Paper DS02

The Significance of Generic Data Transfer Specification for Non-CRF Data in Clinical Trials

Samaneh Sabiani, Merck KGaA, Darmstadt, Germany
Kristina Zweier, Merck KGaA, Darmstadt, Germany

ABSTRACT

Non-CRF data in clinical trials encompasses information not captured in the CRF, including data from medical devices, lab equipment, imaging systems, patient diaries, and other patient-reported outcomes. This data provides vital context, support, and evidence for trial outcomes but is often less structured, posing standardization and integration challenges. Standardizing non-CRF data involves mapping diverse data types to a common format aligned with CDISC standards. A generic data transfer specification is crucial for ensuring data consistency, accuracy, and establishing a mapping roadmap for SDTM compliance. It also provides a structured approach for mapping data to SDTM when direct compliance is unfeasible, enhancing interoperability and harmonization of clinical trial data.

INTRODUCTION

A data transfer specification (DTS) refers to a set of guidelines, rules, and technical specifications that outlines the process of transferring data from one system or platform to another. It defines the structure, format and mapping required for successful data transfer to be used for secure data transmission. Having a data transfer specification is a standardized approach for exchanging data between different systems, applications, or organizations. It helps to identify data inconsistencies, errors, and loss during the transfer process, ensuring that the data remains consistent and usable across different platforms or environments. Data transfer specifications are particularly important in various domains such as healthcare, finance, and research, where accurate and reliable data exchange is critical. Adhering to well-established data transfer specifications enables seamless data integration, sharing, and ensures the accuracy, integrity, and security of the data. In this paper we discuss three different scenarios on how non-CRF data is regularly provided by vendors to Merck and how Merck utilizes its generic DTS template to guide data collection and mapping standardization.

Problem statement:

In clinical trials, it is critical that data adheres to the Standard Data Tabulation Model (SDTM) Controlled Terminology (CT) from the outset. Historically, manual processes have led to late discoveries of non-compliant data, necessitating time-consuming fixes during the SDTM programming phase. This reactive approach not only complicates the integration of external data but also increases the burden on SDTM programming efforts, often resulting in submission delays. To improve this process, we aim to ensure compliance from the beginning, which will facilitate better management of external data and streamline the overall workflow. Our solution involves creating a standardized DTS template that aligns with SDTM CT requirements. This template can also be utilized for validation in Pinnacle 21, a free tool, thereby reducing the effort required for validation and the maintenance of new validation tools. By implementing this approach, we seek to enhance data standardization and ensure better compliance throughout the clinical trial process.

Benefits of a generic DTS template for requesting non-CRF data in an SDTM format from source

The implementation of a generic Data Transfer Specification (DTS) template that can be adapted for all non-Case Report Form (CRF) data/assay types is pivotal for enhancing the efficiency and quality of data integration in clinical studies. By standardizing the request for non-CRF data in a Study Data Tabulation Model (SDTM) format directly from laboratories, we can achieve several key benefits:

1. **Standardization Across Data Types**: A generic DTS template ensures that all non-CRF data, regardless of its source or type, is collected in a consistent manner. This standardization simplifies the integration process and reduces variability, facilitating a smoother aggregation of data from diverse assays and sources.
2. **Streamlined Data Collection**: By utilizing a standardized template, laboratories can more easily understand and fulfill data requests. This leads to quicker turnaround times and minimizes the potential for errors during data submission, ultimately accelerating the clinical study timeline.

3. **Enhanced Data Quality:** Requesting data in an SDTM-compliant format from the outset helps maintain data integrity and quality. A well-defined DTS template promotes clarity in data expectations, reducing the likelihood of errors that can arise from manual data conversion at a later step.
4. **Facilitated Regulatory Compliance:** A generic DTS template that aligns with SDTM requirements ensures that the data is compliant with regulatory standards from the beginning. This proactive approach simplifies the submission process to regulatory authorities, mitigating the risk of delays due to data quality issues.
5. **Improved Data Integration:** A consistent format for non-CRF data allows for more efficient data mapping and integration into the overall clinical study database. This enhances the ability to analyze and interpret data effectively, leading to more robust study outcomes.
6. **Adaptability for Diverse Assays**: The flexibility of a generic DTS template allows it to be tailored for various assay types, ensuring that all relevant data can be captured and integrated without extensive modifications to the initial request.
7. **Compliance from the Beginning:** By ensuring that data follows SDTM CT requirements from the outset, the template helps manage external data more effectively while reducing the burden on SDTM programming by ensuring submission compliance from the start.
8. **Reduced Validation Efforts**: The template can be utilized for validation in Pinnacle 21, a free tool, minimizing the effort required for validation and maintenance of new validation tools, while ensuring standardization and better compliance of data.

