Paper DM03

Practical application of SAS® Clinical Data Integration Server for conversion to SDTM data

Peter Van Reusel, Business & Decision Life Sciences, Brussels, Belgium
Mark Lambrecht, SAS, Tervuren, Belgium

ABSTRACT

Conversion of legacy clinical data into a standardized data model consists of different steps from annotating the CRF, preparing the mapping into the CDISC SDTM model to producing metadata reports and CRT-DDS (define.xml). In order to accelerate the conversion process, we have used the SAS® Clinical Data Integration Server framework, which has been designed to help life sciences organizations visually design conversion processes driven by standardized metadata. We will discuss in this paper our implementation approach and how SAS® Clinical Data Integration Server significantly shortened our throughput time needed to produce SDTM-compliant submission-ready packages. In addition, extensive compliancy and quality control checks ensure that the data and metadata is consistent and according to the defined standards.

INTRODUCTION

Most pharmaceutical organizations have only recently started adapting their internal data models to accommodate the CDISC data models, such as the Study Data Tabulation Model (SDTM) (http://www.cdisc.org). Therefore a core requirement in implementing CDISC standards in clinical development operations is the ability to convert existing or legacy data into CDISC-compliant formats. Business & Decision Life Sciences has been converting legacy clinical data into SDTM-structured data using SAS® Clinical Data Integration Server.

CLINICAL DATA INTEGRATION SERVER

Clinical Data Integration Server is a processing framework provided by SAS to facilitate production of CDISC-compliant data sets and of the associated metadata reports, such as define.xml. A core component of all SAS solutions is the SAS® Open Metadata Architecture (OMA) which stores the CDISC SDTM data model and makes it centrally available to all users and the different clinical projects. The SAS® Clinical Data Integration Server framework contains the target SDTM domains (including the special-purpose relationship datasets), generic job templates, the CDISC controlled terminology and the variables needed to create a custom domain according to the instructions in the SDTM 3.1.1 IG. All related metadata in the SAS OMA is grouped together in repositories. The Foundation repository groups technical and environmental metadata that is needed in other repositories, e.g. about users, groups, security permissions, computing servers, scheduled jobs…

Other repositories are dependant on the Foundation repository. The Global Standards repository (Figure 1) groups the CDISC standard (meta)data and in this configuration acts as the parent repository of the different clinical project repositories. A user opens a metadata profile to a specific clinical project repository (Figure 1) and consequently has access to all parent repositories, given the right permissions. For example, if a user connects to the 712 clinical project repository, the Global Standards repository containing all CDISC standards metadata and the Foundation repository are accessible. This clear separation and dependency of repositories allows for better management of security and access to (meta)data based on the role of the user. Changes to objects requires going through a check in / check out cycle in user repositories whereby changes are captured in an audit trails. The object-locking facility has proven to be very useful when working with a large team of programmers (or data integration programmers) on the same conversion jobs.
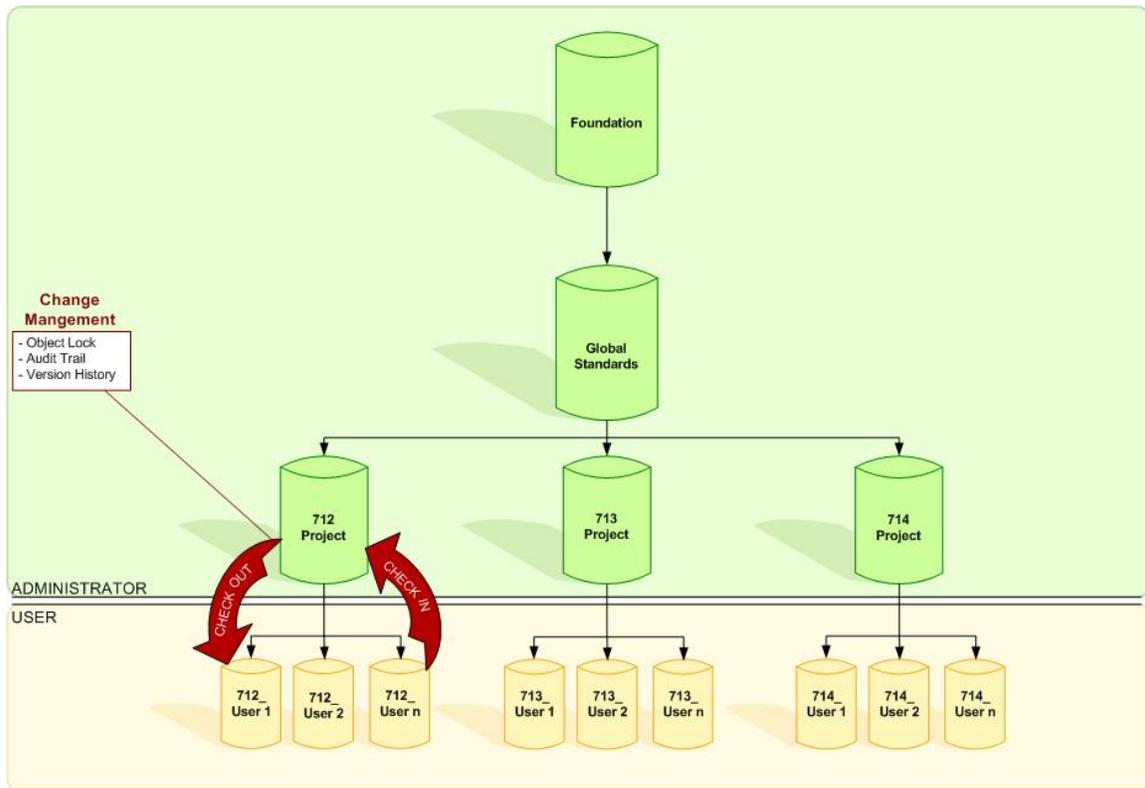
Figure 1 : Structure of Clinical Data Integration Server repositories in the case of 3 clinical projects.

