

Paper 179-2010

Managing Clinical Data Standards: An Introduction to SAS® Clinical Data Integration

Michael Kilhullen, SAS Institute Inc., Cary, NC

ABSTRACT

As the adoption of industry data standards grows, organizations must consider how to effectively implement and manage data standards across a large—often global—user base. SAS® Clinical Data Integration is built upon a centralized metadata repository that is ideal for centrally deploying and managing data standards. Built on SAS® data integration technology, it provides a visual and metadata-driven environment that facilitates the conversion of data to standard formats while collecting more detailed information about the decisions you make during this process. SAS Clinical Data Integration provides specialized interfaces that further leverage the metadata to help you more effectively work with and manage clinical standards. This paper will present SAS Clinical Data Integration features, including how to import, customize, and manage standards metadata, how to monitor and analyze how users consume metadata during the clinical development process, and how metadata is leveraged to provide consistency and reusability across your organization.

INTRODUCTION

SAS Clinical Data Integration 2.1 is a new product offering from SAS that focuses on pharmaceutical industry needs for transforming, managing, and verifying the creation of industry mandated data standards such as the Clinical Data Interchange Standards Consortium (CDISC). The product relies on SAS® Data Integration to provide centralized metadata management using the SAS® Metadata Server and the tools to visually transform data. SAS Clinical Data Integration enhances usability by adding new metadata types, plug-ins, and wizards that assist with clinically oriented tasks such as importing data standards, creating studies and submissions, and adding specialized transformations for transforming clinical data to a standard data model. It also leverages the SAS® Clinical Standards Toolkit to provide validation and conformance checking.

UNDERSTANDING THE CLINICAL METADATA ENVIRONMENT

SAS Clinical Data Integration facilitates the collection and management of metadata across your organization (Figure 1). Most metadata is generated by clinical programmers while they define how the operational data is converted to a data standard. This is a very repeatable process but often needs to be adjusted based on variations in the way that

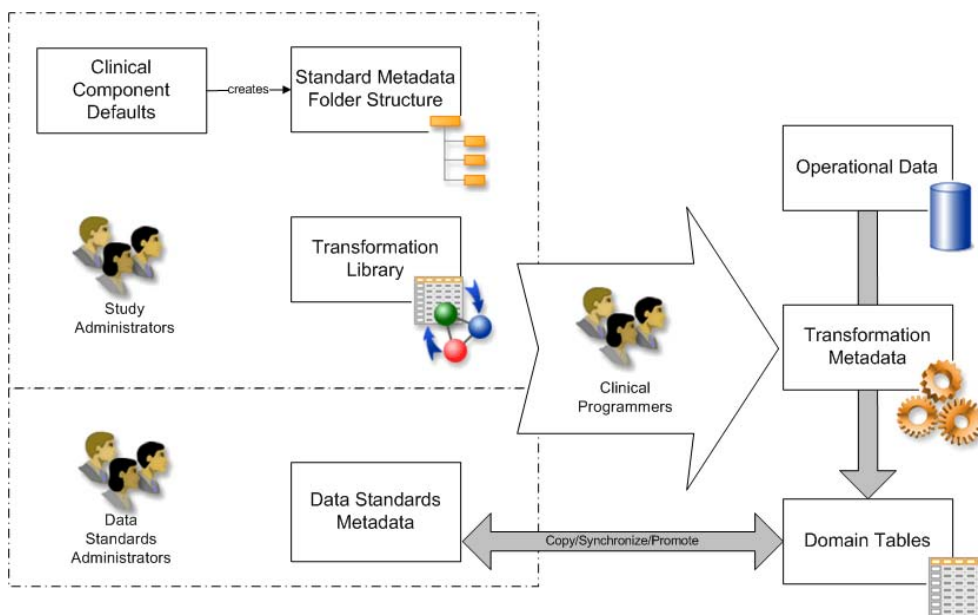


Figure 1: Typical Metadata Collected during a Clinical Trial

data is collected and analyzed between studies. As multiple studies are processed, the study administrators must ensure that the clinical programmers are creating domain tables that comply with a data standard. As new custom domains are defined, or existing domains are modified, the data standards administrator must be able to understand, evaluate, and potentially apply the changes to the centralized data standards. Finally, a study administrator is likely overseeing several studies that might have different clinical programmers assigned. It is important to maintain a consistent work environment, especially if clinical programmers need to work on studies simultaneously, or transition to a new study more quickly. Moreover, because metadata is being collected at all stages of the process, managers can use metadata to monitor the transformation process and generate reports.

In SAS Clinical Data Integration, clinical metadata is stored in the metadata server as properties of metadata objects. Most metadata is related to creating standardized data and is stored as part of the table and column metadata objects. In some cases, new metadata objects are created to specifically address the business needs of managing standards. For example, studies and submissions have their own metadata representation in the SAS Metadata Server that is based on a common metadata object called a clinical component (Figure 2). In addition to having specialized metadata about the study or submission, clinical metadata objects they also catalog the metadata contents (such as jobs, reports, tables, and so on) created by users, define the versions of standards that are allowed for a study or submission, and enable additional processes such as importing, exporting, and archiving. SAS Clinical Data Integration ensures that data standard domains are implemented only as part of a data standard, study, or submission. When data standard domains are implemented in a data standard, we refer to these as *templates* because they are used as the basis for the actual data and metadata collected by a study or submission.

