Dataset-JSON as an Alternative Transport Format for Regulatory Submissions: Final Pilot Report

phuse.global

Contents

Revision History

Version	Date	Summary
1.0	19 June 2024	Finalised report

1. Introduction

1.1 Pilot Background

The pilot project Dataset-JSON as an Alternative Transport Format for Regulatory Submissions began in late July 2023 and concluded at the PHUSE Computational Science Symposium (CSS) in June 2024. The pilot incorporated industry and FDA testing. This pilot is a collaboration between PHUSE, CDISC and FDA. The pilot leads were Stuart Malcolm (PHUSE), Sam Hume (CDISC) and Jesse Anderson (FDA). The project includes four sub-teams: Pilot Testing and Report, Technical Implementation, Business Case, and Strategy for Future Development. The pilot gathered community input and feedback through the PHUSE CSS and the Connect conference workshops held throughout the project.

This pilot report summarises the overall pilot results, the technical findings, and the next steps to address the technical findings. It covers the deliverables from the Pilot Testing and Report and the Technical Implementation sub-teams. The Business Case and the Strategy for Future Development sub-team deliverables will be published (summer 2024) as separate documents as *PHUSE deliverables.*

1.2 Dataset-JSON Background

CDISC Dataset-JSON is a modern dataset format designed to address a broad range of data exchange scenarios, including the regulatory requirements for submission datasets. Its primary purpose is data exchange, and it should not be viewed as a replacement for dataset formats used for analytical purposes. It is inspired by the CDISC Dataset-XML v1.0 specification with important enhancements, including much smaller file sizes, additional metadata and simpler processing. Dataset-JSON supports file and API-based data exchange. JSON-based formats are simple to implement, stable, and widely supported by nearly every technology stack and programming language. Dataset-JSON optionally links to a Define-XML file for more complete metadata. The technical components of Dataset-JSON are published under the Massachusetts Institute of Technology (MIT) open-source licence.

Overcoming the limitations imposed by the SAS Version 5 XPORT Transport Format (XPT) will initially benefit those engaged in dataset exchange since Dataset-JSON is easy to implement in nearly every programming language and many software applications already import and export JSON. In the longer-term, more significant benefits will emerge as the constraints imposed by XPT on the CDISC Foundational Standards and data exchange technologies are removed.

2. Scope

The scope for the Dataset-JSON as an Alternative Transport Format for Regulatory Submissions pilot covers two primary objectives:

1. Demonstrate that Dataset-JSON can transport information with no disruption to business.
2. Demonstrate the viability of Dataset-JSON as the primary transport option.

The pilot consists of four sub-teams:

1. Pilot Testing and Report: consolidate feedback from the pilot submissions and develop report.
2. Technical Implementation: consolidate technical findings identified during the testing and agree on solutions to be implemented at the pilot's conclusion.
3. Business Case: establish a business case for using Dataset-JSON for submissions to justify the investment needed to implement the changes.
4. Strategy for Future Development: define a roadmap for future changes that leverage the benefits of Dataset-JSON once the XPT restrictions have been lifted.

3. Definitions

Abbreviation	Description
ADaM	Analysis Data Model
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
CDER	Center for Drug Evaluation and Research (FDA)
CFR	Code of Federal Regulations
COSA	CDISC Open Source Alliance
FDA	U.S. Food and Drug Administration
IT	Information Technology
JSON	JavaScript Object Notation
MIT	Massachusetts Institute of Technology
OCS	Office of Computational Science CDER FDA
ODM	Operational Data Model
OID	Object Identifier
SDTM	Study Data Tabulation Model
SEND	Standard for Exchange of Nonclinical Data
UG	User Guide
XML	Extensible Markup Language
XPT	XPT is the file extension and shorthand name for the SAS Version 5 (V5) transport file format