For our use of SAS® Clinical Data Integration Server, additional data and metadata was needed. For example, we added datasets with Controlled Terminology (Figure 2) tables, both at a global level (in the Global Standards Repository) and at the clinical project level. Value-level metadata is added in a folder both at the global, project and study levels. When value-level metadata has to be combined with the source data, a "value-level metadata transpose" custom transformation has been created.
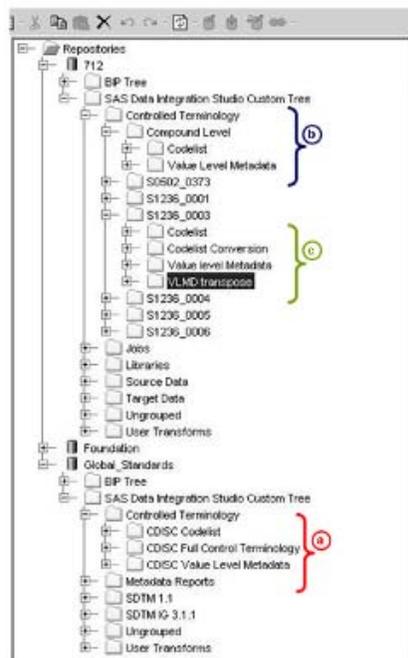
Figure 2 : Folder structure in Clinical Data Integration Server supporting the SDTM conversion and production. (a) Global level – (b) Project level - (c) Study level.

PROCESS

Figure 3 describes the different deliverables in a CDISC conversion process:

- Generation of the SDTM Implementation Guide (IG) 3.1.1 annotated CRF's
- Creation of the SDTM IG 3.1.1 datasets
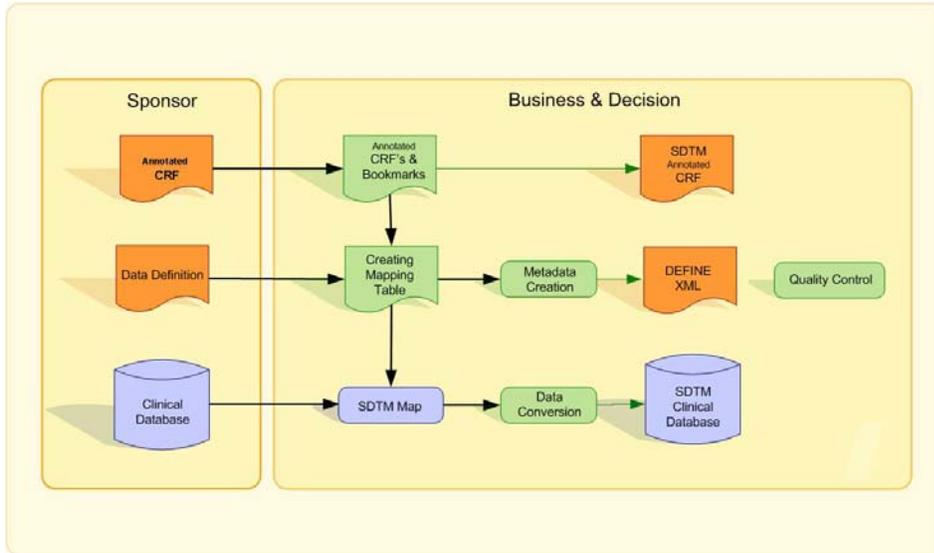- Creation of the define.xml datasets

