PhUSE 2012

Paper PP24

Controlling Controlled Terminology

Ryan Burns, Rho, Inc., Chapel Hill, USA

ABSTRACT

Standards, metadata, and controlled terminology are all important and necessary parts of drug development, especially when it comes to submissions to regulatory agencies. Having a suite of tools and procedures to enter, store, and use metadata is critical for data standards teams to operate efficiently and reliably. When creating a metadata system, one must consider both clinical and database concepts. Difficulties arise when creating a database that integrates controlled terminology with other metadata, because controlled terminology has unique requirements. This paper will discuss solutions for the challenges of managing controlled terminology as a part of data standardization and will do so within the context of an enterprise metadata system.

INTRODUCTION

Controlled terminology is a shared, standard vocabulary and is an important part of submission projects and good database design. It can come from—and be shared by—a variety of groups including standard-setting bodies, such as CDISC or ISO; pharmaceutical companies; CROs; and study teams. Controlled terminology not only has various sources, but also is used during various stages of clinical trials: terminology should be considered when creating CRFs; as part of the clinical, SDTM, and analysis databases; and when creating displays. Managing controlled terminology can be challenging because it must be consistent regardless of its source and where it is used.

There is more to controlled terminology than a simple list of terms. For example, controlled terminology lists from the National Cancer Institute (NCI) of the United States of America contain close to a dozen supplemental attributes, including descriptions, definitions, and identifying codes. Other attributes such as whether a list is extensible should not only be maintained, but are also needed for management of the list. Some attributes are constant for a given term or list while others vary from study to study. One of the principal challenges with managing a controlled terminology database is that it must be integrated with study-specific metadata. This paper will explore how controlled terminology was integrated into a larger metadata system at a CRO.

ENTERPRISE METADATA SYSTEM

Having and following data standards is a necessary aspect of clinical trials—for good reason since doing so allows investigators and reviewers to better answer key questions of efficacy and safety. Having a system in place to manage standards and metadata is crucial. At Rho, a CRO, we have a tool that serves as a metadata database and speccing tool called the Metaspecer.

METASPECER BASICS

The Metaspecer was created originally so that specifications created by statisticians could be machine readable. It's currently used to create dataset specifications, variable attributes, and define files. These three outputs are created by standard SAS® programs that read data users have entered in the Metaspecer. Having specifications, variable attributes, and define files created from a single source results in a more efficient and reliable process than would otherwise be possible.

In its current form the Metaspecer is comprised of an Oracle database in the background and a Microsoft Access user interface. The user interface relies mostly on tabular forms that look similar to Excel worksheets. The main forms are one for dataset-level metadata, one for variable-level metadata, and one for value-level metadata. In addition to the main forms, the interface has several other features including menus, navigation shortcuts, pull down lists, and pop-up auxiliary forms. An example auxiliary form is one that selects terms used from a specified controlled terminology list. There are several other components of the interface aimed at aiding users, including integrity constraints, import wizards, and an automatic audit trail.

METASPECER DATABASE

While only one study at a time can be viewed and changed via the user interface, the back-end Oracle database has all studies' metadata stored together. The database is made up of one table for study-level metadata, two additional tables for analysis databases (dataset- and variable-level metadata), and three additional tables for SDTM databases (domain-, variable-, and value-level metadata). There are also three tables dedicated to controlled terminology that are discussed in detail later in this paper.

The database was designed to be minimally restrictive, leaving most of the constraining to the user interface. This was done to be flexible, allowing restraints to be altered based on specific study requirements and easily adaptable for changing industry standards. Further, as a CRO, we receive all types of requests and odd data, so we have designed a system that will permit metadata we have not yet even dreamed of to be stored if need be.

Although different tables are used, SDTM and analysis/non-SDTM databases are handled similarly. This paper will focus on analysis metadata. The key tables are detailed below.

- Studies: Study-level metadata
  - Descriptive information, for example, sponsor, study title, and investigational product
  - Internal reference information, including network location, study ID, and creation data
  - Processing information, such as what fields are needed in the dataset and variable tables, standards with which the study should comply, and linked studies

