PhUSE 2006

Paper RA01

Converting Data to the SDTM Standard Using SAS® Data Integration Studio

Barry R. Cohen, Octagon Research Solutions, Wayne, PA

ABSTRACT

The CDISC Study Data Tabulation Model (SDTM) is becoming the industry standard for clinical data. Future Marketing Applications to the FDA will require conversion of clinical data from various legacy formats and current internal standards to SDTM. This will be a major undertaking for any organization, and a data integration/data warehouse tool can be of significant help in this process. SAS Data Integration Studio is the new SAS product designed to support data conversion and warehouse building processes. Data Integration Studio works in the SAS 9 architecture and uses SDTM metadata and other metadata to build conversion processes without the user writing program code. Octagon Research Solutions uses Data Integration Studio as the key software application in its process of converting client clinical trials data into SDTM submission-ready datasets. This paper will discuss various aspects of Data Integration Studio and the Octagon SDTM conversion process using this product.

INTRODUCTION

The Clinical Data Interchange Standards Consortium (CDISC) has developed a standard for clinical and preclinical trials data, called the Study Data Tabulation Model (SDTM). The FDA has stated in eCTD guidance that they now would like to receive clinical and preclinical trials data from sponsors according to the SDTM standard. The industry expectation is that within a few years the FDA will require by regulation that sponsors submit their data in the SDTM. It is thus no surprise that the industry is quickly adopting the CDISC SDTM. Many sponsors are already converting clinical trials data from active Marketing Application projects to the SDTM, many others are assessing their SDTM-readiness and currently planning their first SDTM conversions, and still others are exploring ways to integrate the SDTM standard into their full clinical data life cycle.

Throughout these activities, sponsors are beginning to understand that an SDTM conversion project is a major undertaking with significant work required in two top-level areas. The first area regards conversion project/process setup, which includes: process design, process documentation, process validation, staff resource and deployment, and staff training. The second area regards the ongoing conversion work once the project/process is established. The ongoing conversion work includes mapping specifications development and program development and validation per study, for the many SAS datasets per study, the several/many studies per project, and the stream of projects over time. Further, some sponsors have made or are considering a decision to convert their legacy studies to the SDTM standard, too, in order to build a single-standard clinical data warehouse. They will do this both as an effective way to archive their clinical data and as a way to support mining of this data for various purposes. In these cases, the amount of data conversion work involved will be significantly larger than that needed to support individual Marketing Application projects.

Many sponsors will look to data integration technology (sometimes referred to as data warehousing or ETL technology) to help them in the SDTM conversion process. They will obviously look to this technology if they are building a clinical data warehouse. But data integration technology can also help even if they are only looking for an efficient way to do an SDTM conversion for individual Marketing Application projects without warehousing the data. Data integration technology is designed to extract data from a source (e.g., existing clinical trials data in a legacy standard), transform it (e.g., as needed to become SDTM-compliant), and load it to a final target (e.g., to individual SDTM domain datasets for submission to the FDA). Data integration technology can accomplish this process with far less programmer involvement than would be needed if all the programs were developed more traditionally without a data integration tool. In essence, with data integration technology, clinical data management staff use a visual design process to express the data integration and conversion specifications, and the data integration tool then auto-generates the conversion programs from the specifications.

Octagon Research Solutions, Inc. is actively converting clinical trials data to SDTM for sponsors, and is using the data integration technology from SAS, specifically SAS Data Integration Studio, to support its SDTM conversion process. This paper discusses Octagon's SDTM conversion process. It begins with a background discussion about SDTM and how and where it can and will be integrated into the clinical data life cycle across the industry. This will help set the context within which SAS Data Integration Studio can be used to convert clinical data to SDTM. The paper will then discuss our SDTM conversion process and the role of SAS Data Integration Studio in that process. A short list of topics covered includes:

- SDTM – a brief state of the industry
- Where and when the SDTM standard will be integrated into the stages of the clinical data life cycle
- Why late-stage conversion to SDTM will occur for some time to come

