Mapping Corporate Data Standards to the CDISC Model

David Parker, AstraZeneca, Manchester, United Kingdom

ABSTRACT

This paper will discuss the recent publications from the Clinical Data Interchange Standards Consortium (CDISC), relating to data standards including the Case Report Tabulation Data Definition Specification document and the Study Data Tabulation Model document and discuss mapping approaches from a hypothetical corporate standard to the CDISC Operational Data Model (ODM) and Study Data Tabulation Model (STDM).

In recent years CDISC have developed and published recommendations for data standards for Biopharmaceutical Companies, helping to move the industry towards a customer focused standardised way of presenting regulatory submissions. Mapping strategies for transforming existing corporate standards to the CDISC standards are important as some companies do not yet have the confidence that the CDISC models are stable enough to build corporate reporting systems around. For example, some companies adopted version 2 of the CDISC Submission Data Standard only to find it was substantially revised at version 3, although new versions of the SDTM will be backward compatible to version 3.1.

Several strategies are discussed for incorporating the SDTM into the existing submission process and significant mapping incompatibles are identified. For companies not yet prepared to move towards embedding the new standards into their business processes, strategies for mapping do need to be considered.

Finally an XML file with a data structure built around the ODM and ODM extensions is explored as another way of addressing the issue, discussing whether the data can be imported from SAS straight into the relevant sections of the ODM and the final ODM subset by XSLT stylesheets to generate the relevant STDM post SAS processing.

INTRODUCTION

In the past every company developed their own submission data standards and regulators had to learn and understand the complicated company specific data structures in order to conduct a thorough review of a new drug application.

Standard data definitions for the clinical trials industry have long been needed. Too much time is spent in the pharmaceutical industry doing things that in other highly regulated industries such as finance and defence might have been done smarter. The key driver for this inefficiency has been the highly successful ‘blockbuster’ business model which in the past has been to move to market as fast as possible with less regard to end-game costs as the benefits of the potential blockbuster could dwarf such costs. Changes in the external and internal environment has made this model obsolete.

There are also external drivers for change for example the FDA is rapidly moving toward building an information technology infrastructure for using web based technologies in the review process for applications for new drugs, e.g. the JANUS Warehouse (Kubick, 2004) and the adoption of XML as one of the main language of the web. It is therefore logical to surmise that XML is going to become very important in our future submission work. In October 2003 the FDA issued new guidance, which stated that XML format is now acceptable in electronic submissions.

Although XML is going to be important in the transmission and archiving of data between clinical program processes the pre-eminent of SAS for statistical analysis combined with its strong data manipulations capabilities, the large pool of experienced programmers, and the large support infrastructure means that it will continue to be the data processing package of choice for pharmaceutical companies in the research and development of new drugs. The incorporation of an XML engine and new procedures such as PROC CDISC within SAS will strengthen this franchise, but new processes will need to be developed to output the new data structures either in legacy formats such as SAS transport files or the newer technologies such as XML.

XML ADVANTAGES

XML data can be multipurpose; it can be easily transformed to produce different deliverables from one master data set. It has the flexibility to contain repeat values, annotations, codelists, links to other documents e.g. annotated CRFs and has a very comprehensive metadata component.

XML provides a mechanism for data exchange in that it is both platform and package independent, so given an appropriate interface and schema it will read as easily into SAS at it would ORACLE or an other XML enabled package.

XML can easily be converted to other formats such as PDF, HTML or even to another XML document such as a subset of the original XML document.

Some examples of the possible use of XML in regulatory submissions are contained in the CDISC published document, Qubeck 2005. In this document an XML schema is used to define a data definition document derived from the CDISC Operational Data Model (ODM). The data definition document describes an electronic submission component based on the CDISC Study Data Tabulation Model (STDM) structure.

STUDY DATA TABULATION MODEL (SDTM)

The SDTM defines a standardised structure for data tabulations that are required to be submitted as part of a product application to a regulatory authority. The model applies to data tabulations and not to the other data presentations required by regulators. Tabulation datasets are one of the four types of data currently submitted to the FDA. The other types of data are patient profiles, listings and analysis datasets; by submitting the tabulation datasets in the SDTM structure the regulatory need for patient profiles and listings will in future be reduced.

