PharmaSUG 2022 - Paper SI-065

PROC FUTURE PROOF 1.2 - Linked Data

Danfeng Fu, MSD China, Shanghai, CN;
Ellie Norris, Merck & Co., Inc., Kenilworth, NJ, USA;
Susan Kramlik, Merck & Co., Inc., Kenilworth, NJ, USA;
Suhas Sanjee, Merck & Co., Inc., Kenilworth, NJ, USA

ABSTRACT

Throughout the clinical study analysis and reporting (A&R) process, source data goes through several transformation steps in different phases. Although pharmaceutical companies have built streamlined processes to generate data and results in each phase, there is no automated, streamlined way to provide traceability as data moves from SDTM -> ADaM -> TLFs -> study reports. Data traceability is critical to ensure good quality and regulatory compliance. Previously we published a paper that evaluated recent advances in technology and the clinical trial programming skillset to identify opportunities for improved programming efficiencies. Last year, we published an overview of steps and challenges building a linked data proof of concept and a readout to provide data traceability from ADaM to SDTM. This year, we further evaluated table creation processes from ADaM datasets and expanded the use of linked data to enable automated traceability of analysis results in tables, listings, and figures (TLFs) back to ADaM datasets. We also devised a way to link analysis results/datapoints referenced in the study reports to TLFs which in turn can provide traceability back to ADaM & SDTM datasets. By representing all data from SDTM, ADaM and analysis results involved in the entire A&R process as linked data/graph database, we demonstrate end to end traceability from clinical study reports to SDTM data.

INTRODUCTION

In our 2021 paper, we made a case for using linked data in clinical trial analysis and reporting to enhance traceability, efficiency, and quality. Data traceability and quality are regulatory requirements. In the current state, traceability is not fully automated. Manual steps are labor intensive and have the potential to be error-prone. Figure 1 from our 2021 paper, below, shows the complexity and stepwise approach of data transformation for clinical trial analysis and reporting, following the straight arrows from left to right.

Linked data can improve traceability and quality of deliverables in an efficient, reliable and automated fashion. While the first three steps depicted in Figure 1 are automated, the final step, authoring the CSR, is manual. Metadata and carrying forward some variables from SDTM to ADaM data provide traceability between ADaM and SDTM. Standard programing code modules provide traceability connecting analysis data and tables, listings and figures in the step that produces TLF from ADaM data. Linked data, using shared identifiers and references to CDISC standards, could automate the currently manual step of populating the CSR with analysis data and links, in the third step. It also provides back references from the tables, listings, and figures to the analysis datasets (ADaM) and collected data (SDTM).

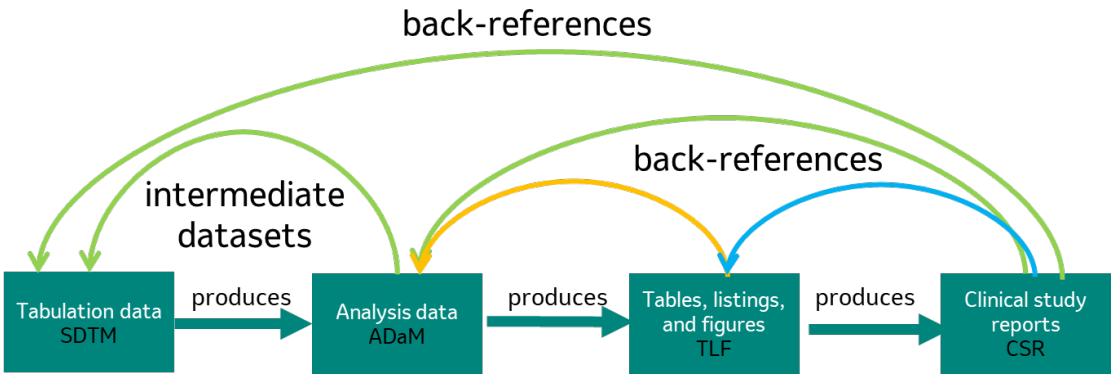


Figure 1. Data transformation steps with back references[1]

Using linked data, we can add back-references to the analysis and reporting (A&R) workflow, and this is represented by the curved arrows in Figure 1. Traceability from the CSR to TLFs (blue arrow) was demonstrated in our 2020 paper.[2] Traceability from TLFs to ADaM datasets (yellow arrow) was demonstrated by the PhUSE working group project "Analysis Results and Metadata in RDF".[3] Our 2021 paper provided details about a proof of concept for automating traceability (green arrows) from

1. ADaM to SDTM by providing users with the contributing variables and observations from SDTM that were used for deriving variables in ADaM
2. CSR narratives to datasets (SDTM/ADaM) by allowing authors to insert URIs from SDTM and ADaM into the reports rather than copy/pasting the results.

Although all of the back-references have been realized from above mentioned papers, we had not implemented TLF to ADaM traceability at our company. In this paper, we start from our last proof of concept (green arrows) and demonstrate TLF to ADaM traceability (yellow arrow) in our internal workflow. We realized this by updating our reporting macros to create TLF metadata to overcome some of the challenges encountered by adopting the approach described in the approach described in the paper from PhUSE working group.[3]

METHODOLOGY

Linked data can be implemented in clinical trial A&R TLF creation process in several ways.