- General issues of a data conversion process
- Overview of Octagon's SDTM conversion process using SAS Data Integration Studio
- Conclusions from our experience doing SDTM conversion using Data Integration Studio

SDTM STANDARD – STATE OF THE INDUSTRY

MOTIVATION FOR ADOPTION – REGULATION AND EFFICIENCY

The immediate motivation for sponsors to adopt the SDTM standard is a regulatory one. Today, the FDA has stated in eCTD guidance that it wants to receive the clinical and preclinical data components of Marketing Applications in SDTM. The FDA has committed to using SDTM and has prepared a set of in-house tools to process SDTM data received from sponsors. The tools include their new internal repository (JANUS clinical data warehouse) that is based upon the SDTM standard, WebSDM to validate and load SDTM data to the JANUS warehouse, and the Patient Profile Viewer (PPV). Today, the FDA desires SDTM data but does not require it. However, the industry generally expects that SDTM will become a requirement within a few years. So today companies are developing SDTM capabilities and SDTM conversion processes within their organizations.

Processing efficiency is the longer-term motivation for sponsors to adopt the industry standard SDTM as their new internal clinical data standard. Much efficiency is possible. Some examples:

- Easier data interchange between sponsors and contract research organizations (CROs)
- Easier data interchange between sponsors and central laboratories
- Easier data interchange during collaboration with research partners
- Use a data standard built and maintained by the industry and thereby avoid building and/or maintaining an internal data standard
- Opportunity to archive all clinical data using a single-standard and trans-study structure in a clinical data warehouse, and the subsequent ability to mine that data for various research purposes

In essence, the industry is recognizing that their proprietary data standards do not have particular advantage vis-à-vis their competitors, and they can gain more from working to the same standard than from developing and maintaining their own standards.

THREE COMPONENTS OF SDTM

The SDTM model is comprised of three components: nomenclature, content, and structure. Knowing the three components is key to understanding both what is involved in converting clinical data to SDTM and where/when the SDTM standard and conversion to SDTM will be integrated into the stages of the clinical data life cycle within an organization.

- Nomenclature concerns the standardization of names used in the SDTM model to identify SDTM domains (datasets) and items (variables) in SDTM domains. For example, the variable named SEX is used in the demography domain, not a variable named GENDER. SEX is the standard nomenclature.
- Content concerns the standardization of the data values of certain variables in the SDTM model. For example, for the variable SEX, the standard values are "M", "F", "U", not "Male", "Female", "Unknown", and not "1", "2", "9". As another example, the standard data values for Yes/No questions are "Y" and "N", not "Yes" and "No". Yet another example is that in a future release of the SDTM model, the values for test names and test codes will be standard. This standard content is also referred to as "Controlled Terminology" in the SDTM model.
- Structure concerns what domains exist, what set of variables exist in each domain, and for these variables, what are the types, lengths, positions, etc. in the dataset. It also concerns the shape of the datasets (i.e., "long and narrow" or "short and wide").

Clinical data can be compliant with SDTM nomenclature and content without being compliant regarding structure. That is, clinical data can be stored in a different structure than the one defined by SDTM, such as in a Clinical Data Management System (CDMS) or in a Clinical Data Warehouse (CDW), and still use SDTM-compliant nomenclature and terminology. This fact is important to understanding how, where, and when the SDTM standard will be implemented throughout the stages of the clinical data life cycle in an organization and where conversion to SDTM will occur in the life cycle.

CLINICAL DATA LIFE CYCLE

Once clinical trials have been designed, the life cycle of the clinical data business process has the following stages:

- Collection – The data is collected in this stage, either on paper-based Case Report Forms (CRF) or using e-CRFs in an Electronic Data Capture (EDC) system.
- Processing – The data is maintained and processed in this stage. All manners of processing needed to define study database tables, define CRFs, and validate the collected data occur here. Data is stored here on an individual study basis, as opposed to trans-study. Database systems are often used in this stage, generically called Clinical Data

Management Systems (CDMS). If an organization is operating according to some internal data standard, then global libraries are built in the CDMS that hold the standard metadata for items used in the various studies.