4. Pilot Report

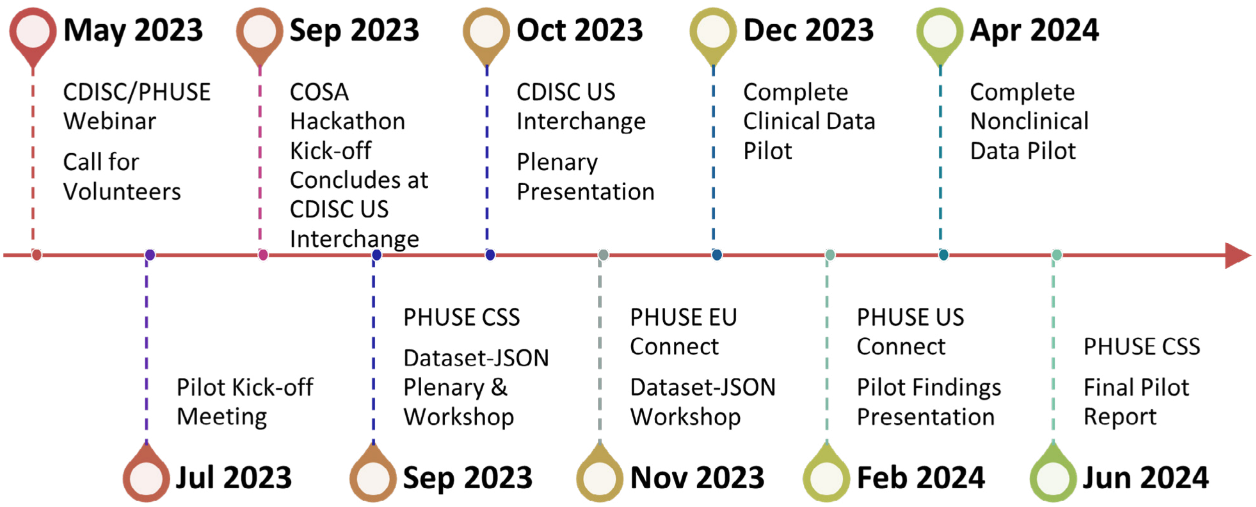
4.1 Pilot Objectives and Goals

The objective of this pilot was to test the feasibility of using Dataset-JSON as a transport format for data from clinical and nonclinical studies submitted with regulatory applications. An additional objective was to identify differences in content when using Dataset-JSON as compared to the traditional XPT v5 transport file. A final objective was to identify any business impacts on sponsors and regulators in creating, managing, submitting and receiving data from clinical and nonclinical studies in the Dataset-JSON format. The goal of this pilot was to demonstrate that Dataset-JSON can serve as a transport file for data from clinical and nonclinical studies with no loss of data and no significant impact on business operations.

4.2 Timeline

The figure below shows the timeline of activities across the PHUSE project. The clinical and nonclinical pilots were completed in December 2023 and April 2024, respectively.

Additional documents detail the findings of each PHUSE project sub-team. The Business Case and the Strategy for Future Development sub-team deliverables will be published (summer 2024) as separate documents as *PHUSE deliverables.*



4.3 Pilot Strategy

There were three testing phases: (a) initial testing (b) test submissions to the Center for Drug Evaluation and Research (CDER) at FDA and (c) evaluation of findings. Each phase aimed to demonstrate that Dataset-JSON can transport the same information with no disruptions. Further details on the strategy for each phase are detailed below.

4.3.1 Phase I: Initial Testing

Initial testing assessed the feasibility of converting sample data from XPT to Dataset-JSON using conversion tools developed during a *CDISC Hackathon*. This enabled testing of internal capabilities by sponsors and regulators, lowering the perceived risk and barrier to entry into pilot participation. Participants chose conversion scenarios (e.g. SAS dataset to Dataset-JSON, R dataframe to Dataset-JSON, SAS XPT to Dataset-JSON) and then converted datasets from their native format into Dataset-JSON using available *tools*. Subsequently, the Dataset-JSON files were converted back into their native dataset format and compared to the original native datasets. This ensured there were no changes in the converted dataset packages. Organisations could perform more than one conversion scenario or combine conversion scenarios. After completing their testing, organisations were requested to complete an online questionnaire. Questionnaires were completed by industry participants and not FDA testers or software product vendors who integrated Dataset-JSON.

