Dataset-JSON Pilot and Hackathon II Update

COSA Quarterly Spotlight – Q3 2023
Sam Hume, D.Sc.
2023-10-05
Dataset-JSON Pilot Project

*Dataset-JSON as an alternative transport format for regulatory submissions*
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

See the PHUSE *Transport for the Next Generation* (2017) White Paper
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
- Used by other healthcare data standards (HL7 FHIR)
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
Dataset-JSON Schemas, Examples, and Specification

• Dataset-JSON example datasets:
  • https://github.com/cdisc-org/DataExchange-DatasetJson/tree/master/examples

• Dataset-JSON schema file:
  • https://github.com/cdisc-org/DataExchange-DatasetJson/tree/master/schema

• Dataset-JSON standard repository:
  • https://github.com/cdisc-org/DataExchange-DatasetJson

• The Dataset-JSON specification
  • https://www.cdisc.org/dataset-json
  • https://wiki.cdisc.org/display/PUB/Dataset-JSON
Conversion Software Tools

• SAS
  • https://github.com/lexjansen/dataset-json-sas (submit issues)
  • 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 (submit issues)
  • 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
8. SUCCESS - Use comments to expand on your conversion success response in Question 7 above.

Enter your answer

9. FORMATS - Were you able to convert Dataset-JSON datasets into SAS datasets or R dataframes or other dataset formats?

- Yes
- No
- Not Sure

10. FORMATS - Use comments to expand on your dataset formats response in Question 9 above.

Enter your answer

11. EXPECTATIONS - Did the conversions to and from Dataset-JSON work as expected?

- Yes
- No
- Not Sure

12. EXPECTATIONS - Use comments to expand on your expectations response in Question 11 above.

Enter your answer

13. CONTENT - What types of datasets content did convert? Select all that apply.

- SEND
- SDTM
- ADaM
- Other

14. CONTENT - Use comments to expand on your content response in Question 13 above.

Enter your answer

15. ACCURACY - Based on your conversion testing, did Dataset-JSON accurately represent your data?

- Yes
- No
- Not Sure
There are 8 unique questions on the questionnaire:
- Comment text boxes exist for most questions
- Contact information is requested (questions 1-4) for possible follow-up questions

The form is on-line and should be completed by each group that tests Dataset-JSON

Anonymized results will be published in the pilot final report
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
• Preparing for clinical data testing
Pilot Subteams

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

- **31 May 2023**: CDISC / PHUSE Webinar Call for Volunteers
- **1 Sept. 2023**: COSA Hackathon Kickoff
- **27 July 2023**: Pilot Kickoff Meeting
- **Q3 SEND Pilot**: Complete Non-clinical Data Pilot
- **Q4 2023**: Complete Clinical Data Pilot
- **17-20 Oct. 2023**: CDISC US Interchange Plenary Presentation
- **5-8 Nov. 2023**: PHUSE EU Connect Dataset-JSON Workshop
- **25-28 Feb. 2024**: PHUSE US Connect Pilot Findings Presentation
- **PHUSE CSS Final Pilot Report**: Q2 2024

**Concludes at US Interchange**

**Interchange**

**PHUSE CSS Conference**

**PHUSE CSS**

**PHUSE EU Connect**

**PHUSE**

**Webinar**

**CDISC**
Dataset-JSON Hackathon II

Creating a draft Dataset-JSON API Specification
Problem Statement

• **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
  • Team will work collaboratively to develop and test the draft specification
  • Will read out the results of the hackathon during the Interchange

• Dates: Sept. 1 – October 13
What do we plan to do with the draft API specification?

• The draft API specification will be released publicly for review and comments

• The API specification will be delivered to the ODM v2.x team for review and publication with an overall ODM v2.0 API specification

• Write-up the results of the hackathon to share with
  • Dataset-JSON Pilot participants
  • FDA representatives interested in Dataset-JSON

• Implement prototypes to demonstrate and test the API specification

• GitHub repository: https://github.com/cdisc-org/DataExchange-DatasetJson-API
Draft Dataset-JSOn API in OpenAPI Specification

/studies/{studyOid}/datasets:
  get:
    summary: Return a list of datasets
    description: Return list of all available datasets for a study
    operationId: read_dataset_list_get
    parameters:
      - name: studyOid
        in: path
        required: true
        schema:
          type: string
    - name: creationDate
      in: query
      required: false
      description: datasets on or after the creation date
      schema:
        type: string
        format: date
    responses:
      '200':
        description: Successful Response

/studies/{studyOid}/datasets/{datasetOid}:
  get:
    summary: Return a dataset
    description: Return a specified Dataset-JSON dataset
    operationId: read_dataset_get
    parameters:
      - name: studyOid
        in: path
        required: true
        schema:
          type: string
      - name: datasetOid
        in: path
        required: true
        schema:
          type: string
      - name: metadataonly
        in: query
Dataset-JSON Pilot: Timeline

- **31 Aug 2023**: Hackathon Kickoff
- **1 Sep 2023**: Hackathon begins
- **13 Oct 2023**: Hackathon ends
- **23 Oct 2023**: Comment period begins
- **13 Oct 2023**: Results published for US Interchange
- **3 Nov 2023**: Comment period ends
- **3 Nov 2023**: Turn API spec over to ODM v2.x team
- **4 Dec 2023**: Turn API spec over to ODM v2.x team
Thank You!
Questions?

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