



Dataset-JSON as Alternative Transport Format for Regulatory Submissions

Jesse Anderson, FDA CDER OCS

Sam Hume, CDISC





Meet the Speakers

Jesse Anderson

Title: Team Lead, Clinical Services

Organization: Office of Computational Science CDER FDA

<https://www.linkedin.com/in/andersonjessep/>



Sam Hume

Title: VP, Data Science

Organization: CDISC

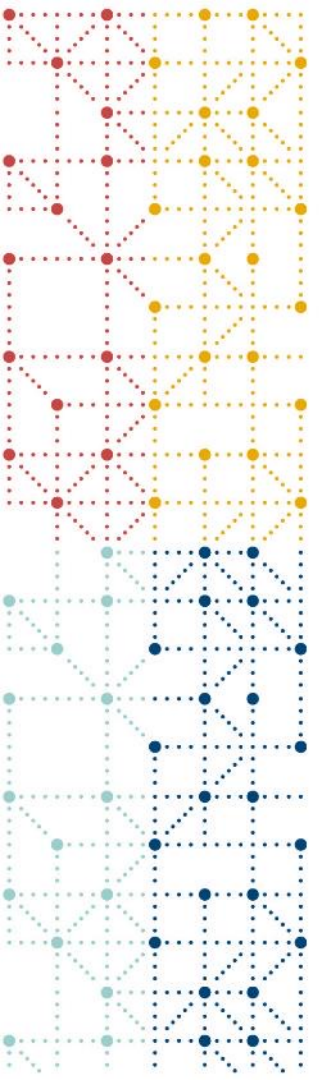
<https://www.linkedin.com/in/sam-hume-dsc>



Disclaimer and Disclosures

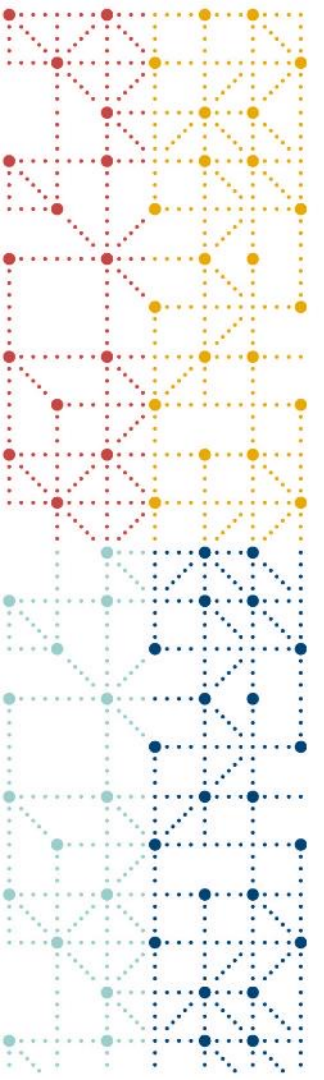
The views and opinions presented here represent those of the speaker and should not be considered to represent advice or guidance on behalf of the U.S. Food and Drug Administration.

The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.



Agenda

1. Pilot Background
2. Preliminary OCS Pilot Results
3. Executing the Pilot
4. Why you should get involved!



Background



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
- Schema supporting any tabular format
- Extensible to support new metadata and new use cases
- Linked to Define-XML for complete metadata
- Created to be used as stand-alone datasets with smaller file sizes

What are the goals of the pilot?

Milestone 1: Short Term

- Pilot 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
- Allow FDA to evaluate how internal tools can support JSON format

➔ **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
- FDA to utilize findings to evaluate tool redevelopment plan to natively consume files in JSON format

