Creating Analysis Data Marts from SDTM Warehouses

Frank L. Roediger, SAS Institute, Inc., Cary, North Carolina

ABSTRACT

Getting clinical trial data into SDTM (Study Data Tabulation Model) domains is only the first hurdle in the race. The next obstacle is being able to combine separate trials' SDTM domains so that they can be analyzed. On the surface, the task of combining SDTM data sets is straightforward – simply stack each trial's domain with the corresponding domain from all the other trials. However, as for any other type of data warehouse that is loaded over time, careful attention needs to be given to assure that strict version control is in place for field attributes, controlled terminologies, and thesaurus-based data such as MedDRA.

This paper describes a two-step approach that has been used to build SDTM warehouses and then to extract analysis data marts from them. The first step is to create two comprehensive warehouses: one for ongoing trials and another for completed trials. The second step is to create analysis data marts by extracting from the warehouses the domains for a set of homogeneous trials. Analysts develop the criteria for homogeneity using information in the trials' Trial domains (TA, TE, TI, TS, & TV). These criteria drive an extract process that creates an analysis data mart of SDTM domains that contain only the records for the trials that meet the analyst's criteria.

INTRODUCTION

After an organization has brought SDTM standards in-house for its clinical trials data, there's a natural tendency to consider the job finished. After all, now that all trials share the same data structures, the SDTM domains can be snapped together like building blocks that have been machined to the same specification. The agonies of having to combine disparate data tables for cross-trial integrations are vestiges of how things used to be done in the old days. Or are they?

Admittedly, working with SDTM domains will not involve the same agonies that we used to face, but there are still some very strong challenges that need to be met, and these challenges cannot be effectively addressed on an ad hoc or impromptu basis.

- Data harmonization: All the trials may share the same structures, but they very likely will not share the same version of thesauruses, such as MedDRA, especially when the trials have been amassed over a span of time.

- Data currency: When can a trial's domains be combined with other trials' domains? Only after it has closed?

- Validation: Even though the individual trials' SDTM domains may be validated, an ad hoc integration of those domains is not.

For us, the most effective way to address these challenges has been to combine the trials' SDTM domains into a data warehouse. Actually we opted for two warehouses, one for closed trials and a separate one for active trials. During our implementation, we began to refer to these warehouses as oceans because of their size and in order to distinguish them from the analysis pools that are derived from them. We have found that labeling the warehouses oceans is an effective way of conveying meaningful information about the warehouses, and I will refer to them in that way throughout this paper.

The ETL that we implemented involves two jobs: the first job copies domains from the trials' SDTM folders and places them in the Oceans, the second job creates Analysis Pools from the Oceans' domains.

Our implementation was done in an instance of SAS Drug Development, but the approach that we used can be accomplished in Base SAS as well.

JOB #1 – CREATING THE ONGOING AND COMPLETE OCEANS

Job #1 consolidates the trials' SDTM domains into the Ongoing and Complete Oceans. The job also has the responsibility for managing the Oceans (moving obsolete data out of them and up-versioning them with MedDRA updates), and for maintaining the metadata tables that document the Ongoing and Complete Oceans' contents.

The job has three modes:

1.   Ongoing. This mode completely re-creates the Ongoing Ocean with the domains from all the currently active trials. Because active trials are in a state of flux, we feel that the soundest approach is to not carry over any of their data. Therefore, we opt to have Job #1 completely drain and re-fill the Ongoing Ocean every time it is executed in Ongoing mode.

2.   Complete. This mode incrementally updates the Complete Ocean with the domains from currently closed studies that have not yet been added to it. Because closed trials are stable, we feel that their data need to be added to the Complete Ocean only once. After a closed trial's domains have been added to the Complete Ocean, they will not need to be added again. In order to accommodate legacy studies that may have used an earlier version of MedDRA, the AE (Adverse Events) and MH (Medical History) domains for all closed trials are converted to the current MedDRA version.

3.   MedDRA. The job needs to be run in this mode whenever an updated version of MedDRA becomes the new corporate standard. This mode up-versions the appropriate fields in the AE and MH domains of both the Ongoing and the Complete Oceans with the updated MedDRA information. This assures that all the rows in the AE and MH domains always reflect the same MedDRA version.