- Storage – This is an emerging stage in the life cycle. Historically, sponsors have extracted the frozen data from the CDMS and stored it in SAS datasets for the Analysis and Reporting stage. The storage repository has simply been an operating-system-controlled file system where the various datasets of individual studies are kept together. Limited consideration has been given to archiving according to a single data standard or to trans-study storage and trans-study analysis and reporting. But storage is changing today as companies begin to think about and develop clinical data warehouses to archive their data in a single-standard, trans-study manner. They are planning to use the warehouse as the data source for the Analysis and Reporting stage for individual Clinical Study Reports (CSRs), as the data source for integrating the data for the Integrated Summary Reports, and for true trans-study data mining for a variety of purposes including clinical research, safety signaling, and others. SDTM is being considered in many organizations as the data standard for the warehouse.
- Analysis and Reporting – The analysis and report programs are developed and executed in this stage to produce the tables, listings, figures, and other analysis outputs for individual CSRs. This tends to be a SAS-centric stage in the life cycle. Data is also integrated across studies in this stage to provide the data input for analysis and reports for Integrated Summary Reports, too. The efforts to integrate the data have historically been large since data has needed to be standardized across studies and since storage structures have not afford an easy integration of data across studies. The situation has improved somewhat over time, especially as sponsors have adopted and enforced internal data standards. However, it is surprising how many sponsors still struggle with this data integration process today.
- Compilation and Submission – The data component of the Marketing Application is assembled and submitted in this stage. The data component includes the SAS datasets, documentation about the datasets, and annotated CRFs that are annotated with the dataset information. Today, if an organization is submitting their data in SDTM-compliant datasets, this is the point at which the final SDTM datasets must be created. It is possible (and may be desirable) to create the SDTM datasets at earlier stages and use them for the activities at those earlier stages, (e.g., the Analysis and Reporting stage). However, this is the last point at which the SDTM datasets can be created. This is referred to as “late-stage” conversion to SDTM.

SDTM IN THE CLINICAL DATA LIFE CYCLE

SDTM can be integrated into any stage of the clinical data life cycle. And yet, for many companies, in the near-term, SDTM is likely to only be integrated into the last stage (Compilation and Submission). This is so because companies have already established sophisticated clinical data business systems and practices that are based upon internal data standards that have their own nomenclature, content, and structure. Thus, a substantial change in business systems and practices will be required to integrate the SDTM standard into any stage of the life cycle. It will be easiest to do this in the last stage where, in essence, all the processing is done and the data is simply being compiled for submission. The data can be converted to SDTM at this stage without requiring a change to upstream systems and practices.

Given the substantial time needed to integrate SDTM into upstream stages of the life cycle, and the desire to submit SDTM-compliant data to the FDA in the near-term, either to meet the current FDA eCTD guidance or to meet the expected FDA regulation, many sponsors will initially do a late-stage conversion of their data to SDTM. This is the first context within which many organizations will need to determine a method to provide data according to the SDTM standard, and the first context within which they might use data integration technology to help integrate the SDTM standard into their clinical data life cycle.

But there are two important incentives to move the integration of the SDTM standard upstream into earlier stages. The first is that the processing efficiencies that come from operating according to the industry’s data standard will only accrue to an organization if they integrate SDTM into the upstream stages of the life cycle. This is where the data exchanges with external partners tend to occur. The second is that an early-stage integration of the SDTM standard (at least nomenclature and content) can avoid or minimize the extra step of late-stage conversion after the work is done, and avoid this at a particularly beneficial point in time (i.e., very late in the project when pressures on the project timeline are greatest).

