Paper DH07

A Robust Framework to Provide and Validate Real-time SDTM Data

Bremer Louw, Bioforum, Bloemfontein, South Africa

ABSTRACT

Clinical trials rely heavily on continuous data monitoring for both patient safety and data quality surveillance. However, using SDTM data for data monitoring has always been a challenge given the obstacles associated with re-running complex SAS/R programs and re-validating the resulting SDTM data on a frequent basis. This white paper presents a framework that explores the potential solutions to make it a reality.

We will explore the major components of a framework required to implement a real-time SDTM data solution and present an implementation used on production studies which include the following –

- Data Acquisition. How to routinely gather the source data needed for SDTM.
- Data Processing. Building a robust conversion mechanism.
- Data Publishing. Routinely disseminate the SDTM data to downstream consumers.
- Data Validation. Automatically running validation checks using both regulatory/CDISC conformance and user-specific rules.
- Issue Detection. Monitoring and reporting issues during the data flow to ensure quick resolution.
- Ensuring that unexpected data issues do not interrupt the flow of data to SDTM.

Validated real-time SDTM data has significant benefits, including providing downstream consumers and decision-makers with up-to-date data to make accurate and timely decisions. The framework presented in this paper could significantly reduce the cost and time needed to obtain SDTM data and ensure the data is validated, reliable and compliant.

Overall, this paper highlights the important components of a robust framework to enable real-time SDTM data. It also emphasizes the potential benefits, including cost and time savings, and better decision-making.

INTRODUCTION

Clinical trials rely heavily on continuous data monitoring for both patient safety and data quality surveillance. However, using Study Data Tabulation Model (SDTM) data for data monitoring has always been a challenge given the obstacles associated with re-running complex SAS® and R® programs and re-validating the resulting SDTM data on a frequent basis. This paper presents a framework that explores the potential solutions to make it a reality and the expected benefits of implementing a real-time SDTM solution.

The most common current industry practice is that the various data owners would send the source data to the SDTM team on a pre-arranged schedule. The timepoints would depend on study deliverables, usually in preparation for major milestones like Data Monitoring Committee (DMC) meetings, Interim Analysis, Database Lock (DBL), etc. This approach results in SDTM data only being available for planned major milestones. The SDTM team would then convert the data into SDTM format, resolve any issues found during the conversion process and transfer the resulting SDTM data to downstream consumers of the data.

This archaic process has many limitations including:

- SDTM data is not readily available to be used for reporting or review,
- Data issues are not reported or resolved timely, only at major milestones,
- Performing a re-run of SDTM is a manual task and hence costly.

Replacing this archaic approach with a modern real-time dataflow approach would resolve the limitations of the current process but necessitates a carefully planned technology framework to ensure robustness and quality SDTM data.

THE FRAMEWORK

The dataflow of SDTM data can be segregated into three sections, namely: acquisition, processing, and publishing. As part of a robust framework, we need to ensure that we detect issues during each process, ensure that the dataflow does not stop when unexpected data is received, validate the resulting SDTM data and report the status of the dataflow to stakeholders.