The job is made up of the following five tasks, each of which is a separate SAS Drug Development process:

1.   Check whether another execution of the job is already running. We perform this check with a signal file: if the signal file already exists, we know that another job is already running, which we signify by assigning a specific value to a global macro variable. Every other task in the job checks the value of this global macro variable to determine if it should execute fully or branch immediately to its termination point.

2.   Provisionally create the new Ocean elements in the Staging library. We decided not to update the Ocean directly because it is harder to identify and clean out "pollution" from the Ocean than it is to avoid putting "polluted" elements into it. Downstream tasks are responsible for determining whether the new provisional elements should actually be added to the Ocean.

3.   Check for "pollution." If the log for Task #2 contains any error messages, notify the Ocean administrators that the job failed and assign a global macro variable with a specific value to communicate this failure to all the downstream Tasks.

4.   Copy the new elements to the Ocean. The new Ocean elements were created in the Staging library by Task #2. Because they have been certified as "pollution-free" by Task #3, they can now be permanently added to the Ocean itself.

5.   Remove the signal file. If the job created the signal file in Task #1, its final responsibility is to delete the signal file in Task #5 to indicate that its work is finished, that the Oceans are stable, and that another job can now execute.

Figure 1 depicts the flow of data throughout the warehouse. The flow of domains into the Oceans is depicted on the left and in the center of the figure. The individual trials' domains are the tributaries. The active trials' domains are tributaries for the Ongoing Ocean, and the closed trials' domains are tributaries for the Complete Ocean. When an active trial is closed, it is removed from the Ongoing Ocean and added to the Complete Ocean. It is also important to check for cross-Ocean duplicates, because the same trial should never have domains in both Oceans at the same time.

JOB #1 – DATA RESOURCES

INPUT DATA TABLES

Enterprise-Wide Trial Inventory

We maintain a single spreadsheet with an enterprise-wide list of all the trials, both historical and ongoing. The spreadsheet has a flag column whose value distinguishes between closed, historical trials and active, ongoing ones. This distinction is important because it determines in which Ocean the trial's SDTM domains will reside. The spreadsheet also contains information that is necessary for locating a trial's folders in the repository hierarchy. Because of the way we have structured our repository hierarchy, the only information we need is a compound identifier and a trial ID; if your hierarchy needs additional information you could set up additional columns in the table to help locate a trial's folders.