It is worth noting that even when the SDTM standard is integrated upstream, and there is no conversion to SDTM per se but rather integrated use of SDTM throughout the life cycle, there will still be several places to use data integration technology in the business process. Following are examples:

- Extract data received in SAS datasets from CROs and central laboratories, transform it, and load it to the CDMS or a clinical data warehouse.
- Extract data from the CDMS that is compliant with SDTM nomenclature and content, transform it by adding derived SDTM items that are not produced and stored in the CDMS (e.g., Age, Arm Codes, Reference Start and End Date, Study Day, and ISO8601 dates), and:
  - (a) Transform its structure and load it to SDTM domain datasets, or
  - (b) Load it to a clinical data warehouse
- Extract data from a clinical data warehouse that is compliant with SDTM nomenclature and content, transform it by deriving standard CDISC ADaM analysis items (e.g., change from baseline values), and load it back into the warehouse (so that the analysis data is archived in the warehouse as well as the collected data)

- Extract data from a clinical data warehouse that is compliant with SDTM and ADaM nomenclature and content, transform its structure and load it to ADaM-compliant analysis datasets for the Analysis and Reporting stage.
- Extract data from a clinical data warehouse that is compliant with SDTM nomenclature and content, transform its structure to SDTM structure, and load it to SDTM domain datasets for the Compilation/Submission stage.

Nonetheless, even if the opportunity for data integration technology is limited to late-stage conversion to SDTM, most sponsors are likely to encounter this opportunity in measurable fashion for several years to come. One reason for this was mentioned above --- that it will take time to implement SDTM as the standard in the upstream stages, and substantial amounts of data will have to be converted to SDTM in the meantime for submission to the FDA. A second reason is that submission projects typically take years to complete, and the study data within them is not often switched from one data standard to another mid-stream. Thus, companies are likely to have a long period of "parallel standards" where the studies of new projects are completed using the SDTM standard, and the studies of existing projects are completed using the earlier internal standard. And these latter studies will need late-stage conversion to SDTM.

DATA CONVERSION – GENERAL PICTURE

I have now described the business context for conversion of clinical data to SDTM, and I turn attention to the data conversion process, which in Octagon's case is centered on the SAS Data Integration Studio. For the remainder of the paper, I will be thinking primarily about late-stage conversions to SDTM, where the non-SDTM datasets used during the Analysis and Reporting stage are now being converted to SDTM. I believe this will be the likely scenario for many/most sponsors in the few or several years just ahead. I will first discuss some general issues of data conversion and I will then describe Octagon's conversion process using the SAS Data Integration Studio product.

GENERAL ISSUES OF DATA CONVERSION

Most generally, a data conversion effort involves extracting data from a source, transforming it in some fashion, and loading the result to a target. In the clinical data world, the source data for a data conversion project is likely to be either SAS datasets or database tables, and the target data is likely to be one of the same. Using SAS datasets as the example for source and target, the conversion process will encounter one or more of the following source-to-target situations, which will very much influence the programs that need to be developed to effect the conversion:

- One-to-one – All the data from one and only one source dataset is going to one and only one target dataset
- One-to-many – The data from one given source dataset is going to multiple target datasets
- Many-to-one – The data from many source datasets is used to build one target dataset

The one-to-one situation is, not surprisingly, the easiest to program. The program to build any target dataset this way would only have to open and read data from one source dataset, make the necessary transformations, and write the results to one target dataset. More complicated programming issues arise if you need to either read from your source datasets multiple times before you have written all their data to your target data sets, or write to your target datasets multiple times before you have finished reading the data from all of your source datasets.