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

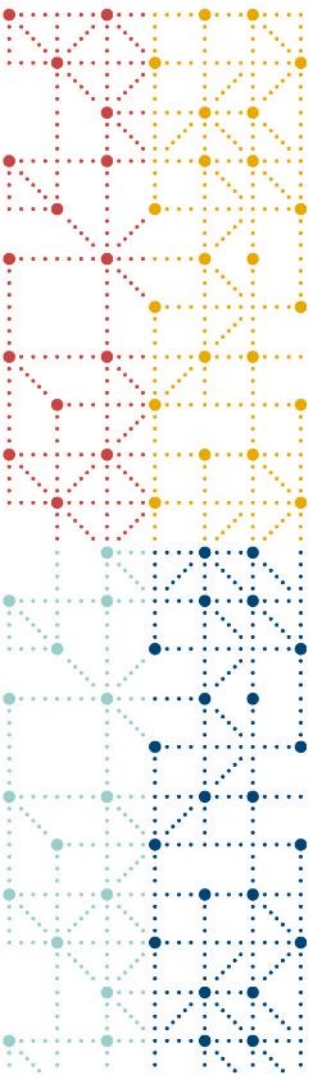


Pilot Subteams

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

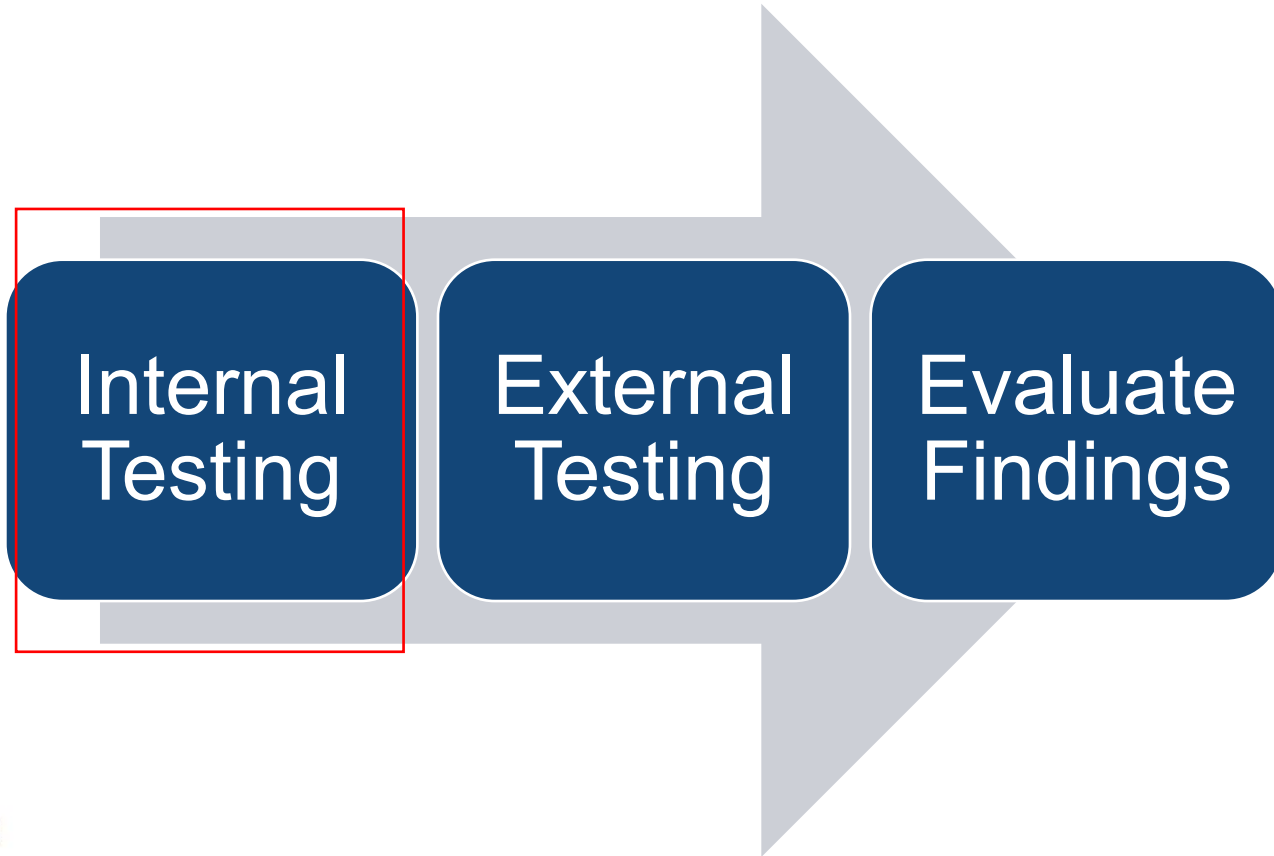
Dataset-JSON Pilot: Timeline





Preliminary OCS Pilot Results

Overall Pilot Strategy



Pilot Strategy

Goal: Confirm data integrity within nonclinical test data submissions

Utilized
Python
Hackathon
Solution

Reviewed
CDISC
Dataset-
JSON
specifications

Converted
Datasets from
XPT to JSON
and back

Performed
Validation
Checks

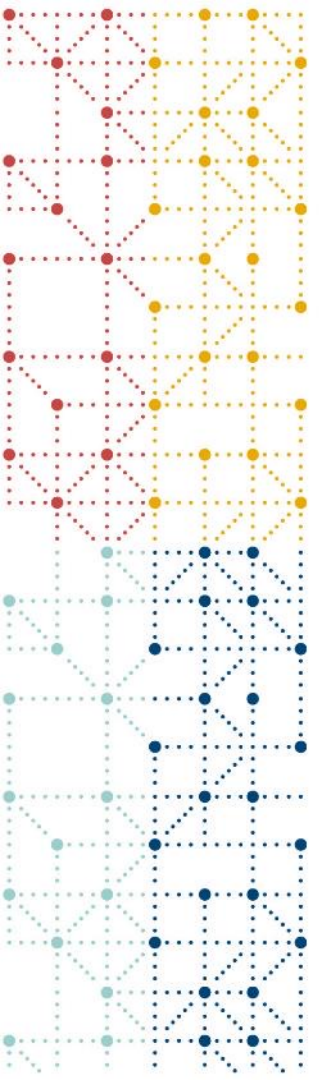


Preliminary Results

- Minimal effort for conversion utilizing Python
- Data integrity maintained throughout conversions
- JSON files comparable to XPT file sizes
- Date issues identified