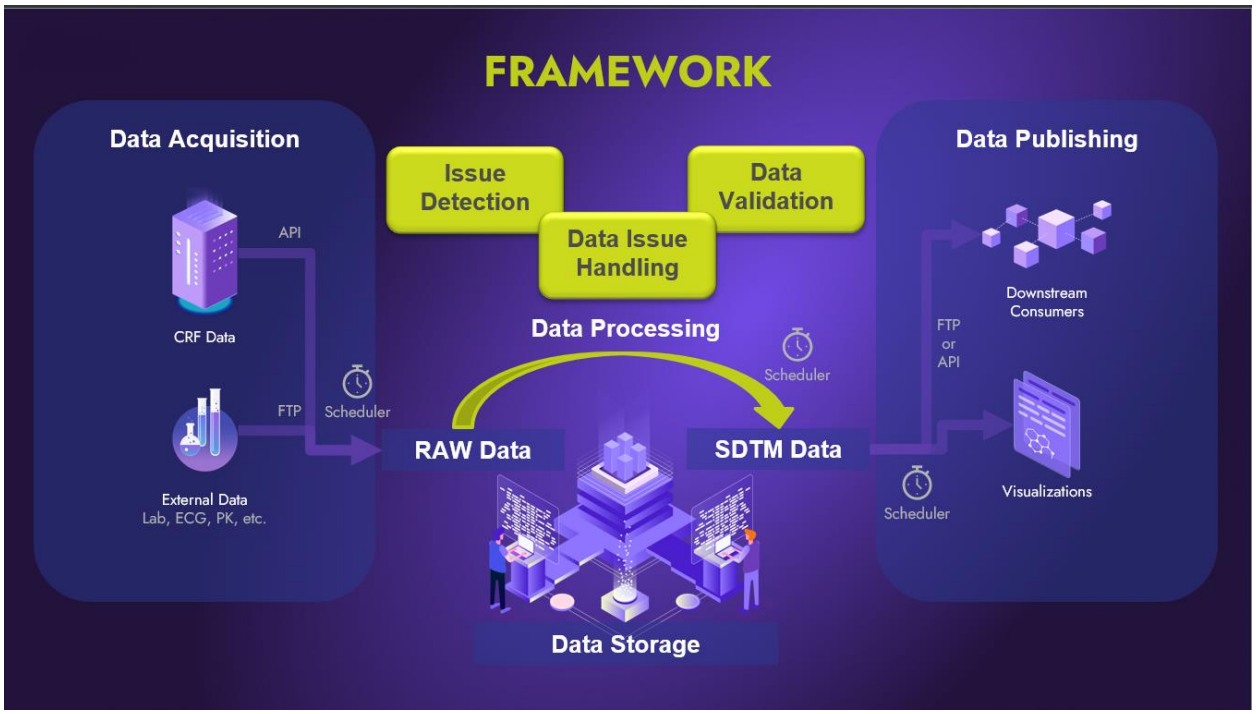

Figure 1 – Real-time SDTM Data Framework

DATA ACQUISITION

Data acquisition is the gathering of all data sources needed to produce SDTM data. This usually includes the Case Report Form (CRF) data from an Electronic Data Capturing (EDC) system and various other data sources not stored in the EDC system, like laboratory results, electrocardiogram (ECG) results, pharmacokinetic (PK) results and many others.

Most modern EDC systems do have Application Programming Interface (API) capabilities which is the most efficient way to gather data. However, older EDC systems and many non-EDC data providers, like local laboratories, do not have these capabilities. In addition, the setup of API connections between internal infrastructure and that of the data vendor is usually costly, because it requires significant manpower from IT teams to create the connection. For these reasons we need to provide alternative methods of data acquisition.

A very suitable alternative is the use of Secure File Transfer Protocol (SFTP). SFTP allows data vendors to upload data to a SFTP location using free software without costly infrastructure requirements. SFTP automation is possible using off-the-shelf software. Various software solutions allow us to monitor a specific SFTP location and as soon as a new file arrives, we can set defined tasks to be performed. For example, as soon as a laboratory places a file on the SFTP the software can copy the file to a network location, extract the contents of the file and start executing SDTM conversion programs.

Scheduler

As mentioned above, some automation software can initiate an action or task when a trigger event occurs. For example, when data is placed on the SFTP, the action of copying the data to an internal location is started. However, depending on the systems and required actions, this might not be a feasible option. In this case a scheduler may solve the problem. A scheduler is software that performs predefined actions at a predefined point in time and it can be scheduled as recurring. Various actions in this framework can be scheduled using such software.

A simple example of using schedulers is setting the data acquisition scheduler (gathering the data) to run at 01:00am, then the data processing scheduler to run at 02:00am and then the data publishing scheduler to run at 03:00am.

It is important to note that because we need to provide a predefined timepoint for each action, we need to ensure there is adequate time for all prerequisite actions to be completed before the next action starts. At the start of a study some actions may occur fast, but as data accumulates the actions may take significantly longer to complete and we need to account for that.

Another important consideration is to manage the data transfer and processing load of our servers, meaning if multiple schedules are run on a server, spread them out over different hours of the day to ensure as little parallel processing as possible. This will ensure the best possible performance of our resources.

DATA PROCESSING

Data processing is the process of creating SDTM datasets from the source data. The most common approach to creating SDTM data is creating programs in SAS® and R® to convert the source data into SDTM data and more recently we have seen the release of SDTM conversion software that are focused only on the conversion of data to SDTM format. Whichever software is used to generate the SDTM data, a way is needed to automate the recurring execution of the conversion process. This can be accomplished by a scheduler as mentioned above or inherent system functionality of the used system.

It is of utmost importance that the conversion mechanism used is robust. If it is very sensitive to data changes, it may cause regular failures of the process and interruptions in the data flow. We will discuss continuous data flow in more detail later in this paper.