Trial-Level SDTM Domains

It is the responsibility of each trial's data management programmers to make sure that validated versions of the trial's SDTM domains are present in the trial's SDMT folder. If a trial's SDTM domains have not yet been validated, they can be withheld from the Ocean by assigning a special value to the flag column in the enterprise-wide trial inventory
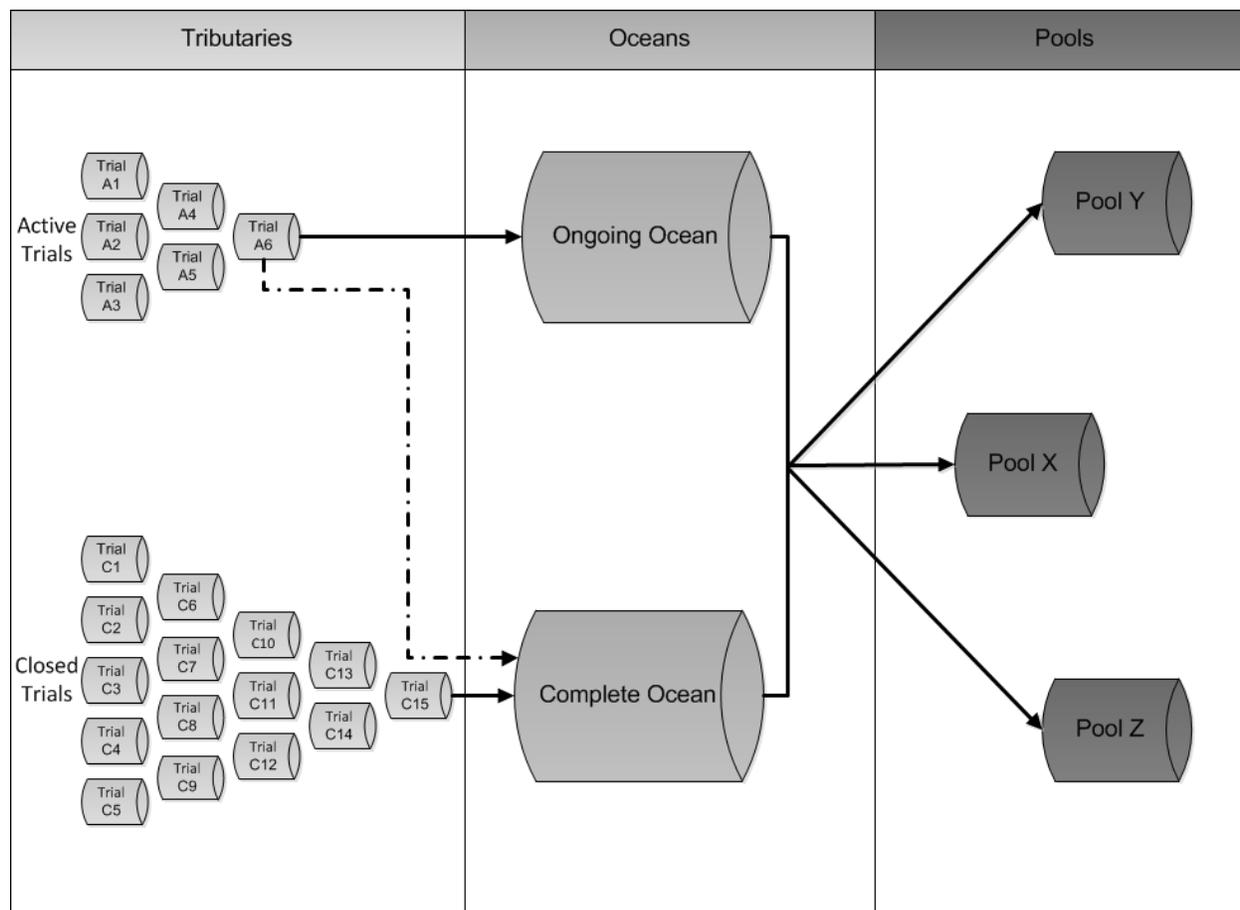
spreadsheet.

Figure 1. Flow of Domains into the Oceans

Enterprise-Wide SDTM Domain Inventory

Not all trial-level SDTM domain folders contain the same set of SAS data sets. Some trials do not use the full set of the standard SDTM domains, others require a set of custom domains, and others contain special-purpose data sets that the trial's data management programmers decided to store in the folder. Because of this variability across trial SDTM folders, it is not advisable to transfer every data set to an Ocean. In order to know which data sets in a trial's SDTM folder to add to an Ocean, we have an "official" list of the enterprise SDTM domains in a spreadsheet: a data set in a trial's SDTM folder is brought into the Ocean only if it is on the list.

OUTPUT DATA TABLES

Ongoing Ocean

The Ongoing Ocean contains an entire set of enterprise SDTM domain tables for every trial that is identified as active in the trial inventory. If no active trial has a validated version of an SDTM domain, that domain will still have a data set in the Ongoing Ocean, but the data set will only be a place-holder and will have zero observations.

Complete Ocean

The Complete Ocean will eventually be larger than the Ongoing Ocean because it contains an entire set of enterprise SDTM domain tables for every closed trial for which there is SDTM data. Because legacy trials may not have used SDTM domains, it will be necessary to make sure that there is a set of validated SDTM domains for every trial that is listed as closed in the enterprise-wide trial inventory.