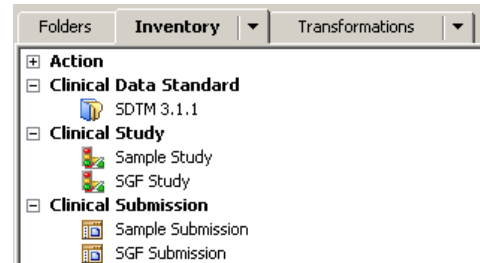


Figure 2: Clinical Metadata Types

MAINTAINING CONSISTENT METADATA CONTENT

SAS Clinical Data Integration can be configured by an administrator to create default metadata content when clinical components are created (Figure 3). This allows you to maintain consistent content between studies or submissions. For example, the administrator can define a standard metadata folder structure for studies. When the new study wizard finishes, it creates the metadata folders (Figure 4). The administrator can also define standard libraries to

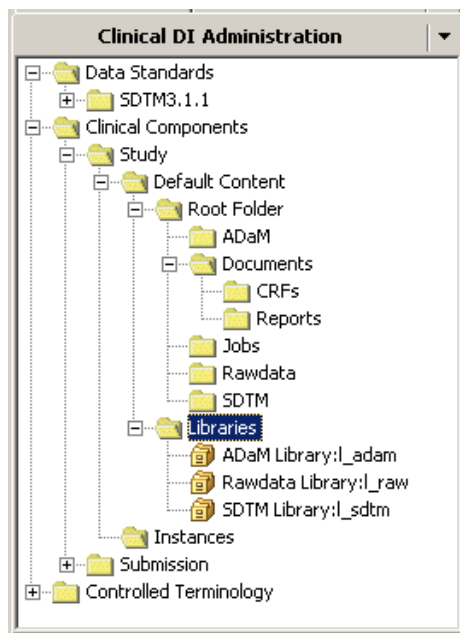


Figure 3: Default Content Defined for Studies

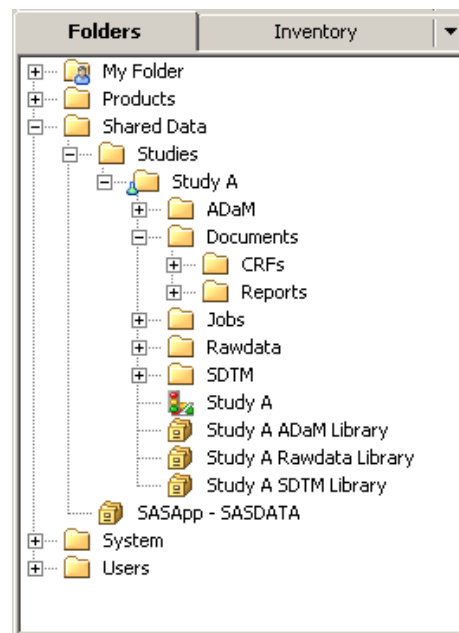


Figure 4: Content Created for a New Study Based on Defaults

ensure that the correct SAS library statements are generated. This scenario is useful when your organization implements standard reporting macros that depend on specific library references to gain access to data and store results. Notice in Figure 4 that the library reference names are updated to include the study name to promote better usability when you need to select a library from a list of all library objects.

In addition to default content, SAS Clinical Data Integration also helps manage the clinical content within a folder. As a best practice, a folder should contain tables with the same type of data, and typically should use the same SAS library for physical storage. This means that you should not find CDISC Study Data Tabulation Model (SDTM) tables mixed with CDISC Analysis Data Model (ADaM) tables in the same folder. (Figure 4 reflects this best practice). SAS Clinical Data Integration plug-ins enforce this rule by defaulting selection windows based on existing content within the selected folder. For example, once a domain is created using the SDTM 3.1.1 data standard, only new domains based on the same standard version can be added to the folder.

WORKING WITH DATA STANDARD METADATA

Data standards are managed exclusively in the SAS Metadata Server. Therefore, every user of SAS Clinical Data Integration has access to the same domain templates, validation checks, and terminology. All data standard related plug-ins are designed to use the centralized metadata and help simplify creating domains during study transformations (Figure 5). Clinical programmers do not need to figure out where the standards are stored or which standards to use for a particular study. Instead, the study manager defines these settings as part of the study