DATA PUBLISHING

Data publishing is a process of routinely disseminating the SDTM data to downstream consumers. Downstream consumers may be the team responsible for creating analysis datasets and reports or it may be a visualization system to visualize the SDTM data for review or many other use cases. The technical method used to disseminate the data would depend on the capabilities of the destination system. For example, if the analysis team needs the data in SAS® dataset format it might be best to use an sFTP to distribute the data to them. However, sending the data to a visualization tool that leverages a database for data storage and has API capabilities it might be better to stream the data to the database via API.

An evaluation should be performed for each downstream consumer to understand the required frequency of updated data, the system capabilities of the downstream system and the use case of the SDTM data for that consumer. This will help determine which data publishing method is most suitable.

DATA STATUS REPORTING

Many unexpected issues may occur during the provision and validation of real-time SDTM data. For this reason, it is of utmost importance to monitor the success of delivery regularly. A major part of this monitoring is the detection of failure points and reporting on validation outcomes. A visualized dashboard that displays various metrics and the status of the different actions in the data flow provides users with a tool to facilitate monitoring. Below figures provide examples of metrics that are valuable to personnel tasked with monitoring the status of real-time data.

Figure 2 – Data Monitoring Dashboard: Data Acquisition

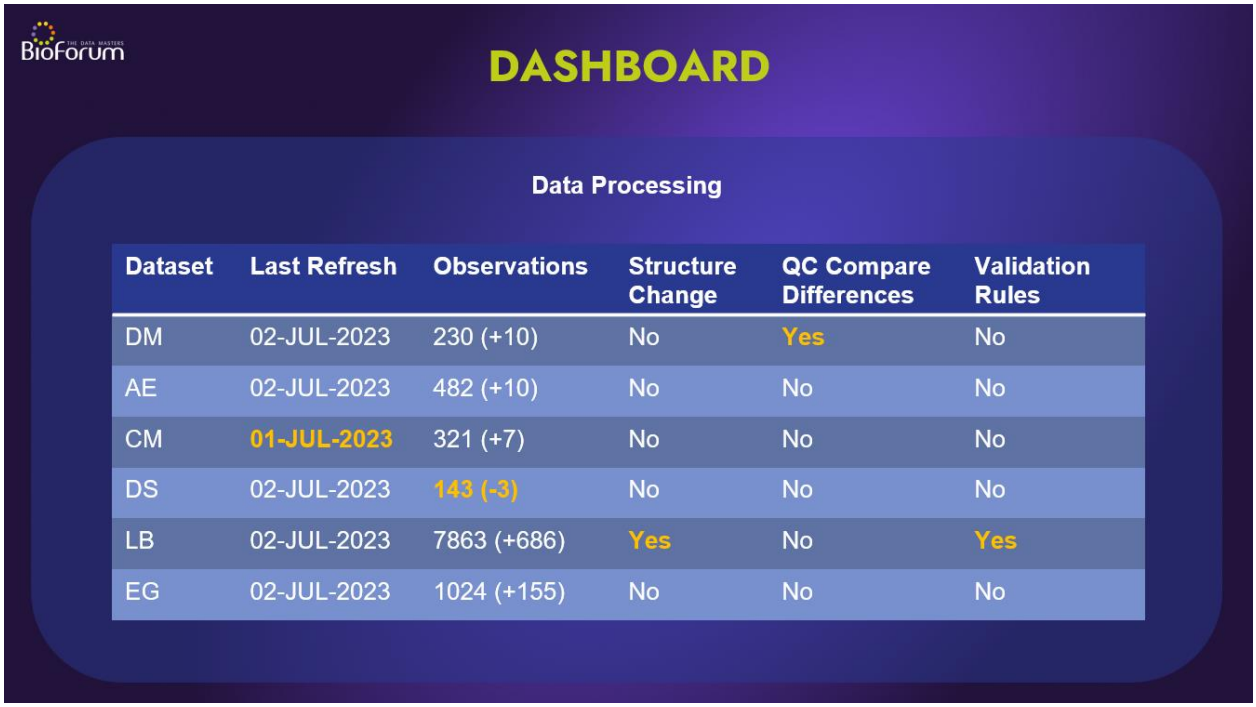

Figure 3 – Data Monitoring Dashboard: Data Processing

Figure 4 – Data Monitoring Dashboard: Data Publishing

The metrics presented in the above figures are derived from the validation and issue detection processes, and hence it is relevant to investigate the process in more detail.