Column Name	Data Type	Comment
IDStudy	NUMBER(10,0)	Primary key, automatically assigned
StudyNum	VARCHAR2(255 BYTE)	Study number for display
StudyTitle	VARCHAR2(255 BYTE)	
Sponsor	VARCHAR2(100 BYTE)	
StudyCreationDatetime	DATE	
ProjRootDir	VARCHAR2(255 BYTE)	
FieldsUsedVariablesTable	CLOB	List of fields from Variables table to display
FieldsUsedDatasetsTable	CLOB	List of fields from Datasets table to display
Lblvars_Custom_Txt1	VARCHAR2(32 BYTE)	Label for custom field in Variables table
Lbldsns_Custom_Txt1	VARCHAR2(32 BYTE)	Label for custom field in Variables table
RelStuds	VARCHAR2(255 BYTE)	List of IDStudy values for related studies

- Datasets: Dataset-level metadata
  - All information about datasets needed in a define file
  - General programming instructions
  - Any other dataset-related metadata

Column Name	Data Type	Comment
IDDatasetName	NUMBER(10,0)	Primary key, automatically assigned
IDStudy	NUMBER(10,0)	Study dataset belongs to
DatasetName	VARCHAR2(255 BYTE)	
DSLabel	VARCHAR2(40 BYTE)	
KeyFields	VARCHAR2(255 BYTE)	
Structure	VARCHAR2(255 BYTE)	
TypeDataset	VARCHAR2(255 BYTE)	Useful when raw and analysis datasets are present
Description	CLOB	
Input_Notes	CLOB	Instructions to programmers
DSSortOrder	FLOAT	Order displayed in interface and outputs
Class	VARCHAR2(40 BYTE)	
SubmitDB	VARCHAR2(3 BYTE)	Used to exclude datasets without deleting metadata
Custom_Txt1	VARCHAR2(255 BYTE)	Custom field as defined in STUDIES table

- Variables: Variable-level metadata
    - Instructions and attributes needed for programming
    - All information needed about variables in a define file
    - Any other variable-related metadata

Column Name	Data Type	Comment
IDVariables	NUMBER(10,0)	Primary key, automatically assigned
IDStudy	NUMBER(10,0)	Study variable belongs to
IDDatasetName	NUMBER(10,0)	Dataset variable belongs to
Name8	VARCHAR2(32 BYTE)	
Label8	VARCHAR2(255 BYTE)	
Definition	CLOB	Programming definition
Type	VARCHAR2(32 BYTE)	SAS variable type
Length	NUMBER(10,0)	
Origin	VARCHAR2(255 BYTE)	
FDADefinition	CLOB	Normally used for comment field of define file
ODMType	VARCHAR2(40 BYTE)	
IDCTNames	NUMBER(10,0)	Controlled Terminology list used
Custom_Txt1	VARCHAR2(255 BYTE)	

CONTROLLED TERMINOLOGY

Incorporating controlled terminology into the Metaspecer was a challenge for several reasons. The terminology needs to be both fixed and flexible: some lists are extensible others are not, the input codes in code lists can vary while the resulting output terminology should be consistent, and the terms used from within a given list are not the same in all studies. Controlled terminology also has to be associated with variables, but not dependent on them—a list must be the same for all variables that use it. Fundamental database concepts including normalization, lack of redundancy, and storage capacity should be considered, for example, having a set of all available controlled terminology repeated for each project is not a good option. Our goal was to create a system to manage controlled terminology across all projects and sponsors that would comply with industry standards, would aid programmers, and would be easy for those creating specifications to use.

DATABASE

Three tables were added to the Metaspecer database to manage controlled terminology. One table stores information about lists as a whole. The second table stores information about each term. The third table stores information that is study-dependent. The first two tables store the list and term information for all studies, so to keep the database from growing unmanageably large, there are no duplicates. Any information that is specific to a study, such as what terms to select from a list, is stored in the third table in a way that minimizes storage space.

- CTNames: List-level metadata
    - List properties, including those assigned by NCI and CDISC
    - Processing information, such as whether a list is extensible or created by the user
    - List name

Column Name	Data Type	Comment
IDCTNames	NUMBER(10,0)	Primary key, automatically assigned
ListName	VARCHAR2(60 BYTE)	
NCIListCode	VARCHAR2(6 BYTE)	
ListVersion	VARCHAR2(50 BYTE)	
ListCustom	NUMBER(1,0)	Indicates user-created list
ListDescription	VARCHAR2(255 BYTE)	
Extensible	VARCHAR2(1 BYTE)	Indicates if a list is extensible
NCIType	VARCHAR2(10 BYTE)	
ListNCIPrefTerm	VARCHAR2(100 BYTE)	
ListPrefTerm	VARCHAR2(100 BYTE)	
ListSynonym	VARCHAR2(255 BYTE)	
FmtName	VARCHAR2(32 BYTE)	

