Dataset-JSON v1.1

Public Review Webinar

Sam Hume, CDISC
Lex Jansen, CDISC

2024-09-03

cdisc

Agenda

1. Dataset-JSON Timeline
2. What is Dataset-JSON v1.1?
3. Dataset-JSON Pilot
4. What Changed in Dataset-JSON v1.1?
5. Public Review

Dataset-JSON v1.1 Timeline

What is Dataset-JSON v1.1

A high-level overview of Dataset-JSON

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

Dataset-JSON Assumptions

- Data exchange scenarios include:
  - Datasets generated by EDC, ePRO, labs, and other data sources
  - APIs will provide the most common means to exchange data
- Aligns with:
  - ODM v2.0 and works with Define-XML
  - DDF USDM, ARS, CORE, CDISC Library, OAK, and other CDISC projects
  - Healthcare data exchange standards like HL7 FHIR
- Infrequent reads/writes to Dataset-JSON files
- Extensible - most vendors extend ODM-based standards
- Many vendors already import/export JSON

Limitations of SAS V5 XPORT Format

Dataset-JSON seeks to address the limitations of SAS V5 XPORT

Data File Format	Storage	Content	Extensibility
• Limited variable types • Limited to US ASCII encoding • 8-character variable names • 40-character labels • 200-character field widths	• Inefficient use of storage space • The inability to compress datasets leads to file logistical issues (e.g., splitting datasets)	• Lacks a robust metadata layer • Only works for 2-dimensional data structures	• Not extensible

From the PHUSE *Transport for the Next Generation* (2017) White Paper

Prior work influencing Dataset-JSON

Test Report for DS-XML Pilot

Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

April 8, 2015

Transport for the Next Generation

Version 1.0
Created 30 Apr 2017

A White Paper by The PhUSE Alternative Transport Format Working Group - Part of the PhUSE Emerging Trends and Technologies Computational Science Symposium Collaboration.

Dataset-JSON Pilot

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

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

- The pilot was a collaboration between CDISC, PHUSE, and the FDA

- The pilot leads were:
  - CDISC: Sam Hume, CDISC
  - PHUSE: Stuart Malcom, Veramed
  - FDA: Jesse Anderson, FDA

- The pilot kickoff was completed on 27 July 2023
  - The final readout occurred at the PHUSE CSS conference, 3-5 June 2024
  - Dataset-JSON as an Alternative Transport Format for Regulatory Submissions: Final Pilot Report

What were 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**

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

Why Dataset-JSON v1.1?

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

Dataset-JSON v1.1 is a response to the findings from the pilot and other feedback.

What Changed in Dataset-JSON v1.1?

Standards Updates

- Flattening the overall structure
- Adding additional datatypes, such as the decimal datatype
- Adding additional metadata to support target datatype conversions
- Converting integer dates to ISO 8601 dates
- Adding support for NDJSON to aid in processing large datasets
- Making most OIDs optional
- Improving the names
- Versioning Dataset-JSON independently of ODM v2.0

Standards Updates

Flattening the overall structure


```
{
    "datasetJSONCreationDateTime": "2023-03-22T11:53:27",
    "datasetJSONVersion": "1.1.0",
    "fileOID": "www.sponsor.xyz.org.project123.final",
    "dbLastModifiedDateTime": "2023-02-15T10:23:15",
    "originator": "Sponsor XYZ",
    "sourceSystem": {
        "name": "Software ABC",
        "version": "1.0.0"
    },
    "studyOID": "xxx",
    "metaDataVersionOID": "xxx",
    "metaDataRef": "https://metadata.location.org/api.link",
    "itemGroupOID": "IG.DM",
    "isReferenceData": false,
    "records": 100,
    "name": "DM",
    "label": "Demographics",
    "columns": [ ··· ],
    "rows": [ ···]
}
```

Standards Updates: datatype conversions

Adding additional metadata to support target datatype conversions

dataType (logical)	JSON Data Type	targetDataType	Comment
string	string		
integer	integer		
decimal	string	decimal	decimal is exchanged as a string and uses a "." as the decimal separator
float	number		
double	number		
boolean	boolean		
datetime	string		ISO 8601 datetime as a string
date	string		ISO 8601 date as a string
time	string		ISO 8601 time as a string
datetime	string	integer	ISO 8601 datetime as an integer (use case: ADaM)
date	string	integer	ISO 8601 date as an integer (use case: ADaM)
time	string	integer	ISO 8601 time as an integer (use case: ADaM)
URI	string		

