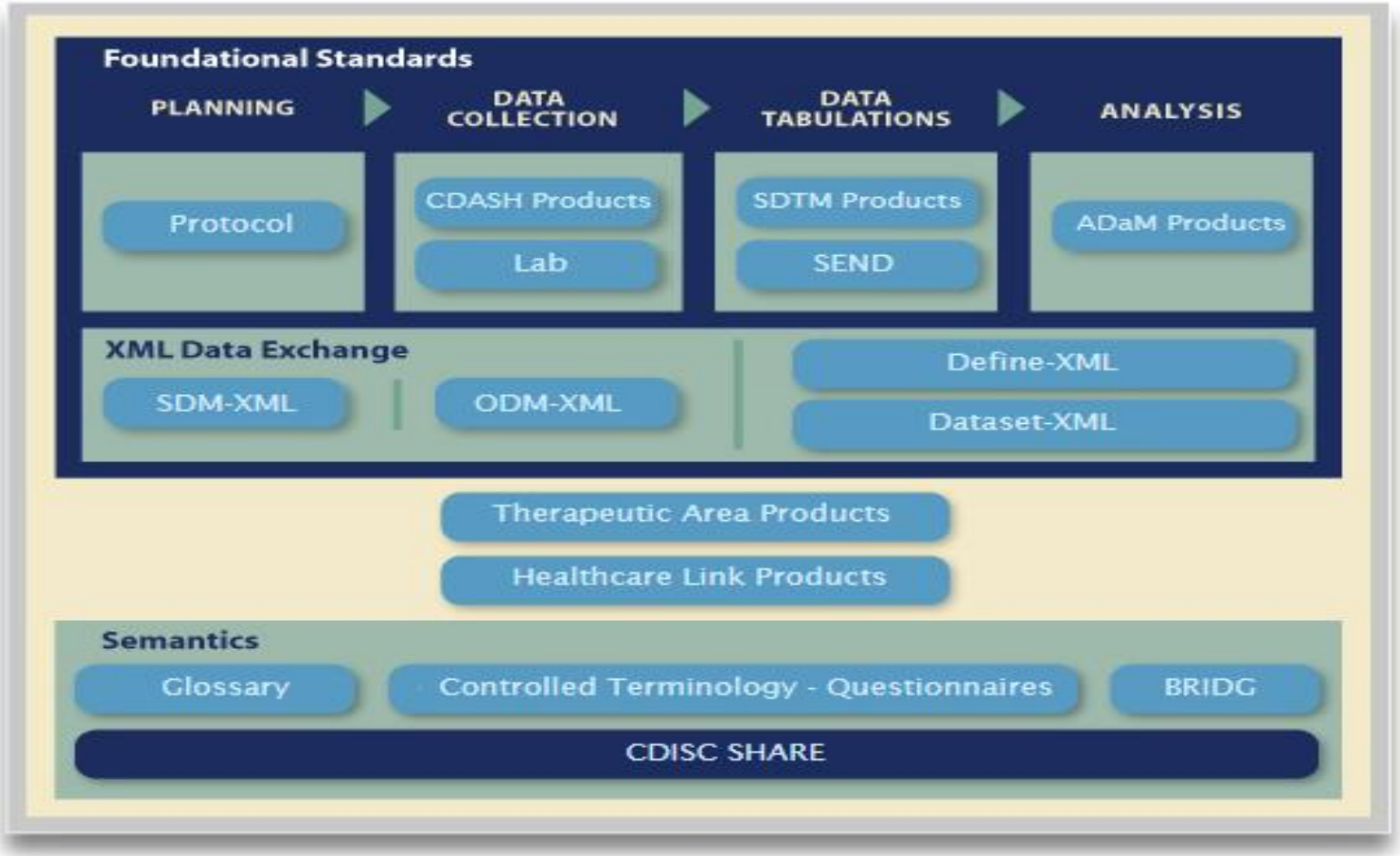
Early Implementation Challenges

- Definition of row-level metadata is required by tall-thin data
  - Initial assumption that TESTCD and ORRES met objective of pooling
  - Test code controlled terms
  - Correlation between test code and variable attributes
- Support programmatic access of standards
  - A single standard metadata design to define all data standards
  - Standard list of data attributes
- Big Pharma Implementation challenges
  - Version control of standards in PDF and excel formats
  - Internal infrastructure (not built on CDISC standards)
  - Maintenance of codelist by NCI – time lag between the use and availability in standards

Current State

- Emerging Technology
  - MDR solutions (SOA; Entimo; Formedix; Onedata; OpenCDISC)
  - Industry level software modules (beginning – focus on define generator is a first baby step)
- Higher CDISC Standards adoption level
  - Pharma companies
  - FDA (availability of better tool and push towards using standards)
  - Non-US regulatory agencies
  - Standards use beyond SDTM and define.xml
- Increased Level of outsourcing / in-sourcing
  - Increased need for data/metadata exchange
  - More consumers than just regulatory agencies

Standards & Implementations



Next Steps in CDISC Standards Evolution

- Industry level Metadata Standard – CDISC SHARE
- Soon to be available standards in SHARE metadata
  - Realization of inconsistencies across different standards.
  - Still need tools to interpret this information in easy to read format.
- Add Value Level Metadata content in CDISC Standards (SDTM and ADaM)
- Analysis Results Metadata
- Study data in XML format (SDS) – remove the 8 character length and other Version 5 transport file limitations
- SUPPQUAL handling
- Define Map Metadata – traceability between different standards
- Software driven by metadata to generate data sets