DATA VALIDATION

Various implementations of data acquisition, processing and publishing have materialized in recent years to deliver on organizational requirements. However, most often these implementations produced unvalidated data because no validation has been performed as part of the process. Various methods of SDTM validation can be utilized to ensure data quality.

These methods can be broken down into three components: creating and reviewing the specification, creating the conversions programs and the validation thereof and validation of the resulting SDTM data. The purpose of validation is to ensure the quality of the SDTM data, and three (3) key quality factors should be considered: Data completeness, Mapping correctness and SDTM conformance. Data completeness is ensuring that all expected source data ends up in SDTM. Mapping correctness is ensuring that the data is mapped correctly to SDTM. SDTM conformance ensures that the SDTM data comply with SDTM and regulatory standards.

LEGACY SDTM VALIDATION

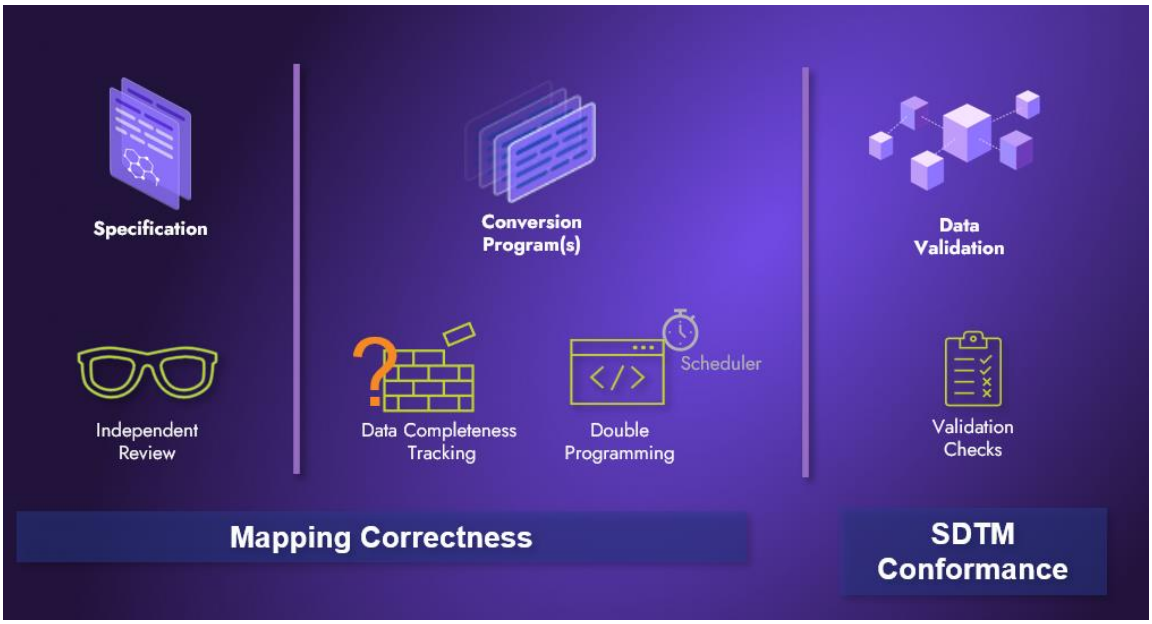

Figure 5 – Legacy SDTM Validation

The most common current validation steps are independent review of the specification, independent double programming for validation of the conversion programs and running validation rules over the resulting SDTM data. Creating and reviewing the specification is to ensure mapping correctness, validation of the conversions programs aims to verify that the specification was implemented as expected and validation of the resulting SDTM data aims to confirm SDTM conformance. Currently there are no objective measure to confirm data completeness in the industry, it is assumed that independent review of the specification will ensure it is accurate and hence all data would be covered in the conversion process, however that is unfortunately not always the case.

Validating real-time data as part of this legacy approach is challenging, however, it is partially possible. The same process used for data processing can be utilized to execute the QC programs after the production SDTM data has been produced. The QC programs should also include the comparison between the production and QC data and the reporting of any differences. The results from the QC compare can then be consumed by the monitoring dashboard to produce the needed status metrics. In scenarios where double programming is not used as the method of QC, for example a manual review of the data, it is impractical to do real-time manual review.

Conformance testing of SDTM data is mostly performed by running a set of validation\conformance rules against the SDTM data and reviewing the resulting report of validation issues. These same tools can be scheduled to run routinely and provide validation results to the monitoring dashboard. The CDISC® CORE engine and set of validation rules are a good example of technology that can be used for this purpose. A big advantage of CDISC® CORE is that we can add custom validation rules to test the quality of our SDTM data.