Backup Oceans

Because the Oceans are such a valuable resource, they are each protected with two backup versions. The first backup contains an exact duplicate of the corresponding Ocean's current version; this version can be promoted to take the place of the Ocean in case the Ocean, for whatever reason, becomes unusable. The second backup contains a copy of the -1 version of the Ocean (that is, the Ocean as it existed before the most recent execution of Job #1).
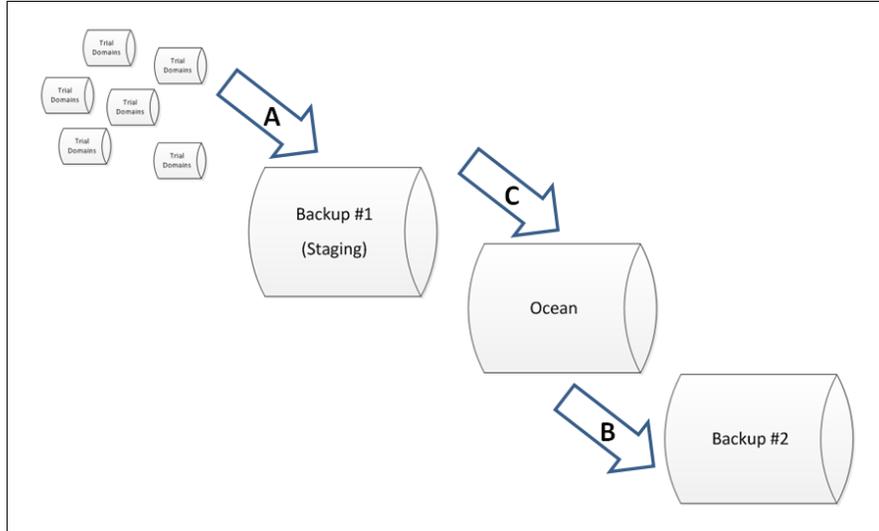

Figure 2. Evolution of the Ocean when Task #2 Executes Successfully

Figure 2 depicts the relationship between the Ocean proper and the Backup Oceans when Task #2 executes successfully.

- Step A – Trials' SDTM domains are consolidated in Backup Library #1 (Staging)

- Step B – Ocean contents are moved to Backup Library #2

- Step C – Backup Library #1 (Staging) copied to Ocean

The first backup also serves as the Staging library, which contains the provisional new Ocean elements. If Task #2 executes successfully (that is, its log does not contain any signs of "pollution"), Task #4 promotes the contents of the Staging library to the Ocean; if, however, Task #2 fails to execute successfully (that is, its log contains indications of "pollution"), Task #4 clears out the Staging library and replaces it with a copy of the Ocean. In either case, at the end of Job #1, the Staging library and the Ocean are duplicates of each other.
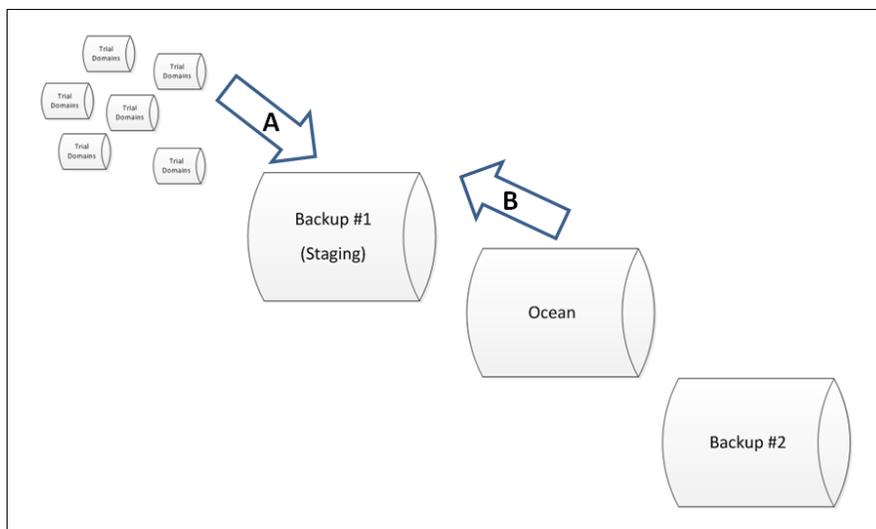
Figure 3. Evolution of the Ocean when Task #2 Fails

Figure 3 depicts the relationship between the Ocean proper and the Backup Oceans when Task #2 fails to execute successfully.

