Paper AD13

# MINT+ Automation and Digitization of SDTM specification

Magdalena Krochmal, Roche, Basel, Switzerland
Adam Foryś, Roche, Basel, Switzerland
Rammprasad Ganapathy, Roche, San Francisco, USA

## ABSTRACT

MINT+ is a web application that automates the creation of study SDTM specifications, including SDTM mappings and Controlled Terminology. A set of generic algorithms to map collected data to SDTM lies at the center of the solution. By comparing study metadata to the metadata in the Roche MDR, MINT+ streamlines the process and reduces the likelihood of errors. Built with React UI, MINT+ enables users to add SDTM mappings for non-standards through an interactive user interface. Study-level mappings are then stored in Saffron Document DB, facilitating metadata sharing and enabling further reuse across other studies using the same data collection elements. This innovative solution significantly reduces the workload and increases the accuracy of SDTM mapping, addressing a significant challenge in the industry.

## INTRODUCTION

The Standard Data Tabulation Model (SDTM) is a standard data model that defines the structure and content of clinical trial data. Established by the Clinical Data Interchange Standards Consortium (CDISC), SDTM provides a standard for organizing and formatting data to streamline data collection, analysis, and reporting processes. SDTM is one of the requirements for clinical data submission to major regulatory agencies like the Food and Drug Administration (USA) and the Pharmaceuticals and Medical Devices Agency (Japan). SDTM datasets summarize information about study subjects in the form of collections called domains. Each domain groups data by topic-specific commonality, e.g., demographics, adverse events, findings, and other relevant clinical trial data. SDTM was declared the standard format for submitting data related to clinical studies in 2004; however, preparing the SDTM datasets remains one of the most time-consuming tasks in the clinical submission workflow.

## PROBLEM STATEMENT

Automating SDTM processes can provide significant advantages to the pharmaceutical and clinical research industry, but it also presents several challenges that must be addressed. One of the challenges is ensuring data quality and consistency, which requires meticulous attention to detail to maintain compliance with SDTM standards. Additionally, complex data mapping and transformations are required to convert raw clinical and nonclinical data into SDTM-compliant datasets. Compliance with evolving regulatory standards is also crucial; automation systems must keep pace with these changes. Integrating disparate systems and data sources into a unified automation workflow can be challenging. Ensuring data accuracy and quality is essential, and automation solutions must incorporate robust validation and quality control checks. Implementing and maintaining SDTM automation systems requires significant resources and specialized skills. Converting existing legacy data into SDTM format can be time-consuming and complex, and managing metadata and updates as standards evolve is critical. Implementing changes while maintaining data integrity can also be challenging. Finally, organizations must weigh the upfront costs of implementing SDTM automation against the long-term benefits. Addressing these challenges requires careful planning, investment, and ongoing monitoring to ensure data complies with evolving standards and regulatory requirements.

### BENEFITS OF AUTOMATED SDTM GENERATION

Automated SDTM generation provides several benefits to clinical data management, helping to streamline the process of transforming raw clinical trial data into the standardized SDTM format. One of the significant benefits of automated SDTM generation is the increased efficiency it brings to clinical trials. By reducing the time required to map raw clinical trial data into SDTM datasets, valuable time is saved, and the overall clinical trial process can be made more efficient. Additionally, automated data mapping ensures that the generated SDTM datasets are more accurate, eliminating the potential for human error in the data mapping process. This, in turn, ensures data consistency and integrity. Automated SDTM generation also helps to ensure compliance with regulatory standards. Since SDTM datasets are used to support regulatory submissions, it is essential to meet specific regulatory requirements. Automated SDTM generation can ensure that generated datasets are in compliance with regulatory standards, saving time and reducing the risk of errors. Finally, automation can result in cost savings by reducing the need for manual data management, increasing efficiency, and reducing the risk of errors. Automated SDTM generation provides several benefits for clinical trial data management and analysis, making the process more efficient, accurate, compliant, and cost-effective.

## METADATA-DRIVEN SDTM AUTOMATION

Metadata-driven SDTM automation is a powerful approach that relies on data definitions to automate the creation of standardized datasets for clinical and nonclinical research. It offers efficiency, consistency, flexibility, and compliance benefits while reducing errors and promoting data traceability. As the pharmaceutical and clinical research industry evolves, metadata-driven automation will likely play a crucial role in enhancing data management processes. We have successfully developed an R-based solution to automate the creation of SDTM submission deliverables. Closely linked to Data Standards, our metadata-driven automation solution can automate ~80% of the SDTM domains based on the concept of Reusable Algorithms.