One way is breaking the analysis and reporting into different layers and storing the values of sub-setting criteria along with the analysis results as proposed by PhUSE CS Semantic Technology Working Group.[3] In this method, there are the analysis layer and the reporting layer. In the analysis layer, analysis results are generated which are then stored in linked data as RDF data cubes. In the reporting layer, analysis results are retrieved from the analysis layer RDF and displayed in the desired format.

In this method, since all the analysis results are stored in the RDF format upfront, the same results used in different TLFs need not be derived again. They can be re-used by retrieving the results from RDF in the presentation layer. For instance, if number of participants within the population is used in 20 tables, it is only derived once and stored in a centralized RDF dataset rather than being derived in 20 table programs. The linkage of analysis results to ADaM is established in this approach by storing the data required to replicate the sub-setting criteria along with the analysis results.

While this method has a benefit of reducing the computational time and promoting consistency across the TLFs, it is not consistent with the design principles of the existing A&R infrastructure at our company. Hence it is very resource intensive to implement. In addition, storing the information required to replicate sub-setting criteria applied to ADaM is not efficient since it results in storing redundant information. This involves storing the selected values of the variables used to subset for each analysis result. Hence if there are "m"

analysis results and “n” variables used for sub-setting, we are looking at a “m by n” matrix along with the analysis results resulting in a very large dataset.

The approach proposed here involves storing the sub-setting criteria as separate metadata rather than with the analysis results to overcome some of the challenges with the approach adopted by the PhUSE CS Semantic Technology Working Group. To be able to establish the links to underlying data used from ADaM, the metadata needs to capture all the critical information in creating the A&R table. This is explained in detail in the following section.

IMPLEMENTATION

RDF OVERVIEW

The World Wide Web Consortium (W3C) originally designed the Resource Description Framework, commonly referred to as RDF, as a data model for metadata in 2003. RDF has since become a general method for description and exchange of graph data, as it provides a mechanism for allowing anyone to make a basic statement about anything. The basic building block for RDF is called the triple, which corresponds to the "subject-predicate-object" pattern. In comparison to the more common tabular form which represents data with rows, columns and cells, the RDF identifier for the row is called the subject of the triple, the column identifier is referred to as the predicate of the triple, and the value in the cell is named the object of the triple. A single node represents the subject or object while an edge between the two nodes represents the predicate.

Uniform Resource Identifiers (URIs) that distinguish one resource (i.e., node or edge) from another merge RDF triples, or statements, into a single model. An ontology, which is a formal specification of the concepts, types, properties, and interrelationships of entities within a domain of the real world, groups these single models together. Ontologies describe and link disparate and complex data so that it is understandable, useful, and durable. In 2004, the World Wide Web Consortium (W3C) derived the Web Ontology Language (OWL) and published it as a standard knowledge representation language for authoring ontologies as it provides both (1) expressive and flexible data modeling and (2) efficient automated reasoning.[4,5]

RDF IMPLEMENTATION

To provide traceability between the TLF summary data and ADaM observation details, the first step was to design an ontology for the relevant vocabulary terms with the OWL knowledge representation language. We generated the ontology with Apache Jena, an open-source Java framework for building Linked Data applications, to represent TLF metadata, TLF data and TLF mapping. The URI with the prefix "https://data.gin.merck.com/ontologies/TLFOntology identified each resource.

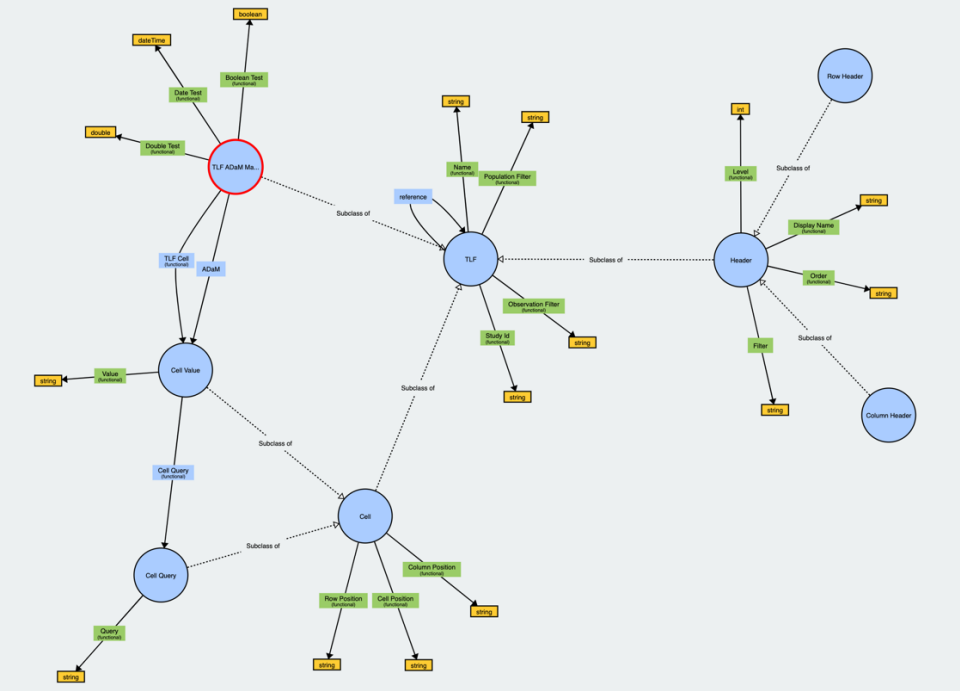


Figure 2. Graph visualization of TLF ADaM mapping (Credit: https://service.tib.eu/webvowl/#)