- Step A – Trials' SDTM domains are consolidated in Backup Library #1 (Staging)

- Step B – Ocean contents are copied to Backup Library #1 (Staging)

- Backup Library #2 is unchanged

Metadata

Metadata is crucial for the Oceans because it establishes a documented trail of when and from where every data element came into the Oceans. Without this metadata, it would not be possible to have a validated warehouse. We opted to have two metadata tables, one at the trial-level and one at the domain-level.

Ocean Trials Inventory Table

The Ocean Trials Inventory table contains one row for every trial that is in either Ocean. It contains the following set of fields:

- Compound – the trial's compound identifier

- Trial ID – the trial's identifier

- Ocean – the Ocean in which the trial currently resides

- Ocean Date/Time – the datetime when the trial was added to the Ocean

- MedDRA version – the MedDRA version that was used to encode the trial's AE and MH domains

Figure 4 shows a section of a sample Ocean Trials Inventory table.

#	Compound	TrialID	ocean	ocean_dttm	meddra_version
1	Compound A	a00_01	complete	2012-02-29T15:07:35	99.1
2	Compound A	a00_02	complete	2012-02-29T15:07:35	99.1
3	Compound A	a00_03	complete	2012-02-29T15:07:35	99.1
4	Compound A	a00_04	complete	2012-02-29T15:07:35	99.1
5	Compound A	a00_05	complete	2012-02-29T15:07:35	99.1
6	Compound B	b00_11	ongoing	2012-02-29T14:44:03	99.1
7	Compound B	b00_12	ongoing	2012-02-29T14:44:03	99.1
8	Compound B	b00_13	ongoing	2012-02-29T14:44:03	99.1
9	Compound B	b00_14	ongoing	2012-02-29T14:44:03	99.1
10	Compound B	b00_15	ongoing	2012-02-29T14:44:03	99.1

Figure 4. Ocean Trials Inventory table

Ocean Domains Inventory

There are two Ocean Domains Inventory tables, one for each Ocean (the two Ocean Domains Inventory tables can be consolidated into a single table, but we opted to store them separately). Both tables contain the following columns (if you consolidate the two Ocean Trials Inventory tables, you will need an additional column to identify the domain's Ocean):

- Domain – the standard SDTM name of the domain

- Compound – the trial's compound identifier

- Trial Identifier – the trial's ID

- Path – the path to the location of the trial's SDTM folder in the repository

- Trial Date/Time – the datetime when the SDTM domain was created in the trial's SDTM folder

- Ocean Date/Time – the datetime when the SDTM domain was added to the Ocean

Figure 5 shows a section of the Ocean Trials Inventory table.

#	domain	Compound	TrialID	TrialPath	study_dttm	ocean_dttm
1	ae	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2012-02-22T21:09:58	2012-02-29T14:44:03
2	ce	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2012-01-28T14:38:59	2012-02-29T14:44:03
3	dm	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2012-01-28T14:39:18	2012-02-29T14:44:03
4	mh	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2012-01-28T14:39:40	2012-02-29T14:44:03
5	suppae	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2012-01-28T14:40:12	2012-02-29T14:44:03
6	suppmh	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2012-01-28T14:40:44	2012-02-29T14:44:03
7	ta	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2011-11-15T03:14:16	2012-02-29T14:44:03
8	te	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2011-11-15T03:14:18	2012-02-29T14:44:03
9	ti	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2011-11-15T03:14:22	2012-02-29T14:44:03
10	ts	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2011-11-15T03:14:25	2012-02-29T14:44:03
11	tv	Compound...	b00_11	/Data/Compound B/b00_11/Active/SDTM	2011-11-15T03:14:28	2012-02-29T14:44:03

Figure 5. Ongoing Ocean Domains Inventory table

JOB #2 – CREATING ANALYSIS POOLS

Job #2 creates an analysis pool of the observations from the Ongoing and Complete Oceans' SDTM domains. The user defines the selection criteria in a filter text file, and only trials that meet those criteria will be included in the analysis pool. A common metadata table contains information about every execution of Job #2.

Each analysis pool requires its own set of folders. Before Job #2 can be run, the user will need to make sure that the folders have been created.

