



Dataset-JSON

An Alternative Transport Format for Regulatory Submissions

BUSINESS CASE

02-Jun-2025

How We Submit Data Today



SAS .XPT files, 'flat' data structure, version controlled



30+ year-old technology



Limited file size, restricted widths, lengths and formats of variables



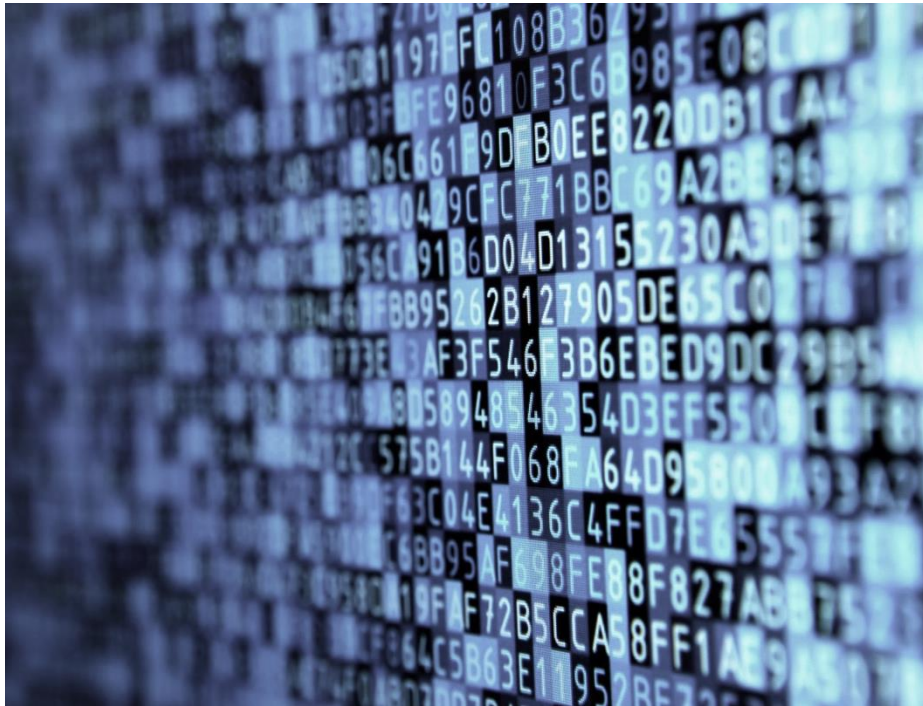
Only supports US ASCII character format



A Case for Change

Dataset-JSON is a widely accepted data format that will streamline data exchange between organisations. Moving to Dataset-JSON will prepare us for the current digital world and enable companies to more accurately, completely and efficiently represent their clinical research data, to better serve their patients/customers. Dataset-JSON supports current data standards but also allows for standards to evolve and removes current limitations imposed by xpt format.

Dataset-JSON



Dataset-JSON is:

Part of the ODM v2.0 standard

An open-source MIT licence

Schema supporting any tabular format

Extensible to support integrated metadata and new use cases

Linked to Define-XML for complete metadata

Integrated with CORE for conformance checking

Dataset-JSON advantages:

Based on the JSON standard used worldwide

Open-source and truly human-readable

Same or smaller file sizes relative to current required format

Removes variable naming, width or format limitations

Simple transformation to/from SAS data

Dataset-JSON – Business Case

Feature	Description	Benefit
Vendor-neutral	Dataset-JSON is an open, vendor-neutral format based on the widely implemented JSON standard and developed via a consensus-based standards development process.	Dataset-JSON is an independent standard that works well with almost any programming language and technology stack, including R and Python. XPT, though an open specification, works best with SAS.
Unicode support	Supports Unicode encoding, the most widely implemented encoding scheme.	Enables study datasets to include local language content such as Japanese or Chinese. XPT does not support Unicode encoding and is limited to US ASCII.
JSON-based	The JSON standard is the most common format for data exchange today and is the default format for data exchange via APIs.	Enables more modern methods of data exchange, including web-based APIs. JSON is supported by nearly every programming language and by most technology platforms.
Includes metadata	Dataset-JSON includes the metadata necessary to process, transform and view datasets.	Most processing tasks work with the metadata included in Dataset-JSON, simplifying how tools process datasets. Today, metadata is also in Define-XML in another file. * Teams have asked if it would be better to keep the metadata separate from the data, but we do not believe this is a good industry approach when looking to the future. There was a concern that if kept separate, it could be out of sync with the data.