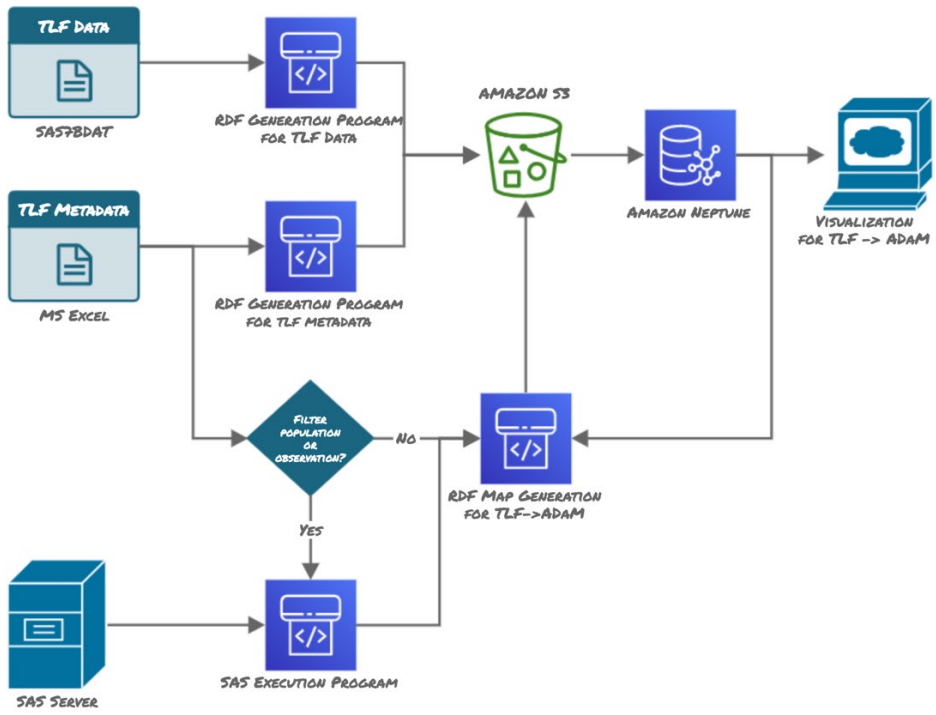


Figure 3. System Architecture & Data Processing Pipeline

We anonymized the input clinical trial data for use in creating tables, datasets and TLF metadata, which includes population and observation information. The three data sources encompassed SAS datasets (*i.e.,* SAS7BDAT files) with the binary encoded TLF data, Microsoft Excel files with the TLF metadata, and a SAS database containing the ADaM datasets. Here, we present the methods in which TLF metadata information traces back to source data.

Part I

During the first stage, we linked the ADaM dataset records to the previously generated ADaM URIs, using the unique key columns, by executing SAS procedure statements (i.e., SAS PROC) as a precursor for the subset creation based on population and observation conditions. A 2021 PoC[1] that experimented with establishing traceability between SDTM and ADaM records by transforming them into RDF produced these unique ADaM URIs.

Part II

In the second stage, we generated RDF triples for the (1) TLF metadata, (2) TLF data, and (3) the newly created subset dataset.

As a first step**,** we extracted the tabular TLF Metadata from the individual spreadsheets pertaining to studies, metadata, populations, and observations. We then mapped the tabular information listed in Table 1 to the ontology, as RDF graphs, among the categories of Metadata Basic RDF, Column Header, Row Header and Cell Link. Based on the population and observation sub-setting conditions present in the TLF metadata workbook's respective spreadsheets, we dynamically created two SAS programs so that the SAS datasets values were subset with the conditions specified in the metadata. Next, we extracted and parsed the TLF data from SAS datasets before transforming it into RDF graphs within the relevant portions of the ontology.

We loaded the resulting graph structures for each data source into an Amazon Simple Storage Service (S3) bucket as objects, and used a line-based, plain text format for encoding an RDF dataset known as N-quads.[6] We then uploaded the RDF data to Amazon Neptune, which is a fully managed graph database service for highly connected datasets.[7] Later, we retrieved the TLF metadata graphs from Neptune through the SPARQL Protocol and RDF Query Language (SPARQL),[8] which is designed to query information from data sources that can be mapped to RDF.