REAL-TIME VALIDATION

Bioforum has developed an SDTM automation platform and as part of the development had to reimagine the way SDTM data is validated.

Mapping Correctness: Specification Review Quality

The first concern with the current process was that the initial review of the SDTM specification was not always of good quality. The specification reviewer only had the source data and the specification to refer to and could not imagine all potential outcomes of the logic, especially when we think about data issues. Early during the conduct of a study, it is not possible for the reviewer to predict all potential variations of the data and data issues, hence the specification may be lacking to cater for these.

For this reason, the process was changed so that the review for mapping correctness is not done at the start as a specification, but rather an independent review of mappings already performed. That way we can provide the review with comprehensive review tools that provide the mapping logic, the source and SDTM data all in the system and linked to each other.

This approach resulted in us finding issues previously not possible with specification review and we found a significant reduction in review hours and decreases in timelines.

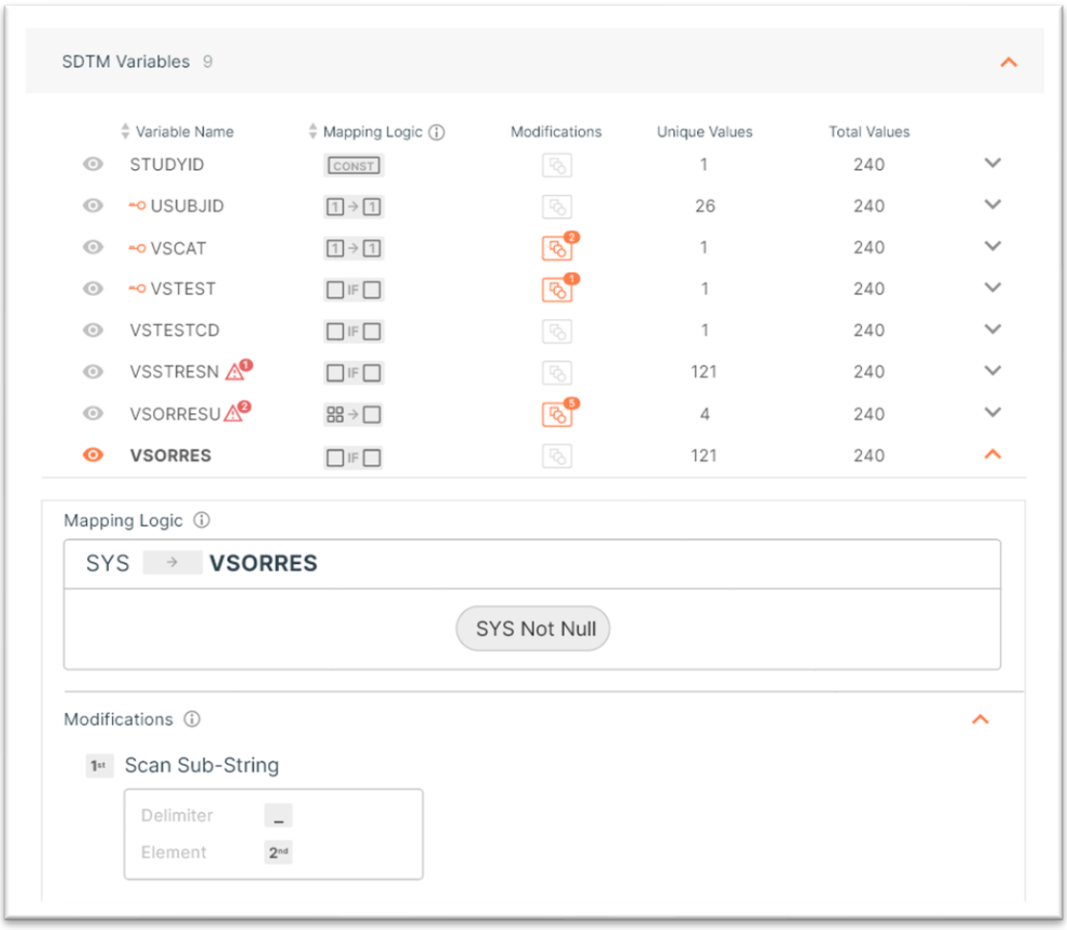

Figure 6 – SDTM Review Module

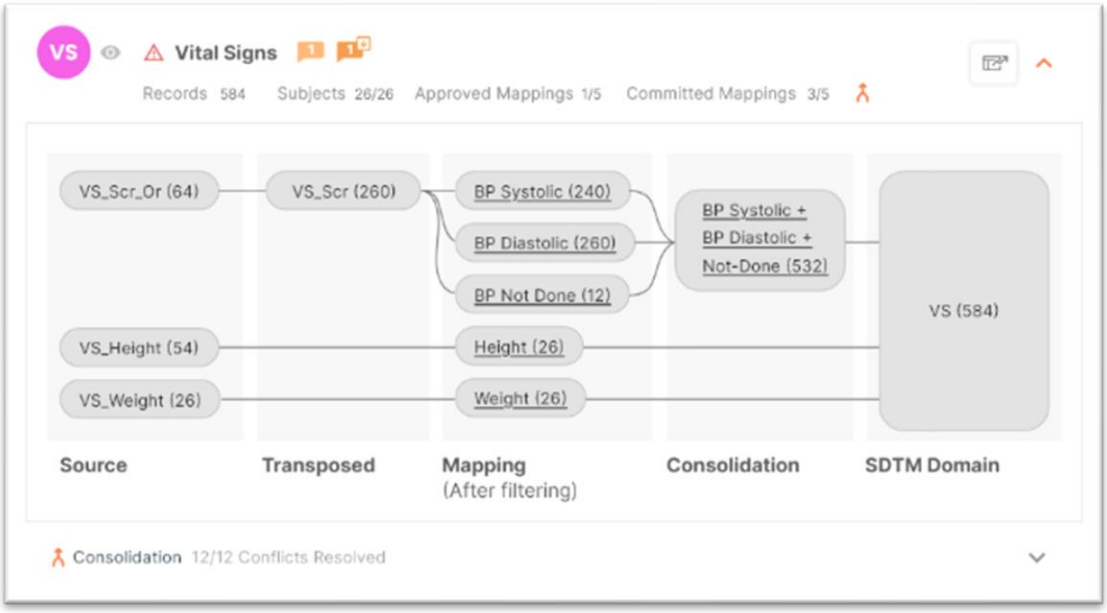

Figure 7 – Data Flow Diagrams

Why double programming?

