Dataset-JSON Pilot Report and Next Steps

Jesse Anderson, FDA-CDER
Sam Hume, CDISC

Meet the Speakers

Jesse Anderson

Title: Data Standards Lead

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.

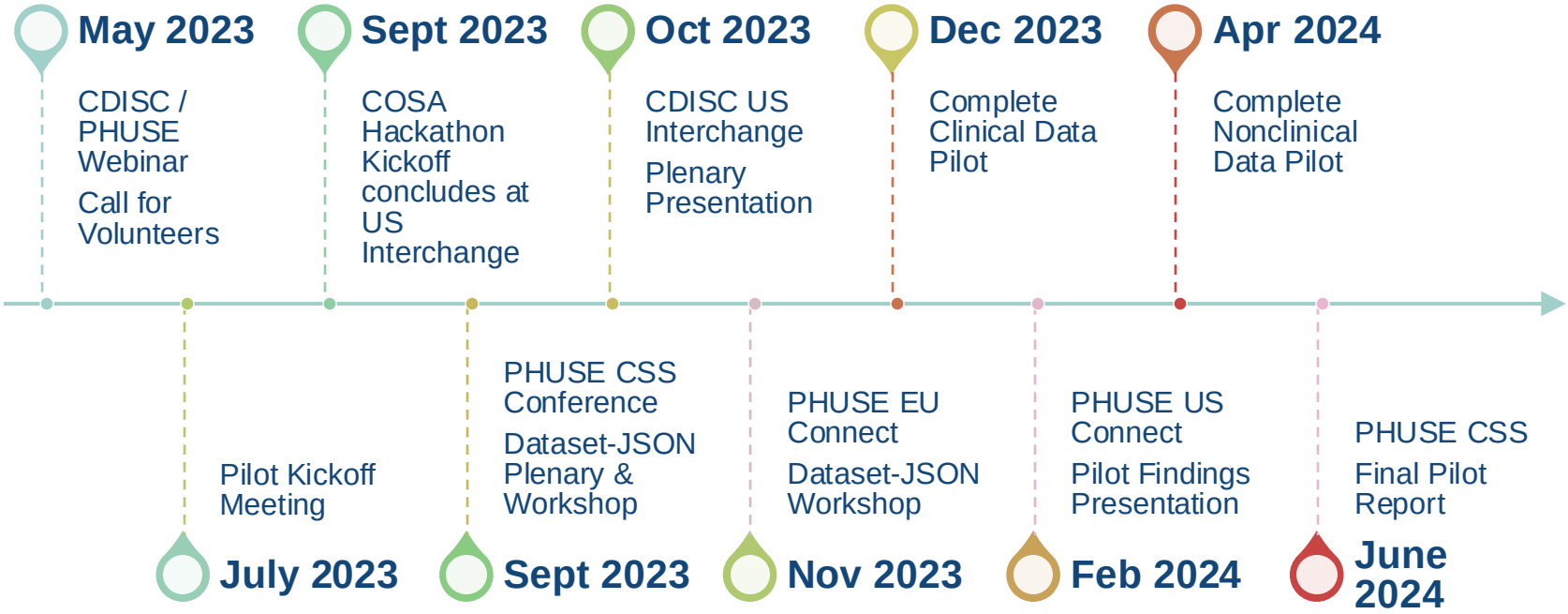
Agenda

1. The Dataset-JSON Pilot
2. Pilot Results
3. Technical Findings
4. Next Steps

The Dataset-JSON Pilot

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

Project Timeline



Project Subteams

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

Pilot Goals

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**

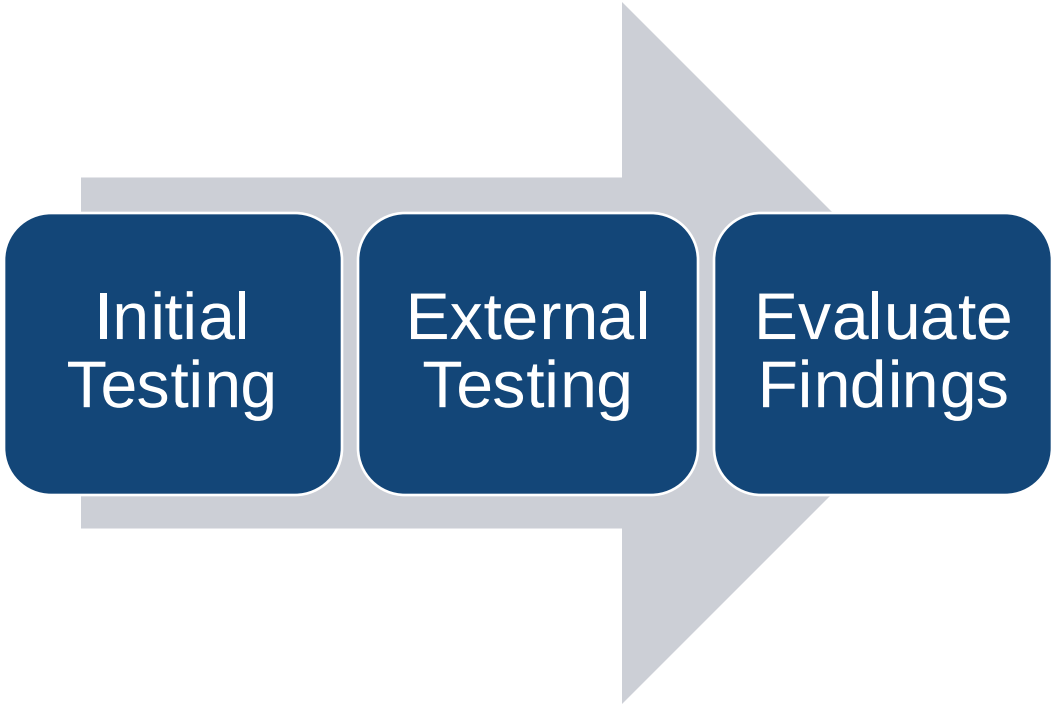
Milestone 2: Development of future strategy

- 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**

cdisc

Pilot Strategy



Pilot Strategy

- **Initial Testing**: Industry and FDA complete internal testing utilizing CDISC Hackathon Tools
- **External Testing**: Test regulatory JSON submissions via test Electronic Study Gateway
- **Evaluate Findings**: Team to review findings from questionnaire and FDA testing
- Report out on findings to industry and address issues in Dataset-JSON v1.1

Pilot Results

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

Summary of Findings

Overall, results showed minor date representation, display format issues, and precision concerns (full findings listing here)

Findings can be addressed with: (1) standards updates, (2) User Guide content, and (3) tool updates and enhancements

Many findings related to the conversion tools and interoperability testing

Most issues (e.g., date representation, precision) related to conversion tools and interoperability testing across different tools

cdisc

Pilot Conclusions

Milestone 1 satisfied: Dataset-JSON can transport information with no disruption to business and is viable as the primary transport option

FDA testing noted that, to date, COTS analytical tools do not accept JSON files. Updates necessary for full implementation of Dataset-JSON as transport file for regulatory data.

Dataset-JSON files successfully tested and submitted to FDA via test Electronic Study Gateway with no data integrity issues

Dataset-JSON v1.1 to address the pilot findings

Technical Implementation

A high-level overview of some key technical findings and solutions covered in the pilot report.