Part III

In the third stage, we established the links between TLF and ADaM datasets. We did this by creating explicit links between the TLF data and the corresponding ADaM data from which it was derived. A SAS program integrated the two datasets and established the linkage by first adding the ADaM URI information from Neptune to the ADaM SAS dataset. Then, we placed the ADaM SAS dataset into the SAS environment and ran a dynamic SAS program with the population and observation sub-setting conditions against this file. After deriving the population and observation subsets from the matching conditions, we retrieved the URIs and converted the data into RDF for upload into Neptune.

TLF Metadata Basic RDF	TLF Column Header	TLF Row Header	TLF Link SAS program
• name	• order	• order	• cellPosition

TLF Metadata Basic RDF	TLF Column Header	TLF Row Header	TLF Link SAS program
• tlfName • studyId • typeId • populationFilter • populationTable • observationFilter • observationTable	• name • displayName • level • filter • typeId • reference • tlfName • studyId	• name • displayName • level • filter • typeId • reference • tlfName • studyId	• sasProc • tlfCell • typeId • reference • tlfName • studyId

Table 1. TLF Metadata Attributes

Subject	Predicate	Object
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<https://data.gin.merck.com/ontologies/TLFOntology#CellValue>
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#cellPosition>	"0005-0020"
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#RowHeader>	<https://data.gin.merck.com/RowHeader/P111MK9999/asr0baseline0characteristics/0020>
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#ColumnHeader>	<https://data.gin.merck.com/ColumnHeader/P111MK9999/asr0baseline0characteristics/0005>
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#value>	"21"
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#typeId>	"CellValue"
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#reference>	<https://data.gin.merck.com/Tlf/P111MK9999/asr0baseline0characteristics>

Subject	Predicate	Object
<https://data.gin.merck.com/CellValue/P111MK9999/asr0baseline0characteristics/0005-0020>	<https://data.gin.merck.com/ontologies/TLFOntology#tlfName>	"asr0baseline0characteristics"

Table 2. TLF RDF Examples

TLF DATA AND ADAM/SDTM DATA

The ADaM and SDTM data used in this proof of concept has already been uploaded to database since the earlier proof of concept last year.[1]

The browser displayed the TLF data for the users to interact and benefit from the enhanced traceability provided from linked data implementation. TLF data had same values as in the created analysis table in RTF format. All the variable names in the TLF data also exist in TLF metadata as a column header so that every value in the TLF data can find a corresponding cell in TLF metadata.

Design of TLF Metadata Structure

The TLF metadata design is based on two aspects of our company's internal A&R process. One is the analysis table structure, and the other is the structure of existing standard SAS macros. The standard reporting macros were updated to create the TLF metadata in MS Excel format. The TLF metadata consists of three major components. They are the population criteria elements to reflect analyzed population, the observation criteria elements to reflect records retrieved from input dataset for analysis, and the criteria for each cell from corresponding row and column. More detailed descriptions of each of the metadata components are listed below. An example is presented below in Figure 4.

1. Population criteria elements include the dataset name for population records and criteria (in SAS/SQL syntax) for retrieving records that represent the target population from source dataset. Another important information we needed for population part is how table contents are arranged in the designated table layout. In most cases, a typical table presents analysis results for treatments and/or groups of population in different rows or columns. Therefore, providing the type of table structure in metadata, which is "column/row", is essential for dealing with data from different columns and rows. Refer to Figures 6 & 7 for examples for each type of table.
2. Observation criteria elements include the dataset name and criteria (in SAS/SQL syntax) for retrieving records that contribute to the results in the calculation and analysis.
3. Criteria for each cell includes row criteria and column criteria and the corresponding level information for the criteria in each row or column. Criteria for each cell in the table is combined from criteria for row and criteria for column. For inferential statistics, the metadata contained information about the SAS procedure used for calculation. Level information is included to indicate how the criteria in row dimension and column dimension are to be combined.

   Level indicators are used to differentiate the criteria for usage:

   - level=0 indicates the row/column for population;
   - level=1 indicates the criteria should be applied for a given row/column;
   - level in 2-5 indicates the criteria should also combine criteria from other rows/columns;
   - level=99 means procedure information should be displayed instead of contributing records from

the source dataset.

The final criterion for each value in the table is a combination of population, observation, row and the column criteria, in the way controlled by level indicator. Executing the final criterion for each value of the table displayed in the user interface (UI), can return the records contributing to the calculation of a given value.

TLF METADATA PREPARATION

Providing traceability from TLF to ADaM relies heavily on the TLF metadata. How to make the information provided in the metadata both correct and accurate is the biggest challenge we had in the POC. The approach we took is to make updates in standard table creation macros so that the metadata can be created at the same time when users generate the tables. As we have mentioned that the TLF metadata we created has the same layout (Figure 4) as the output table, each kind of description in the table (i.e., treatment group, description for rows) were analyzed and located in the table generation SAS macro. In the macro when creating the descriptive variables, we add other two variables to provide the information we need (criteria and level). And finally, we output all TLF metadata information to 3 sheets (table value sheet, population criteria sheet and observation criteria sheet) in an MS Excel file.