The purpose of double programming is twofold, to pick up mistakes in the SDTM specification and to ensure that the programmer implemented the requirements of the specification correctly. With the improved review tools mentioned above in place we can guarantee accurate mapping logic, so the only reason for double programming left then is to prove that the system did the conversion per the specification\logic. For this purpose, we performed Computer System Validation and tested all features of the system thoroughly to prove that they perform exactly as intended every time.

That brought us to the point where double programming was redundant and a comprehensive review of the mapping logic provided much better quality than the legacy double programming process.

Real-Time Data Validation

Executing validation checks on SDTM data in the legacy SDTM approach is a waterfall process. The SDTM data needs to be created first, then the validation checks should be executed over the data and then any issues detected should be updated in the programming (production and\or QC), and the process should be repeated until there are no more issues present. These repetitive cycles are time consuming and do not facilitate first time quality.

The re-imagined approach is actual real-time validation, where the user gets validation feedback immediately when they perform a mapping. The SDTM automation platform has a validation engine that executes the relevant validation rules in real time, as the user performs actions and provides immediate feedback. The validation engine was built to use the CDISC CORE syntax to enable the use of open-source validation rules like the CDISC CORE library and industry developed rules. It also allows the user to program custom rules in CDISC CORE syntax and add it to the real-time validation library.