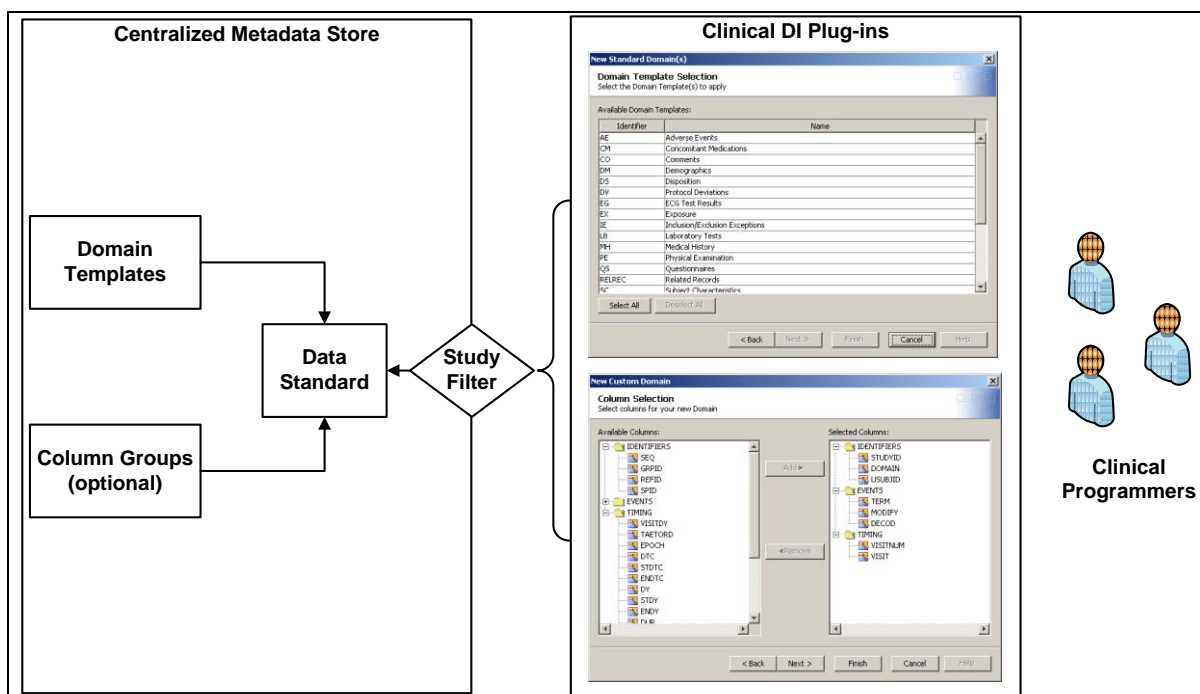


Figure 5: Example of Leveraging Centralized Data Standards with Plug-ins

definition, and the plug-ins ensure that only applicable standards are displayed to programmers. For example, if a Study Manager defines the study to use SDTM version 3.1.2, the SAS Clinical Data Integration plug-ins will *not* display any other SDTM versions.

Accessing the domain or domain template metadata is the same for administrators and programmers. Clinical metadata is accessible through standard SAS® Data Integration Studio dialog boxes that are enhanced by SAS Clinical Data Integration plug-ins. Additional tabs are added where the clinical metadata is displayed (Figure 6). If you have edit permissions, then you can change the values of the clinical metadata. Changes are considered study or submission specific. That is, changes do not propagate back to the data standard template. The same is true of changes to data standard templates; they do not propagate to registered domains. Other tools are available in SAS Clinical Data Integration to compare domains to templates and selectively apply changes if needed.

Clinical properties are displayed on property tabs as a table (Figure 7). Depending on how properties are defined, they

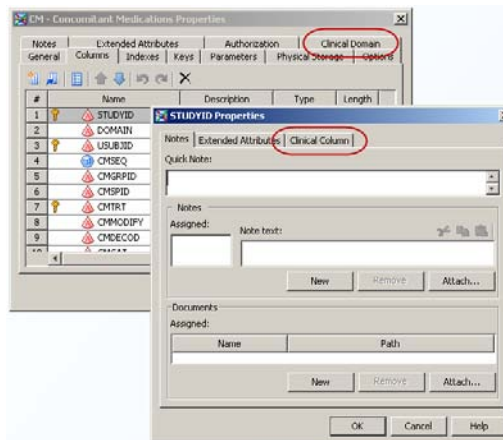


Figure 6: Table and Column Clinical Properties

might be displayed as a text or a selection list. Properties will have default settings, but you can adjust them depending on how you use the standards in your organization.

| Name | Value |
|-------------------|------------------------|
| Archive | ..ydm.xpt |
| Archive Title | |
| Classification | Special Purpose |
| Domain Identifier | DM |
| Is Custom | False |
| Purpose | Tabulation |
| Structure | One record per subject |

Figure 7: Clinical Properties Tab

SAS Clinical Data Integration supports a common property model for data standards, studies, submissions, and domains. The property model is largely based on the SDTM. However, it has been generalized so that you can add your own modified SDTM implementations or internal data standards. The properties collected for clinical objects are summarized in Table 1. In addition, you can add custom properties and notes to any metadata object.

| Clinical Metadata Type | Additional Properties Managed by SAS Clinical Data Integration |
|------------------------|--|
| Data Standard | Formal name, version, type, base standard type, vendor, supports toolkit validation |
| Study | Study title, short study title, phase |
| Submission | Submission title, short submission title |
| Domain | Identifier, purpose, classification, structure, archive location, archive title |
| Domain column | Term, origin, role, core, display format, qualifiers, XML type, XML codelist, algorithm, whether the column contributes to the key |