Figure 3 : Conversion of legacy clinical data into the SDTM format, including define.xml, for submission purposes.   Metadata is indicated in orange, data and data mapping in blue and processes in green.

The process (Figure 3) is carried out by the conversion team consisting of data mappers who define the mapping and of programmers who create the conversion programs. The different steps in the whole process are :
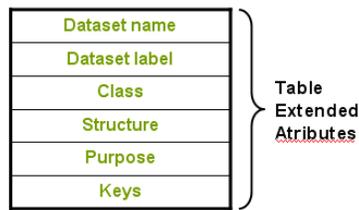
1. Annotating the CRF's
2. Set-up of a new clinical project in Clinical DI
3. Creation of conversion processes in Data Integration Studio
4. Management of metadata and production of define.xml
5. Compliancy checks

STEP 1: ANNOTATING THE CRF

First the SDTM mappers annotate the CRF's, either paper or electronic, and add the SDTM variables onto the forms. (Figure 4). The described annotations allow to quickly understand how source data in the original format are mapped to SDTM target data. The define.xml will contain links to the respective annotated CRF's.

Figure 4 : CDISC annotated CRF's.

STEP 2: SET-UP OF A NEW CLINICAL PROJECT

When a new clinical project is set up, an external Microsoft® Excel® file is constructed that contains the source datasets, and the information needed to map source variables onto target variables. This separate source of metadata makes the setup at the start of the conversion process a flexible process. All the mapping information and the preparation of all needed metadata are done by the mappers in this Excel sheet.

A new Clinical Data Integration Server study folder or clinical project repository is then created with the appropriate source, target and job folders. The target SDTM dataset structures are copied from the Global Standards repository. Project- or study-specific metadata about the datasets or variables is then programmatically added to the Clinical Data Integration Server metadata in the project or study level repositories (Figure 5) from the central mapping file. This metadata is later re-used for compliancy checks, for the production of value-level metadata, and for the creation of the define.xml. The time needed to set up a new study is kept to a minimum because of the automated procedure.

Figure 5 : Table and variable metadata as extended attributes and notes.

The mapping sheet also contains advanced metadata such as a code list conversion table and an automatic transpose table. This metadata will be used later in the process to create both the define.xml, and to automate big parts of the conversion processes.

STEP 3: CREATION OF CONVERSION PROCESSES IN SAS® DATA INTEGRATION STUDIO
The programmers then start working using the mapping logic produced by the data mappers. SAS Data Integration Studio provides a drag-and-drop interface which allows the programmers to visually create Data Integration Studio jobs (Figure 6) that auto-generate robust SAS code. Because one job per SDTM domain is created in each study, it is possible to reuse a large part of the transformation logic over different studies if the source dataset structure is stable. A large number of transformations are available in the Process Library of Data Integration Studio. The ability to generate custom transformations using the transformation generator allowed us to add specific clinical transformations such as ISO8601 date-time conversions, a look-up and transpose transformation …These custom transformations can be parameterized and therefore easily deployed in different clinical projects and studies.
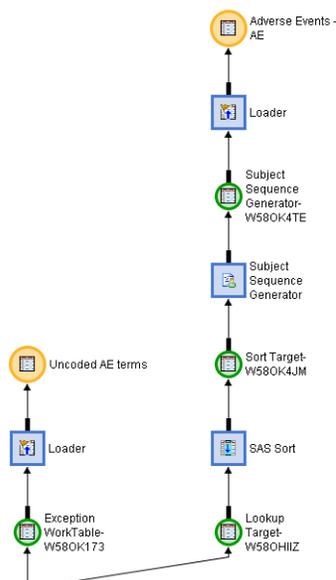
Figure 6 : Part of a job flow in Data Integration Studio resulting in the creation of the AE SDTM domain