Standards Updates: numeric dates

Converting numeric dates to ISO 8601 dates for data exchange

```
"columns": [
  {"itemOID": "ITEMGROUPDATASEQ", "name": "ITEMGROUPDATASEQ",
   "label": "Record Identifier", "dataType": "integer"},

...

  {"itemOID": "IT.ADAE.TRTSDT", "name": "TRTSDT",
   "label": "Date of First Exposure to Treatment", "dataType": "date",
   "targetDataType": "integer", "displayFormat": "E8601DA."},
  {"itemOID": "IT.ADAE.TRTEDT", "name": "TRTEDT",
   "label": "Date of Last Exposure to Treatment", "dataType": "date",
   "targetDataType": "integer", "displayFormat": "E8601DA."},
  {"itemOID": "IT.ADAE.ASTDT", "name": "ASTDT",
   "label": "Analysis Start Date", dataType": "date",
   "targetDataType": "integer", "displayFormat": "E8601DA.", keySequence": 3}]


"rows": [

  [1, "CDISCPILOT01", 701",  "01-701-1015", "2014-01-02", "2014-07-02", "2014-01-03",
...]

  ...

]
```

Standards Updates: NDJSON

Support for NDJSON

```
{"datasetJSONCreationDateTime": "2024-08-01T16:35:31", "datasetJSONVersion": "1.1.0"  ...}
[1, "CDISCPILOT01",  "DM",  "CDISC001",  84, ... ]
[2, "CDISCPILOT01",  "DM",  "CDISC002",  76, ... ]
[3, "CDISCPILOT01",  "DM",  "CDISC003",  61, ... ]
```

- All the metadata is contained in a JSON object in line 1
  - Includes dataset metadata and column definitions
- Each data row is written as an array in a single line of JSON
- NDJSON files use .ndjson as the extension
  - JSON files use .json as the extension

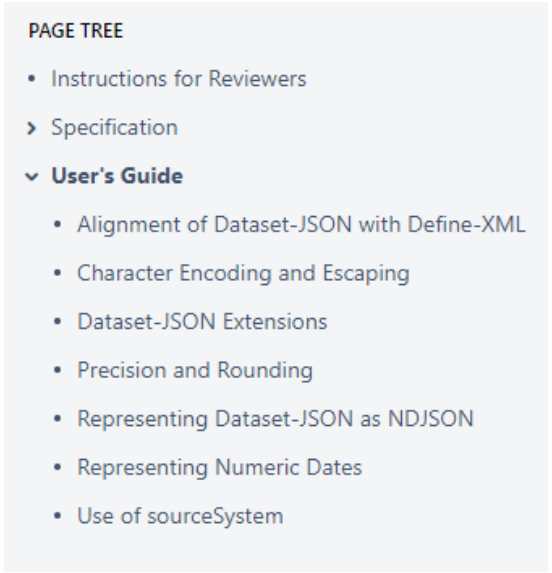
User's Guide (UG) Content

- Addresses the findings that were questions about implementing Dataset-JSON
- UG will be a living document first authored in the wiki
- Dataset-JSON has many more datatypes than SAS XPT
- Dataset-JSON has more metadata than SAS XPT
- Data exchange scenarios not widely understood
- Dataset-JSON removes the SAS XPT restrictions

User’s Guide (UG) Content

Current User's Guide articles on the Wiki:

- Alignment of Dataset-JSON with Define-XML
- Character Encoding and Escaping
- Dataset-JSON Extensions
- Precision and Rounding
- Representing Dataset-JSON as NDJSON
- Representing Numeric Dates
- Use of sourceSystem



Tool Updates and Enhancements

- Updates needed to support Dataset-JSON v1.1
- Additional documentation and test cases
- Possible NDJSON support
- Better large dataset processing capabilities
- Support for Parquet conversions
- Dataset-JSON Viewer software
  - Hackathon tentatively scheduled for October 1st - November 12th (COSA Webinar Sept. 24th)
- Generate usage metrics
- Open-source software used for the pilot
  (SAS tool used in the pilot has been updated for Dataset-JSON v1.1:
  https://github.com/lexjansen/dataset-json-sas)

Features to complete

Feature	Due Date
• Review and decide on the solutions to the pilot finding	✓
• Update the standards specification	✓
• Author a User's Guide	✓
• Update the schema	✓
• Update the examples	✓
• Develop an NDJSON version based on the draft	✓
• Update NDJSON examples	✓
• Internal Review	✓
• **Public Review**	10 September 2024
• JSONX development - compressed archive of datasets	TBD
• Review draft API specification	TBD
• Work with conversion software developers to update tools	TBD
• Work with view software developers to develop/update tools	TBD