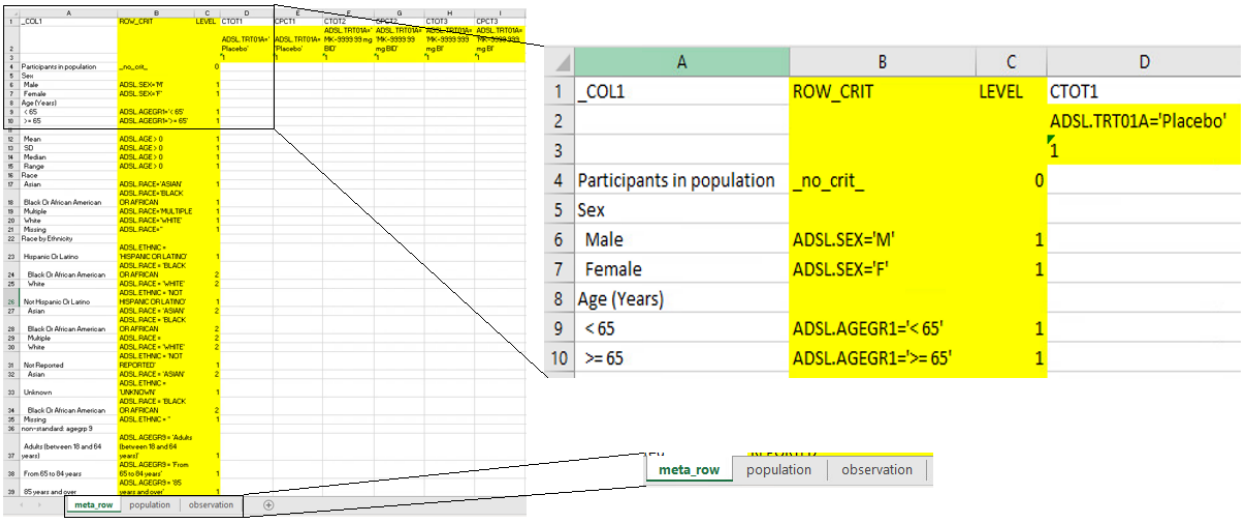


Figure 4. Example of TLF metadata with added information highlighted

READOUT

Figure 5 shows the screenshot of the data browser when you start. The user can select the study of choice and table within that study to browse.

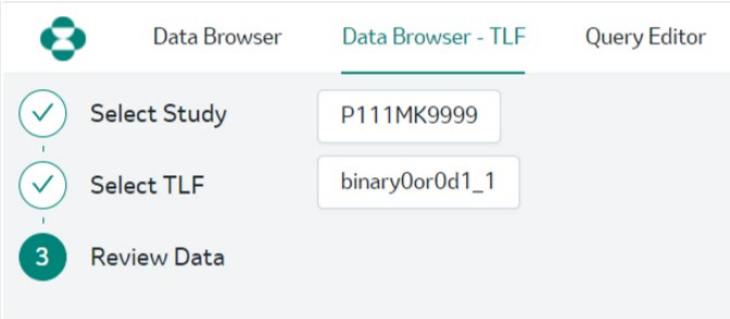
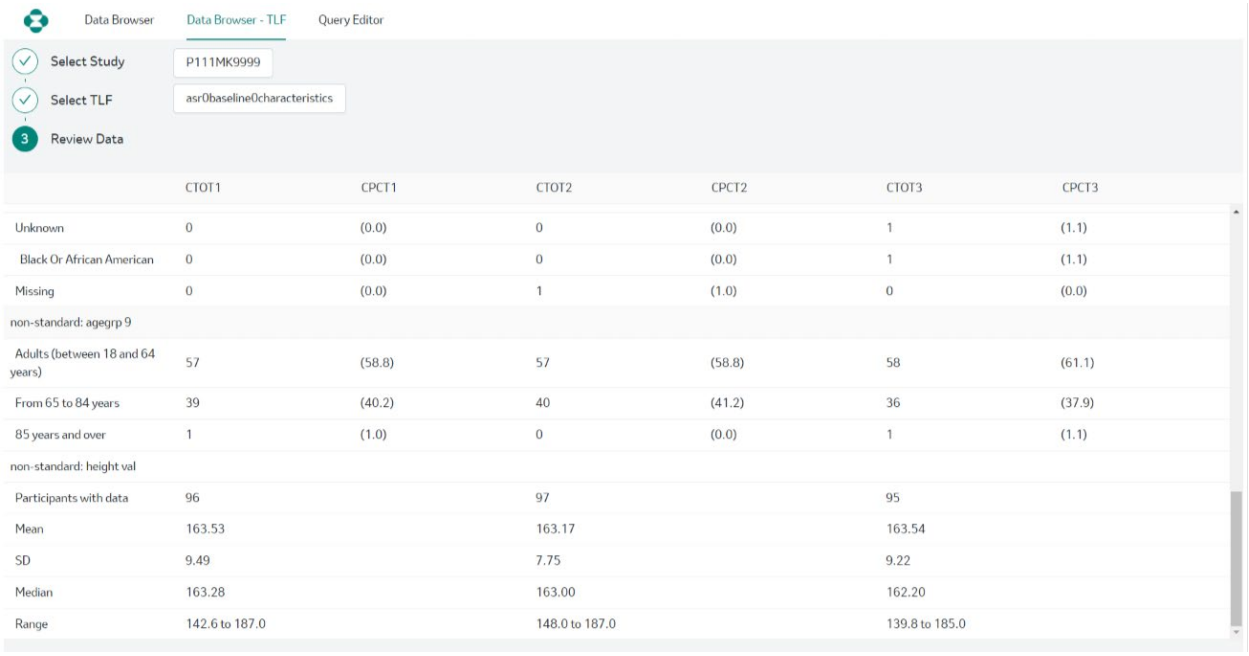