Those reusable mapping Algorithms are the backbone of SDTM automation, where SDTM mappings are defined as a set of rules that transform the collected (eCRF, nonCRF) source data into the target Tabulation Data Model and can be re-used across different Domains (see Fig.1). Algorithms can be pre-specified for Data Standards. Users can reuse the algorithms for extension to standards or new data types. Algorithms are programming language-agnostic; the concept does not rely on a specific programming language for implementation.

| Mapping Algorithm | Description | Example mapping application |
|---|---|---|
| **01_ASSIGN_NO_CT** | One-to-one mapping with no controlled terms. | MH.MHTERM<br>VS.VSORRES<br>AE.AETERM |
| **02_ASSIGN_CT** | One-to-one mapping with controlled terms. | AE.AEOUT |
| **05_HARDCODE_CT** | Hard code the target based on the source with controlled terms. | FA.FATEST (*FA.FATEST = 'Verbatim Term'*) |
| **06_HARDCODE_NO_CT** | Hard code the target based on the source without controlled terminology. | CM.CMINDC (*CM.CMINDC = 'HYPOXIA'*) |
| **09_IF_THEN_ELSE** | Conditionally check a condition and apply a mapping | AE.AERELNST (*If checked then AE.AERELNST = 'CONCURRENT ILLNESS'*) |

Fig.1. Example of reusable algorithms and application in SDTM mappings.

### MINT+

MINT+ is a web application that enables Statistical programmers or Data Scientists to create the Study SDTMv specification. MINT+ serves as the user interface having REST API (`rsaffron.api`) in the backend and Saffron document store for data storage. High-level data flow is presented in Fig.2.

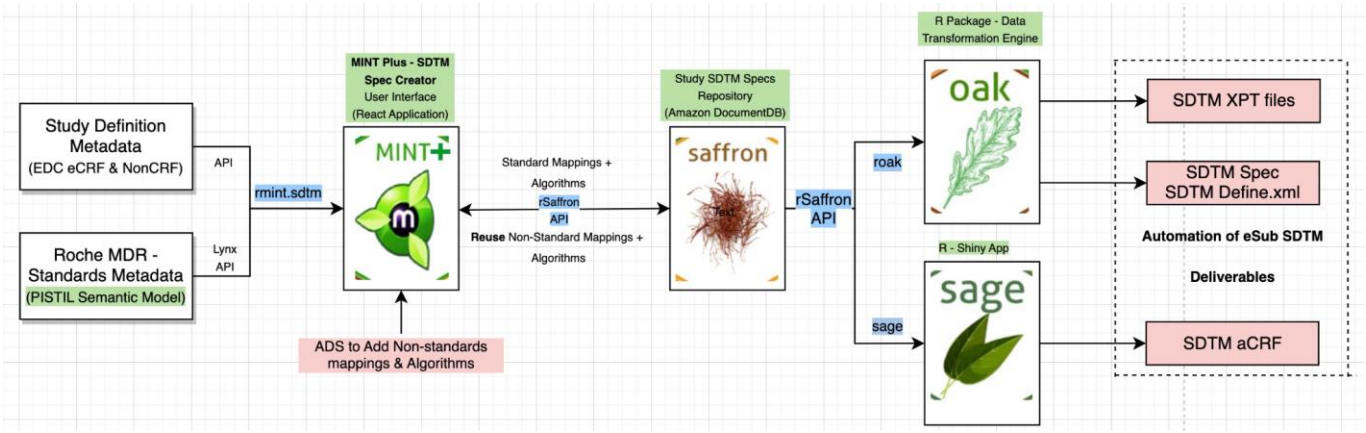


Fig.2. MINT+ data flow.

Our custom solution brings several important functionalities to Data Scientists when working on study SDTM annotations (Fig. 3). One of the key functionalities is the ability to create and access study SDTMv specifications online. This feature allows users to develop and maintain the SDTMv specification for their study in a digital format, making it easier to manage and share. To prevent users from overriding each other's work, MINT+ also has a locking capability to ensure that only one user can make changes to a specific specification at any time. This feature helps maintain data consistency and ensures all changes are tracked and documented. Another essential functionality of MINT+ is the annotation status management workflow, which helps manage specification development processes. This feature allows users to track the progress of specification development and ensures that all changes are thoroughly reviewed and approved before being used in production. MINT+ also offers an interactive user interface that allows users to add non-standard mappings and algorithms for non-standards. This feature supports users with the creation of an appropriate mapping for a given source dataset and variable, guiding the flow and prompting them to fill out information essential for successful mapping execution.

It is important to highlight that having the annotations stored in the underlying database, MINT+ offers a suggestions feature that enables users to reuse past specifications and reduces the time and effort required to create new specifications from scratch. As such, MINT+ promotes and supports molecule-level consistency of mappings. In the long term, this will enable users to identify and resolve inconsistencies in non-standard mappings, improving the accuracy and consistency of the generated SDTM datasets across studies.