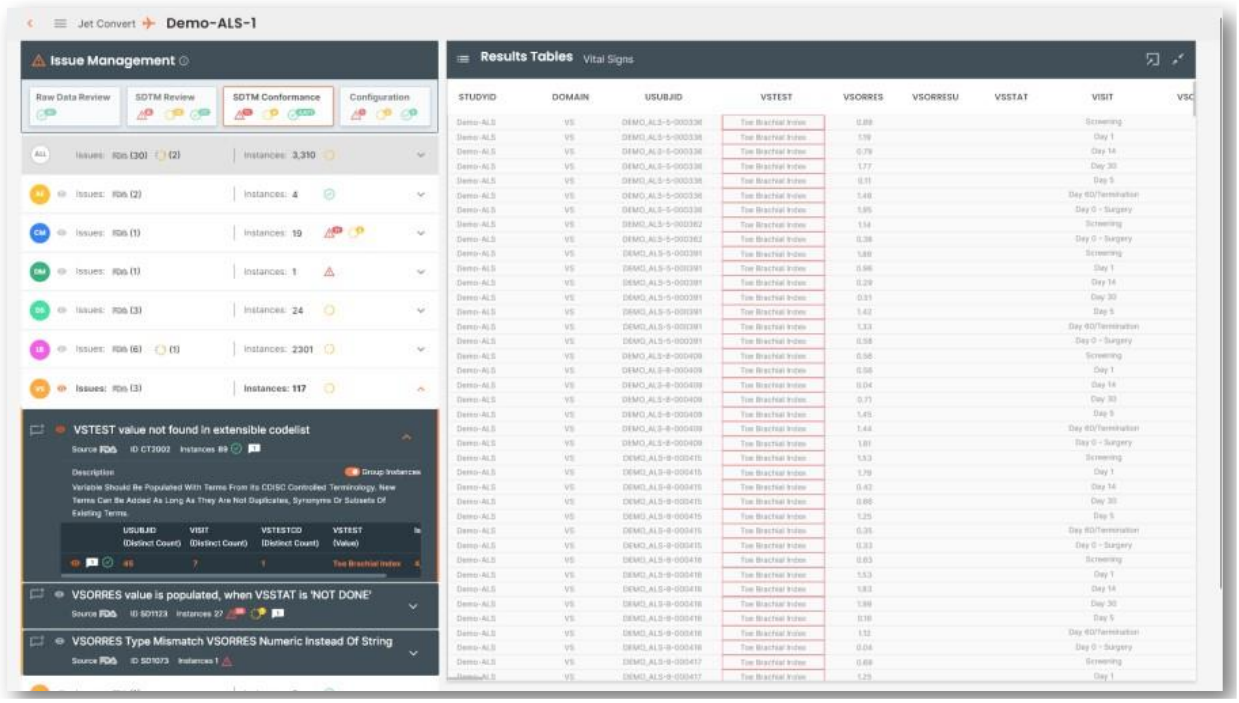

Figure 8 – Real-Time Validation

Data Completeness Tracking

As mentioned, there are currently no objective measures to confirm data completeness in the industry. In this SDTM automation platform each source data point, not variable or records, is tracked from the source data to the SDTM data. Because of this traceability the system provides a "% Complete" metric that will indicate how much of your source data was mapped to SDTM, how much was marked as NOT SUBMITTED and how much has not been used in SDTM. This feature provides the user with complete certainty that all the source data was handled during the SDTM conversion process.

