Paper DS13

Innovative Storage of SDTM Mappings as JSON: Boosting Efficiency and Data Consistency

Rohan Parmar, Roche, Mississauga, Canada
Edgar Manukyan, Roche, Mississauga, Canada
Rammprasad Ganapathy, Genentech, San Francisco, USA

ABSTRACT

The creation of Study Data Tabulation Model (SDTM) datasets can be a time-consuming process hindered by individual interpretations of data mappings. This variability results in non-uniform SDTM datasets across studies, complicating organizational efforts for downstream pooling both across and within molecules. Traditionally, SDTM mappings are scattered across spreadsheets and PDF documents, which are not machine-readable, making it challenging for study teams to search and reuse them efficiently.

At Roche, we have developed an innovative technique to store SDTM specifications as JavaScript Object Notation (JSON) files in a document store. This approach enables users to easily search and reuse SDTM mappings across studies, significantly enhancing efficiency and consistency.

In this paper, we will discuss the benefits of having SDTM specifications as JSON files in a machine readable format, including improved searchability and reusability. We will also delve into the nuances of implementing this technique, highlighting the practical advantages and potential challenges.

INTRODUCTION

The SDTM is a standardized framework specifying clinical trial structure and content. Developed by the Clinical Data Interchange Standards Consortium (CDISC), SDTM ensures consistency in organizing and formatting data, simplifying the processes of data collection, analysis, and reporting. It is a mandatory format for clinical data submissions to key regulatory bodies, such as the Food and Drug Administration (FDA) in the United States and the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan. SDTM datasets capture information about study participants within thematic collections known as domains. These domains categorize data based on shared topics, such as demographics, adverse events, findings, and other critical clinical trial details. Although SDTM was adopted as the standard for clinical study data submissions in 2004, creating these datasets remains one of the most labor-intensive tasks in the clinical submission process.

JavaScript Object Notation (JSON) is a lightweight, text-based data interchange format widely recognized for its simplicity and flexibility. Designed to be both human-readable and machine-readable, JSON is commonly used to store and exchange structured data. Its key-value pair architecture makes it ideal for representing complex, hierarchical datasets in a concise and intuitive manner. In the context of SDTM mappings, JSON provides a modern, standardized way to store specifications, allowing for easier integration with automated tools and enhancing reusability. By leveraging JSON to organize SDTM mappings, it becomes possible to address inefficiencies in the traditional approaches, which rely on static formats like spreadsheets and PDFs.

This paper explores the innovative use of JSON for storing SDTM specifications, demonstrating how it can improve efficiency, consistency, and accessibility across studies while addressing practical implementation challenges.

BACKGROUND AND CURRENT CHALLENGES

The process of creating SDTM datasets is both labor-intensive and prone to inconsistencies. Despite the standardized framework established by CDISC, many organizations struggle to achieve uniformity in their SDTM datasets across studies. This lack of uniformity arises from varying interpretations of mapping rules and specifications, often influenced by individual preferences and study-specific nuances.

THE TRADITIONAL APPROACH TO SDTM MAPPINGS

Historically, SDTM mappings have been documented in formats like spreadsheets and PDF documents. These static formats, while familiar, present several limitations. They are not machine-readable, making it difficult to automate processes or integrate them with advanced analytical tools. Moreover, their rigid structure complicates the task of searching, updating, and reusing mappings across multiple studies. This reliance on static documentation fosters inefficiencies, as teams must repeatedly recreate or modify mappings for each study.

LIMITATIONS OF STATIC FORMATS (SPREADSHEETS AND PDFS)

Static formats like spreadsheets and PDFs are ill-suited for the dynamic and collaborative nature of modern data management. Searching for specific mappings or related details often involves manual effort, slowing workflows. Furthermore, version control becomes a significant challenge, as changes to mappings are not automatically tracked, leading to discrepancies and errors. Collaboration across teams is also hampered. Static files are typically shared through email or other non-integrated platforms, increasing the risk of miscommunication and redundant efforts.

STRUGGLE WITH POOLING

Storing SDTM mappings in disparate spreadsheets and PDF files invariably leads to redundancy. This redundancy introduces another level of variability in the mappings causing difficulties with pooling. The inherent variability in interpretation of mappings by SDTM programmers causes new mappings to be created within studies. Mappings that may have already been created are created again as there is manual work involved in searching and copying mappings across spreadsheets and PDF files so users may not copy something that is preexisting. When it comes to pooling, these 'novel' mappings cause additional programming effort to reconcile mappings that are actually the same.

THE CASE FOR INNOVATION

In light of these challenges, there is a compelling need for innovation in the way SDTM mappings are stored and managed. Transitioning to a machine-readable, flexible format such as JSON represents a significant step forward. JSON offers a structured yet lightweight solution, enabling organizations to enhance the accessibility, consistency, and reusability of their SDTM mappings. By addressing the limitations of traditional approaches, JSON paves the way for more efficient and scalable data management practices.