By using this model sponsors enable regulatory reviewers to use standardized software tools and become very experience in handling industry wide standardised datasets. It can be used for both clinical and non-clinical trial data across different therapeutic areas. The are currently 23 data domains described in the SDTM and each defines a superset of variables that can be used, it is possible to drop some of these variables but several core variables are mandatory and must be included in every dataset submitted. It is not necessary to include data in all domains; this is driven by the details of the trial protocol and negotiated between the sponsor and the regulatory agency. A complication is that the required data specified by the SDTM is a mixture of both raw (as collected) and derived (transformed in some way e.g. to standard units). In modeling the implementation of the SDTM it is noticed the data specified in one domain may have been collected on one or more case report forms and conversely that data collected on one case report form does not necessarily map to one domain. For example, previous drugs/therapy may be collected on several CRF pages, previous chemotherapy, previous hormone therapy and previous drug therapy, may all be mapped to the one SDTM domain, possibly Concomitant Medication or SUPPQUAL.

DATA DEFINITION DOCUMENT

Qubeck, 2005, describe a methodology of producing an XML rendition of the data definition document (define.xml) that contains the meta data needed to support regulators reviewing submission datasets. The original format of the data definition document was PDF (define.pdf) and had two main parts, a table of contents (TOC) and a collection of data definition tables.

In a define.xml, an ODM element known as the 'ItemGroupDef' is used to provide TOC domain level metadata, whereas the element 'ItemDef' is used to define the main metadata of the variables included within the datasets.

The data definition document lists the datasets the sponsor has included in the submission and describes the content of each dataset. The XML version allows the data definition document to be both human-readable and machine-readable.

The Case Report Tabulation Data Definition (CRT DD) specification provides the FDA reviewers with a clear description of the usage, structure, key fields, and content of each dataset and the origin and role of each variable which helps reviewers reproduce any analyses, tables, graphs and listings needed to support the submission.

OPERATIONAL DATA MODEL

The Operation Data Model (ODM) was adopted by CDISC as a standardised specification of a XML schema to be used for the data interchange and archiving of clinical trial data and metadata. The advantage of being in XML is that it is both vendor neutral and platform independent. (Palmer et al, 2003).

The data definition document described above has been incorporated into the ODM structure but to do this vendor

extensions as provided for by the ODM usage guidelines were introduced. This combination enables the electronic transmission of metadata to support the CDISC Study Data Tabulation Model (STDM), CDISC Analysis Dataset Model (AdaM) and the Standard for Exchange of Non-clinical Data (SEND).