Figure 5. Data Browser for TLF



	CTOT1	CPCT1	CTOT2	CPCT2	CTOT3	CPCT3
Unknown	0	(0.0)	0	(0.0)	1	(1.1)
Black Or African American	0	(0.0)	0	(0.0)	1	(1.1)
Missing	0	(0.0)	1	(1.0)	0	(0.0)
non-standard: agegrp 9						
Adults (between 18 and 64 years)	57	(58.8)	57	(58.8)	58	(61.1)
From 65 to 84 years	39	(40.2)	40	(41.2)	36	(37.9)
85 years and over	1	(1.0)	0	(0.0)	1	(1.1)
non-standard: height val						
Participants with data	96		97		95	
Mean	163.53		163.17		163.54	
SD	9.49		7.75		9.22	
Median	163.28		163.00		162.20	
Range	142.6 to 187.0		148.0 to 187.0		139.8 to 185.0	

Figure 6. UI presenting data from table

Once the user selects the study and TLF, same values as the ones in the created A&R table (in RTF format) will be displayed in the UI. As you can see in Figure 6, our proof of concept supports a table layout that presents results for different treatments across different columns. It also supports another kind of table layout (Figure 7), in which results for different treatments present across different rows.

	_NVAR	_SVAR	CI	PVALC
Placebo	32	13	4.10 (1.33, 12.61)	0.0391
MK-9999 99 mg BID	34	10	2.31 (0.74, 7.19)	0.2246
MK-9999 999 mg BI	35	7		

Figure 7. Table with treatment groups in rows

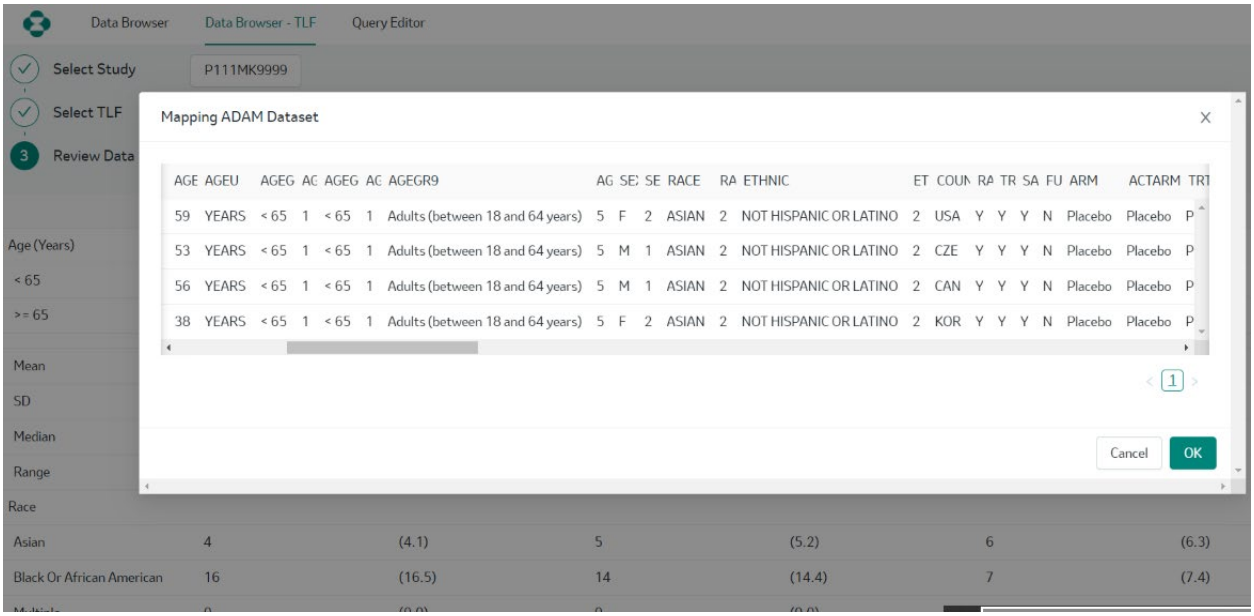


Figure 8. UI providing contributing records from ADaM dataset

When user clicks on the value of “4” from one treatment group in the section of “Race - Asian” group, which was created in a typical baseline characteristic table, the UI can provide 4 records from ADSL that contributed to this value i.e. variable RACE= “ASIAN” in the pop-up window.