Project Subteams

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

Processing Large Datasets

Findings

- Some conversion solutions were unable to convert very large datasets effectively
  - Took too long
  - Failed to complete
- Conversion tools are uneven in their ability to process large datasets

Solutions

- Standards: Add NDJSON as an alternative Dataset-JSON representation to allow any JSON library to process large datasets
- Tools: Update conversion software tools to use JSON libraries that support streaming and add support for NDJSON
- Docs: Test and capture tool processing metrics

cdisc

Date Representations: Date Epochs

Findings

- Date epochs are different for SAS (1/1/1960) and R (1/1/1970)
  - Interoperability issue
  - Impacts generation of dates as integers

Solutions

- Standards: Represent dates as ISO 8601 datetimes. Add metadata to inform the conversion tools to convert the dates to an integer where appropriate.
- Tools: The conversion tools will manage converting dates to and from the ISO format transparently.
- Docs: Add to User's Guide (UG). Using ISO 8610 date formats is considered a JSON best practice.

Numbers and Precision

Findings

- Precision mismatches sometimes occurred when comparing floating point values with many digits after the decimal.
  - Various JSON libraries apply floating point and rounding strategies
  - Interoperability issue

Solutions

- Standards: Add a decimal datatype that stores a number as a JSON string and converts it back to a numeric decimal datatype with no rounding or loss of precision. Add new metadata to describe the technology that generated the data
- Tools: Store decimal numbers as a string to be converted by the conversion tool instead of the JSON library. Document the rounding strategy used
- Docs: UG will document how to work with and compare floating-point numbers and the fact that minor rounding differences exist when using different technologies

Datatypes and Associated Conversions

Findings

- For languages not using display formats, there is no indicator that an integer should be interpreted as a date
- Precision may be impacted by rounding that occurs in the JSON libraries
- It is unclear when to use specific datatypes.
- Expand the available datatypes

Solutions

- Standards: Add additional data types. Add additional metadata to represent the target data type so that, for example, the receiver knows that an integer represents a date
- Tools: Store decimal datatype numbers as a string to be converted by the conversion tool and not by the JSON library. Add support for additional datatypes and conversion metadata
- Docs: UG will document the new target datatype metadata as well as when and how to use the datatypes supported by Dataset-JSON

Unicode and Encoding

Findings

- Dataset-JSON supports Unicode, while SAS XPT is ASCII-based
- Non-ASCII characters in the original dataset were not supported in the target SAS dataset.
- How can we ensure we do not encounter encoding and decoding issues when working with non-ASCII characters?

Solutions

- Tools: Request that conversion software flag characters that don't match your intended encoding scheme.
- Docs: UG will document best practices for dealing with encoding. For example, the Dataset-JSON team recommends using UTF-8 encoding, the default encoding scheme for JSON. Document improved global language support.

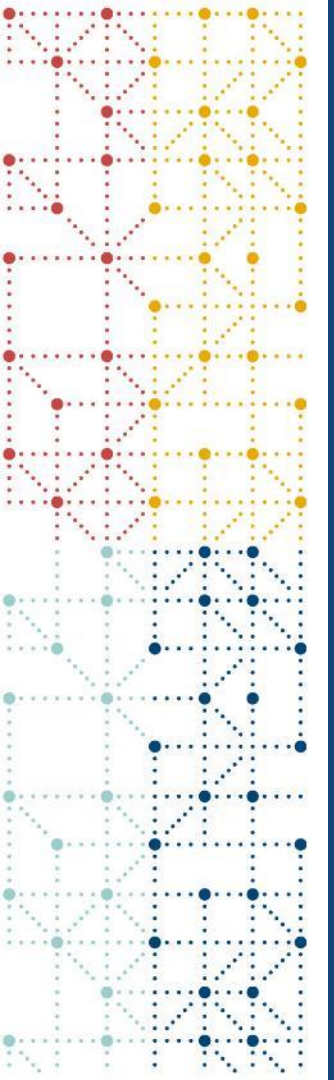
Define-XML Metadata and OIDs

Findings

- Is a Define-XML required to generate Dataset-JSON?
- OID metadata is not an XPT requirement; not everyone knows how to use them
- Can the conversion software generate ITEMGROUPDATASEQ?
- Dataset-JSON requires metadata not needed for XPT

Solutions

- Tools: Add support for generating the needed metadata without a Define-XML, including auto-generating OIDs and ITEMGROUPDATASEQ. Make OIDs optional.
- Docs: Define-XML remains a submission requirement but is not a Dataset-JSON requirement. Dataset-JSON optionally references Define-XML. UG will provide best practices for creating and managing Dataset-JSON metadata, including OID and ITEMGROUPDATASEQ generation.