Dataset-JSON – Business Case

Feature	Description	Benefit
Simple to implement	The 2022 Dataset-JSON Hackathon demonstrated that Dataset-JSON is easy to process from SAS, R, Python, Java, C++ and JavaScript. Dataset-JSON is simple to transform into SAS datasets, R data frames or Python data frames.	Since Dataset-JSON is simple to implement, most software vendors will have no trouble supporting Dataset-JSON, and many open-source software tools will be available to work with Dataset-JSON datasets.
No artificial length constraints	Does not have artificial length constraints on variable names, labels and data fields.	Removing the XPT constraints will allow the CDISC standards to use more descriptive variable names, easing learning and comprehension of the standards as well as allowing dataset variables to contain data longer than 200 characters.
Precise datatypes	Dataset-JSON works with a subset of the more precise ODM/Define-XML datatypes and removes the XPT restrictions to Num and Char datatypes.	Aligns Dataset-JSON with a subset of the datatypes in Define-XML and provides datatypes that align with modern software and data storage technologies.
Extensible	Like ODM, Dataset-JSON can be extended to add new features.	Extensibility means that developers can add new metadata and data to Dataset-JSON to meet the unique needs of their applications.

Dataset-JSON – Business Case

Feature	Description	Benefit
Small file sizes	Although not smaller than Parquet or some other formats, Dataset-JSON files sizes are smaller, on average, than XPT or Dataset-XML.	Ensures the Electronic Submissions Gateway submissions and related dataset storage are not impacted by file size and are better off than they would be with XPT.
Storage efficiency	Dataset-JSON stores data more efficiently than SAS XPT because it uses variable field widths instead of fixed field widths.	Improved storage efficiency means programmers don't need to set field lengths and that file sizes are not inflated by incorrectly sized fields.
Open source	Dataset-JSON is published under the MIT licence to allow developers to embed the schemas into their software and extend the schemas.	This allows developers to use and extend the Dataset-JSON schemas to support their requirements without cost.
Open-source tools	Since the 2022 Dataset-JSON Hackathon, developers have begun creating a set of open-source software tools for converting Dataset-JSON into native dataset formats. Additional applications include viewers, REST APIs and streaming.	Free tools to convert datasets to and from Dataset-JSON eases the change process and makes it simple for programmers to get the tools for working with Dataset-JSON. Supports the FDA goal of encouraging open-source reviewer tool development.



Dataset-JSON – Business Case

Feature	Description	Benefit
Works with Define-XML	Dataset-JSON includes the basic metadata needed to process each dataset, while Define-XML contains metadata for all study datasets. Dataset-JSON does not include everything that is in a Define-XML – such as code list definitions and value-level metadata, and links to the reviewer’s guide. That is why Define-XML complements Dataset-JSON.	Dataset-JSON contains enough metadata to process the datasets, simplifying how software processes dataset metadata. Define-XML can be used as a specification to generate shell Dataset-JSON files. CDISC plans to create a Define-JSON to complement Dataset-JSON.
Human-readable	Dataset-JSON, and JSON in general, is a text-based, human-readable standard.	Dataset-JSON can be viewed by off-the-shelf editors and viewers, in addition to purpose-built viewers, making it easy for anyone to read.
Streaming support	Streaming allows the dataset to be processed incrementally rather than being read into memory beforehand.	Streaming makes it possible to process very large datasets on normally configured laptops or servers. Many JSON libraries support streaming, meaning the full dataset need not be read into memory for processing. These libraries support incremental reading of datasets, enabling applications to consume very large datasets without running out of memory.
Part of ODM v2.0	Dataset-JSON is part of ODM v2.0, which now supports serialisations in XML and JSON.	ODM v2.0 and Dataset-JSON together cover a wide range of data exchange scenarios because tabular and hierarchical datasets are supported.