The questionnaire included the following questions:

1. Demographic information (e.g. Company, Respondent's Name)
2. What conversion tool did your company use? (R/SAS/Both/Other)
3. Were you able to convert existing datasets to Dataset-JSON? (Yes/No/Not Sure)
4. Were you able to convert Dataset-JSON into SAS, R or another dataset format? (Yes/No/Not Sure)
5. Did the conversion work as expected? (Yes/No/Not Sure)
6. What types of datasets did you use for the conversion? (SDTM/ADaM/Other)
7. Did Dataset-JSON represent data as expected? (Yes/No/Not Sure)
8. Takeaways and other feedback you can provide (Free Text)
9. What level impact is needed to change your existing workflow to use Dataset-JSON? (Free Text)

4.3.2 Phase II: Test Submissions to FDA/CDER

The next step was to test receipt of data from clinical and nonclinical studies from across sectors of industry. FDA/CDER engaged with industry volunteers for five clinical studies and three nonclinical studies to submit three versions of dataset packages: the original XPT v5 dataset packages, the converted JSON dataset packages, and the dataset packages converted back into their XPT v5 format. The goals were to confirm receipt of these dataset packages and that data integrity had been maintained across all three versions. This would demonstrate that there had been no disruption to business.

Additionally, FDA/CDER performed independent conversions of the XPT files to confirm the data integrity had been maintained and that the submitted XPT and JSON files matched the received and converted datasets. The team reviewed the CDISC Dataset-JSON specification and used the Python hackathon solution. The team then performed automated and manual validation checks using DataFit and SAS and noted differences between the submitted and the converted datasets. Subsequently, loading of Dataset-JSON datasets into data management and analytical tools was tested.

4.4 Phase III: Evaluate Findings

The third phase of the pilot was to evaluate the findings from the previous two phases. Overall, the results of each phase contained similar feedback from both industry and FDA. Specific findings from each of the previous phases are provided below.

4.5 Results from Phase I

A request for participation in Phase I testing of the *PHUSE project* was announced in June 2023. In February 2024, the sub-team reviewed questionnaire responses from 19 respondents. It is possible that participants conducted additional testing without answering a questionnaire. The table below shows the distribution of conversion tools used by the survey respondents.

Topic	Yes (n, %)	No (n, %)	Not Sure (n, %)
Conversion Tools Used			
(Q2) Used SAS as conversion tool*	17 (89%)	2 (11%)	0
(Q2) Used R as conversion tool*	4 (21%)	15 (79%)	0
(Q2) Used Python as conversion tool*	1 (5%)	18 (95%)	0
(Q5) Did the conversion work as expected?	12 (63%)	4 (21%)	3 (16%)
Overall Findings			
(Q1) Were you able to successfully convert datasets to Dataset-JSON?	18 (95%)	1 (5%)	0
(Q4) Were you able to convert Dataset-JSON into SAS, R or another dataset format?	18 (95%)	1 (5%)	0
(Q7) Did Dataset-JSON represent the data as expected?	10 (52%)	1 (5%)	8 (43%)

*Three organisations used both SAS and R.

Responses for dataset conversion in Phase I were limited to clinical study data (e.g. SDTM only, ADaM only, or both SDTM and ADaM). The data show that most of the respondents used SAS as a conversion tool (n=17, 89%). Also, the majority successfully converted datasets (n=18, 95%) into JSON and back into XPT. However, almost half the respondents (n=9, 47%) were uncertain if Dataset-JSON accurately represented the data. Respondents identified several reasons for this uncertainty, including rounding differences that resulted in a potential loss of precision, metadata differences, and the lack of a tool to check the dataset for accuracy.

Respondents raised the following questions and concerns.

- Respondents raised concerns about data accuracy (decimal places, rounding data formats, and variable lengths) that appear to be based on the conversion tool used. The technical implementation team are aware of these differences and will address them in the Technical Implementation portion of this report.
- Participants suggested including additional messaging to confirm that existing SAS and R conversion tools have been validated.
- Participants pointed out that using required and optional metadata within the JSON datasets remains unclear and should be clarified before implementation. These concerns were documented by the Technical Implementation team, with proposed solutions planned to be included in the next release of the Dataset-JSON specification expected later in 2024.
- Participants raised additional questions and concerns related to variable length, character length and non-ASCII characters. Specifically, Dataset-JSON dataset packages did not successfully convert back to XPT when non-ASCII characters were included in files. Participants noted similar issues when the variable and character lengths exceeded their limits.
- Participants pointed out the importance of 'DisplayFormat' for the system to correctly represent numeric and date data types in SAS and R, respectively, and this needs to be explicitly stated in the Dataset-JSON documentation.