The Job #2 can be run in two modes:

1. Filter Check. This mode only runs a series of error checks on the criteria in the filter text file; it does not create any data sets in the Analysis Pool's SDTM domains folder. Filter Check mode allows users to preview selection criteria before using them to actually create the Analysis Pool. Especially after the Oceans have become large, the process of filling the Analysis Pool can take a considerable time; because running only the Filter Checks requires much less time, the Filter Check mode is an efficient way for users to determine whether they have set up the selection criteria in the filter text file correctly.

2. Create Pool. This mode runs the error checks as in the Filter Check mode, but it also creates the resulting data sets in the Analysis Pool's SDTM folder.

Job #2 is made up of the following three tasks, each of which is a separate SAS Drug Development process:

1. Check whether another execution of the job is already running (this is the same as Task #1 in Job #1).

2. Run error checks for the Filter Text File and create data sets in the analysis pool's SDTM folder.

3. Remove the signal file (this is the same as Task #5 in Job #1).

The creation of Analysis Pools from the Oceans is depicted on the right side of Figure 1 above. It is important to note that we did not opt to save versions of the Analysis Pools because we did not want users to become confused about which version of an Analysis Pool is the current one. Consequently, every time a user runs Job #2 in Create Pool mode, a completely new set of data sets is created in the Analysis Pool's SDTM folder. If a user wants to compare the effects of different versions of selection criteria, they will need to create a set of Analysis Pool folders for each set of selection criteria.

JOB #2 – DATA RESOURCES

INPUT DATA TABLES

Filter Text File

The Filter Text File contains the criteria for selecting a trial from the Oceans to be included in the Analysis Pool. Each line in the Filter Text File contains a valid expression that defines a condition that a trial must meet in order to be selected. The Filter Text File can contain as many lines as the user wishes. However, in order to be selected, a trial must meet the condition for each of the expressions (that is, the expressions are all connected with a logical AND operator).

We have chosen to require that all the expressions be based on fields that are in one of the Trial Design domains (TA – Trial Arms; TE – Trial Elements; TI – Trial Inclusion/Exclusion Criteria; TS – Trial Summary; and TV – Trial Visits). The reason for this requirement is that our Ocean-to-Analysis Pool extracts are all trial-based (that is, if a trial meets all the selection criteria, then all its rows in every SDTM Ocean domain is selected to be in the Analysis Pool; there is no subject-level selection, only trial-level).

The Filter Text File contains two sub-fields: 1) the two-character Trial Design domain; and 2) the conditional expression. The conditional expression must follow the syntax rules for a WHERE clause. Job #2 brackets each conditional expression in open- and close-parentheses and inserts a logical AND between each one.

Users create a Filter Text File by selecting the template, which contains only a "ruler row". The "ruler row," which is retained as the first row in the filter text file, helps users make sure that they correctly align the two sub-fields in each row that they enter and that they do not extend either of the fields beyond its specified limits. The following is a set of sample lines from a Filter Text File:

```
ts tsparmcd = 'AGEU' & tsval = 'YEARS'
ts tsparmcd = 'AGEMIN' & INPUT(tsval, best4.) >= 12
ts tsparmcd = 'AGEMAX' & INPUT(tsval, best4.) <= 64
```

In this example, the criteria are selecting all trials that screened for subjects between the ages of 12 and 64 years old, inclusive.

Job #2 performs some rudimentary validation checks on the Filter Text File, and issues a report of any check failures. If the Filter Text File fails any of the validation checks, Job #2 stops executing.

OUTPUT DATA TABLES

Analysis Pool

The Analysis Pool is a folder with a SAS data set for each of the "official" enterprise SDTM domains. Within each SAS data set are observations from the corresponding Ongoing or Complete Ocean SDTM domain where the STUDYID field contains the value of a trial that meets all the selection criteria from the Filter Text File. If none of the selected trials has any rows for an "official" enterprise SDTM domain, the Analysis Pool will have a place-holding zero-obs SAS data set in its place. This is like the blank pages labeled, "This page was intentionally left blank" in manuals: we didn't was a subsequent user to wonder if the absence of an "official" enterprise SDTM domain was intended or inadvertent.