What does DMMI* Standards Generic DTS Template include?

The DMMI Standards generic DTS Template is an Excel file which includes at least the following mandatory sheets:

1. Cover
2. Contact Information
3. Transfer Specifications
4. Data Structure
5. Value Level Metadata (VLM), in case the vendor provides data in adherence with Merck's non-CRF collection data standards or VLM-lookup, in case the vendor cannot provide the data in adherence with Merck's non-CRF collection data standards
6. Visits

* Data Monitoring, Management & Innovation

DTS solutions and SDTM validations for different data structure scenarios

Depending on the vendor's internal processes, provided data can vary in structure and terminology. No matter how the data can be provided, the Generic DTS Template is conceptualized as a tool to guide the correct data standardization. The following three scenarios are examples of data structures Merck has encountered. The presented tables show the data structure, VLM/VLM-Lookup sheets of a typical DTS used within Merck. Merck's Data Integration managers are setting up the DTS based on study needs and Merck's Standard experts provide SDTM validation.

Scenario 1: Data belongs to one SDTM domain and can be provided with full adherence to Merck's data collection standards

For the creation of the data structure for a specific assay type, it is recommended to use the given Data Structure in the Generic DTS template (which is based on the Findings Structure according to the SDTMIG) and fill the variables/columns according to the SDTMIG requirements for the respective domain. The below table shows an example for a data structure sheet of genetic findings DTS.

Figure 1.1: Data Structure sheet

Variable Name	*Variable Label*	*Type*	*Length*	*Codelist or format*	*SUPP--*	*Comments*
STUDYID	Study Identifier	Char	20			XXXXXXXX
SUBJID	Subject Identifier	Char	20			e.g. 1001001
GFTESTCD	Short Name of Genomic Measurement	Char	8	(TESTCD)		See VLM tab!
GFTEST	Name of Genomic Measurement	Char	40	(TEST)		See VLM tab!
GFTSTDTL	Measurement, Test, or Examination Detail	Char	200	(GFTSDTL)		See VLM tab!
GFORRES	Result or Finding in Original Units	Char	200			See VLM tab!

GFORRESU	Original Units	Char	15	(UNIT)		See VLM tab!
GFSPEC	Specimen Material Type	Char	20	(GENSMP)		See VLM tab!
GFMETHOD	Method of Test or Examination	Char	30	(METHOD)		See VLM tab!

Value Level Metadata (VLM) is a component of data standards, such as CDISC (Clinical Data Interchange Standards Consortium) standards, that provides detailed information about individual variable values. VLM includes additional contextual information and annotations related to the measurement values captured in a dataset. The below table shows an example for a VLM sheet of genetic findings DTS.