Table 1: Additional Properties Available for Clinical Objects

CONTROLLING HOW PROPERTIES ARE USED

While the CDISC data standards dictate what you need to collect, and in some cases what a property should contain, how your company uses a data standard will vary depending on how well the data standard integrates into your established business processes. In some cases, the values are strictly implemented. That is, clinical programmers can select only from a list of allowable values. If a value is not found, you need to contact the data standards administrator to add the content. When the data standards administrator adds the value, you can then edit the properties and select the value. Other companies prefer a less strict approach. If a value is not available in a list, you can type in a new value. The value you provide is considered study specific and does not propagate to the data standard. Rather, the data standards administrator periodically evaluates the values used for a property across many studies and adjusts the standard values accordingly. Other companies prefer to simply have the user type in values without a list. Whichever approach fits your business needs, SAS Clinical Data Integration provides the flexibility to allow you to define how a property is collected, including whether the property is required, dependent on a lookup list, customizable by the user, and so on (Figure 8).

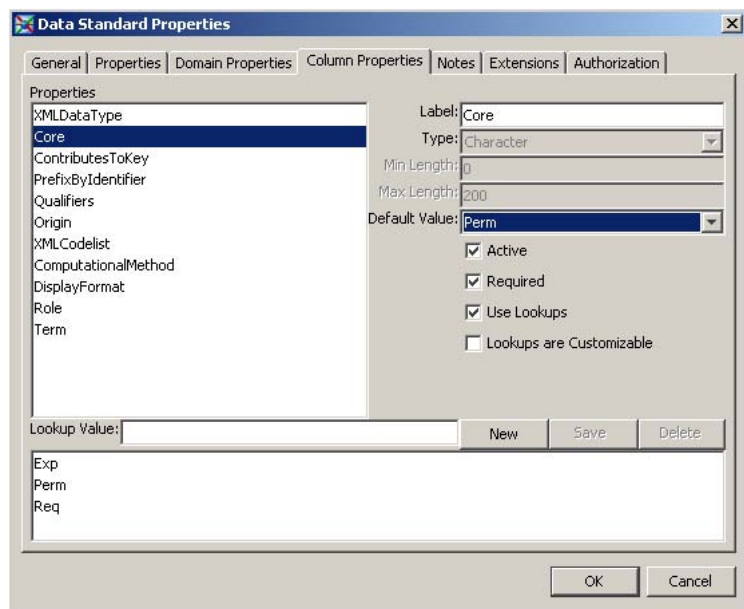


Figure 8: Column Property Model Editor

CREATING CUSTOM DOMAINS

Certain CDISC standards, such as SDTM, support user-defined domains, which are domains needed for clinical data that is not defined as a standard domain in the SDTM implementation guide. In this case, the underlying data model is used to assemble the new domain (Figure 9). Data models are generalized into *column groups* that can support both industry standards and vendor-specific data standards. Column groups are simply a logical grouping of columns that are combined and organized into a domain. The columns within a column group will also contain default metadata values, which will be copied to the new domain when created.

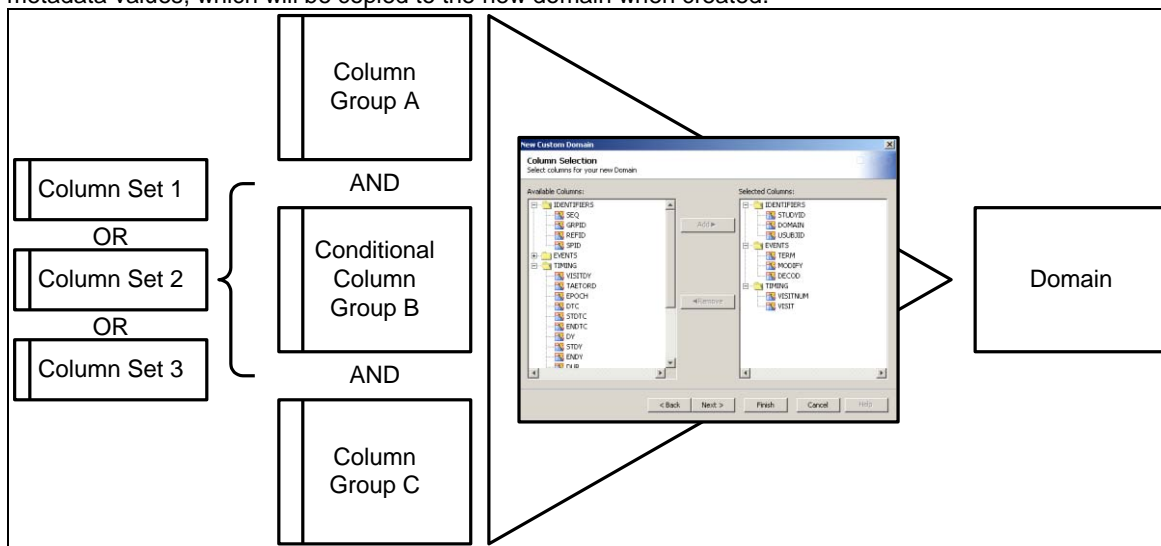


Figure 9: Assembling Column Groups into a Domain

Column groups can also be considered conditional. That is, columns can be selected from one column set within the conditional column group but not another column group. This would be the case for SDTM where a domain can contain only intervention, event, or finding columns, in addition to identifier and timing columns.

