COSA Dataset-JSON Hackathon II Kickoff

Sam Hume

2023-08-31

cdisc

Agenda

1. Welcome
2. Hackathon objectives
3. Background
4. API specification development
5. Timeline & next steps

Welcome to the COSA Dataset-JSON Hackathon II

~ 100 registered participants

~ 6 weeks in duration

Report outcomes prior to the US Interchange

Collaborative hackathon

Kickstart standards development

Aligned with the Dataset-JSON Pilot

cdisc

Hackathon Objectives

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
- Participants will be added to an authors list in the repo
- Implement prototypes to demonstrate and test the API specification

cdisc

Hackathon Background

What is Dataset-JSON and Advantages

What is JSON?
An open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays

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
- An open-source MIT license
- Schema supporting any tabular format
- Extensible to support integrated metadata and new use cases
- Linked to Define-XML for complete metadata
- Integrated with CORE for conformance checking

Dataset-JSON advantages…

- Based on the JSON standard used worldwide
- Open-source and truly human readable
- Same or smaller file sizes relative to current required format
- Remove variable naming, width, or format limitations
- Simple transformation to/from SAS data

cdisc

COSA Dataset-JSON Hackathon 2022

- Dataset-JSON Hackathon open-source solutions available in the COSA Repository Directory
- Solutions created include:
  - Conversion to and from different dataset formats
  - Dataset browsers / viewers
  - Methods for handling large datasets
  - RESTful Web Services
- Overall Impressions of Dataset-JSON:
  - Works as a general data exchange
  - Works as a general dataset format
  - Works with web-based APIs
  - Works with a wide-range of programming languages and technology stacks
  - Simple to process
  - Easy to transform into SAS datasets, R or Python dataframes, and CSV
  - File sizes smaller than SAS XPORT v5 and Dataset-XML
  - A language, platform independent data exchange format

Language	# Solutions
R	5
SAS	4
Python	5
JavaScript	4
Java	1
Swift	1
XSLT	1

https://cosa.cdisc.org/

Dataset-JSON 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 how FDA toolset can support JSON format and identify tool upgrade roadmap

➔ **Success Criteria: Accept Dataset-JSON as a transport format option (in addition to existing XPT format)**

Milestone 2: Long Term

- Enhance the CDISC SDTM and ADaM standards beyond XPT limitations (e.g. Variable names > 8, labels > 40, data > 200
- New Define-XML / Define-JSON based on ODM v2.0
- Enhanced conformance rules
- Collaborate with FDA to develop plan to retool their environment to natively consume JSON

➔ **Success Criteria: accept advanced Dataset-JSON as the only transport format option and deprecate XPT**

cdisc

Dataset-JSON Conference Activities

2023 PHUSE/FDA CSS

- Hands-on Workshop
- Subteam working sessions
- Plenary presentations

2023 CDISC US Interchange

- Dataset-JSON Pilot plenary presentation

2023 PHUSE EU Connect

- Dataset-JSON Workshop

2024 PHUSE US Connect

- Dataset-JSON Workshop
- Presentation to cover FDA pilot findings and next steps

2024 PHUSE CSS

- Presentation to cover final pilot report

API Specification Development

Example Dataset-JSON User Stories

#	User Story
1	As a sponsor, I want to read all dataset data from my EDC Vendor or CRO for a specified study so that I can create datasets in SAS or R formats
2	As a sponsor, I want to get a listing of all available datasets for a specified study so that I can retrieve each dataset individually
3	As a sponsor, I want to get a listing of all datasets that have changed as of a certain date so that I can retrieve datasets that have updates
4	As a sponsor, I want to retrieve datasets in chunks so that I do not have to transfer all records of a large dataset in one request
5	As a sponsor, I want to retrieve a Define-XML that's associated with the Dataset-JSON datasets
6	As a sponsor, I want to retrieve dataset metadata for a specified dataset so that I can prepare to retrieve the dataset data