The level of impact on internal regulatory processes was reported as low (n=6, 32%), medium (n=8, 42%) and high (n=5, 26%). For those responding with a low level of impact, participants commented that it is straightforward to use JSON as a data transport file. Participants needed additional time to review these dataset packages. Overall, it will be beneficial to clarify required metadata and provide comprehensive guidelines and well-documented tools, including a JSON viewer. Participants responding with a medium level of impact commented that it was difficult to determine impact since the current processes are in flux with limited IT capabilities to process these dataset packages. Additional documentation on GitHub and/or the specification would be helpful, and there would be a need to define a new internal review process of these data. The high-impact respondents provided similar feedback to the other two impact levels. Additional concerns are related to using a validator tool to confirm the datasets represent information appropriately, and the process for creating the Define-XML file in conjunction with the metadata that is included in the JSON datasets.

4.6 Phase II: Test Submissions to FDA/CDER

Upon receipt of submissions from industry, FDA/CDER performed initial data integrity testing and tool loading testing for both clinical and nonclinical submissions. As expected, applying pretty-print formatting rather than single-line formatting increased the size of the files. Single-line formatting for Dataset-JSON dataset packages is recommended for submission to FDA/CDER. There were no issues with parsing or ingesting JSON files into existing data management tools. Further, OCS confirmed that the datasets maintained their integrity across the submitted JSON files, XPT v5, and the converted JSON files by performing automated and manual checks on these dataset packages.

When testing dataset conversions with five clinical studies, FDA/CDER observed similar issues as reported by the pilot participants. For example, JSON headers listing "referenceData" instead of "clinicalData" caused conversion errors that were due to using the available hackathon conversion program. This therefore highlighted the need to validate the conversion tools. Additionally, like the industry feedback, it was found that the "displayFormat" field in either the Define-XML or the JSON file is crucial for ensuring proper date and numeric format. The Dataset-JSON specification should clarify expectations regarding headers and fields. FDA/CDER also tested the loading of Dataset-JSON files into analytical tools used for regulatory review. The clinical datasets failed to fully load into SAS, JMP and JMP Clinical. Adjustments to each tool would be needed to support the header and label information available in Dataset-JSON files. Using JMP, the team managed to get parts of dataset files to load (e.g. single rows, columns) but was unable to get full dataset files loaded with all labelling and data formatting information. While it may be possible to use SAS for data quality checks, significant adjustments to existing code would be needed to accommodate header and label information to run analytical script packages. Currently, it is not possible to use JMP Clinical since the system does not recognise .json as an acceptable format to load data into the tool.

FDA/CDER also tested Dataset-JSON on three nonclinical studies. One participant used a different Python library to read the XPT files as compared to both the hackathon solutions, thereby limiting further evaluation. As with the findings from the clinical studies, this suggests further clarity is needed in the Dataset-JSON specification to recommend a preferred validated method to create Dataset-JSON packages or convert existing dataset packages to JSON. Additionally, non-ASCII characters, “clinicalData” headers and capitalised column headers resulted in failed conversions from JSON to XPT due to the limitations of the XPT format. Development should consider incorporating inclusive headers for nonclinical study data in conversion scripts. FDA/CDER also tested the loading of Dataset-JSON files into analytical tools used for nonclinical regulatory review. These included SAS and SEND Explorer Warehouse. As with the clinical studies, these study datasets failed to fully load into SAS or SEND Explorer Warehouse. While the team confirmed it is possible to use SAS for data quality checks after significant modifications to the code, SAS was unable to load the header and label information. SEND Explorer Warehouse was unable to ingest or load data despite multiple attempts and adjustments to the dataset packages.

Across both clinical and nonclinical datasets received, data was successfully parsed and loaded into data management tools. However, there were no successful loads of Dataset-JSON files onto analytical tools used for regulatory review. Further, conversions from Dataset=JSON to XPT failed when features unique to JSON (e.g. metadata, capitalised column headers) were included. While it is possible to convert JSON to XPT files by removing the extra metadata, the added metadata and flexibility offered by the JSON format will be lost. This limits the use of a more modern transport technology with increased functionality for regulatory submission information exchange. This strongly supports the modernisation of tools and systems used for regulatory review to the more flexible and modern Dataset-JSON exchange format. This also suggests that for successful adoption, analytical tools will have to be updated to allow loading of the Dataset-JSON format.

4.7 Conclusion

Overall, the findings from the pilot support using Dataset-JSON as an alternative transport format for XPT v5. Most organisations successfully converted datasets from XPT to JSON using existing hackathon solutions. The findings were consistent across industry and FDA. FDA/CDER testing strongly suggests that analytical tools used for regulatory review are updated to take advantage of the metadata and flexibility provided by the Dataset-JSON format.