Metadata

A metadata table contains all the information about every successful execution of Job #2. We opted to include rows for Filter Check runs as well as Create Pool runs. The Job #2 metadata table contains the following fields:

- Pool ID – the name of the root folder that contains all of the Analysis Pool's sub-folders.

- Pool Date/Time – the datetime when Job #2 was executed to create the Analysis Pool SDTM domains in the Analysis Pool's SDTM folder

- Row Type – the descriptor (either "FILTER" or "DOMAIN") or the type of information that is contained in the VALUE field

- Value

    o when Row Type="FILTER" – the expression from the Filter Text File

    o when Row Type="DOMAIN" – the following sub-fields (from Job #1's Ocean Domain Inventory metadata table)

- ▪ Domain – the standard SDTM name of the domain

- ▪ Path – the path to the location of the trial's SDTM folder in the repository

- ▪ Trial Date/Time – the datetime when the SDTM domain was created in the trial's SDTM folder

- ▪ Ocean Date/Time – the datetime when the SDTM domain was added to the Ocean

- Sequence/Trial Identifier

    - o when Row Type="FILTER" – sequence of the expression in the Filter Text File

    - o when Row Type="DOMAIN" – the Trial ID

The Pool Date/Time field contains the same datetime value for every metadata row for an execution of Job #2. The first set of rows within Pool Date/Time is always for Row Type="FILTER"; if no Ocean trials meet all the selection criteria, the Pool Date/Time will have not have any rows where Row Type="DOMAIN". If any trial meets all the selection criteria, all the "official" enterprise SDTM domains will have a row for each selected trial; if a trial does not have any rows in the corresponding Oceans' SDTM domains, the Value field will contain only the SDTM domain's name and null values in the other Value sub-fields.

Figure 6 shows a section of the Job #2 metadata table.

#	PoolID	Pool_DTTM	Row_Type	value	Seq_TrialID
1	QA_Trial	2012-02-29T16:35:31	FILTER	ta studyid eq 'b00_15'	1
2	QA_Trial	2012-02-29T16:35:31	FILTER	te studyid eq 'b00_15'	2
3	QA_Trial	2012-02-29T16:35:31	FILTER	ti studyid eq 'b00_15'	3
4	QA_Trial	2012-02-29T16:35:31	FILTER	ts studyid eq 'b00_15'	4
5	QA_Trial	2012-02-29T16:35:31	FILTER	tv studyid eq 'b00_15'	5
6	QA_Trial	2012-02-29T16:35:31	DOMAIN	ae|2012-01-28T14:54:24|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
7	QA_Trial	2012-02-29T16:35:31	DOMAIN	ce|2012-01-28T14:54:28|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
8	QA_Trial	2012-02-29T16:35:31	DOMAIN	cm|||	b00_15
9	QA_Trial	2012-02-29T16:35:31	DOMAIN	co|||	b00_15
10	QA_Trial	2012-02-29T16:35:31	DOMAIN	dm|2012-01-28T14:54:47|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
11	QA_Trial	2012-02-29T16:35:31	DOMAIN	ds|2012-01-28T14:54:51|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
12	QA_Trial	2012-02-29T16:35:31	DOMAIN	ex|2012-01-28T14:54:56|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
13	QA_Trial	2012-02-29T16:35:31	DOMAIN	ho|||	b00_15
14	QA_Trial	2012-02-29T16:35:31	DOMAIN	ie|||	b00_15
15	QA_Trial	2012-02-29T16:35:31	DOMAIN	lb|2012-01-28T14:55:02|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
16	QA_Trial	2012-02-29T16:35:31	DOMAIN	mh|2012-01-28T14:55:10|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
17	QA_Trial	2012-02-29T16:35:31	DOMAIN	relrec|||	b00_15
18	QA_Trial	2012-02-29T16:35:31	DOMAIN	sc|||	b00_15
19	QA_Trial	2012-02-29T16:35:31	DOMAIN	se|||	b00_15
20	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppae|2012-01-28T14:55:33|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
21	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppce|||	b00_15
22	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppcm|||	b00_15
23	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppdm|||	b00_15
24	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppds|||	b00_15
25	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppex|||	b00_15
26	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppho|||	b00_15
27	QA_Trial	2012-02-29T16:35:31	DOMAIN	supplb|||	b00_15
28	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppmh|2012-01-28T14:56:06|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
29	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppxc|||	b00_15
30	QA_Trial	2012-02-29T16:35:31	DOMAIN	suppxp|||	b00_15
31	QA_Trial	2012-02-29T16:35:31	DOMAIN	sv|||	b00_15
32	QA_Trial	2012-02-29T16:35:31	DOMAIN	ta|2011-11-15T03:21:47|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
33	QA_Trial	2012-02-29T16:35:31	DOMAIN	te|2011-11-15T03:21:50|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
34	QA_Trial	2012-02-29T16:35:31	DOMAIN	ti|2011-11-15T03:21:53|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
35	QA_Trial	2012-02-29T16:35:31	DOMAIN	ts|2011-11-15T03:21:56|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
36	QA_Trial	2012-02-29T16:35:31	DOMAIN	tv|2011-11-15T03:21:58|2012-02-29T14:44:03|/Data/Compound B/b00_15/Active/SDTM	b00_15
37	QA_Trial	2012-02-29T16:35:31	DOMAIN	vs|||	b00_15
38	QA_Trial	2012-02-29T16:35:31	DOMAIN	xc|||	b00_15
39	QA_Trial	2012-02-29T16:35:31	DOMAIN	xp|||	b00_15