STEP 4 : METADATA MANAGEMENT

Once the Data Integration Studio jobs are created and the SDTM datasets are populated with values, define.xml can be produced. Metadata stored in the SAS Metadata Server can be queried from Base SAS, Java, or .NET programs. For our purposes, the Base SAS syntax was meeting our requirements. The information needed in the define.xml is easily extracted from the SAS metadata repositories (see Figure 7). The define.xml is produced in SAS Clinical Data Integration Server using a user-written custom macro that makes use of a SAS template. The define.xml includes the code lists, the value-level metadata and the pseudo-code used for the computed variables (Figure 8). By using specific style sheets, it is possible to surface the built-in links to annotated CRF's, to the code lists and computed variables.
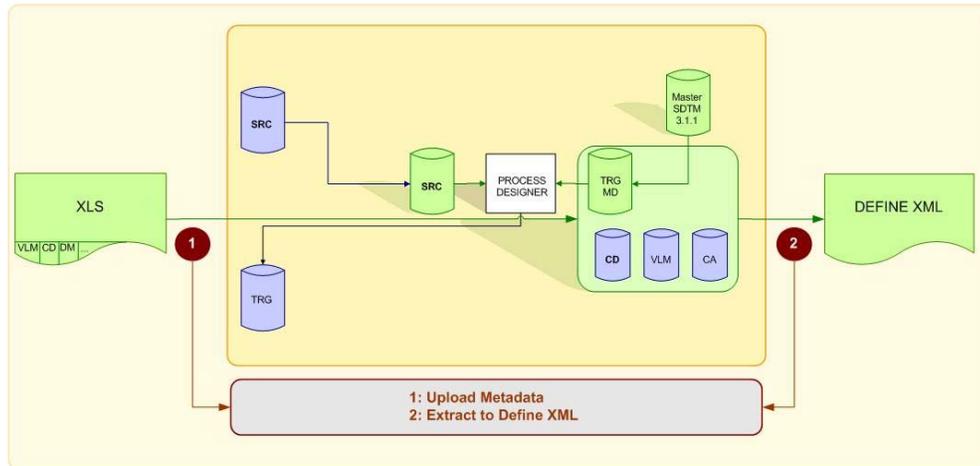
Figure 7 : Production of define.xml

STEP 5: SDTM COMPLIANCE CHECKS

The metadata about source, SDTM target datasets and job conversion metadata, is automatically collected in the SAS® metadata server and enables further exploitation. Extensive compliance checks allow for cross-checking the produced SDTM datasets with the CDISC standards contained in the Global Standards repository, and with the produced define.xml for consistency. Quality issues that are identified are added to a large exception table, and in a second step a check report generator generates from this table a set of quality check reports.