Transformations can be seen in three groups, as follows:

- Changes to nomenclature – This largely involves changes to dataset names, variable names and variable labels, to what things are called if you will.
- Changes to content – This largely concerns changes to the values of the data to meet an accepted standard. For example, a source variable with the values 1 and 2, (which correspond to Male and Female), is changed to have the standard values M and F in the target variable. Or a set of non-standard names for ECG tests in the source dataset is changed to a set of standard test names in the target dataset. Or a date variable that is numeric in the source dataset is changed to a character representation in the target dataset. Or the units for the results of a particular test in the source dataset are changed to standard units in the target dataset.
- Changes to structure – This basically concerns the set of datasets used, the particular variables found in the particular datasets and their types, lengths, and positions, and the organization of data by rows and columns in the datasets. The structural issue of mapping source datasets and variables to target datasets and variables was alluded to above in the one-to-many and many-to-one situations. Another structural change can occur if the organization of the data is changing from "short and wide" datasets to "long and narrow" datasets. For example, a source dataset structure where test results for multiple tests done at the same visit are on the same record (i.e., short and wide) is changed to a target structure where each test at a given visit has its own record and that record holds both the test name and the result (long and narrow). Another structural change can occur if there are items needed in a target dataset than do not directly exist in any of the source datasets.

Another conversion issue concerns the approach chosen to develop the programs that do the actual conversion. There are situations where the conversion is needed only once or a limited number of times, and for a small number of datasets, and the programming will thus be done on an ad hoc basis. But when there are large amounts of data to convert, and the program development effort will consequently be large, the conversion will likely involve some kind of software application. Such applications will have these two key features: (1) The conversion programs for a given set of data are auto-generated by the application, which reads and interprets conversion specifications; (2) Specifications for tasks (such as transformations) that re-occur across studies being converted are developed once and reused many times. These conversion software applications can either be developed in-house or acquired in the marketplace. The commercially available products are often referred to as data integration applications, or sometimes ETL applications where ETL stands for “Extract, Transform, Load”. SAS Data Integration Studio is such an application.

Yet another conversion issue concerns the exact set of steps that are needed in the process of conversion and what staff skills are needed for each step. The exact steps may vary based upon whether you will develop ad hoc conversion programs or use a conversion application. And if you use an application, the exact steps may vary by the design and features of the application. For example:

- For ad hoc program development: Do you need a separate mapping specifications step, where someone specifies the details of how the source data will map into the target datasets? Is this step done by the programmer and if not, then by what type of staff?
- For conversion applications: Do you still need a separate mapping specifications step when using a conversion application as opposed to ad hoc program development? With a data integration application, is the mapping specifications step done inside the application? What staff skills are needed for working with the data integration application? Are they programmer skills or data management skills? Are these skills different from the mapping specification skills? Can a person develop the conversion jobs or processes inside the data integration application without writing any program code? Is this the same if the conversion application is built in-house or acquired commercially?

One more issue concerns system and process validation. In a regulated industry, such as the pharmaceutical clinical trials industry, where the programs and processes must be validated, there are implications for ad hoc development of conversion programs versus using a conversion application. Specifically, all ad hoc programs must be validated. The more programs you write this way, the more validation you must do. In contrast, if you use a conversion application, the application must be validated by you if you build it or by the third party vendor if you purchase it. Then you must validate the process you conduct using the application. But you do not need to validate the programs generated by the application.

OCTAGON’S SDTM CONVERSION PROCESS

In this section, Octagon’s SDTM conversion process, which uses SAS Data Integration Studio as the conversion application, is discussed. The overall process is described, discussing how the process works in regard to the various general conversion issues raised in the section above.

OVERVIEW

There are four steps to Octagon’s SDTM conversion process, as follows. There is a fifth step, validation of the work, but the validation is seen as integrated into each of the other four steps. The steps are identified here, and then the first two, which are the ones needed to produce the SDTM datasets, are discussed in separate subsections below.