Figure 1.2: VLM sheet

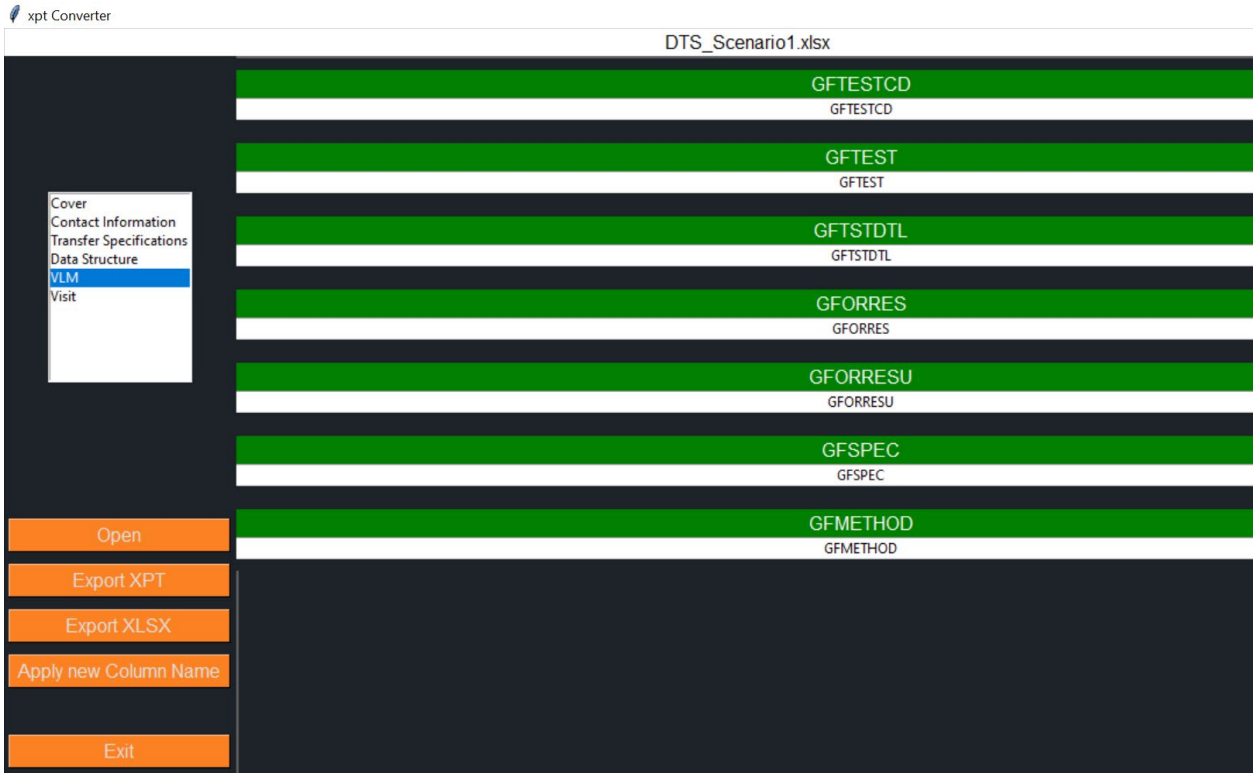
GFTESTCD	**GFTEST**	**GFTSTDTL**	**GFORRES**	**GFORRESU**	**GFSPEC**	**GFMETHOD**
MICRISTB	Microsatellite Instability	MICROSATELLITE INSTABILITY OVERALL STATUS	HIGH		DNA	POLYMERASE CHAIN REACTION
MICRISTB	Microsatellite Instability	MICROSATELLITE INSTABILITY OVERALL STATUS	STABLE		DNA	POLYMERASE CHAIN REACTION
TMB	Tumor Mutation Burden	NORMALIZED NUMBER OF SEQUENCE VARIANTS	*numeric*	/10^6 bp	ctDNA	NEXT GENERATION TARGETED SEQUENCING
TMB	Tumor Mutation Burden	VARIANT SEQUENCE BURDEN INTERPRETATION	HIGH		ctDNA	NEXT GENERATION TARGETED SEQUENCING
TMB	Tumor Mutation Burden	VARIANT SEQUENCE BURDEN INTERPRETATION	LOW		ctDNA	NEXT GENERATION TARGETED SEQUENCING

SDTM Validation for Scenario 1

Once a DTS file is prepared, it is reviewed by Standards experts for SDTM alignment. The Standards expert reviews the data structure sheet to make sure that naming conventions used in the data structure for variable names, variable labels and controlled terminology align with SDTM naming conventions.

In order to validate an excel file based DTS according to SDTM rules, the Standards experts group is converting the VLM sheet of the DTS to an xpt file with a Python-language based tool. Afterwards we are running the xpt file through the Pinnacle 21 Community version. This way we are making sure that non-CRF data is truly SDTM/SDTM-like from the beginning.

The screenshot below shows our xpt converter tool. The DTS file is read in and the VLM sheet is selected. The tool shows that the variable names are SDTM compliant.



By clicking “Export XPT”, the VLM sheet can be saved to one’s PC. For a smooth validation run with Pinnacle 21 Community version, the xpt file is renamed to the respective SDTM domain it represents, in this example “gf.xpt”.