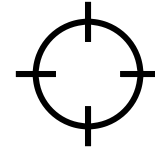
Dataset-JSON – Business Case

Feature	Description	Benefit
Text-based standard	Dataset-JSON, and JSON in general, is a text-based standard.	As a text-based standard, it is easy to view, process and exchange. Text-based formats work well for archives because they do not require proprietary tools to work with them.
Drop-in XPT replacement	Dataset-XML worked as a drop-in replacement and resolved the limitations in SAS V5 XPORT, negatively impacting on the CDISC standards – as was demonstrated in the 2014 Dataset-XML FDA Pilot.	Working as a drop-in replacement for XPT means existing processes, including electronic submissions, and tools are not impacted by the change. This makes the change easier to implement.
Compatible with existing standards	Dataset-JSON can be used with any tabular dataset, including existing CDISC standards datasets.	Works with existing standards to minimise the impact of changing standard formats for data exchange and submissions, while being extensible to support future needs. Together, ODM and Dataset-JSON cover a wide range of data exchange scenarios.

Concerns Considerations

Food for Thought

The following concerns should be taken into consideration when looking into converting to Dataset-JSON.



Dataset-JSON should be thought of as an exchange format, but, internally, companies may prefer to build system integrations using other formats



Effect on contracts with current vendors and the cost of doing business with them (increasing and/or decreasing accordingly)



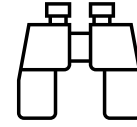
Effect on current in-house systems and cost of conversions



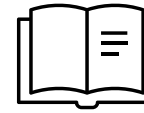
Businesses trending towards other technologies such as APIs.

Concerns Considerations

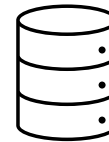
Continued



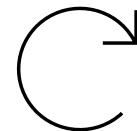
Need a Dataset-JSON viewer



Need to develop a Dataset-JSON user guide,
with conventions on data handling



Data storage – with an API, an implementer
might never store data as a Dataset-JSON file



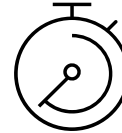
Dataset-JSON standard to be updated to
address findings from the pilot

Concerns Considerations

Continued



CDISC Core are working on developing the capacity to import/export Dataset-JSON

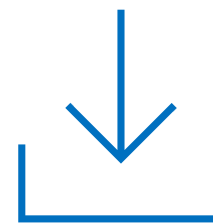


When to develop/create Dataset-JSON datasets? Throughout the study or when creating the final submission package?

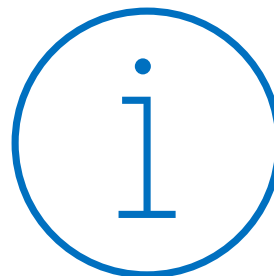
Resources Information



[Dataset-JSON as Alternative Transport Format for Regulatory Submissions](#)



[CSV vs Parquet vs JSON for Data Science by Stephen](#)



[Working with Dataset-JSON using SAS – Lex Jansen](#)



[Why JSON for Datasets?](#)



The Global Healthcare Data Science Community

Contact Channels

- 📞 UK +44 1843 609600
- ✉️ office@phuse.global
- 🌐 phuse.global

Social Media

- 🐦 [@phusetwitta](https://twitter.com/phusetwitta)
- 📘 [/phusebook](https://facebook.com/phusebook)
- 📺 [/phusetube](https://youtube.com/phusetube)
- 🌐 [/company/phuse](https://linkedin.com/company/phuse)

Disclaimer

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

Project Contact Information

Email: workinggroups@phuse.global

Acknowledgements

Project Sponsors: Chris Price (PHUSE), Peter Van Reusel (CDISC)

Project Leads: Stuart Malcolm (PHUSE), Sam Hume (CDISC)

Sub-team Leads: Nate Blevins (GSK), Marguerite Kolb (J&J)