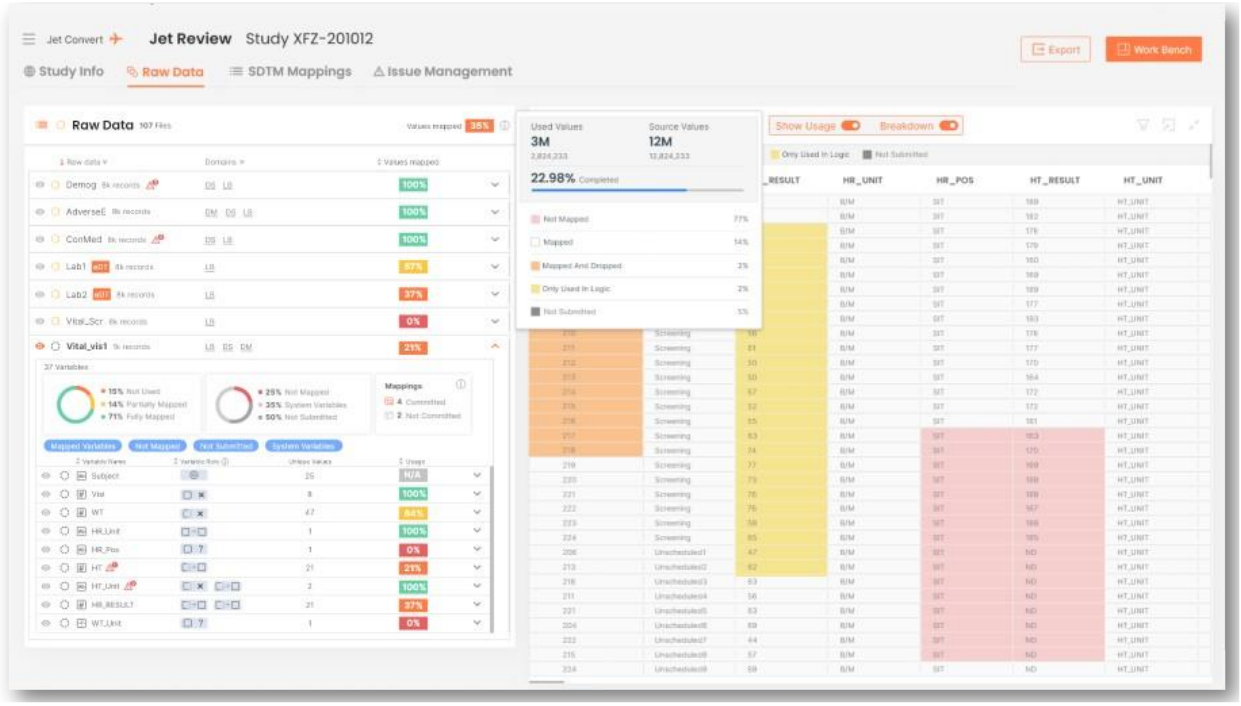

Figure 9 – Raw Data Review Module \ % Complete Tracking

How does this relate to real-time data validation?

Figure 6 – Real-time Validation

Mapping correctness is ensured by a robust review module which is updated every time new data flows into the system, hence the monitoring of conversion issues in real-time is very efficient with no manual effort needed to prepare the metrics for review.

The auto generated data completeness metrics ensure that as data changes through the conduct of a study it is very easy to monitor and ensure that all source data ends up in SDTM. For example, if a CRF change occurs where 3 new variables are added to a form, that form will immediately not show as 100% complete and raise an alert to the user. The user can then drill down to see exactly what data is unhandled and take the needed action to correct.

The elimination of double programming removes the requirement to run separate production and QC programs and compare them to find potential issues, rather the issues will be highlighted in the review tools focusing on identifying data anomalies and corrective actions rather than tedious back and forth updates between production and QC programs to make sure their results match.

Lastly, every time new data flows into the system the set of validation rules, not only the FDA business rules, but also CDISC CORE and any custom rules are run over the data and up to date validation results are provided with all the comments from the user given previously. Any new validation issues are alerted to the user immediately for review and action. This makes the review of validation issues very effective and efficient in real-time.

CONTINUOUS DATA FLOW

There are various issues that may cause failure in one or more components of the data flow and therefore stop the data from flowing as expected. The most common cause of such failures are structural changes to the source data and unexpected\unhandled data values.

Structural changes to the source data may occur during protocol amendments or it may be unexpected due to changes to the source system. It is recommended that during data acquisition a comparison is made to the previous version of the source data to detect any structural changes. If any structural changes occurred an alert should be raised immediately to prompt an investigation into the origin of the change and corrective actions required.

To avoid unexpected data values, it is recommended to use defensive programming in both production and QC programs to avoid the programs from failing, but rather identify, handle, and report the specific values in a predefined method. Reporting of these issues is essential to ensure corrective action is taken to resolve them. An example of unexpected data values are invalid dates. For example, dates should be programmed to check the validity of the data before processing it to SDTM, if the date is invalid it will not be processed to SDTM but rather reported for correction.

BENEFITS

Validated real-time SDTM data has significant benefits, including providing downstream consumers and decision-makers with up-to-date data to enable them to make accurate and timely decisions. SDTM data is used to generate analysis data and the analysis outputs (tables, listings, and figures). Having real-time SDTM data enables downstream users to more frequently refresh the analysis outputs without the dependency on the SDTM team to deliver the data manually. The SDTM data can also be used in standardized visualizations for data surveillance, having real-time data for data surveillance is critical.

Daily delivery of SDTM data will enable all downstream users to have access to the data on demand, rather than requesting an SDTM data transfer and having to wait for the resources to be available and the time to produce the data transfer. Such a request can take from 3 days to several weeks depending on resource availability.

This framework required minimal to no human effort to refresh the SDTM data daily which eliminates the costs associated with each SDTM re-run. If a SDTM re-run is performed manually and regularly for a project these costs can be significant due to the amount of human effort required.

CONCLUSION

Validated real-time SDTM data has significant benefits, including providing downstream consumers and decision-makers with up-to-date data to make accurate and timely decisions. The framework presented in this paper could significantly reduce the cost and time needed to obtain SDTM data and ensure the data is validated, reliable and compliant. A robust framework to provide validated real-time SDTM data will ensure readily available SDTM data, timely data issues resolution and no human effort to refresh the SDTM data.

CONTACT INFORMATION

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

Bremer Louw
Bioforum
bremer.louw@bioforumgroup.com
https://bioforumgroup.com/