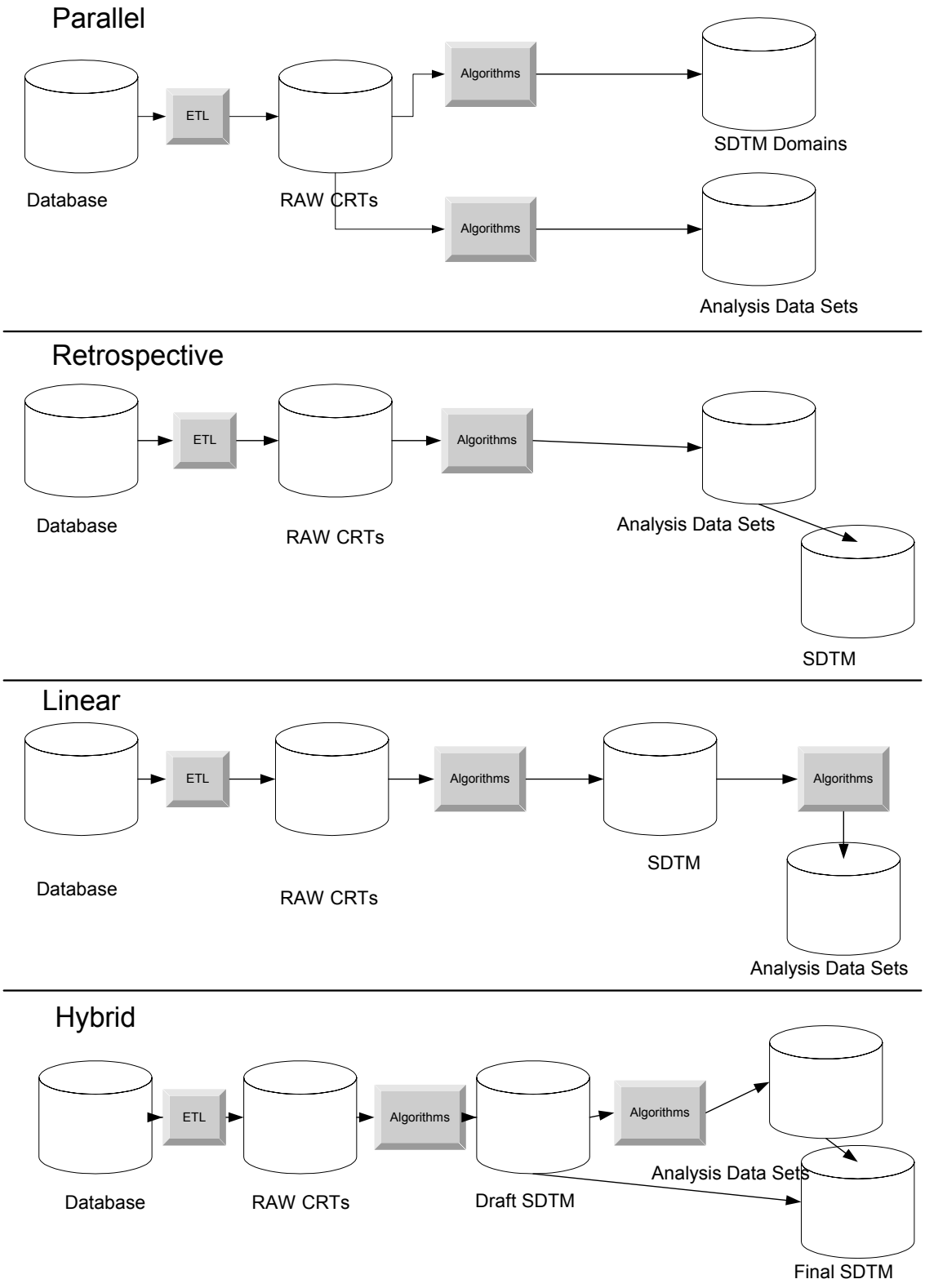
WHEN TO MAP

Kenny et al, 2005 asked the question of how and when to incorporate the SDTM datasets during the clinical trial process and considered four methods of implementing the SDTM, identifying advantages and disadvantages of each strategy. The methods are classified as Parallel, Retrospective, Linear and Hybrid. Table 1 lists the advantages and disadvantages and Figure 1 represents the strategies pictorially. They assume that prior to the implementation of the SDTM an existing scenario for the creation of clinical trial datasets was to extract from the DBMS and use the extracted data as the submission tabulation files and to build analysis datasets from this extract.

Table 1: Advantages and Disadvantages of various mapping Strategies

	Parallel Method	Retrospective Development	Linear Method	Hybrid Method
Advantages				
STDM and AdaM independent	X			
SDTM created at time of submission	X	X		X
Parallel Project Teams	X			
Minimum re-engineering of existing processes	X			
Enhancements/new releases to SDTM model are not affected		X		
Analysis programs utilise the STDM			X	
SDTM domains facilitate standardisation of Analysis Datasets			X	X
Logical flow of software development A to B to C			X	
Analysis programs submitted to the Agency are useable and informative to the reviewer as SDTM input.			X	X
Creation of e.g. baseline or population flags done in harmony SDTM vs. AdaM				X
Derived records can be added to STDM domains if required				
Disadvantages				
Documentation different for each dataset and decreased efficiency	X			
AdaM derived variables do not reference SDTM	X			
Regulators does not have original extracts for AdaM derived variables	X	X		
Analysis Programs submitted to Agency does not point to raw data	X	X		
Validation between SDTM and AdaM needed	X			
Date imputation have to be undone to match SDTM date standard		X		
All CRF variables in the SDTM would have to be retained in the analysis datasets		X		
Validation between the SDTM and source data necessary		X		
Development of Analysis Datasets depend on completion of SDTM			X	
Development of SDTM depends on completion of Analysis Datasets		X		X
STDM domain created for all studies regardless of whether part of a submission			X	X
Potential Outsourcing Problems			X	X

FIGURE 1 FOUR MAPPING STRATEGIES IDENTIFIED BY KENNY ET AL.



MAPPING ISSUES

To model a mapping process, a comparison of some data standards used in a legacy study to the SDTM was undertaken using Microsoft ACCESS 2000 database. The data standards were copied to database tables and mapped to the nearest domain in the CDISC SDTM Table. The SDTM was copied from the Microsoft EXCEL spreadsheet provided in the CDISC members area.