Example Dataset-JSON User Stories

#	User Story
7	As a sponsor, I want a listing of all studies for which datasets exists so that I can select a study to request datasets from
8	As an EDC vendor, I want to send datasets for a study to a sponsor so that they can begin to refine them for use in analysis
9	As an EDC vendor, I want to create a dataset repository from the raw data store using the API so that the dataset data is available for the sponsor to retrieve
10	As a sponsor, I want to submit Dataset-JSON datasets to the regulatory authorities so that I can complete a submission

Collaborative Work

- Will setup a GitHub repository to collaboratively develop the API spec
  - OpenAPI 3.x will be used as the standard for the machine-readable spec
  - Project documentation will be maintained in Markdown
  - Project communication will take place in Slack and GitHub
- An OpenAPI Specification schema is available to aid in development
  - https://github.com/OAI/OpenAPI-Specification
- Numerous tools, including many free ones, can be used to edit the API spec
  - https://openapi.tools/
- Separate repos may house prototype API implementations

Example REST API Implementations

- Ideally, the API spec will be implemented by 1 or more prototypes

- A prototype will
  - Aid others in reviewing and commenting on the API
  - Provide a mechanism to demonstrate the API
  - Allow us to test the API specification
  - Possibly evolve to be a reference implementation

DDF-RA / Deliverables / API / **USDM_API.yaml**

Code | Blame | 1428 lines (1428 loc) · 35.1 KB

```yaml
            content:
              application/json:
                schema:
                  $ref: '#/components/schemas/HTTPValidationError'
    '/v3/studyDefinitions/{studyId}':
      put:
        tags:
          - Production
        summary: Update a study
        description: Update an entire study including all child element wi
        operationId: update_study_v3_studyDefinitions__studyId__put
        parameters:
          - name: studyId
            in: path
            required: true
            schema:
              type: string
              title: Studyid
        requestBody:
          required: true
          content:
            application/json:
              schema:
                $ref: '#/components/schemas/Wrapper'
        responses:
          '200':
            description: Successful Response
            content:
              application/json:
                schema:
                  type: string
                  format: uuid
```

RESTful Web Service using Dataset-JSON

- A simple prototype RESTful Web Service for querying submissions from a repository, using Dataset-JSON for the response, has been implemented
- Try it out at: http://xml4pharmaserver.com/WebServices/Submission_Services_Dataset-JSON/

SubmissionService / **Get all VS records for which VSTESTCD=SYSBP and VSORRES<= 100 mmHg**

Save

GET http://localhost:8080/SubmissionService/rest/SingleDataSet/CDISCPILOT01/dataset/VS?variable=VSTESTCD&variablevalue=SYSBP&resultvariable=VSORRES&comparator=le&value=100 ...

Params • Authorization Headers (6) Body Pre-request Script Tests Settings

Query Params

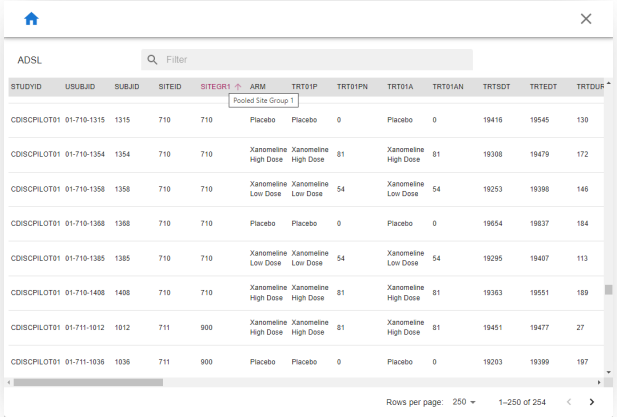
	KEY	VALUE	DESCRIPTION
☑	variable	VSTESTCD	
☑	variablevalue	SYSBP	
☑	resultvariable	VSORRES	
☑	comparator	le	
☑	value	100	

Author: Jozef Aerts

cdisc

stream/serve/view-dataset-json

- **Authors:** Parexel (Juan Abdon, Ivan Osipov, Mauro Bringas, Dmitry Kolosov)
- **Repository:** stream/serve/view-dataset-json
- **Description:** This solution includes 3 subprojects
  - stream-dataset-json - Python library to read Dataset-JSON files as a stream
  - serve-dataset-json - Python library to serve Dataset-JSON files via API
  - view-dataset-json - TypeScript project implementing a viewer for Dataset-JSON files