If a data standard uses an underlying model, SAS Clinical Data Integration provides a wizard to guide you in defining the custom domain. Data standard administrators can use the same wizard to design new domain templates. Alternatively, they can identify and promote custom domains defined by users in studies and submissions to the data standard. Finally, you might know of a custom domain created for another study that has not been promoted to a data standard. Rather than recreate it, SAS Clinical Data Integration can copy it as long as it is based on the same

data standard version used by your study. In doing this, metadata settings are preserved and the necessary metadata relationships to the new study are automatically adjusted.

SYNCHRONIZING DOMAIN METADATA

When domains are created in a study or submission, the metadata stored in the data standard templates are copied. This allows users to modify the metadata according to the study requirements without impacting the metadata that others might be using for other studies. This, of course, means that clinical programmers can change any attribute of the domain, including labels, lengths, formats, types, and column order. Furthermore, after the domain is copied, the data standards administrator might change the template metadata based on trends in how it was used in other studies. For example, it is common to adjust character column lengths to avoid truncation in a study that collected more text. If the standard administrators find that programmers are frequently changing the length, they can change the length in the domain template so that manual adjustments are no longer

necessary. SAS Clinical Data Integration provides a refresh domain plug-in that compares a domain with the data standard template used to create it (Figure 10). It will identify differences found and allow the clinical programmer to choose which changes to apply to the domain.

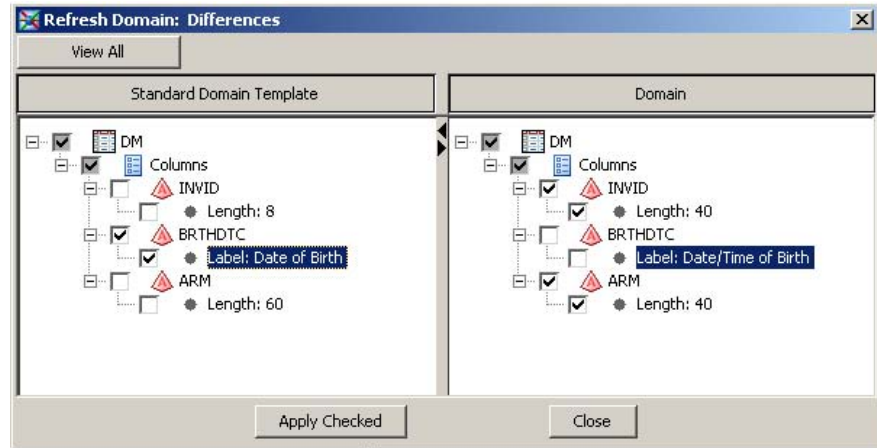


Figure 10: Refresh Domain Metadata

ANALYZING THE USE OF CLINICAL DATA STANDARDS

As data standards are consumed by your organization, a data standards administrator can monitor how the data standards are being used. SAS Clinical Data Integration makes this easy because of the wealth of metadata collected about standards and how they are used. The data standards administrator can select from a list of clinical components that are based on the data standard selected. The metadata about the domains created are compared and displayed in a table that shows you what columns were used in the domains across the clinical components. This will be useful when clinical programmers are consistently adding new columns and you need to make a decision about adding one to the domain template. Finally, custom domains are also displayed in the comparison so that you can see which custom domains are being added frequently to clinical components. In this case, the data standards administrator can choose to promote the custom domain. This will add the metadata for the selected custom domain to the data standard, making it easier for other programmers to create it.

LEVERAGING THE SAS CLINICAL STANDARDS TOOLKIT

The SAS Clinical Standards Toolkit is a SAS macro approach to supporting clinical data standards in Base SAS®. It supports defining data standard domains, conversion of domains between CDISC models, and validation and conformance checks. It provides periodic updates when new standards and new versions of standards are released. Once the updates are applied to the toolkit, they are

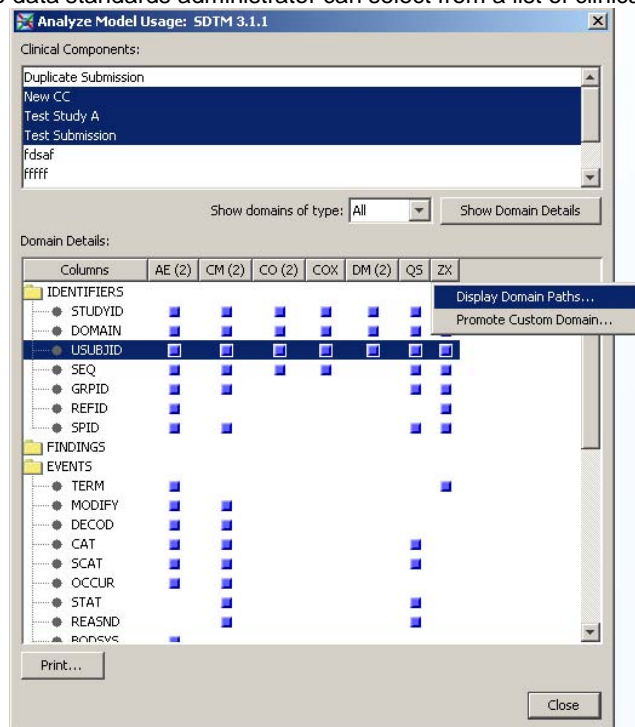


Figure 11: Analyzing How Standards Are Being Used

automatically detected by SAS Clinical Data Integration during the import and validation processes.

