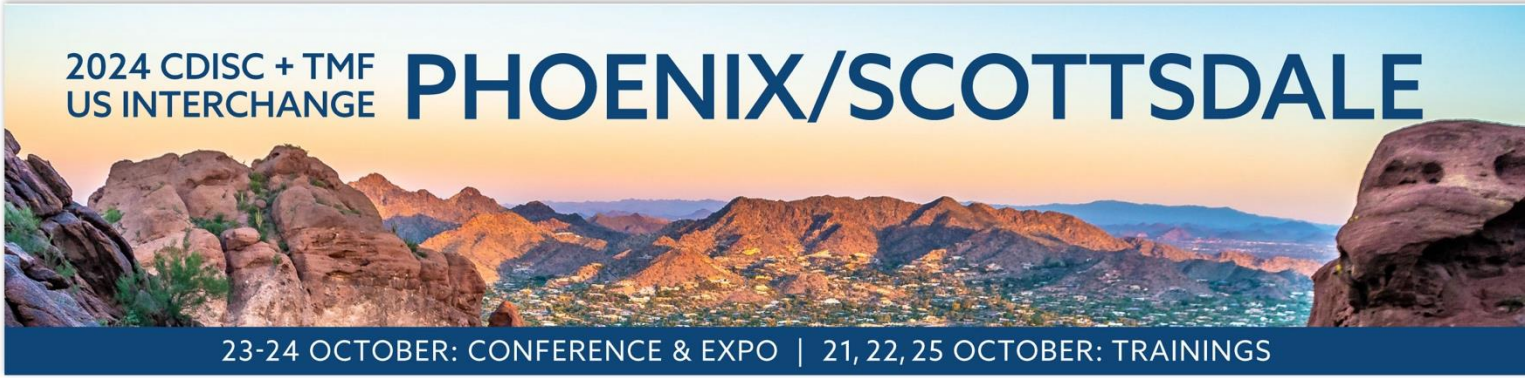



# Dataset-JSON Update

Sam Hume, D.Sc., CDISC
Session 7A: Data Science
October 24, 2024

# Meet the Speaker

## Sam Hume

Title: Principal Consultant

Organization: CDISC

Sam Hume co-leads the CDISC Data Exchange Standards team, advises CDISC leadership on strategy and technical topics, and contributes to COSA, CORE, CDISC Library, and other CDISC projects. Sam formerly served as the CDISC VP of Data Science. During his 30 years in the biopharmaceutical industry, he has held several senior-level technology positions. Sam is an active PHUSE contributor. He holds a doctorate in Information Systems.

# Introducing Dataset-JSON v1.1

This is a quick introduction to Dataset-JSON assuming that most of you have at least a basic understanding of what the standard is and why we're working on it.