Next Steps

Using the pilot findings to improve the standard, documentation, and tools

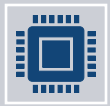
Open-Source Conversion Software Tools

SAS

- The SAS conversion software by Lex Jansen
- Includes a macro for comparing libraries with SAS datasets
- Documentation is included

R

- R conversion package by Atorus Research and Johnson & Johnson
- Documentation is included

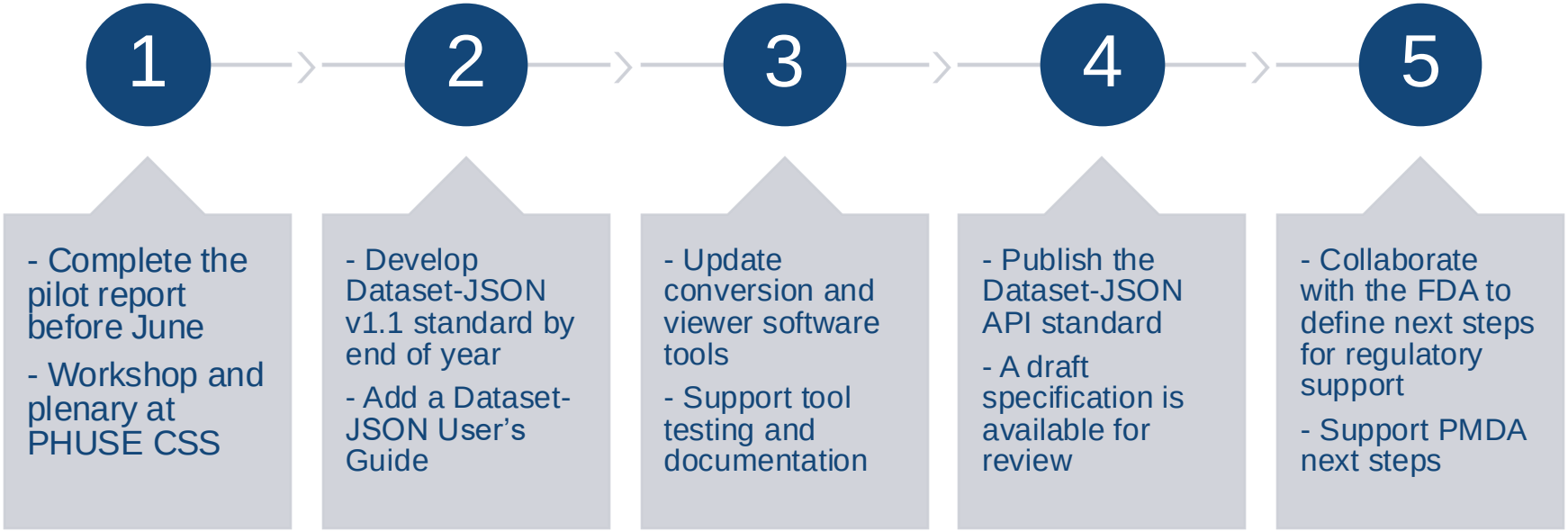


Python

- Multiple Python conversion software tools
- Documentation is included
- Covers multiple dataset formats, including Parquet and SAS

- Volunteers committed to ensuring a Parquet conversion tool is available
- Open-source software teams need contributors

Next Steps



Much of this activity is happening concurrently – volunteers needed

Thank You!

Jesse Anderson:

- jesse.anderson@fda.hhs.gov
- https://www.linkedin.com/in/andersonjessep/

Sam Hume:

- shume@cdisc.org
- https://www.linkedin.com/in/sam-hume-dsc