Key Takeaways from the Dataset-JSON Pilot

Sam Hume, CDISC

Meet the Speaker

Sam Hume, DSc.

Title: VP, Data Science

Organization: CDISC

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

I co-lead the PHUSE/CDISC/FDA Dataset-JSON as an Alternative Transport Format for Regulatory Submissions Pilot with Stuart Malcolm (Veramed) and Jesse Anderson (FDA). I also co-lead the Dataset-JSON v1.1 standards development team.

Agenda

1. Background
2. Conclusion
3. Findings
4. Next Steps

Background

A brief introduction to the pilot to baseline folks before we get into the findings and next steps

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

- The pilot is a collaboration between CDISC, PHUSE, and the FDA
- The pilot leads are:
  - CDISC: Sam Hume, CDISC
  - PHUSE: Stuart Malcom, Veramed
  - FDA: Jesse Anderson, FDA
- The pilot kickoff was completed on 27 July 2023
  - The final readout will occur at the PHUSE CSS conference, 3-5 June 2024
- Pilot testing is open to anyone
  - The official pilot testing is complete, but there are still opportunities to participate in testing

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 eliminate the limitations of legacy formats

Dataset-JSON is…

- Designed to support a broad range of data exchange scenarios
- Supports API and file-based data exchange
- JSON is simple to implement, very stable, and widely supported
- Open-source schema supports any tabular format
- Extensible to support new metadata and new use cases
- Linked to Define-XML for complete metadata

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**

Conclusion

Before we get into the detailed findings, let's summarize the pilot's preliminary high-level takeaway. This is preliminary since the pilot team is still working on the final report.

Dataset-JSON (DSJ) Preliminary Pilot Outcomes

DSJ can function as an XPT alternative

DSJ met the pilot objectives:

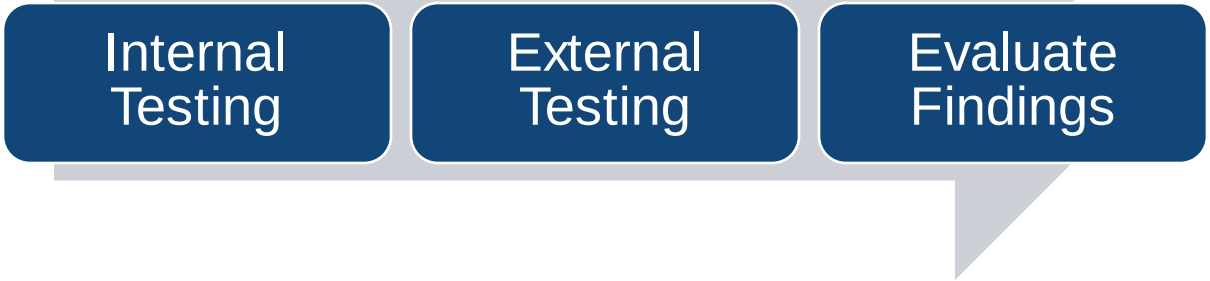
- No show-stoppers were identified
- Milestone 1 was satisfied
- Demonstrated that Dataset-JSON can transport information with no disruption to business
- FDA testing was successful

Improvements Needed

Three categories of improvements are needed:

1. Update the standard
2. Create a User's Guide
3. Update and enhance tools

FDA CDER OCS Pilot Testing



- Minimal effort required to convert datasets utilizing Python conversion software
- Internal testing: Used the Python conversion software. Data integrity confirmed.
- External testing: Includes submissions through test ESG with CDER and CBER data.
- No showstopper findings. Integer date conversion findings encountered.

Findings

A summary of the key findings from the pilot testing

Summary of Findings

After analyzing the reported results, we categorized them into 21 distinct findings.

Findings have solutions that include: (1) standards updates, (2) User's Guide content, and (3) tool updates and enhancements

Many findings related to the conversion tools and interoperability testing

The remainder of this section will highlight the most interesting findings

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

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

cdisc

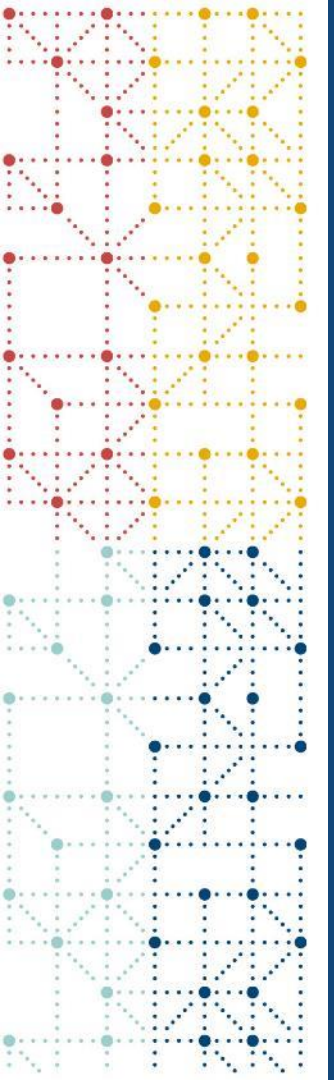
Define-XML Metadata and OIDs

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

What do we need to do to improve the standard and tools?