Next Steps

Internal



Executing the Pilot

Open-Source Conversion Software Tools



SAS

High-level Process Overview



1. Select the software conversion tool(s) to use (one or more)



2. Select study datasets and convert them to Dataset-JSON



3. Convert Dataset-JSON datasets into a native dataset form



4. Compare the native datasets back to the original format and ensure there are no unexpected changes in converted dataset



5. Optionally, test other conversion scenarios (e.g., interoperability tests)



6. Complete the on-line questionnaire to record your findings

Example Conversion Scenarios

Pilot Conversion Scenarios

1. SAS dataset -> Dataset-JSON -> SAS dataset
2. R dataframe -> Dataset-JSON -> R dataframe
3. Dataset-JSON -> native format -> Dataset-JSON



Complete at least 1 scenario from the list above or use your own scenario

Example Interoperability Scenarios

Interoperability Conversion Scenarios

1. SAS dataset -> Dataset-JSON -> R dataframe
2. R dataframe -> Dataset-JSON -> SAS dataset



Interoperability scenarios are optional and may be tested in addition to the Pilot Conversion Scenarios. Improved interoperability is a benefit of Dataset-JSON.

Technical Implementation Observations



File Size Observations from Initial Testing

		SAS v5 XPT	Dataset-XML	Dataset-JSON
SDTM	FT	5,917	4,287	858
	LB	2,699	4,104	640
	VS	784	1,372	229
ADaM	ADLBC	33,441	145,575	24,942
	ADQSNPIX	12,840	61,561	8,404
	ADVS	13,313	51,646	8,257

- These are numbers about uncompressed files
- These are exchange formats, not operational formats
- File sizes are in KB
- CDISC pilot datasets were used for the comparison



Dataset-JSON Hackathon II

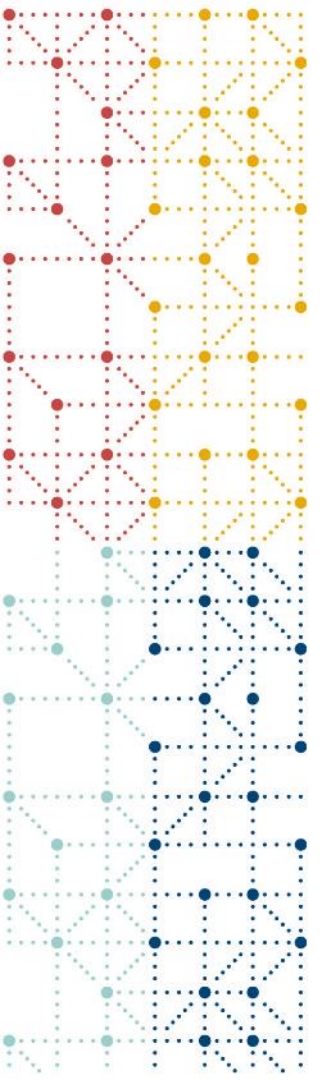
- Primary objective: Create a draft REST API specification for Dataset-JSON
- Secondary objective: Proof-of-concept implementations to demonstrate and test the API specification
- Virtual hackathon
 - <https://github.com/cdisc-org/DataExchange-DatasetJson-API>
- Draft API specification will be used as an input to:
 - Dataset-JSON Pilot future strategy and business case teams
 - ODM v2.x API development



Why JSON?

JSON is...

- The most widely used data interchange format
- An open, stable standard that's human-readable
- Supported by nearly every programming language
- Simple to implement - easy to read and write
- The de facto standard for retrieving data from APIs
- Part of ODM v2.0 that's now available as JSON
- Used by other healthcare data standards (HL7 FHIR)



Interested? Get Involved!

The pilot is open to anyone that wants to participate.

Thank You!

Questions?

Interested in Participating?

Jesse Anderson:

- jesse.anderson@fda.hhs.gov

Sam Hume:

- shume@cdisc.org



Open-Source Conversion Software Tools

- SAS

- <https://github.com/lexjansen/dataset-json-sas>
- The SAS conversion software by Lex Jansen
- Dataset-JSON example files are included in the repository
- Includes a macro for comparing libraries with SAS datasets
- Includes a Python script for validating Dataset-JSON
- Documentation is included

- R

- <https://github.com/atorus-research/datasetjson>
- <https://atorus-research.github.io/datasetjson/index.html>
- <https://cran.r-project.org/web/packages/datasetjson/index.html>
- R conversion package by Atorus Research and Johnson & Johnson
- Documentation is included

- Python

- <https://github.com/dostiep/Dataset-JSON-Python> ([submit issues](#))
- Python conversion software by Pierre Dostie
- We will not cover the Python tooling in the workshop
- The Dataset-JSON Pilot will focus on SAS and R, but any conversion tool can be used