OVERVIEW OF JSON AS A STORAGE SOLUTION

The adoption of JSON as a storage solution for SDTM mappings offers numerous advantages. JSON is a lightweight, text-based format designed for data interchange. Its key-value pair structure, combined with support for hierarchical organization, makes it particularly well-suited for representing complex datasets like SDTM mappings.

INTRODUCTION TO JSON: KEY FEATURES

JSON's design emphasizes simplicity and flexibility. Its syntax is easy to read and write for humans while remaining machine-readable. This dual compatibility allows for seamless integration with a variety of programming languages and tools. JSON also supports nested structures, enabling users to represent complex relationships between data elements in a concise and organized manner.

ADVANTAGES OF JSON FOR SDTM MAPPINGS

Using JSON for SDTM mappings addresses many of the limitations inherent in static formats. JSON's machine-readable nature facilitates automation, allowing organizations to streamline workflows and reduce manual effort. Additionally, its hierarchical structure supports the organization of mappings in a logical and intuitive way, improving accessibility and reusability. JSON files can be easily queried and updated, further enhancing their utility in dynamic data environments.

COMPARISON WITH ALTERNATIVE STORAGE FORMATS

Compared to traditional formats like spreadsheets, PDFs, or XML, JSON offers a balance of simplicity and functionality. While XML is also a structured format, it is often verbose and complex, making it less practical for certain applications. Spreadsheets, on the other hand, lack the flexibility and machine-readability of JSON, making them unsuitable for automation and dynamic querying. PDFs, though widely used for documentation, are particularly limited in the context of SDTM mappings. PDFs are static and primarily designed for presentation rather than data storage. They lack inherent structure for querying or machine processing, which makes extracting and reusing information from PDFs a tedious and error-prone task. Additionally, PDFs do not support hierarchical data representation, which is critical for organizing complex mapping specifications. By contrast, JSON provides a dynamic, machine-readable format that supports both scalability and ease of access, making it a superior choice for managing SDTM mappings.

IMPLEMENTATION OF JSON FOR SDTM MAPPINGS

At Roche the backend implementation of the SDTM specifications provides a robust and efficient framework for managing SDTM mappings using JSON. At the core of this solution lies a REST API {rsaffron.api} and Amazon DocumentDB Saffron as the document store.

BACKEND ARCHITECTURE

The REST API, developed using the R plumber package, serves as the intermediary between the front end and the database. It facilitates actions such as fetching and writing data to the Saffron database. This API layer enables collaborative work by allowing users to edit study SDTMv specifications while setting a lock on the study data to prevent concurrent edits. The API also assembles final views required for downstream applications, such as the {roak} data transformation engine, which utilizes the mappings for final SDTM(v) creation.

SAFFRON DOCUMENTDB STRUCTURE

The Saffron database stores study SDTMv specifications in a machine-readable JSON-LD format and is structured into five key collections:

- study_metadata: Tracks the current status of each study, including study name, analysis status, input data sources, and document lock information.

- study_metadata_audit: Maintains a history of changes to the study metadata, ensuring data integrity over time.

- study_gdsr_versions: Stores information about the Global Data Standards Repository (GDSR) setups, including versions and modifications.

- mint_views: Contains the output of mapping analyses, supporting future review and analysis.

- custom_anno: Captures user-provided annotations, allowing additional context to be associated with study metadata.

- custom_anno_audit: Maintains a complete history of all user-provided annotations, ensuring transparency and traceability.

Every user action, including adding or modifying SDTMv mappings, is automatically stored in Saffron. This structure ensures that study-level annotations, non-standard algorithms, and audit trail information are preserved for future use and can even serve as training data for machine learning models aimed at predicting SDTM annotations.

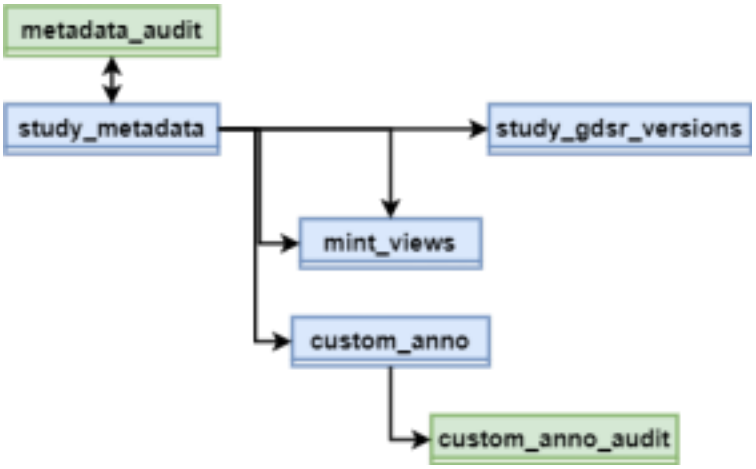

Fig.1. Simplified Saffron documentDB schema.