MAPPING PROCEDURE EXPLORED USING MICROSOFT ACCESS

- Read in the corporate data standards into a database table
- Assign a CDISC domain prefix to each database module.
- Attach a combo box containing the SDTM variable for the selected domain to a new mapping variable field
- Search each module, and within each module select the most appropriate CDISC variable
- Search for variables mapped to the wrong type Character not equal to Character; Numeric not equal to Numeric.
- Review the mapping to see if any conflicts are resolvable by mapping to a more appropriate variable.
- Next search by role and verify that the mapped variable is appropriate for each role.
- Ensure all 'required' variables are present.
- Make recommendations to change corporate variables to the new type.

COMMON ISSUES IN MAPPING DUMMY CORPORATE STANDARDS TO CDISC STANDARDS

- Character variables defined as Numeric
- Numeric Variables defined as Character
- Variables collected without an obvious corresponding domain in the CDISC SDTM mapping. So must go into SUPPQUAL
- Several corporate modules that map to one corresponding domain in CDISC SDTM.
- Dictionary codes not in SDTM parent module, so if needed must be collected in SUPPQUAL.
- Core SDTM is a subset of the existing corporate standards
- Different structure of Lab CDISC Domain e.g. baseline flag.
- Vertical versus Horizontal structure, (e.g. Vitals)
- Additional Metadata needed to describe the source in SUPPQUAL
- Dates – combining date and times; partial dates.
- Data collapsing issues e.g. Adverse Events and Concomitant Medications.
- Adverse Events maximum intensity
- Metadata needed to laboratory data standardization.

Of course some of these issues can be easily programmed around, so mapping in this exercise was relatively easy.

DISCUSION OF ODM

Another proposal for dealing with incorporating the SDTM into the existing process is to map directly to the ODM. As the ODM can be used to bring in diverse data from external sources such as CROs using different systems, could the ODM not be used to import existing SAS corporate structures? This may necessitate using vendor extensions to the ODM but would enable the master dataset to keep more metadata such as the original variable name, the original CRF question etc.

XMLSpy was used to create a sub set of an ODM, mapping some dummy demography data and then using a XSLT stylesheet to output to the SDTM structure to test this proposition. SAS 9 may offer a more subtle way forward to deal with such a problem in that it could be used to build an ODM directly from SAS datasets. It would then be necessary to write a stylesheet outside of SAS to convert the ODM structure to the subset of the SDTM. However it may be acceptable to send the data to regulators with appropriate stylsheets and allow the regulars to load the data into their systems using the sylesheets as interfaces between the ODM and their web-based system.

CONCLUSION

This paper has reviewed some current discussions around the implementation of the CDISC standards focusing on issues which companies might have while in transition to embedding the standards into their business processes. A mapping exercise was undertaken which highlighted some issues, which may be common in other implementations. Four strategies identified by Kenny et.al. were summarised and a further strategy was explored creating an XML data structure using the ODM schema as a template. From this data structure it was proposed that the SDTM could be

created using XSLT stylesheets to load the data directly into IT systems such as JANUS.

REFERENCES

- Strategies for Implementing SDTM and AdaM Standards, Susan J. Kenny, Michael A.Litzinger Paper FC03 PharmaSug 2005
- Implementation of the CDISC SDTM at the Duke Clinical Research Institute, Jack Shostak Paper FC01 PharmaSug 2005
- Defining and Validating CDISC Data Standards to XML in SAS Technology. Proceedings of the PharmaSUG 2004 .
- CDISC Procedure for the CDISC SDTM 3.1 Format, http://support.sas.com/rnd/base/topics/sxle82/cdiscsdtm.html
- Testing CDISC's Operational Data Model in SAS, Michael Palmer & Julie Evans, www.cdisc.org
- Data Management: The Clinical Research and Regulatory Perspective, NIH BECON/BISTIC Symposium 21st June 2004, Wayne Kubick.
- Case Report Tabulation Data Definition Specification, Qubeck, CDISC 2005

CONTACT INFORMATION

David Parker
AstraZeneca
Alderley Park
Macclesfiedl SK10 4TG, United Kingdom
Work Phone: +1625 231834
Email:david.t.parker@astrazeneca.com