5. Technical Implementation

5.1 Background

The Technical Implementation section describes the key pilot findings and the proposed solutions to address the findings. These technical findings were originally documented by:

- The questionnaire the pilot participants completed after testing
- FDA testers
- Open-source conversion software developers
- Commercial software developers importing Dataset-JSON
- Commentary from testers delivered outside of the questionnaire.

The Dataset-JSON PHUSE Workshop attendees and CDISC Dataset-JSON standards development team proposed solutions to the pilot findings. These solutions will be implemented in Dataset-JSON v1.1, a version dedicated to addressing the pilot findings. The solutions to the pilot findings require three categories of change:

1. Update the standard specification, the schema, and the example datasets.
2. Create a User Guide (UG) to provide additional documentation.
3. Update and improve the open-source conversion tools.

The sub-sections that follow summarise the specific findings and their proposed solutions.

5.2 Dataset-JSON Structure and Naming

Software developers building tools to convert or import Dataset-JSON noted that the schema could be flattened to make it simpler to process. They also recommended that the standard not use the dataset OID as a JSON name (in the JSON name-value pair), but that all metadata names be fixed in the standard and all dataset metadata contain the values.

In response to this feedback, a flatter structure that is simpler to create and process has been proposed by the Dataset-JSON v1.1 standards development team. Several objects with hierarchical structures were flattened such that the entire standard is based on simple name-value pairs except for the column metadata and the data records, which are represented as lists.

Some names were simplified by replacing the ODM-based names with more general ones. The column metadata was named *items* and is now named *columns.* The data records were named *itemData* and are now named *rows.*

These changes make Dataset-JSON v1.1 flatter, easier to understand, and simpler to process than the v1.0 version.

5.3 Processing Large Datasets

While most pilot testers did not encounter performance issues, some did note slow performance when processing large datasets. In this case, a large dataset is one that is too large to fit into the memory of the computer processing the conversion. Each conversion tool handled large datasets differently.

Most programming languages have at least one JSON library that can stream large JSON files, but a few JSON libraries don't handle streaming efficiently, which impacts on their ability to process large datasets. The Dataset-JSON v1.1 standards development team will work with the open-source conversion tool authors to test the tools and capture performance metrics. Conversion tools may need to use a different JSON library that's proven to work well with large datasets.

In addition to testing the conversion tools, the Dataset-JSON v1.1 standards development team will add an alternative JSON format, called NDJSON, that supports processing large datasets with any JSON library. NDJSON, or newline delimited JSON, creates a file where each line can be processed as JSON independently of the rest of the file. This permits easy processing of NDJSON files in chunks or as streams. Dataset-JSON v1.1 datasets may be formatted using JSON or NDJSON.

5.4 Date Epochs

Several testers reported interoperability findings. These findings appear when the sender and receiver use different programming languages or technologies, such as when a SAS programmer sends a Dataset-JSON dataset to an R programmer. In this case, testers noted that SAS and R use different date epochs to generate integer-based date representations. SAS generates an integer date based on the number of seconds since 01-Jan-1960, while R uses 01-Jan-1970. In this case, there will be a decade of seconds difference between the SAS and the R integer-based dates.

To address this issue, Dataset-JSON v1.1 will represent dates as ISO 8601 datetimes. An additional metadata column attribute, called *targetDataType*, will be added to represent the *dataType* into which the datetime should be transformed. In the case of an integer-based date, the *dataType* will be datetime and the *targetDataType* will be integer. The conversion tools will convert integer-based dates into ISO 8601 datetimes and then back into the integer *targetDataType* using the appropriate date epoch. The UG will include a section on datetime datatypes to help users better understand how to represent and process them in Dataset-JSON v1.1.

5.5 Numbers and Precision

Some testers noted that rounding or slight differences in precision occurred, especially during interoperability testing.

Some of these differences may be accounted for by differences in the technologies converting Dataset-JSON back into native datasets, such as with SAS or R. Workshop attendees noted that differences in rounding or precision are expected when working in different programming languages. A pharmaverse blog post on *rounding differences between SAS and R* and another blog post on *working with floating point numbers in {admiral}* provide details. Workshop attendees stated that these are not problems to be solved by Dataset-JSON, but ones that should be addressed through documentation and education. For this reason, the UG will include a section that addresses precision and rounding.