SUPPQUAL

Variable Name	Variable Label
STUDYID	Study Identifier
RDOMAIN	Related Domain Abbreviation
USUBJID	Unique Subject Identifier
IDVAR	Identifying Variable
IDVARVAL	Identifying Variable Value
QNAM	Qualifier Variable Name
QLABEL	Qualifier Variable Label
QVAL	Qualifier Variable Value (must not be null)
QORIG	Origin
QEVAL	Evaluator

SUPPQUAL

- Why we need SUPPQUAL?
  - Requirement for JANUS data warehouse
- SUPPQUAL Implementation Challenges
  - Mix of data and metadata
  - Metadata information lost during transformation
  - Moving information back and forth for data submission (SDTM) and data analysis (ADaM)
- SUPPQUAL solution
  - Define the suppqual element in its parent domain
  - Include a suppqual flag variable in your element-level metadata

Period4: Application of the Principles to Data Flow

- Epoch 1 of defining "**Map Metadata**" (MAPMD)
  - Map Metadata is standardized in structure
  - Map Metadata is standardized in content
  - Stores the traceability from data state to data state
- Epoch 2 where "**Map Metaprogramming**" (MAPMP) automates data flow
  - a Data Transformation Engine (DTE) that creates target databases from source databases based on simple mappings defined in map metadata.
  - Source DBMD, target DBMD, and MAPMD between them enable metaprogramming of data flow. Data flows from raw to SDTM to ADaM to IDB to reports and each data state has corresponding DBMD and MAPMD.

MAPMD Structure Example

MAPMD			
SOURCE_DOMAIN	**SOURCE_ELEMENT**	**TARGET_DOMAIN**	**TARGET_ELEMENT**
DM	SUBJID	DM	SUBJID
DM	GENDER	DM	SEX
DM	STARTDT	DM	STARTDT

Source DBMD					
DOMAIN	**ELEMENT**	**LABEL**	**TYPE**	**FORMAT**	**LENGTH**
DM	SUBJID	Subject ID	C		
DM	GENDER	Subject Gender	C		
DM	STARTDT	Start Date	N	DATE9.	

Target DBMD					
DOMAIN	**ELEMENT**	**LABEL**	**TYPE**	**FORMAT**	**LENGTH**
DM	SUBJID	Patient Identifier	C		
DM	SEX	Patient Sex	N		
DM	STARTDT	Start Date	C		

MAPMD and MAPMP Can Enable

- Transform a database from one structure to another, with the DTE and map metadata
  - Change values of variables and paramrels
  - Change values of parameter variables
  - Rename variables and data sets
  - Move variables between same-keyed data sets
  - Rekey variables!
  - Re-attribute variables and data sets
  - Keep variables
  - Drop variables
  - Derivations
- Move variables between short-wide and tall-thin data structures

MAPMD and MAPMP Can Enable

- Greatly improve data flow transparency, for the FDA and other reviewers of data.
  - The MAPMD content can be inserted into the define file in a formatted manner and it is not a translation between English and SAS code.
  - Same contents for define file and programmer instructions
- Create integrated databases from individual study databases
- Subset standards metadata to study metadata, keeping only domains and elements that apply to the study
- Define tiers of data standard – maps can define view-like subsets. Map metadata can create target DBMD from source DBMD

Define File Example

Data Set: DM Column: AGE

Target Description:
Age (Computed in years)

Map Metadata Content:
the source data set is DM the source variable is BRTHDAT

the source data set is EXPO the source variable is EXSTDAT

Map Code:

```
%ut_age_years(fromvar=BRTHDAT,tovar=EXSTDAT,decimal=0);
```

Map Description:
ut_age_years macro calculates the age between two specified dates. This macro is called in a data step containing the two date variables. This macro can calculate age as an integer or a decimal value.

Reference formula:

```
age=floor((intck('month',&fromvar,&tovar)-(day(&tovar) <
day(&fromvar)))/12);
```

This macro enhances this reference formula by adding support for decimal ages and negative ages. The reference formula was taken from www.sas.com technical tips.

Recap

- No Data Standards
- Data Standards resident in documents
- Data Standards resident in “database metadata” (implemented at a wide variety of levels of design)
- Specs in metadata does not do any more than Spec in documents without metaprogramming
- “Database Metaprogramming” paired with database metadata to automate database attributes
- Industry standard of a robust metadata design (SHARE)
- The pairing of metadata and metaprogramming is what really delivers value in all types of metadata
- **The next step is to define “map metadata” between CDISC standards and its paired metaprogramming to automate data flow**

Contact Information

Gregory Steffens
Novartis Pharmaceuticals
Work Phone: 862-778-6550
Email: Gregory.steffens@novarits.com
LinkedIn: http://www.linkedin.com/pub/gregory-steffens/a/bb4/6ab/

Praveen Garg
ICON
7740 Milestone Pkwy,
Hanover MD 21076
Work Phone: 410-696-3218
Email: Praveen.Garg@iconplc.com
LinkedIn: http://www.linkedin.com/pub/praveen-garg/b/3b4/a84

BACKUP Slides

Example

- Mfile Example

- Define Example