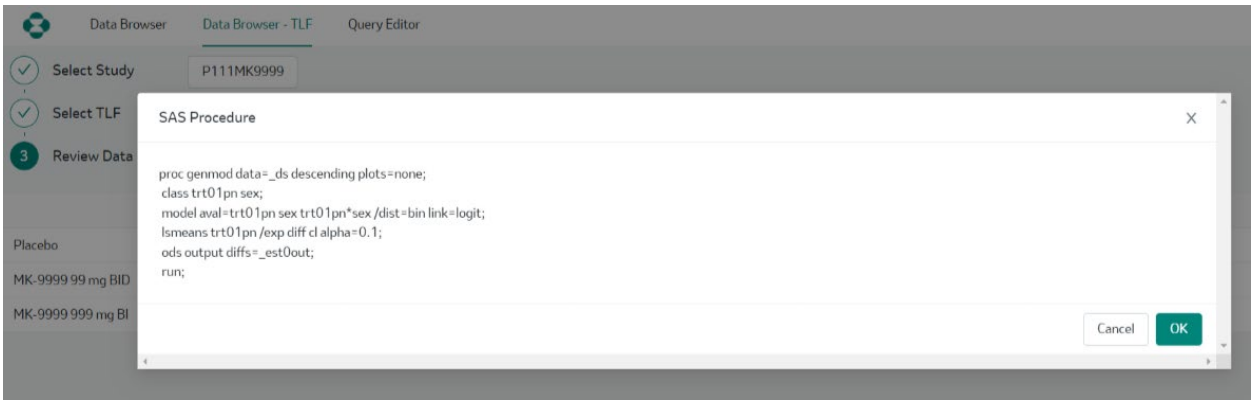


Figure 9. UI providing SAS procedure used to derive analysis results

When user clicks on any value with inferential statistics such as p-value in Figure 7, information about the SAS procedure used in the calculation will be provided in the pop-up window (Figure 9).

Discussion

In this section, we discuss several challenges encountered during the POC. We also discuss the benefits and cost resulting from implementation of linked data in clinical trial A&R workflow.

The Challenge

Providing data traceability is like building a bridge between analysis results in the TLFs (output) with ADaM and SDTM data used for derivation of analysis results (inputs). The bridge needs fundamental construction to support it. The construction of the bridge should meet the requirement of providing all critical elements from the input side of the bridge and information needed from the output side. Though it is not too difficult to sum up the things from both sides, there are different ways to build the bridge (TLF metadata layout) and different kinds of material to use (how to create the metadata). And the challenges we had were to choose from different ways to build the bridge and to choose the materials

Below is a summary of the benefits and cost of linked data implementation.

Benefits

1. Easy to understand and to create metadata

Having metadata with the same layout as the analysis table makes it easier for users to know the final criterion to be used for each value. The convenience of quickly locating criteria for each row and column helps users not only to check the accuracy of the criteria but also to make modifications in the criteria creation process.

2. Providing detailed and robust SAS procedure information

For some statistical results in an A&R table, detailed SAS procedure information appears to have the same importance in traceability as contributing dataset records. Utilizing the approach used in the proof of concept to update standard SAS macros for creating TLF metadata makes it possible to obtain exactly same SAS procedure information as was used in the analysis.

3. Accuracy

Deciding to create TLF metadata by updating table generation SAS macros is also due to the accuracy of the criteria. Although it is possible to utilize descriptive text in created TLF to lookup values in ADaM datasets, i.e. a descriptive text “>65” can be found in variable AGEGRP1 from ADSL, so that many criteria can be derived based on this kind of rationale, we still cannot guarantee the accuracy of all the derived criteria in this approach. While using updated table generation SAS macros, we can make sure every created criterion follows the same logic as the one used to create the analysis results.

Cost

While linked data adoption in clinical trial A&R has several benefits, it comes with a cost which is described in this section.

1. Effort to update and maintain existing SAS macros

The approach we adopted requires modifications to existing table generation SAS macros to include additional information in the metadata. These modifications inevitably increase the complexity of the macros. Meanwhile, making updates also require comprehensive understanding of the macro. Gaining this level of familiarity and understanding is time consuming.

2. Resolved code

Commonly, if a clinical study needs to be submitted to regulatory agency, the resolved code from the SAS macro to generate the analysis results presented in the TLF should also be included in the submission package. Since the SAS macros were updated to generate the TLF metadata, additional code to support metadata creation shows up in the resolved SAS program. This additional code unrelated to creating the analysis results presented in the TLFs, burdens reviewers in understanding the program. One of the options to deal with this situation is to add a post-processing step to remove the pieces of code used for creating the metadata.

3. Dealing with non-standard tables

When running tables using standard SAS macros developed by most pharmaceutical companies, making updates to the standard macros can be a one-time job. However, almost every clinical study has some study specific tables that cannot be covered by the standard macros. For creating TLF metadata for these tables, study programmers need to learn and practice to build each metadata in correct layout and content, according to the approach we designed. A potential area to explore in our future work is to improve and/or automate the metadata creation process.

