



CSS 2023

The Computational Science Symposium

18–20 September
Silver Spring, MD



phuse.global



Introducing Dataset-JSON

What is Dataset-JSON?

A dataset exchange standard for exchanging tabular data leveraging JSON designed to meet the regulatory submission needs and eliminating limitations of legacy formats

Dataset-JSON is...

- Part of the ODM v2.0 standard and based on the JSON standard
- Open-source and truly human readable
- An open-source MIT license
- Schema supporting any tabular format
- Extensible to support new metadata and new use cases
- Linked to Define-XML for complete metadata



Why change transport formats?

- Numerous SAS V5 XPORT (XPT) limitations
 - Limited variable types
 - Limited to US ASCII encoding
 - 8-character variable names
 - 40-character labels
 - 200-character field widths
 - Lacks robust metadata
 - Not extensible
- Not truly vendor neutral
- Not broadly supported by new technologies



Why JSON?

- JSON is...
 - The most widely used data interchange format
 - An open standard that's human-readable
 - Supported by nearly every programming language and technology framework
 - Simple to implement - easy to read and write
 - Best alternative for a standard data exchange format



Why not include Define-XML?

- Dataset-JSON
 - provides metadata for viewing, converting, and loading
 - seeks to maintain small file sizes
- Define-XML
 - includes metadata not needed for many use cases
 - would significantly increase dataset file sizes
 - complements Dataset-JSON



Why don't we have Define-JSON?

- Plan to work on a Define-JSON based on ODM v2.x
- Currently working on ODM v2.0 supplements
 - API
 - ODMPATH
 - JSON schemas
 - Compressed ODM
 - Streaming ODM



Why create a Dataset-JSON API?

- JSON is the de facto standard for API data exchange
- Defining a standard API for the exchange of datasets creates an opportunity to begin moving away from static file transfers



Why is Dataset-JSON row-based?

- Dataset-JSON was created primarily as a data exchange standard
- Some analysis tools are column-based
- Row-based datasets are easy to read and write
- Convert Dataset-JSON to native datasets for analysis



What about non-tabular datasets?

- Dataset-JSON was designed to represent tabular datasets
- Dataset-JSON is an ODM-based standard
- ODM represents non-tabular data



What are the goals of the pilot?

Milestone 1: Short Term

- Pilot submissions using JSON format with existing XPT ingress/egress to carry the same data
- Same content, different suitcase, no disruption to business process on either side
- In parallel, evaluate with FDA how their toolset can support JSON format and identify tool upgrade roadmap

➔ **Success Criteria: Demonstrate that Dataset-JSON can transport information with no disruption to business**

Milestone 2: Development of future strategy

- Evaluate how current and future industry standards can benefit without XPT limitations
 - e.g., Variable names > 8, labels > 40, data > 200
- Evaluate combining metadata with data
 - e.g., Define-XML / Define-JSON based
- Enhanced conformance rules
- Collaborate with FDA to develop plan to retool their environment to natively consume JSON

➔ **Success Criteria: Demonstrate the viability of Dataset-JSON as the primary transport option**



Current Status of Pilots

- Internal testing within FDA CDER OCS environment complete
- Successful conversion of nonclinical datasets from XPT to JSON without information loss
- Next steps: Accept test submissions from PHUSE Dataset-JSON working group members



Pilot Subteams

1. Pilot Submissions Report
2. The Dataset-JSON Business Case
3. Technical Implementation
4. Strategy for Future Development



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)