In addition, the Dataset-JSON v1.1 standards development team decided to add support for the *decimal* datatype. This datatype exists in ODM and represents numbers more exactly than the floating-point datatype. *Decimal* datatypes represent non-repeating decimal fractions without rounding. Decimal numbers will be represented in Dataset-JSON as a string surrounded by quotes and have a *targetDataType* of decimal. Representing decimal numbers as a string prevents JSON libraries from automatically interpreting the number as a floating point and delegates the datatype type casting to the conversion software. Adding support for the *decimal* datatype provides an alternative for users who desire high levels of precision and exact representations of numbers.

5.6 Datatypes

The Dataset-JSON v1.1 standards development team agreed to add additional ODM datatypes to accommodate data representation needs identified during testing. Examples of these datatypes include datetime, decimal and Boolean.

Testers noted that for languages not using the Dataset-JSON *displayFormat* attribute, no attribute exists that indicates, for example, that an integer should be interpreted as a date. The Dataset-JSON v1.1 standards development team agreed to add the *targetDataType* attribute to capture datatype conversions, such as when a datetime should be represented as an integer.

The UG will provide additional details and examples describing how and when to use specific Dataset-JSON *dataTypes* and *targetDataTypes*. This is in response to testers' requests for guidance, noting that Dataset-JSON has many more datatypes than SAS XPT.

5.7 Nulls and Empty Strings

Testers asked whether empty strings ("") should be replaced with *null* in Dataset-JSON. The specification states, *"Missing values are represented by null in the case of numeric variables, and an empty string in case of character variables."* Workshop participants and the Dataset-JSON v1.1 standards development team recommend keeping the current specification and adding documentation and examples in the UG.

5.8 Non-ASCII Characters and Quoting Strings

Some testers reported dataset mismatches due to non-ASCII characters in the original dataset during conversion. While these non-ASCII characters are displayed correctly in Dataset-JSON, the issue arose during the conversion back to an SAS dataset. Similarly, participants noted issues with converting to SAS XPT format if the variable and character lengths exceeded the established limits.

Encoding best practices will be documented in the UG. For example, the Dataset-JSON v1.1 standards development team recommends using UTF-8 encoding, the default encoding scheme for JSON. Since Dataset-JSON supports Unicode and SAS XPT uses ASCII, the conversion software should identify cases where the target dataset's encoding scheme does not support characters in the source dataset. When the conversion software encounters unsupported characters in the Dataset-JSON dataset, an error should be thrown.

One tester could not convert datasets due to quote imbalance errors, but the PHUSE pilot team could not recreate the error. The JSON standard requires that strings be encapsulated in double quotes. Double quotes within strings must be escaped with a backslash; single quotes within strings will not be escaped. If the conversion software encounters imbalanced quotes or opening quotes without a closing quote, it should throw an error. The rules for using quotes for strings will be documented in the UG.

In summary, since JSON uses UTF-8 encoding, many non-ASCII characters are legal values. Ideally, the software processing Dataset-JSON will ensure correct encoding and inform the user if the dataset contains characters not supported by the receiving technology.

5.9 Metadata

Testers asked if a Define-XML file is required when using Dataset-JSON. Define-XML remains a submission requirement but not a Dataset-JSON requirement. Many data exchange scenarios do not require a Define-XML file. Dataset-JSON optionally references Define-XML. When a Define-XML is present, the OIDs in the Dataset-JSON datasets must match those in the Define-XML.

Dataset-JSON requires more metadata than SAS XPT datasets. This provides the metadata needed to convert Dataset-JSON datasets to other formats, support Dataset-JSON viewers, and enable software to import Dataset-JSON. Incorporating the metadata in the dataset ensures it will always be available with the dataset.

The UG will provide best practices for creating and managing Dataset-JSON metadata, including OID and ITEMGROUPDATASEQ generation.

5.10 Binary File Formats

Though not formally reported by testers, some testers noted that certain binary dataset formats are smaller and have faster read/write times than JSON-based formats. JSON is a lightweight, easy-to-implement data exchange format. It is the most widely used, broadly supported exchange format and is the de facto standard for data exchange via APIs.