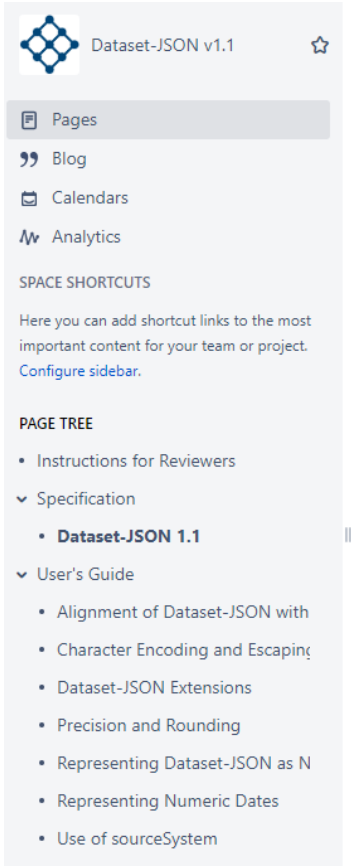
Public Review

Public Review

- Public Review starts on September 10, 2024
- Public Review lasts 30 days and closes on October 10, 2024
- Reviewers are requested to provide comments via JIRA
- The JIRA project is Dataset-JSON review comments (DSJSONCT), located at: https://jira.cdisc.org/projects/DSJSONCT)
- Instructions for Reviewers: https://wiki.cdisc.org/display/DSJSON1DOT1/Instructions+for+Reviewers

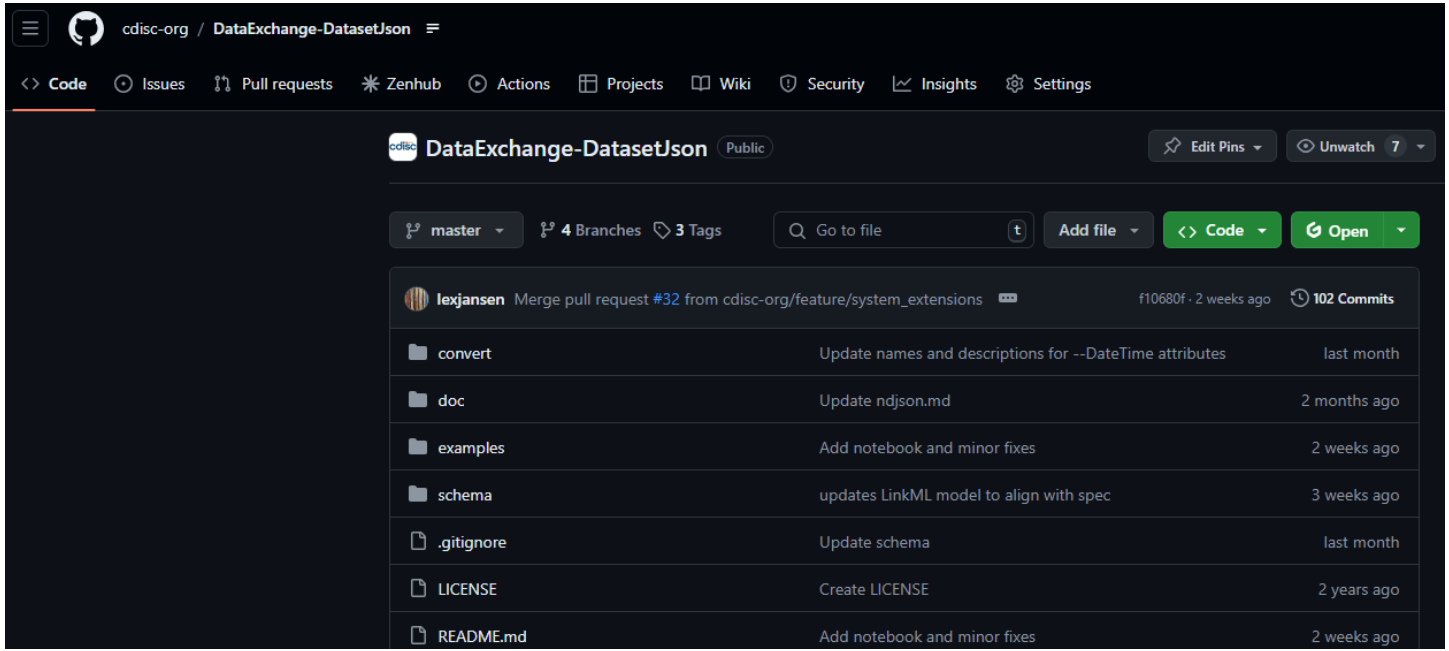
CDISC Wiki

Contains the Dataset-JSON v1.1 Specification and a User's Guide with articles



GitHub: JSON Schema and Examples

- The JSON schema, both for JSON and NDJSON representations, and examples can be found at the GitHub repository for the Dataset-JSON Version 1.1 standard: https://github.com/cdisc-org/DataExchange-DatasetJson



Thank You!

Questions?

shume@cdisc.org

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

ljansen@cdisc.org

https://www.linkedin.com/in/lexjansen/