BENEFITS OF JSON-BASED SDTM MAPPINGS

Adopting JSON as the foundational format for SDTM mappings offers a range of tangible benefits for both efficiency and data integrity.

ENHANCED SEARCHABILITY AND REUSABILITY

JSON's hierarchical and machine-readable structure allows users to quickly query and locate specific mappings across multiple studies. This streamlined search capability reduces manual effort and promotes the reuse of mappings, enabling consistent application of standards across studies.

IMPROVED CONSISTENCY ACROSS STUDIES

By centralizing mappings in a document database and using a standardized JSON-LD format, the risk of inconsistencies due to human error or miscommunication is significantly reduced. The audit trail functionality further reinforces consistency by maintaining a comprehensive history of changes.

SUPPORT FOR AUTOMATION AND SCALABILITY

JSON's compatibility with modern programming languages and tools facilitates the automation of tasks such as mapping generation, validation, and downstream transformations. This automation not only improves efficiency but also ensures scalability as the volume of studies and complexity of data increase.

REUSABLE MAPPINGS AND JSON SCHEMA

SDTM mappings are structured as algorithms that transform collected source data, including CRF and non-CRF data, into the target tabulation data model. These mapping algorithms define the syntax of SDTM mappings and ensure consistency and reusability across studies. With 22 unique mapping algorithms designed to cover approximately 80% of therapeutic area standards and non-CRF data models, JSON enables a structured and efficient approach. Algorithms can be reused across different domains, pre-specified for standards, and extended by users to accommodate non-standard data types or new study requirements. This approach supports both standard and non-standard mappings, ensuring flexibility while maintaining data integrity.

```
{
        raw_dataset: 'ACTRAGM',
        raw_dataset_label: 'Actual Cohort and Treatment Arm',
        raw_variable: 'COHRT1',
        raw_variable_label: 'Previous Cohort',
        raw_variable_type: 'CheckBox',
        annotation_text: 'SC.SCORRES when SC.SCTESTCD = \'PCOHORT\'',
        annotation_ordinal: 1,
        target_sdtm_variable_domain: 'SC',
        target_sdtm_variable: 'SCORRES',
        target_sdtm_variable_role: 'Result Qualifier',
        mapping_algorithm: '01_ASSIGN_NO_CT',
        @id: 'e4acbfe6-a3c3-4bd2-a857-024990c58faa'
}
```

Fig.2. An Abbreviated JSON Mapping Example.

FLEXIBILITY OF JSON SCHEMA

Compared to traditional SQL data, No-SQL JSON schemas support a high level of flexibility. The JSON structure allows for additional properties to be added seamlessly without affecting a large pre-defined table structure. This allows the JSON based SDTM mappings to support future complexity if required.

CHALLENGES AND LESSONS LEARNED

Validation requirements can introduce some programming overhead when using JSON and no-SQL databases. The strict typing and structure provided by tabular data formats ensures data consistency which is a big part of validation requirements and no-SQL databases require some additional work to accomplish the same.

HANDLING DATA COMPLEXITY

The biggest impact of data complexity with the flexibility afforded by JSON is that applications need to account for the additional properties that may have been introduced for some mappings. SQL-like databases have strict conformance but this is not the case with no-SQL databases. Therefore it is important to account for a fluid data schema when building applications. This is felt particularly when searching for mappings. Having deeply nested JSON mappings makes querying for them difficult.

VALIDATING JSON SCHEMAS

Due to the fluid nature of JSON schemas, validation steps can become complex. In order to ensure that all mappings will work when creating SDTMs there is frequent testing involved. Every part of the application has to be tested such that users cannot introduce errors within mappings that will cause mappings to fail.

CONCLUSION

The adoption of JSON as the primary format for storing and managing SDTM mappings represents a significant advancement in clinical data management. By replacing static, manual processes with a structured, machine-readable solution, organizations can achieve greater efficiency, consistency, and scalability. The integration of JSON with modern data architectures enables streamlined automation, enhances collaboration, and lays the groundwork for future advancements in machine learning and data-driven decision-making. As clinical data continues to grow in complexity, embracing JSON ensures that SDTM mappings remain adaptable, reusable, and aligned with evolving industry standards.

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:

Author Name: Rohan Parmar
Company: Roche
Address: 7070 Mississauga Rd, Mississauga, ON L5N 5M8, Canada
Work Phone: +1 800-561-1759
Email: rohan.parmar@roche.com
Website: https://www.rochecanada.com

Author Name: Edgar Manukyan
Company: Roche
Address: 7070 Mississauga Rd, Mississauga, ON L5N 5M8, Canada
Work Phone: +1 800-561-1759
Email: edgar.manukyan@roche.com
Website: https://www.rochecanada.com

Author Name: Rammprasad Ganapathy
Company: Genentech
Address: 7070 Mississauga Rd, Mississauga, ON L5N 5M8, Canada
Work Phone: +1 650-225-1000
Email: ganapathy.rammprasad@gene.com
Website: https://www.gene.com