Figure 8 : A detail of the define.xml file

CONCLUSIONS

Data from multiple phase Is to Phase IIIs and mega-trials with tens of thousands of subjects have been successfully converted into an SDTM-compliant submission package using the above described process. A similar process can be applied for ADaM dataset production. While the preparation of the mapping process is done in a central mapping file, the actual mapping processes and metadata exploitation is carried forward in SAS Clinical Data Integration Server. The main advantage of Clinical Data Integration Server is that existing business processes do not have to be redrafted, making them more efficient and transparent. The SAS® Open Metadata Architecture has a flexible metadata model that drives standardization and enforces documentation. The user-specific project repositories impose collaboration and an audit trail of the development process. Clinical Data Integration Server is executed on scalable server-client architecture and a process that used to run for hours on an interactive SAS Windows session is now executed in minutes. This contributes to a fast and efficient optimization of the data mapping process, an added advantage that was not clear to us at the start of the project.
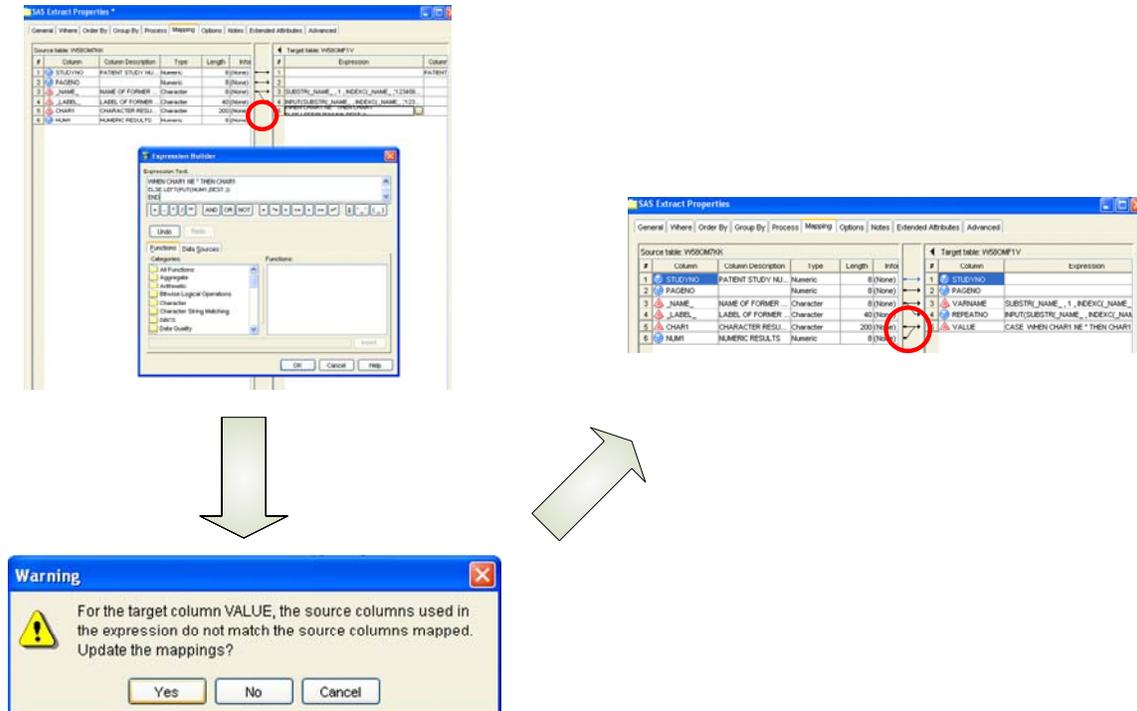
Figure 9 : Data Integration Studio automatically captures variable mapping inconsistencies

The ability to rapidly translate the mapping requirements into robust Data Integration Studio-generated SAS® code significantly reduces the time needed in a conversion project. Moreover, because the jobs are metadata-driven, Data Integration Studio captures the link between the mapped datasets and columns (impact analysis) and automatically checks for variable mapping inconsistencies (Figure 9).

The flexibility of SAS Clinical Data Integration Server, the integrated metadata that is centrally accessible and the increased collaboration between developers have highly improved the efficiency of our team in executing CDISC conversion projects. The proven process methodologies, the CDISC expertise in our team, and the SAS Clinical Data Integration Server technology have all contributed to this success.

REFERENCES

Kilhullen, Michael. "Implementing CDISC Data Models in the SAS® Metadata Server" Proceedings for the Pharmaceutical Industry SAS® Users Group Conference, 2006 & 2007

ACKNOWLEDGMENTS

BIOGRAPHY

Peter Van Reusel, manager CRO services
Peter Van Reusel is Business & Decision Life Sciences' CDISC expert and representative in the CDISC bodies and teams as well as one of the few CDISC SDTM trainers providing the CDISC SDTM courses in Europe.
Mark Lambrecht, SAS Institute consultant
Mark Lambrecht, PhD, is Principal Consultant at SAS and promotes and implements SAS' solutions for clinical data integration, genomics, statistics and reporting in the life sciences industry.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Peter Van Reusel	Mark Lambrecht
Business & Decision, Life Sciences	SAS Institute
Rue de la Révolution 8	Hertenbergstraat 6
B-1000 Brussels	B-3080 Tervuren
Belgium	Belgium
Office phone: +32 2 510 05 40	Office phone: +32 2 766 07 00
http://www.businessdecision-lifesciences.com	http://www.sas.com/industry/pharma/

Brand and product names are trademarks of their respective companies.