The Dataset-JSON standard targets a wide range of data exchange scenarios involving tabular datasets. It is optimised for ease of sharing tabular data between information systems. File size, read/write speeds, and ease of querying, while important, are secondary to support for data exchange. Dataset-JSON provides reasonable file sizes and processing speeds such that it functions well for data exchange. It does not need to be the optimal dataset format for big data or analytical processing. In many cases, Dataset-JSON data exchange will be API-based. Many data exchange scenarios may never store Dataset-JSON as a file, but, instead, the data is retrieved using an API and stored in a database or native dataset format. Dataset-JSON may also be compressed, as is commonly implemented in APIs, which reduces the data size significantly.

This *LinkedIn post* highlights key reasons for selecting a JSON-based format.

5.11 Software Tools

Software data conversion tools functioned as the user interface to the Dataset-JSON-based data exchange during the pilot. Conversion software tools convert native datasets into Dataset-JSON and convert Dataset-JSON back into native dataset formats. The pilot primarily focused on three open-source conversion tools originally created during the 2022 COSA Dataset-JSON Hackathon. Currently, these software tools support Dataset-JSON v1.0, and once the CDISC team has completed the v1.1 specification we will work with the software developers to update the tools.
Less technical testers asked for additional documentation and examples to help them use the tools successfully. The Dataset-JSON v1.1 standards development team will develop additional documentation and example datasets and help test the tools. Volunteers will develop a new tool for converting Parquet datasets.

Testers also asked for Dataset-JSON viewer tools. The UG will list the currently available viewer tools, as many testers were unaware of them. The development of new viewer tools will also be encouraged.

5.12 Pretty Printing

By default, most programs that generate Dataset-JSON, and JSON in general, do not include the line breaks and tabs that make JSON easy to read. These extra characters increase the dataset size and are not needed by software tools. However, adding these characters to format Dataset-JSON – called pretty printing – supports human readability. Many editors and other tools optionally format JSON, including Dataset-JSON. Workshop participants and the Dataset-JSON v1.1 standards development team recommend that conversion software provide a pretty-print option. Dataset-JSON viewer software applications also address the need for human readability.

5.13 Split Datasets

The dataset-splitting requirements for regulatory submissions remain unchanged and are the same for Dataset-JSON and SAS XPT.

The Dataset-JSON v1.1 standards development team will create a JSONX format that provides an archive file which contains all the split dataset subsets and may be compressed to minimise the file sizes. JSONX archival files may not be an acceptable format for regulatory submissions.

5.14 Commercial Software Support

For successful use as a data exchange standard, it is important that commercial software tools support Dataset-JSON. Testers want CDISC CORE and Pinnacle 21 to support Dataset-JSON for conformance checking. Currently, CORE supports Dataset-JSON. Other software tools used for regulatory review should support the ability to import Dataset-JSON. Software tool developers who participated in the pilot workshops noted that Dataset-JSON support was simple to implement.

References

1. *Dataset-JSON as Alternative Transport Format for Regulatory Submissions PHUSE Working Group project web page*

2. *CDISC Dataset-JSON web page*

3. *No More XPT? Piloting New Dataset-JSON For FDA Submissions*

4. *COSA Dataset-JSON Hackathon Solutions*

5. *Why JSON for Datasets?*

6. *pharmaverse blog post: Rounding*

7. *pharmaverse blog post: Floating Point*

7. Disclaimer

The opinions expressed in this document are those of PHUSE and CDISC based on the findings of the pilot. Although FDA participated in the pilot, the findings and solutions in this report should not be construed as the regulator's policies nor should they be viewed as regulatory authority requirements.

8. Project Contact Information

- Email: workinggroups@phuse.global

9. Acknowledgements

Project Leads:
- Stuart Malcolm (PHUSE)
- Sam Hume (CDISC)
- Jesse Anderson (FDA)

Project Sponsors:
- Chris Price (PHUSE)
- Peter Van Reusel (CDISC)
- Lilliam Rosario, PhD (FDA)

Sub-team Leads:
- Kathy Brown (Sanofi)
- Nate Blevins (GSK)
- Marguerite Kolb (J&J)
- Hui Liu (Merck)
- Nicole Thorne (BMS)
- Eli Miller (Atorus Research)
- Eddy Foster (Roche)

FDA CBER Representatives:
- Lisa Lin (FDA)
- Gabriela Lopez Mitnik (FDA)

Note: Business Case and Strategy for Future Developments acknowledgements will feature in their respective publications.