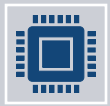
Open-Source Conversion Software Tools

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

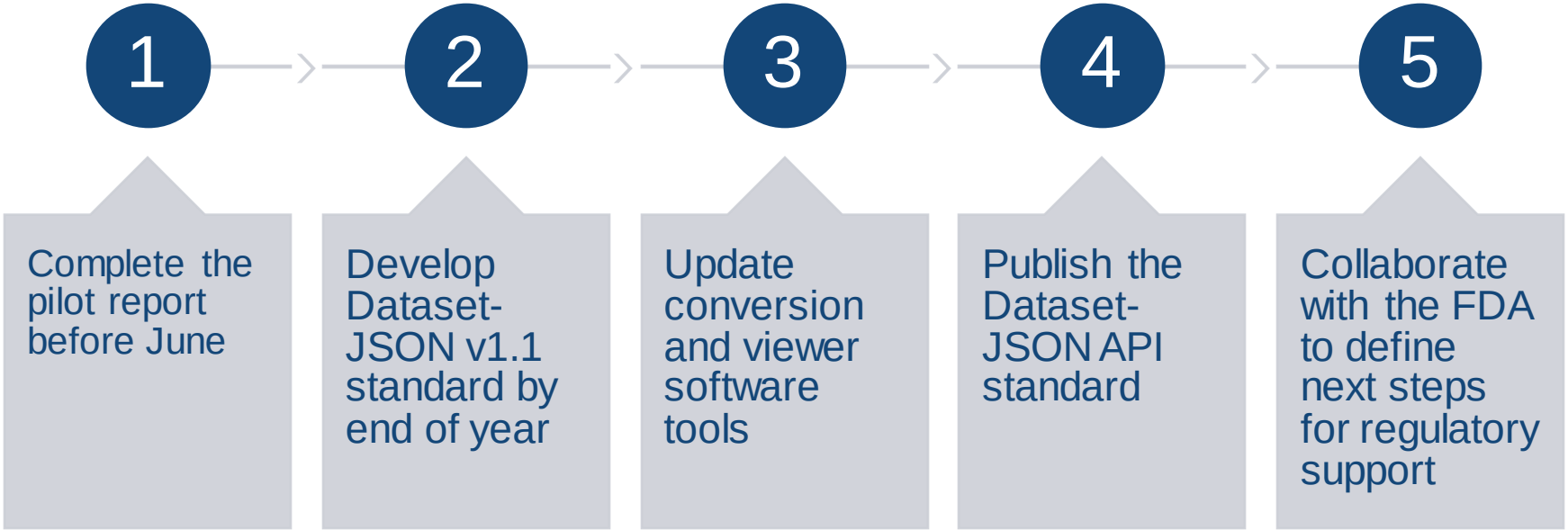


Python

- 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!

Questions?

Interested in volunteering for Dataset-JSON v1.1?

https://www.cdisc.org/volunteer/form

shume@cdisc.org

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

Resources

- Dataset-JSON API draft specification
  - https://github.com/cdisc-org/DataExchange-DatasetJson-API
- Dataset-JSON specification
  - https://www.cdisc.org/dataset-json
  - https://wiki.cdisc.org/display/PUB/Dataset-JSON
- Dataset-JSON GitHub repository
  - https://github.com/cdisc-org/DataExchange-DatasetJson
- Dataset-JSON v1.1 wiki
  - https://wiki.cdisc.org/display/DSJSON1DOT1/Dataset-JSON+v1.1
- COSA Directory Dataset-JSON Hackathon I projects
  - https://cosa.cdisc.org/hackathons/datasetJson
- SAS conversion software:
  - https://github.com/lexjansen/dataset-json-sas
- R conversion software:
  - https://github.com/atorus-research/datasetjson
- Python conversion software:
  - https://github.com/swhume/dataset-json
  - https://github.com/dostiep/Dataset-JSON-Python