Impact of linked data adoption in clinical trial A&R

Traceability is critical in clinical trial A&R. By representing SDTM, ADaM and TLF results as linked data, end-to-end traceability from CSR all the way back to SDTM data can be realized.

To achieve traceability from ADaM to SDTM, the programs generating these deliverables need to be updated, refer to [1] for more details.

To achieve traceability from TLF to ADaM, the TLF macros need to be updated to generate the metadata as explained in the methodology section and to include the sub-setting criteria required to retrieve the ADaM data used for the TLF. The metadata also must contain information about the SAS procedure used to generate the statistics. This can be used to automate generation of Analysis Results metadata (ARM) which is a regulatory submission deliverable.

In addition to updating the TLF macros, programs generating ADaM datasets need to be updated to generate intermediate datasets to facilitate links between ADaM and SDTM.

There is no straightforward way in the current state to link the statistics in the study reports to the TLFs delivered by clinical trial programmers. Authors of study reports and manuscripts often review hundreds of TLFs to identify which statistics to include in reports and publications. When the TLFs are refreshed the author must make sure the statistics used in the reports are still accurate and consistent with the updated TLFs. This process is manual, cumbersome, and hence error prone. Linked data can be used to provide better traceability between TLFs and study reports.

Once the TLF data are converted to RDF, each datapoint/result will have a URI. As shown in the previous section, the URI of these results can be referenced in the study reports and the results can be retrieved by querying the RDF rather than copying the statistics from TLFs. This will make sure that the results used in study reports are consistent with those in the TLFs. This will also reduce manual effort required for performing quality checks on the study documents. This can be extended to referencing data points as well, e.g. most frequently occurring AEs.

Conclusion

In the phase 2 of linked data proof of concept, we successfully integrated TLF to ADaM and analysis results to CSR into the application we built so that end-to-end traceability from CSR back to SDTM was established.

Key elements for success of this proof of concept were

- a reliable and close partnership and alignment between our company's BARDS (Biostatistics and Research Decision Sciences) and IT organizations,
- a carefully designed project framework, carefully designed and constructed metadata,
- the use of knowledge graphs.

The alignment and partnership between the BARDS and IT organizations have been necessary for the obvious reason of acting in concert to move the project forward, but also due to specialized knowledge in each of these organizations. This project has been ongoing for 3 years, with continual alignment and partnership. It was important to listen to one another to understand needs and limitations and to have a framework and project plan to keep everyone focused on the goals and upcoming steps within the project.

The metadata contains the information that drives the content and format of the output. This was important to overcome some of the challenges of implementing the approach in the PhUSE working group project[3] in our company's A&R framework.

The advantages of linked data and metadata are the automated traceability from final CSR back to tables listings and figures, to ADaM data, and from ADaM data to SDTM. However, the cost of metadata construction and the added complexity to the existing global A&R standards is considerable. Expanding the linked data and metadata methodology and construction beyond the global A&R standards to therapeutic area standards and project-specific TLFs would take additional work, and could actually increase time needed to produce deliverables. It will be important to determine if the cost is worth the benefit of automated traceability. In the meantime, we can continue to explore ways to reduce the cost.

References

1. Gillespie, A., Kramlik, S., Mynarz, J., Fu, D., Sanjee, SR. 2021 "PROC Future Proof v1.1- Linked Data" Presented at PharmaSUG Virtual conference. Available at https://www.pharmasug.org/proceedings/2021/SI/PharmaSUG-2021-SI-046.pdf
2. Gillespie, A., Kramlik, S., Sanjee, SR. 2020 "PROC Future Proof" Presented at PharmaSUG Virtual conference 2020. Available at https://www.lexjansen.com/pharmasug/2020/SI/PharmaSUG-2020-SI-173.pdf
3. Andersen, M., Hungria, M., Sanjee, S. (2016) Generating Analysis Results and Metadata. PhUSE EU Connect, 2016. Available at https://www.lexjansen.com/phuse/2016/tt/TT05.pdf
4. Allemang, D. and Hendler, J. (2009). Semantic Web for the Working Ontologist. Morgan Kaufmann.
5. Learn OWL and RDFS 2022, Cambridge Semantics, accessed 4 March 2022, https://cambridgesemantics.com/blog/semantic-university/learn-owl-rdfs/owl-101/
6. RDF 1.1 N-Quads 2022, W3C, accessed 4 March 2022, https://www.w3.org/TR/n-quads/
7. What Is Amazon Neptune? 2022, AWS, accessed 7 March 2022, https://docs.aws.amazon.com/neptune/latest/userguide/intro.html
8. What is SPARQL? 2022, Ontotext, accessed 7 March 2022, https://www.ontotext.com/knowledgehub/fundamentals/what-is-sparql/

Acknowledgments

The authors would like to recognize the following subject matter experts from our IT and statistical programming departments who partnered with us and worked diligently on this proof of concept: Amy Gillespie, Ananth Kumarasamy, Dinesh Kalawadia, Jan Bárta, Guowei Wu, Eric Qi, Murali Talanayar, Richa Shukla and Martyna Kolacka.