- Development of mapping specifications
- Development of conversion jobs in Data Integration Studio
- Development of the Define document
- Development of the Annotated CRF

DEVELOPMENT OF MAPPING SPECIFICATIONS

This is the first step in the conversion process. A mapping specialist conducts the step. The mapping specialist examines each item in each source dataset of the study and determines if the item should be migrated to the target SDTM datasets, and to which SDTM domain dataset and variable the item will migrate. Some items in the source datasets are not collected clinical data and will not migrate to the target SDTM datasets. During this step the mapping specialist also identifies additional data that is needed in particular SDTM domain datasets being created but is not coming directly from the source datasets. There is a fair amount of this situation in SDTM conversion, and it is one reason that the mapping specifications step is more involved than just specifying the disposition of existing source dataset items. The mapping specialist needs strong clinical data knowledge and strong SDTM knowledge to accomplish this work.

The mapping specialist starts by reviewing the protocol and source datasets to gain an overall understanding of the data being converted to SDTM. The specialist then runs a custom SAS program that reads the contents of all the SAS datasets in the source library and generates an Excel spreadsheet with one row for each source dataset-variable combination in the source data and populates this row with a variety of metadata about the item in that row. Additional columns on each row are then used to indicate to which target SDTM dataset-variable combination the source variable will migrate, and any transformation that is necessary. Finally, the specialist adds rows to the mapping specs spreadsheet for items that are needed in the various target SDTM datasets that are not directly in the collected data in the source datasets.

Many issues tend to arise during the mapping step that must be resolved before it is completed. They can range, for example, from clarification about the source data, to decisions on the mapping of particular source data in the SDTM model, to decisions on sponsor-defined standard terminology. The number of issues can be measurable and their resolution is critical to a correct and complete conversion to SDTM. Thus, Octagon uses its own in-house developed (and commercially available) process management software, ViewPoint®, to manage this process.

There is a Quality Control (QC) review when the mapping specs spreadsheet is completed. A second mapping specialist reviews the work in the spreadsheet and reports errors back to the first specialist for correction. In matters of judgment, where the two specialists do not agree, the issue is brought to Octagon's resident SDTM experts for resolution.

DEVELOPMENT OF CONVERSION JOBS IN DATA INTEGRATION STUDIO

The Data Integration Studio environment is one where the user (whom we refer to as the "developer") defines a job that loads one or more source datasets, maps source variables to variables in target datasets, transforms the source variables where needed, derives new variables where needed, and finally loads the finished variables to the target dataset(s). The developer executes the completed job in the Data Integration Studio environment. When the job executes, Data Integration Studio generates a SAS program according to the definition of the job, and submits that program to the SAS environment for execution, with the results (i.e., Log) returned to the Data Integration Studio environment and the resultant dataset(s) placed in a specified SAS library.

Note that this is not traditional SAS program development, nor is a SAS programmer doing it. Rather, the developer is using the Data Integration Studio environment (menus, selection lists, graphic schematics, etc.) to define a job, and Data Integration Studio then writes the SAS program from this job definition. The developer needs the following skills in this role: (1) some clinical data knowledge (but not as much as the mapping specialist); (2) some SDTM knowledge (but not as much as the mapping specialist); (3) some data processing/management/programming experience, (but not as much as a fully experienced programmer, and it does not have to be with SAS data and SAS programming).

There is one exception to the statement that the developer does not write any SAS program code. It concerns defining SAS variable transformations/derivations. These are done using Expressions and the Expression Builder inside Data Integration Studio. Expressions are snippets of SAS program code. Examples include SAS statements that use the PUT function to convert the data type of a source variable and SAS statements that use the CASE statement (i.e., similar to the SELECT statement) to decode the values of a coded source variable. These code snippets can be written by the developer, or created using the Expression Builder that is quite similar to a menu-driven Query Builder.

There is also one role for a SAS programmer in the Data Integration Studio environment when using it for SDTM conversion. Occasionally, there are processes that you cannot define inside Data Integration Studio, or cannot define easily, with the native toolset. If so, a SAS programmer will program this process as a SAS program outside the Data Integration Studio environment. The program will be placed in the Data Integration Studio process library, and then engaged by the developer as he/she defines Data Integration Studio jobs. Such programs are called custom transformations or custom processes in the Data Integration Studio environment.