SAS Clinical Data Integration greatly simplifies the use of the SAS Clinical Standards Toolkit. First, it assumes control of all clinical metadata. This means that once metadata is imported from the SAS Clinical Standards Toolkit, the powerful features found in SAS Clinical Data Integration are used to manage the data standard definitions. When the SAS Clinical Standards Toolkit macros are used in the transformation process, the metadata needed by the SAS Clinical Standards Toolkit is automatically exported and restructured for execution. Secondly, SAS Clinical Data Integration adds graphical user interfaces to SAS Clinical Standards Toolkit macros. This adds a metadata-driven approach to defining SAS Clinical Standards Toolkit tasks. Finally, validation checks and terminology are also imported into the metadata server and are displayed as manageable objects in the clinical administration interfaces.

IMPORTING DATA STANDARDS METADATA

SAS Clinical Data Integration provides a Data Standards Metadata Import Wizard to help data standards administrators select and load metadata. This is a one-time process per model version; once metadata loaded, the SAS Metadata Server manages changes and additions to the data standards. The wizard prompts you to select the standard and version from the toolkit, displays the metadata content in detail for verification, and then imports the metadata (Figure 12). Once imported, the standard and domain templates are surfaced through SAS Clinical Data Integration Server plug-ins. Custom versions of standards can also be imported from the SAS Clinical Standards Toolkit. Assuming that the custom data standard is registered to SAS Clinical Standards Toolkit, when the metadata import wizard is run, the custom data standard will automatically appear in the data standard selection lists.

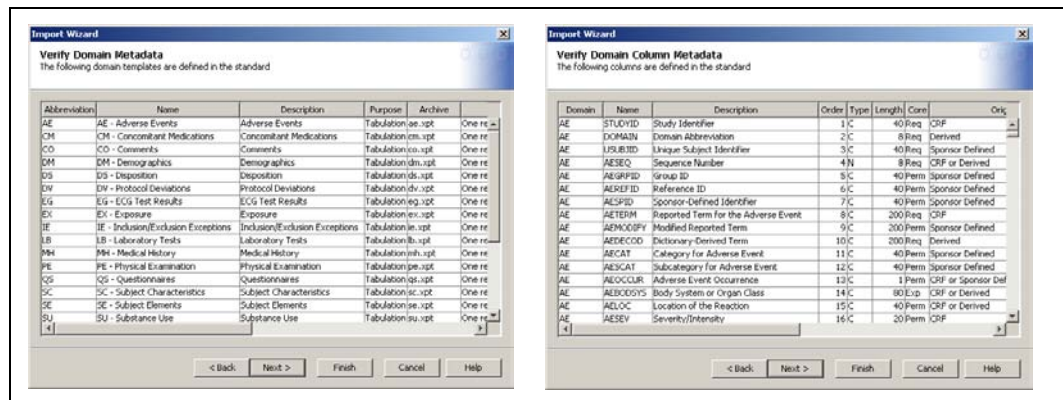


Figure 12: Sample Displays from the Metadata Importer

VALIDATING DOMAIN CONTENT AND STRUCTURE

As data standard domains are implemented in studies and submissions, the structure and content can vary from the data standard. You must periodically verify that the domains maintain conformance to the data standard. SAS Clinical Data Integration provides a transformation to run validation checks and generate reports through the SAS Clinical Standards Toolkit. The SAS Clinical Standards Toolkit provides 143 unique SDTM 3.1.1 validation checks. These checks are derived from three sources: the CDISC-SDTM WebSDM™ documented checks, checks supporting loads into the Janus study data repository, and checks added by SAS which are based on data management and cleansing experiences building CDISC-SDTM domains using SAS products and solutions. The validation checks are designed to enable an assessment of the consistency of data values within a specific column, between columns, across records within a specific data set, and across data sets. In addition to the provided validation checks, the data standards administrators can create their own customized compliance checks using the Manage Compliance Checks dialog box (Figure 13).