Afterwards, the xpt file is run through the Pinnacle 21 Community version. During the review, only domain specific findings are regarded. The below screenshot shows domain specific findings from the Pinnacle 21 report.

Variables	Values	Pinnacle 21	Message	Category	Severit	Standards Expert Comment
GFORRESU	/10^6 bp	CT2002	GFORRESU value not found in 'Unit' extensible codelist	Terminology		Please use the SDTM submission value of "/MBP" as a unit. The listed unit in the DTS is only a synonym for the submission and not do be used.
VARIABLE, DATASET	DOMAIN, GF	SD0056	SDTM Required variable DOMAIN not found	Metadata		okay
VARIABLE, DATASET	GFSEQ, GF	SD0056	SDTM Required variable GFSEQ not found	Metadata		okay
VARIABLE, DATASET	STUDYID, GF	SD0056	SDTM Required variable STUDYID not found	Metadata		okay
VARIABLE, DATASET	USUBJID, GF	SD0056	SDTM Required variable USUBJID not found	Metadata		okay
VARIABLE	GFDTC	SD0057	SDTM Expected variable GFDTC not found	Metadata		okay
VARIABLE	GFREFID	SD0057	SDTM Expected variable GFREFID not found	Metadata		okay
VARIABLE	GFSTRESC	SD0057	SDTM Expected variable GFSTRESC not found	Metadata		okay
VARIABLE	VISITNUM	SD0057	SDTM Expected variable VISITNUM not found	Metadata		okay

The Pinnacle 21 report shows that a unit was used in the DTS file, which is not an SDTM submission value from the SDTM CT codelist for Units (C71620). The Standards expert advises the usage of the correct SDTM submission value. The other listed findings for the GF domain are irrelevant in the context of the DTS validation and can therefore be ignored.

The Standards expert provides the DTS creator with the feedback from the Pinnacle 21 report. The DTS is updated and reviewed again by the Standards expert until no SDTM mapping issues are present anymore. Once the DTS receives the green light from the Standards expert, it can be used in the further processes of setting up data transfers from the vendor as well as being the source for SDTM mapping instructing for future SDTM domain programming.

Scenario 2: Data belongs to different SDTM Domains and can be provided with full adherence to Merck’s data collection standards

The table below provides an example of a data structure sheet for data from various SDTM domains (MB, LB, IS, CP). It is important to avoid including the SDTM Domain Abbreviation in the variable names. Instead, you can include

a separate column for the abbreviation in the dataset (Figure 2.1) and specify the correct assignments in the VLM/VLM-Lookup table (Figure 2.2).

Figure 2.1: Data Structure sheet

Variable Name	Variable Label	Type	Length	Codelist or Format	SUPP--	Comments
Domain	Domain Abbreviation	Char	2			See VLM tab!
STUDYID	Study Identifier	Char	20			XXXXXXXX
SUBJID	Subject Identifier	Char	20			e.g. 1011002010
TESTCD	Short Name of Measurement, Test or Examination	Char	8	(TESTCD)		See VLM tab!
TEST	Name of Measurement, Test or Examination	Char	40	(TEST)		See VLM tab!
BDAGNT	Binding Agent	Char	200	(ISBDAGT)		See VLM tab!
MRKSTR	Marker String	Char	200			See VLM tab!
TSTPNL	Test Panel	Char	200			See VLM tab!
CAT	Category	Char	200			See VLM tab!
ORRES	Result or Finding in Original Units	Char	200			See VLM tab!
ORRESU	Original Units	Char	15	(UNIT)		See VLM tab!
SPEC	Specimen Material Type	Char	20	(SPECTYPE)		See VLM tab!
METHOD	Method of Test or Examination	Char	30	(METHOD)		See VLM tab!