ADSL

Filter

STUDYID	USUBJID	SUBJID	SITEID	SITEGR1	ARM	TRT01P	TRT01PN	TRT01A	TRT01AN	TRTSDT	TRTEDT	TRTDUR
				Pooled Site Group 1								
CDISCPILOT01	01-710-1315	1315	710	710	Placebo	Placebo	0	Placebo	0	19416	19545	130
CDISCPILOT01	01-710-1354	1354	710	710	Xanomeline High Dose	Xanomeline High Dose	81	Xanomeline High Dose	81	19308	19479	172
CDISCPILOT01	01-710-1358	1358	710	710	Xanomeline Low Dose	Xanomeline Low Dose	54	Xanomeline Low Dose	54	19253	19398	146
CDISCPILOT01	01-710-1368	1368	710	710	Placebo	Placebo	0	Placebo	0	19654	19837	184
CDISCPILOT01	01-710-1385	1385	710	710	Xanomeline Low Dose	Xanomeline Low Dose	54	Xanomeline Low Dose	54	19295	19407	113
CDISCPILOT01	01-710-1408	1408	710	710	Xanomeline High Dose	Xanomeline High Dose	81	Xanomeline High Dose	81	19363	19551	189
CDISCPILOT01	01-711-1012	1012	711	900	Xanomeline High Dose	Xanomeline High Dose	81	Xanomeline High Dose	81	19451	19477	27
CDISCPILOT01	01-711-1036	1036	711	900	Placebo	Placebo	0	Placebo	0	19203	19399	197

Rows per page: 250 1–250 of 254

- **Purpose:** The goal of the project is to write a library which allows to efficiently read Dataset-JSON files (including huge file sizes) and show to how it can be utilized for different purposes.
- **License: MIT**

Timeline and Next Steps

Dataset-JSON Pilot: draft Timeline

31 Aug 2023
Hackathon Kickoff

1 Sep 2023
Hackathon begins

13 Oct 2023
Hackathon ends
Results published for US Interchange

23 Oct 2023
Comment period begins

3 Nov 2023
Comment period ends

4 Dec 2023
Turn API spec over to ODM v2.x team

cdisc

Next Steps

- Meeting date, time, cadence
  - How often do we need to meet to discuss topics vs. hashing them out in slack or GitHub?
- Use existing Dataset-JSON Hackathon Slack workspace
  - Will send out invites to participants not already on it
- Setup GitHub repo
  - cdisc-org repository named DataExchange-DatasetJson-API
  - https://github.com/cdisc-org/DataExchange-DatasetJson-API
- Use previous Dataset-JSON Hackathon wiki space
  - Do we need the Wiki or can we stick to GitHub and Slack?

Open Questions

- Use OAS 3.0 or 3.1?
- Will we ever create a GraphQL API specification?
- Track use cases and requirements using GitHub Issues instead of the wiki?
- Recommended editors for authoring an OAS specification?

cdisc

Additional Resources

Additional Resources

- 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
- ODM v2.0 specification
  - https://wiki.cdisc.org/display/PUB/ODM+v2.0
- COSA Directory Dataset-JSON Hackathon I projects
  - https://cosa.cdisc.org/hackathons/datasetJson
- 2022 Working with Dataset-JSON using SAS (Lex Jansen)
  - https://www.lexjansen.com/cgi-bin/xsl_transform.php?x=pharmasug2022#pharmasug2022.ad150

Additional Resources

- Open API Specification 3.0 Tutorial
  - https://support.smartbear.com/swaggerhub/docs/tutorials/openapi-3-tutorial.html

- CDISC OAS specification examples
  - DDF: https://github.com/cdisc-org/DDF-RA/blob/main/Deliverables/API/USDM_API.yaml
  - ARS: https://github.com/cdisc-org/analysis-results-standard-api/blob/main/openapi/ars.yaml

cdisc

Thank You!

Questions?

shume@cdisc.org

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