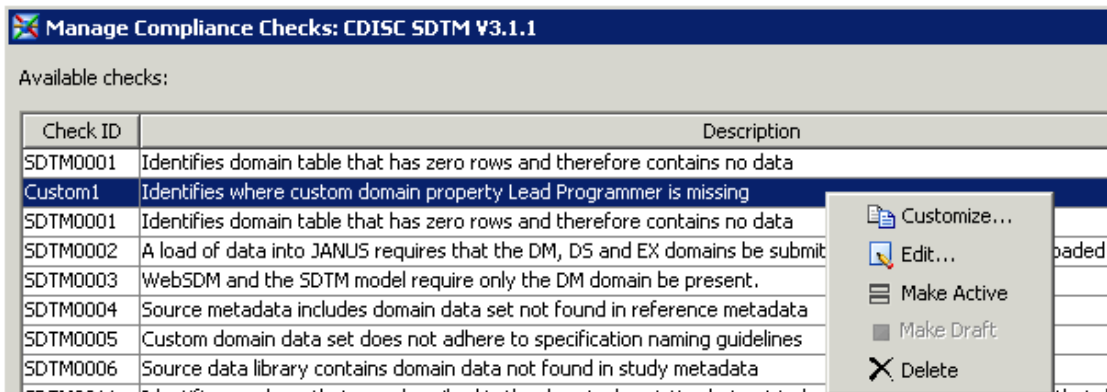


Figure 13: Customizing Compliance Checks

The CDISC-SDTM Compliance transformation is provided for use in building validation jobs. You specify the data standard you wish to validate against, and select the domains to be assessed and the set of compliance checks you wish to run (Figure 14). The model compliance transformation allows you to add as many checks as desired. It also offers filtering capabilities to help find the necessary checks.

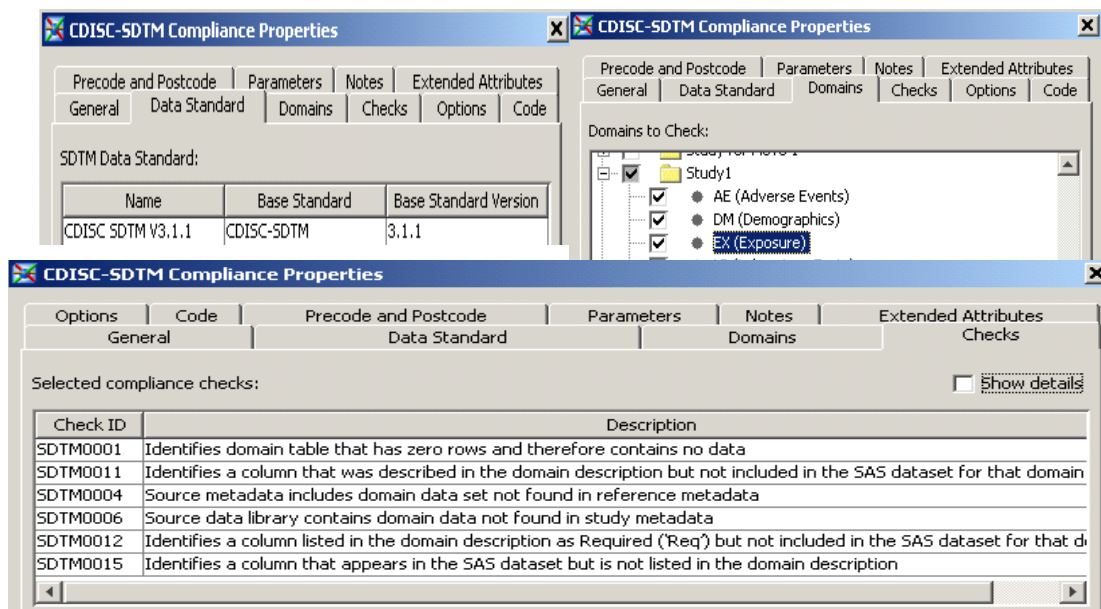


Figure 14: Examples of the CDISC-SDTM Compliance Transformation

After the job runs, two data sets are produced. The *results* data set contains the findings of the compliance assessment and the *metrics* data set contains summary statistics on the validation process (Figure 15).

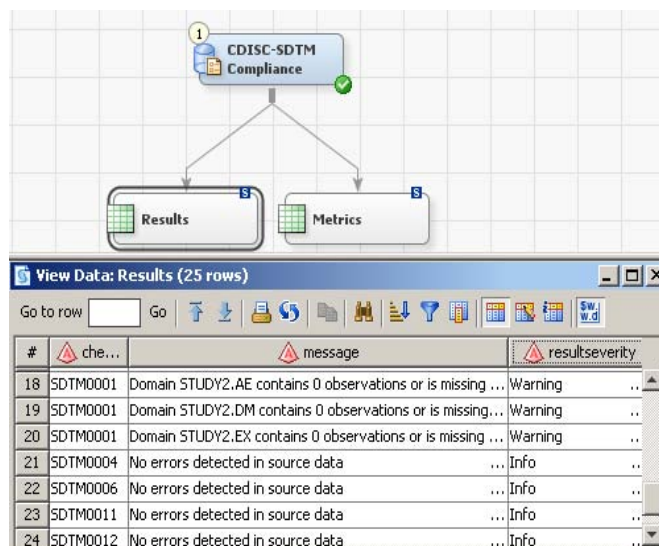


Figure 15: CDISC-SDTM Compliance Transformation and Results Data Set

GENERATING CRT-DDS

The metadata managed by SAS Clinical Data Integration can be published to CRT-DDS using the CRT-DDS Transformation (Figure 16). This transformation extracts metadata and passes it on to the SAS Clinical Data Standards Toolkit for define.xml creation. The transformation allows you to specify properties to control encoding, log level processing, and an output style sheet (Figure 17).