Figure 2.2: VLM sheet

Domain	TESTCD	TEST	BDAGNT	MRKSTR	TSTPNL	CAT	ORRES	ORRESU	SPEC	METHOD
MB	MTB	Mycobacterium tuberculosis				TUBERCULOSIS SCREENING	Positive		BLOOD	MICROSCOPY
MB	MTB	Mycobacterium tuberculosis				TUBERCULOSIS SCREENING	Negative		BLOOD	MICROSCOPY
MB	MTB	Mycobacterium tuberculosis				TUBERCULOSIS SCREENING	Indeterminate		BLOOD	MICROSCOPY
LB	WBC	Leukocytes				HEMATOLOGY	*numeric*	10^9/L	BLOOD	AUTOMATED COUNT
IS	MBAB	Microbial-induced Antibody	HEPATITIS B VIRUS CORE ANTIGEN			SEROLOGY	Reactive		SERUM	ELECTROCHEMILUMINESCENCE
IS	MBAB	Microbial-induced Antibody	HEPATITIS B VIRUS CORE ANTIGEN			SEROLOGY	Non-reactive		SERUM	ELECTROCHEMILUMINESCENCE
CP	TLYCE	T-Lymphocytes		CD3+ ABS	A167 QTBNK Assay	IMMUNOPHENOTYPING	*numeric*	10^6/L	BLOOD	FLOW CYTOMETRY

SDTM Validation for Scenario 2

In Scenario 2, the VLM sheet presents with a first column that shows the respective SDTM domain the content is supposed to be mapped to. The columns with the variable names are domain independent without the respective two letter domain prefix. Furthermore, not all listed variables are always used in each SDTM domain, which needs to be taken into consideration when doing the SDTM validation of the DTS.

In order to create valid SDTM domain xpt files, we are using another Python-language based tool to extract the information from the VLM sheet and create several xpt files for each SDTM domain listed in the Domain column.

The below screenshot shows this tool with the information about the different domains the tool detected in the VM file.

When clicking on the “Select” button, a new toplevel window pops up that lists the individual variables in the VLM sheet.

Clicking on the “Split” button at the top of the new window, creates an excel file with the respective SDTM domain information in individual sheets. The screenshots below show the MB and IS sheets with the various VLM variables now with the domain prefixes.

	MBDomain	MBTESTCD	MBTEST	MBBDAGNT	MBMRKSTR	MBTSTPNL	MBCAT	MBORRES	MBORRESU	MBSPEC	MBMETHOD
2	MB	MTB	Mycobacterium tuberculosis				TUBERCULOSIS SCREENING	Positive		BLOOD	MICROSCOPY
3	MB	MTB	Mycobacterium tuberculosis				TUBERCULOSIS SCREENING	Negative		BLOOD	MICROSCOPY
4	MB	MTB	Mycobacterium tuberculosis				TUBERCULOSIS SCREENING	Indeterminate		BLOOD	MICROSCOPY

Cover | MB | LB | IS | CP

	ISDomain	ISTESTCD	ISTEST	ISBDAGNT	ISMRKSTR	ISTSTPNL	ISCAT	ISORRES	ISORRESU	ISSPEC	ISMETHOD
2	IS	MBAB	Microbial-induced Antibody	HEPATITIS B VIRUS CORE ANTIGEN			SEROLOGY	Reactive		SERUM	ELECTROCHEMILUMINESCENCE
3	IS	MBAB	Microbial-induced Antibody	HEPATITIS B VIRUS CORE ANTIGEN			SEROLOGY	Non-reactive		SERUM	ELECTROCHEMILUMINESCENCE

Cover | MB | LB | IS | CP

The SDTM expert can now prepare the individual sheets for conversion to xpt files by updating variable names like the domain column that now has a name of “LBDomain” or by removing variables/columns that are not used in this specific domain.

Afterwards the same steps are repeated as in scenario 1. The excel file with the individual SDTM domain sheets is read into the xpt conversion tool, the individual xpt files are generated and then run through the Pinnacle 21 Community version for the SDTM validation.

Scenario 3: Data cannot be provided in adherence to Merck’s data collection standards

The table below represents an example of a data set which does not follow any existing standards for structure and terminology. It includes various columns capturing sample information and horizontally provided Lab Tests which might follow by their associated units along with their corresponding results. Further relevant columns such as Specimen Type, Specimen Condition, Method, Vendor Evaluation etc. which are typically included in a dataset, and for which controlled terminology is available are represented by the column “Other Variables” for a simplified representation.