Our process also includes a step where we check our work done in Data Integration Studio. We are not validating the Data Integration Studio product because it has been validated by SAS. Rather, we are checking that we have produced correct SDTM datasets by following our validated work procedures correctly. Some key points:

- During our Data Integration Studio development work:
    - We have SOP's and Guidelines that we wrote for the Data Integration Studio developers' work, and we train all our developers in them as well as in general use of the Data Integration Studio application. We have a validated process for our developers to follow.
    - We use validated SDTM template datasets as the definition for the target SDTM datasets we produce. These templates are available by downloading the Excel spreadsheet with the SDTM model definition from the members-only area of the CDISC website. This model metadata can be used to build empty SAS datasets for the target SDTM domains. (Note: If you license Data Integration Studio, it is worth checking with SAS to see if these template datasets are now available to you from SAS).

- At the end of the Data Integration Studio development work:
    - SDTM-compliance checking – Here we check the structure, nomenclature, and standard content (i.e., controlled terminology) of the SDTM datasets. We do these checks programmatically.
    - Validation of target data against source data – Here we check that the right source data wound up in the right place in the right target dataset. In essence, an SDTM domain dataset built via conversion from source data might be SDTM-compliant and still have the wrong data values in it. You need to check the target data against the source data to insure complete correctness. This type of checking is harder to do programmatically, especially in our case because the source data changes from client study to client study. However, we are progressing on some automation of this type of checking.

CONCLUSION

The Clinical Data Interchange Standards Consortium (CDISC) has developed a standard for clinical and preclinical trials data. The standard is called the Study Data Tabulation Model (SDTM). The FDA has stated in eCTD guidance that they now would like to receive clinical and preclinical trials data from sponsors in the SDTM. And the industry expectation is that within a few years the FDA will require by regulation that sponsors submit their data in the SDTM. The industry is quickly adopting the CDISC SDTM. Many sponsors are already converting clinical trials data from active Marketing Application projects to the SDTM, many others are assessing their SDTM-readiness and currently planning their first SDTM conversions, and still others are exploring ways to integrate the SDTM standard into their full clinical data life cycle. The initial motivation for this is regulatory but there are many opportunities for processing efficiency, particularly surrounding data interchange among the various organizations involved in the clinical data life cycle.

The SDTM is comprised of nomenclature, content, and structure. This view of the SDTM is important to understanding where and when SDTM can and will be integrated into the various stages of the clinical data life cycle. The stages are: Collection, Processing, Storage, Analysis/Reporting, and Compilation/Submission. For many sponsors, for some time to come, SDTM will only be integrated into the last stage. This is because much time is needed to integrate a new data standard throughout the whole life cycle. Many sponsors will continue to handle their study data according to their current standard and do this "late-stage conversion" while they work in parallel on integrating SDTM as their new standard throughout their clinical data life cycle. The primary reasons for integrating SDTM fully into the life cycle are (1) if your whole process is based upon the SDTM data standard, you do not have to convert your data to SDTM at the end of the process, and (2) to fully capitalize on the significant opportunities for processing efficiency throughout the life cycle.

SDTM conversion projects are major undertakings, with significant work required in two top-level areas. The first area regards project/process setup, which includes process design, process validation, process documentation, staff resourcing and deployment, and staff training. The second area regards the ongoing conversion work once the project/process is established. The conversion work includes mapping specifications development and conversion program development and validation per study, for the many datasets per study, several/many studies per project, and the stream of projects over time. Data integration technology is designed to support the data conversion process. SAS Data Integration Studio is a member of this technology class.