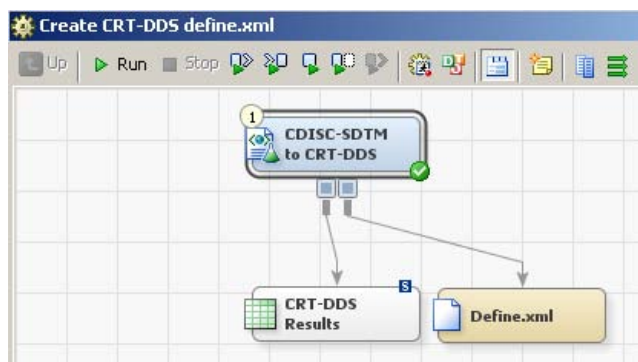


Figure 16: CRT-DDS Define.xml Process Flow

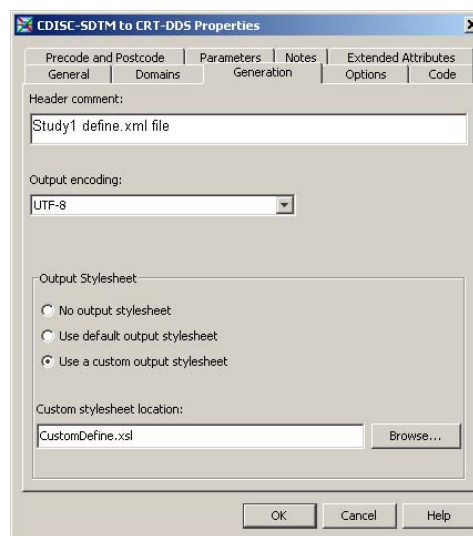


Figure 17: CDISC-SDTM to CRT-DDS Transformation Properties

After the job runs, a define.xml file and a results data set are produced. The Results data set documents the results of the generation of the CRT-DDS file. The define.xml file contains summary information about each of the domain data sets, detailed information about each column in each domain, and code list information (Figure 18).

| Datasets for Study Study1 | | | | | | | |
|---------------------------|----------------------------------|--|-----------------------------------|-----------------|-----------------------------------|-----------------|-----------------------------|
| Dataset | Description | Structure | Purpose | Keys | | | |
| DM | Demographics | Special Purpose - One record per subject | Tabulation | STUDYID USUBJID | | | |
| AE | Adverse Events | Events - One record per event | Tabulation | STUDYID USUBJID | | | |
| EX | Exposure | Demographics Dataset (DM) | | | | | |
| LB | Laboratory Tests | Controlled Terminology (Code Lists) Section | | | | | |
| | | Variable | Label | Type | Controlled Terms or Format | Origin | Role |
| | | STUDYID | Study Identifier | text | | CRF | Identifier |
| | | DOMAIN | Domain Abbreviation | text | | Derived | Identifier |
| | | USUBJID | Unique Subject Identifier | text | | Sponsor Defined | Identifier |
| | | SUBJID | Subject Identifier for the Study | text | | CRF | Topic |
| | | RFSTDTTC | Subject Reference Start Date/Time | text | | Sponsor Defined | |
| | | RFENDTTC | Subject Reference End Date/Time | text | | Sponsor Defined | |
| | | SITEID | Study Site Identifier | text | | CRF or Derived | |
| | | INVID | Investigator Identifier | text | | CRF or Derived | |
| | | INVNAM | Investigator Name | text | | CRF or Derived | |
| | | BRTHDTC | Date/Time of Birth | text | | CRF or Derived | |
| | | AGE | Age in AGEU at RFSTDTTC | float | | CRF or Derived | |
| | | AGEU | Age Units | text | 1 | CRF or Derived | |
| | | SEX | Sex | text | 19 | CRF | |
| | | | | | | | Code Value |
| | | | | | | | Code Text |
| | | | | | | | AGEU, Reference Name (1) |
| | | | | | | | DAYS |
| | | | | | | | HOURS |
| | | | | | | | MONTHS |
| | | | | | | | WEEKS |
| | | | | | | | YEARS |
| | | | | | | | COUNTRY, Reference Name (2) |
| | | | | | | | ABW |
| | | | | | | | AFG |
| | | | | | | | AGO |
| | | | | | | | ATA |

Figure 18: Examples of CRT-DDS

CONCLUSION

This paper has shown several key features of SAS Clinical Data Integration 2.1 related to implementing and managing CDISC standards. By centrally collecting and managing metadata, it can be used to automate setup and transformation processes, reuse metadata objects to expedite data standardization, feed validation and conformance checking, and improve the administration, consistency, and use of standards within an organization.

RECOMMENDED READING

- Hunley, Eric, Gary Mehler, and Nancy Rausch. 2009. "What's New in SAS® Data Integration Studio 4.2" *Proceedings of the SAS Global Forum 2009 Conference*. Cary, NC: SAS Institute Inc.
- Villiers, Peter. 2009. "Supporting CDISC Standards in Base SAS Using the SAS Clinical Standards Toolkit." *Proceedings of the SAS Global Forum 2009 Conference*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Michael Kilhullen
 SAS Institute Inc.
 908-760-6528
michael.kilhullen@sas.com
<http://www.sas.com>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.