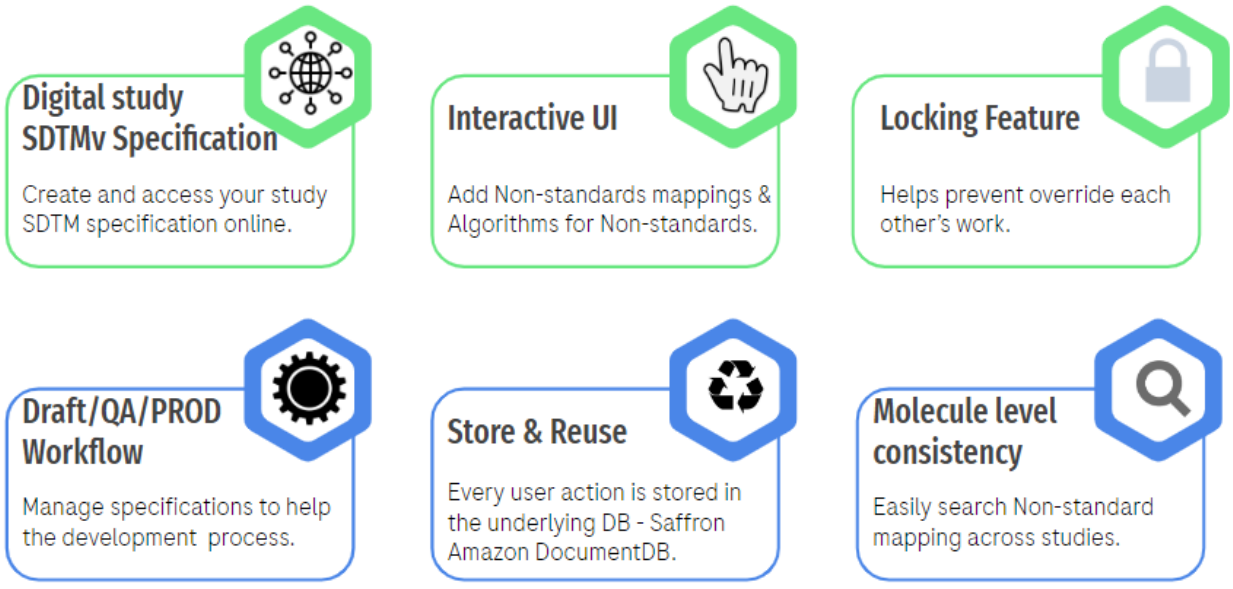


Fig. 3. MINT+ functionalities.

**FRONTEND**

MINT+ frontend is developed with the ReactJS framework. General user path flow when navigating in MINT+ Application includes initial analysis configuration followed by creating custom, non-standard mappings and relevant updates on the derived SDTM variables.

In the analysis configuration step, study metadata is processed to annotate Standards with pre-specified Algorithms. Analysis inputs are sourced from internal systems via APIs, including Study Definition Metadata (eCRF and NonCRF) and Standards Metadata, where Data Standards with pre-specified Algorithms live. Multiple analyses can exist for a study depending on study-specific needs (e.g. branch-out analysis). Each analysis has a development status (DRAFT/QA/PROD) depending on the maturity of the study annotations.

Created study mappings are presented in MINT+ content tabs, including:

- SDTM Spec View - target to source view of the study mappings with the possibility to edit or inactivate some of the derived variables.
- Non Standards - annotated source datasets view where users can modify or add new custom mappings. The built-in search feature enables users to copy non-standard annotations from other studies (e.g. useful in case of repeated forms) or other studies.
- Discrepancies - listing all discrepancies in the input study metadata compared to Data Standards such as label, data format, and log discrepancies.
- Data Tabulation - source to target view of the existing annotations.

In addition, the F.A.I.R. Dashboard tab supports the management of the creation of study annotations, presenting users with statistics regarding the percentage of standard vs. non-standards in the study and progress in annotating study-specific variables.

**BACKEND**

The backend solution of the MINT+ App consists of REST API (`rsaffron.api`) and a document store, Amazon DocumentDB Saffron. REST API, developed with the R plumber package, is a layer responsible for handling any action triggered on the front end, such as fetching from and writing data to the Saffron database. The backend layer of MINT+ enables collaborative working on creating study SDTMv specifications, allowing users to set a lock on the study data during the editing process. This mechanism prevents other users from overriding their work. `rsaffron.api` is also responsible for the assembly of final views to be used by downstream applications like the `roak` data transformation engine, where mappings are being used for final SDTM(v) creation.