# 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

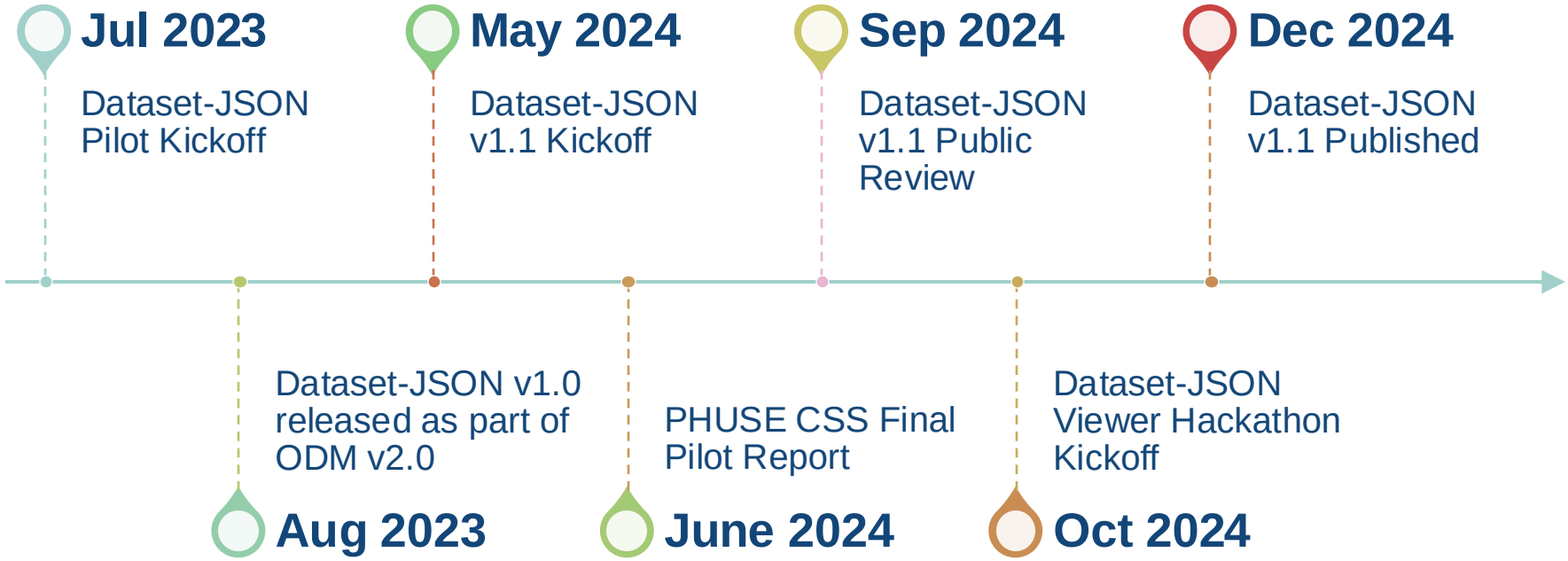
Dataset-JSON is…

- 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
- 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, Define-XML, DDF USDM, ARS, CORE, CDISC Library, OAK, etc.
  - Healthcare data exchange standards like HL7 FHIR
- Addresses SAS V5 XPORT limitations
- Infrequent reads/writes to Dataset-JSON files
- Extensible - most vendors extend ODM-based standards

# Dataset-JSON v1.1 Timeline



# Dataset-JSON v1.1 Training


## Dataset-JSON v1.1 Hands-On Implementation Training

- First Dataset-JSON training class
- Offered at US Interchange in October
- Covers background on SAS XPT, JSON, and the Pilot

## A 4-hour class featuring hands-on exercises

- Hands-on exercises using SAS and Python

# Public Review

Public Review started on Sept. 10, 2024

Public Review closed on Oct. 10, 2024

No major changes expected from the Public Review comments

On track for a Dec. 2024 final release of Dataset-JSON v1.1

# 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 will occur at the PHUSE CSS conference, 3-5 June 2024
  - Dataset-JSON as an Alternative Transport Format for Regulatory Submissions: Final Pilot Report

# What was the primary goal 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**

# The Pilot Results and Dataset-JSON v1.1

The Pilot achieved its primary objective and demonstrated that Dataset-JSON can transport information with no disruption to business

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

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

Dataset-JSON v1.1 addresses the pilot findings

# What Changed in Dataset-JSON v1.1?

What’s new and different in Dataset-JSON v1.1

# Changes in Dataset-JSON v1.1

1. Standards Updates
2. User Guide
3. Software Updates and Enhancements

# 1. Standards Updates: Flatter Structure & Naming

Dataset-JSON v1.1 has a flatter, simpler 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": [ ···]
}
```

# 1. Standards Updates: Flatter Structure & Naming

Dataset-JSON v1.0 structure for comparison