- CTValues: Term-level metadata
  - Term properties, including those assigned by NCI and CDISC
  - Input codes and sort order when there is an inherent order to terms within a list
  - The terms

Column Name	Data Type	Comment
IDCTValues	NUMBER(10,0)	Primary key, automatically assigned
IDCTNames	NUMBER(10,0)	List term belongs to
NCICode	VARCHAR2(6 BYTE)	
CvalValue	VARCHAR2(255 BYTE)	Term
CvalCode	VARCHAR2(40 BYTE)	Input code*
CvalSortOrder	NUMBER	Sort order if needed*
CvalCustom	NUMBER(1,0)	Indicates user-created term
CvalDefinition	VARCHAR2(255 BYTE)	
CvalNCIPrefTerm	VARCHAR2(100 BYTE)	
CvalPrefTerm	VARCHAR2(100 BYTE)	

* These are default values that can be overwritten in the CTUsed table.

- CTUsed: Study specific term-level metadata
  - All terms that are selected by study
  - Study-specific input codes and sort order

Column Name	Data Type	Comment
IDCTUsed	NUMBER(10,0)	Primary key, automatically assigned
IDStudy	NUMBER(10,0)	Study row/term belongs to
IDCTValues	NUMBER(10,0)	Selected term
UsedCode	VARCHAR2(40 BYTE)	Input code, overwriting default
UsedSort	NUMBER	Sort order, overwriting default

USAGE

Having a place to store controlled terminology and associated metadata is only a start. Populating the database and integrating it with a larger metadata system is necessary to have a useful way of handling controlled terminology. Initially populating the database is done by copying data from NCI, other external sources, or existing internal lists to the CTNames and CTValues tables. All fields except for IDs, list name, and term in those two tables can be left blank. Additionally, fields can be given custom values, for example, CTNames.ListCustom can be 0 for NCI lists, 1 for corporate standard lists, and 2 for any user-created lists. If any supplemental information is required, such as who added a list to the database and when, additional fields could be added.

Integrating the controlled terminology database with the Metaspecer is more a challenge. Users should be able to assign lists to variables, create new lists, create new terms, modify existing user-created terms, select and deselect terms to be used from within a list, and change input codes and sort order. All of these actions will be done using the Metaspecer form that is used for variable metadata, as well as controlled-terminology specific forms that can be invoked from the Variables form.

Assigning a controlled terminology list to a variable is done with a pull-down list in the Variables form. When a list is selected, the ID value of that list will be stored in the Variables table. The pull-down list is populated with the names of all standard lists, any lists that have been created for the current study, and any lists that have been created for any related studies (as defined in the Studies table). In addition to list names there is also an option for "New", which is selected to create a new list. When a user selects "New" a form will pop up allowing the user to enter the new list name and other attributes. The only restriction is that the name of the new list cannot be the same as the name of any standard list or list used in the current or a related study.

All other actions are performed in a controlled terminology form that is opened from the Variables form, specifically from the cell that contains the name of the controlled terminology list associated with a variable. The form displays attributes of the selected list in the header; has a row for each term within the list; and has columns for term, input code, sort order, and a check box indicating if the term is selected for use or not. Restrictions are built into the form so users cannot alter standard terms, add to non-extensible lists, or leave a list with no associated terms. The table below shows what happens in the database when a user performs the specified action.

User Action	Database Result
Add term	Row inserted into CTValues with new term and associated list ID; row inserted into CTUsed with new term and study IDs
Change term	Term updated in CTValues
Select term for use	Row inserted into CTUsed with term and study IDs
Deselect term	Corresponding row in CTUsed deleted
Change input code	Corresponding row in CTUsed updated with new input code
Change sort order	Corresponding row in CTUsed updated with new sort order value

CONCLUSION

Controlled terminology is one of many pieces of metadata associated with a variable; however, since it is not variable-dependent there are unique challenges that must be managed. The way we have integrated controlled terminology in the Metaspecer at Rho is not a static solution—it evolves as standards change and new ideas are thought of or suggested. It is my hope that others will be able to take ideas from this paper to use at their companies and will also suggest ideas of other ways controlled terminology can be controlled.

ACKNOWLEDGMENTS

I would like to thank Jeff Abolafia and Susan Boyer for their encouragement and support.

CONTACT INFORMATION

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

Ryan Burns
Rho, Inc.
6330 Quadrangle Dr
Chapel Hill, NC 27517
Work Phone: +1 (919) 408-8000
Fax: +1 (919) 408-0999
Email: rburns@rhoworld.com
Web: www.rhoworld.com