The Saffron database stores the Study SDTMv specification created by Data Scientists in a machine-readable JSON-LD format. The Saffron database structure consists of five collections (Fig. 4.), with *study_metadata* being a central collection designed to keep track of the current status of each study processed in MINT+. This collection maintains information such as study name, analysis status, input data sources, and document lock information. The *study_metadata_audit* collection keeps track of old entries of study_metadata, capturing changes to the metadata over time and ensuring that data integrity is maintained. The *study_gdsr_versions* collection maintains information about current and past setups of GDSR (Global Data Standards Repository) for any given study. This collection keeps track of the GDSR version, date, and any relevant modifications. The *mint_views* collection keeps the actual mapping analysis outputs. This collection provides a record of the output generated by the data analysis tool and allows for future analysis or review of the output data.

The *custom_anno* collection keeps information about user-provided annotations for the study, allowing for additional context to be added to the study metadata. Finally, the *custom_anno_audit* collection keeps full information about user-provided annotations for the study, including old and outdated annotations. This collection ensures that a comprehensive record of all user-provided annotations is maintained and that changes to annotations over time are well-documented.

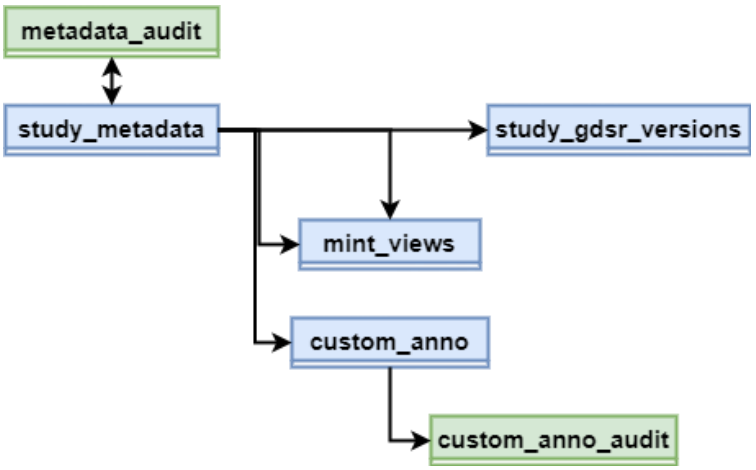


Fig.4. Simplified Saffron documentDB schema.

Overall, the database structure for these collections is designed to ensure that data integrity is maintained and all relevant information about the study is captured accurately. These collections provide a comprehensive record of the study metadata and associated data, facilitating efficient study management and analysis. Every user action in MINT+ is effectively stored in the Saffron database. When the user adds or changes SDTMv mappings or Algorithms of non-standards, those modifications are automatically saved in Saffron for future reference.

Therefore, the Saffron database serves as:

- study-level annotation storage
- source of non-standards for further reuse or promotion to standards
- source of audit trail information
- source of training data for machine learning algorithms aiming at SDTM annotation prediction.

#### END-TO-END WORKFLOW

MINT+ provides automated creation of Study SDTMv Specifications, Controlled Terminology, Value Level Metadata, and SDTM Define.xml using created mappings metadata. Those mappings are a source for downstream systems in the SDTM automation pipeline, providing study-specific information regarding the algorithms necessary for the final creation of SDTM submission deliverables.

## CONCLUSION AND OUTLOOK

SDTM automation is becoming increasingly popular in the pharmaceutical industry as it helps to streamline the process of creating SDTM datasets. This can help reduce costs, save time, and improve compliance with regulatory requirements. Our automation tool, MINT+, offers users a seamless experience connecting to source CRF and non-CRF data and Global Data Standards. Relying on the concept of reusable algorithms, study mappings can be created and executed with downstream tools to create SDTM submission deliverables. Collected mappings metadata is easily transformed into Study SDTM(v) Specification and SDTM Define.xml.

In conclusion, MINT+ offers several functionalities that can streamline the clinical trial data management process, improve data quality, and ensure compliance with regulatory requirements. These functionalities include creating and accessing study SDTMv specifications online, managing specification development processes, storing and reusing past specifications, preventing users from overriding each other's work, adding non-standard mappings and algorithms, and enabling molecule-level consistency. The level of automation is always proportional to the adherence to Global Data Standards. The manual programming effort might vary depending on the amount of non-standard datasets and variables at CRF / non-CRF source. However, the possibility of automation should encourage global standards adoption or reuse of existing mappings on a molecule level.

## ACKNOWLEDGMENTS

The work presented in this paper is the result of the collaborative efforts of the authors together with the entire Oak Garden Team.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the authors at:

magdalena.krochmal@roche.com, adam.forys@roche.com, ganapathy.rammprasad@gene.com