Figure 6. Job #2 metadata table

OUTPUT REPORTS

Job #2 creates a .lst file with the following set of sub-reports for both Filter Check and Create Pool executions.

Expression Listing

Because it is helpful for users to see the expressions that they submitted (they might think that they had referred to a different Filter Text File than the one that was actually used!), we have Job #2 create an expression listing sub-report that displays each of the expressions in the Filter Text File.

Ocean Trial Listing/Non-Ocean Trial Listing

Users may not realize that a particular trial has not yet been added to the Oceans. The Ocean Trial Listing identifies all the trials that are in the Ongoing and Complete Oceans – these trials will be the only ones that can be brought into the Analysis Pool. The Non-Ocean Trial Listing identifies the trials that are not yet in either of the Oceans. These could be legacy trials whose data has not yet been converted into validated SDTM domains or active trials whose SDTM data management processing is not yet validated.

Non-Selected Ocean Trials Listing

If an Ocean trial does not meet all the selection criteria in the Filter Text File, it will not be selected for the Analysis Pool. The Non-Selected Ocean Trials Listing displays the ID for non-selected trials, the Ocean in which each trial is stored, how many selection expressions that each trial met, and the expression sequence number (from the Expression Listing) of each selection expression that the trial did meet.

Selected Ocean Trials Listing

This is a listing of the Ocean trial that meet all the selection expressions in the Filter Text File. It contains the same set of information as the Non-Selected Ocean Trials Listing.

CONCLUSION

My purpose has been to describe the features of a full-scale SDTM warehouse implementation that I have helped to develop. If you have already tackled such an implementation, I hope that I have given you some ideas that you can use to evaluate the design choices that you have made and either confirm that yours were the optimal ones for your warehouse or give you some points to consider for enhancements. If you haven't implemented an SDTM warehouse yet, I hope I have given you some issues to consider when you begin your design and some suggestions to consider when you develop your implementation.

ACKNOWLEDGMENTS

I would like to acknowledge a deep indebtedness to Pantaleo Nacci of Novartis Vaccines and Diagnostics for his conceptualizations of and practical insights about SDTM warehouses and data marts; and to my colleagues at SAS, Sandeep Juneja and Ben Bocchicchio, for their imaginative and inventive insight into SAS programming in general and this development project in particular.

CONTACT INFORMATION

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

Name:	Frank Roediger
Enterprise:	SAS Institute, Inc.
Address:	720 SAS Campus Drive, U3102
City, State ZIP:	Cary, NC 27513
Work Phone:	919 531-0519
E-mail:	frank.roediger@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.