Figure 3.1: Vendor’s data structure specification

Study_ID	Patient_ID	MSI_High	TMB_score (10^6 bp)	TMB_category	Specimen	Test Method
XXXXXXXX	1001001	Detected	*numeric*	Low	DNA	PCR

The table below illustrates an example of a data structure sheet. It shows how the vendor's data specifications, including "Field Name," "Type," and "Field Description" (on the left), are transformed into the data structure format of Merck's Non-CRF Data Collection Standards (on the right).

Figure 3.2: Data Structure sheet

Vendor's Data Specification			*Merck's Non-CRF Data Collection Standards*						
Field Name	**Type**	**Field description**	**Variable Name**	**Variable Label**	**Type**	**Length**	**Codelist or format**	**SUPP--**	**Comments**
Study_ID	varchar	Study identifier from customer	STUDYID	Study Identifier	Char	20			XXXXXXXX
Patient_ID	varchar	Patient identifier from customer	SUBJID	Subject Identifier	Char	20			e.g. 1001001
MSI_High	varchar	MSI status indicates whether the sample is MSI-High "Detected" or "Not Detected"	GFTEST	Name of Genomic Measurement	Char	40	(TEST)		See VLM-Lookup tab!
TMB_score (10^6 bp)	double precision	TMP score	GFTEST	Name of Genomic Measurement	Char	40	(TEST)		See VLM-Lookup tab!
TMB_category	varchar	TMP score categories are "High" and "Low"	GFTEST	Name of Genomic Measurement	Char	40	(TEST)		See VLM-Lookup tab!
Specimen	varchar	Specimen Material	GFSPEC	Specimen Material Type	Char	20	(SPECTYPE)		See VLM-Lookup tab!
Test Method	varchar	Test Method	GFMETHOD	Method of Test or Examination	Char	30	(METHOD)		See VLM-Lookup tab!

The table below provides an example of a VLM-Lookup sheet. It demonstrates how the vendor's lab test specifications, including the "Vendor Lab Test Name" (on the left), are transformed into the data structure format of Merck's Non-CRF Data Collection Standards (on the right).

Figure 3.3: VLM-Lookup sheet

Vendor Lab Test name	**GFTESTCD**	**GFTEST**	**GFTSTDTL**	**GFORRES**	**GFORRESU**	**GFSPEC**	**GFMETHOD**
MSI_High	MICRISTB	Microsatellite Instability	MICROSATELLITE INSTABILITY OVERALL STATUS	HIGH		DNA	POLYMERASE CHAIN REACTION
MSI_High	MICRISTB	Microsatellite Instability	MICROSATELLITE INSTABILITY OVERALL STATUS	STABLE		DNA	POLYMERASE CHAIN REACTION
TMB_score (10^6 bp)	TMB	Tumor Mutation Burden	NORMALIZED NUMBER OF SEQUENCE VARIANTS	*numeric*	/10^6 bp	ctDNA	NEXT GENERATION TARGETED SEQUENCING
TMB_category	TMB	Tumor Mutation Burden	VARIANT SEQUENCE BURDEN INTERPRETATION	HIGH		ctDNA	NEXT GENERATION TARGETED SEQUENCING
TMB_category	TMB	Tumor Mutation Burden	VARIANT SEQUENCE BURDEN INTERPRETATION	LOW		ctDNA	NEXT GENERATION TARGETED SEQUENCING

SDTM Validation for Scenario 3

The SDTM validation of a DTS in scenario 3 follows the same steps as in scenarios 1 or 2 with the exception that the Standards expert extracts only the SDTM information from the VLM-Lookup sheet into a new excel file before running the domain splitting tool (if necessary) and the xpt conversion tool. The created xpt files are then also validated through the Pinnacle 21 Community version.

CONCLUSION

The significance of a generic DTS template for requesting non-CRF data in an SDTM format from laboratories lies in its ability to standardize, streamline, and enhance the quality of data integration in clinical studies. By fostering consistency and compliance, this approach ultimately contributes to more efficient and effective clinical research outcomes.

ACKNOWLEDGMENTS

We would like to thank our colleague Francis de la Poza Arredondo, who is responsible for the creation of the xpt converter and the domain splitter tool.

CONTACT INFORMATION

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

Author Name: Samaneh Sabiani, Kristina Zweier
Company: Merck KGaA
Email: Samaneh.sabiani@merckgroup.com, kristina.zweier@merckgroup.com