```
{
    "creationDateTime": "2023-06-28T15:38:42",
    "datasetJSONVersion": "1.0.0",
    "fileOID": "www.cdisc.org/StudyMSGv2/1/Define-XML_2.1.0/2023-06-28/cm",
    "asOfDateTime": "2023-05-31T00:00:00",
    "originator": "CDISC SDTM MSG Team",
    "sourceSystem": "Sponsor System",
    "sourceSystemVersion": "1.0",
    "clinicalData": {
        "studyOID": "cdisc.com/CDISCPILOT01",
        "metaDataVersionOID": "MDV.MSGv2.0.SDTMIG.3.3.SDTM.1.7",
        "metaDataRef": "https://metadata.location.org/CDISCPILOT01/define.xml",
        "itemGroupData": {
            "IG.CM": {
                "records": 68,
                "name": "CM",
                "label": "Concomitant Medications",
                "items": [
                    {"OID": "ITEMGROUPDATASEQ", "name": "ITEMGROUPDATASEQ", "label": "Rec
                    {"OID": "IT.CM.STUDYID", "name": "STUDYID", "label": "Study Identifie
                "itemData": [
                    [1, "CDISCPILOT01", "CM", "CDISC001", 1, "ASPIRIN", "PROPHYLAXIS OR N
                    [2, "CDISCPILOT01", "CM", "CDISC001", 2, "METAMUCIL", "PROPHYLAXIS OR
                ]
            }
        }
    }
}
```

# 1. Standards Updates: Data Types

## 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 | | |

# 1. Standards Updates: Date/Time Variables

Timing variables (datetime, date, time) are stored as ISO 8601 strings in JSON.

The *targetDataType* attribute needs to be specified when different from *dataType* attribute or the JSON data type.

For ADaM datasets, the *targetDataType* must be set to integer for date and datetime variables.

User Guide article on representing numeric dates.

# 1. Standards Updates: NDJSON

## NDJSON = Newline Delimited JSON

- NDJSON uses newlines to separate and define the structure of JSON data
- Each line in an NDJSON file is a valid JSON value, such as an object or array

## NDJSON is a variation of JSON that's designed for bulk data transfer

- Can be used to store structured data like datasets

## Dataset-JSON NDJSON and JSON datasets have the same content

- NDJSON content is written or read as 1 line of valid JSON at a time.
- The NDJSON and JSON formats are both part of the Dataset-JSON standard.

# 1. Standards Updates: NDJSON

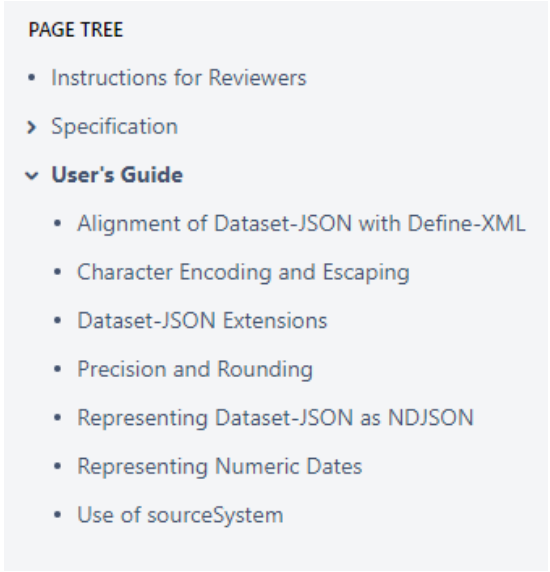
```
{"datasetJSONCreationDateTime": "2024-08-01T16:35:31", "datasetJSONVersion": "1.1.0"  ...}
[1, "CDISCPILOT01",  "DM",  "CDISC001",  84, ··· ]
[2, "CDISCPILOT01",  "DM",  "CDISC002",  76, ··· ]
[3, "CDISCPILOT01",  "DM",  "CDISC003",  61, ··· ]
```

- 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

# 2. User Guide (UG) Content

Current User 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




**UG will be a living document first authored in the wiki**

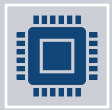
# 3. Tool Updates and Enhancements

## 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



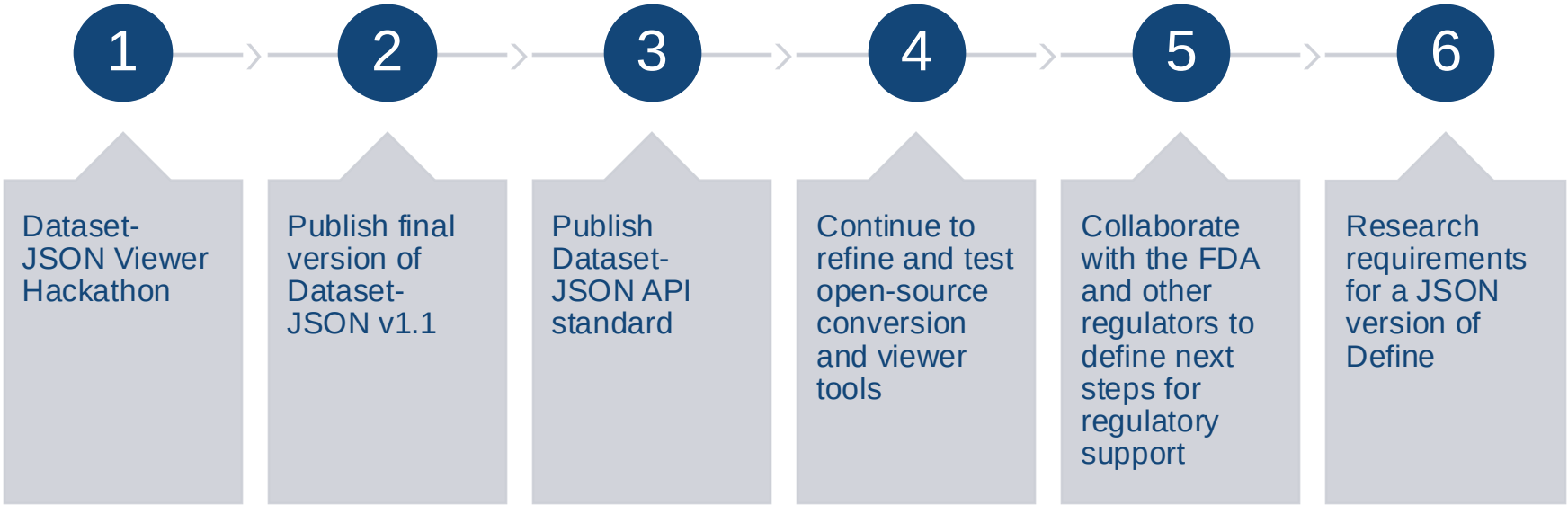
## Python

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

Dataset-JSON Viewer Hackathon In Progress

# Next Steps

# Next Steps




Much of this activity is happening concurrently

# COSA Dataset-JSON Viewer Hackathon

- Primary objective: Create Dataset-JSON Viewer software
  - Projects can be new or extensions to existing software
  - Projects must support Dataset-JSON v1.1
- Significant and broad interest in viewer software for Dataset-JSON
  - A finding from the *Dataset-JSON as an Alternative Transport Format for Regulatory Submissions Pilot*
- Virtual hackathon
  - Individuals or teams will work virtually
  - Project demonstrations will be scheduled for a COSA Spotlight Webinar
- Dates: Oct. 8th – Jan. 6th

# Dataset-JSON REST API Specification

## API Scope & Assumptions

- Focused on data exchange
- Supports the basics
- Uses compression and pagination to support larger datasets
- Supports a broad range of data exchange scenarios

# Thank You!

Questions?

Sam Hume, DSc

shume@cdisc.org

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