Octagon Research Solutions is a consulting organization with deep domain knowledge regarding CDISC standards and their use in the industry. With its domain knowledge and Data Integration Studio, the company has designed an SDTM conversion process that has addressed the key challenges of conversion to SDTM, including:

- Handling all extraction scenarios, with source data residing in SAS datasets, database tables, and other sources
- Handling all required data transformations due to changes in nomenclature, content, and structure to achieve the SDTM standard
- Handling all the integration scenarios that arise because of the one-to-many and many-to-one relationships between source datasets and SDTM domain datasets
- Assigning subject-specific record sequence numbers and specifying relationships between records in parent SDTM domains and related child SDTM datasets (e.g., Comments and SUPPQUAL datasets), based upon those sequence numbers
- Table lookups to convert legacy data to SDTM controlled terminology and other sponsor-defined standard terminology
- Automated quality control checking for each SDTM dataset created

Data Integration Studio has been an important contributor to Octagon's success in its conversion project objectives. Some particular strengths of the product that have contributed to this success are:

- The conversion process specifications are expressed inside Data Integration Studio using the tool's visual design process. Thus, Octagon is able to primarily use clinical data management staff instead of SAS programmer staff for this work. This in turn has allowed Octagon to reduce the cost of data conversion services for clients and to scale the process more quickly as work volume has changed.
- In Octagon's process, users do not write programs but rather express specifications in the visual design process. The tool then generates the conversion programs from the mapping specifications. The SAS application and our work process are validated so the programs produced by the application do not need to be separately validated.
- Data Integration Studio works with the clinical data and with clinical metadata in the SAS metadata repository. Thus, the SDTM model metadata can be stored in the SAS metadata repository and then be used both to correctly define the target datasets and to then provide "truth" during automated validation of the final created datasets.
- Data Integration Studio allows reuse of workflows and SDTM metadata across studies.
- Data Integration Studio is extensible. Custom transformations can be written that are used as utility modules to easily accomplish particular data conversion and integration tasks that are encountered regularly.

Octagon's SDTM conversion process is centered on and has benefited from the use of Data Integration Studio. The company's process is today validated, stable, and robust. In the first year of running the process, Octagon converted over 60 studies covering multiple sponsors. Each of these studies was in a unique legacy form. Octagon built a department of 20 staff in its first year of SDTM conversion services, of which 12+ actively used the Data Integration Studio application. The company expects to convert between two and three times more studies in its second year, and increase the number of staff using the Data Integration Studio application in a corresponding fashion.

Nonetheless, Octagon's process is still young and evolving. The company expects to leverage this SAS technology further as it expands its conversion services in response to the pharmaceutical/biotechnical industry's steadily increasing adoption of the CDISC SDTM standard. The following are the primary focus right now of the next steps in this evolution:

- Extending the Data Integration Studio visual interface so the mapping specifications development, which is now done in Excel spreadsheets, can be done inside Data Integration Studio and be part of the DI Studio job specification work.
- Extending our use of Data Integration Studio to include loading the SDTM data produced to clinical data warehouses in addition to the present individual SDTM datasets.
- Developing additional custom transformations to further automate the dataset validation process.

REFERENCES

SAS Institute Inc. 2004. "SAS® 9.1.3 ETL Studio: User's Guide". Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2004. "Using SAS® ETL Studio to Integrate Your Data Course Notes". Cary, NC: SAS Institute Inc.

Susan Kenny and Michael Litzinger: "Strategies for Implementing SDTM and ADaM standards", Paper FC03 at PharmaSUG 2005, http://www.pharmasug.org/2005/FC03.pdf

ACKNOWLEDGEMENTS

I would like to thank David Evans, Chief Information Officer, Octagon Research Solutions, for his insights into the current state of adoption of SDTM in the industry and the likely paths and timeframes of future adoption. I would also like to thank Tom Guinter, Vice President of Clinical Data Strategies, Octagon Research Solutions, for his review of this paper and his insights into the evolving role that Data Integration Studio can play in the SDTM conversion process.

CONTACT INFORMATION

Barry R. Cohen
Director, Clinical Data Strategies
Octagon Research Solutions, Inc.
Wayne, PA 19087
610 535 6500 x635
bcohen@octagonresearch.com
www.octagonresearch.com