

AD09 PHUSE EU CONNECT 2025

**No Patients? No Problem!
Define your Study Data with
Python at the CORE**



Agenda

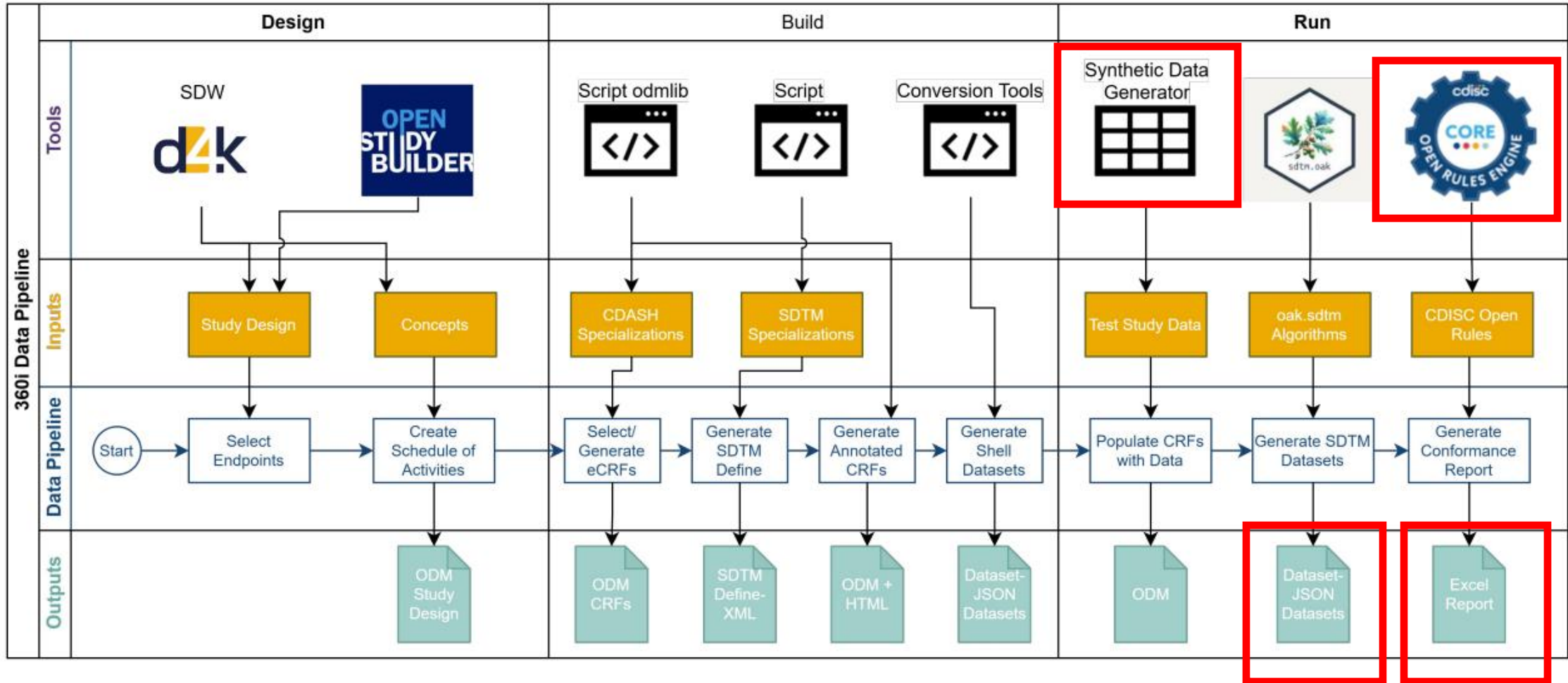
- Objectives
- Synthetic Data
- Define XML
- Dataset JSON
- CDISC CORE
- Summary
- Q&A

Objectives

- Pull together recent industry initiatives
- High quality synthetic data
- Interoperability / Future proofing



Objectives



Synthetic Data – Why?

- Control timelines / resource
- Quality of EDC test data
- Training, Tool Development
- Customisable

Synthetic Data Challenges

- Easy to create random data in single columns
- Harder to create compliant data across a full row
- Harder still to create compliant data across all domains

INFORMED
CONSENT

AE
OUTCOMES

EXPOSURE
DATES

STUDY
ARMS

STUDY
VISITS

MEDDRA

REFERENCE
RANGES

Define.xml

- Metadata rich!
- Available at start of study
- Yes really!!!

Standard	SDTM-IG 3.4
Study Name	User_John_McDade
Study Description	Training Study XYZ
Protocol Name	TEST_STUDY
Metadata Name	Data Definitions for User_John_McDade, SDTM-IG 3.4
Metadata Description	Data Definitions for User_John_McDade, SDTM-IG 3.4

Datasets					
Dataset	Description	Class	Structure	Purpose	Keys
TA	Trial Arms	TRIAL DESIGN	One record per planned element per arm	Tabulation	STUDYID, ARMCD, TAETORD
TE	Trial Elements	TRIAL DESIGN	One record per planned element	Tabulation	STUDYID, ETCD
TI	Trial Inclusion/Exclusion Criteria	TRIAL DESIGN	One record per I/E criterion	Tabulation	STUDYID, ITESTCD
TS	Trial Summary	TRIAL DESIGN	One record per trial summary parameter value	Tabulation	STUDYID, TSPARMCD, TSSEQ
TV	Trial Visits	TRIAL DESIGN	One record per planned visit per arm	Tabulation	STUDYID, ARMCD, VISITNUM
CO	Comments	SPECIAL PURPOSE	One record per comment per subject	Tabulation	STUDYID, USUBJID, RDOMAIN, IDVAR, IDVARVAL, COREF, COV
DM	Demographics	SPECIAL PURPOSE	One record per subject	Tabulation	STUDYID, USUBJID
SE	Subject Elements	SPECIAL PURPOSE	One record per actual element per subject	Tabulation	STUDYID, USUBJID, SESTDTC
SV	Subject Visits	SPECIAL PURPOSE	One record per actual visit per subject	Tabulation	STUDYID, USUBJID, SVSTDTC, VISITNUM

VS (Vital Signs) - FINDINGS						
Related Supplemental Qualifiers Dataset: SUPPVVS (Supplemental Qualifiers for VS)						
Variable	Where Condition	Label / Description	Type	Length or Display Format	Controlled Terms or ISO Format	Origin / Source / Method / Comment
STUDYID		Study Identifier	text	11		Protocol
DOMAIN		Domain Abbreviation	text	2	SDTM Domain Abbreviation • "VS" = "Vital Signs Domain"	Assigned
USUBJID		Unique Subject Identifier	text	20		Derived STUDYID concatenated with '/' concatenated with
VSSEQ		Sequence Number	integer	3		Derived Sequential number identifying records within each
VSTESTCD		Vital Signs Test Short Name	text	6	Vital Signs Test Code [8 Terms]	Assigned
VSTEST		Vital Signs Test Name	text	24	Vital Signs Test Name [8 Terms]	RF Annotated Case Report Form [51 97 97]
VSCAT		Category for Vital Signs	text	10	Category for Vital Signs • "PULSE AND BLOOD PRESSURE" • "BODY MEASUREMENT" • "RESPIRATORY" • "BODY TEMPERATURE"	Assigned
VSPOS		Vital Signs Position of Subject	text	20	Position • "SITTING" = "Sitting" • "SEMI-RECUMBENT" = "Semi-recumbent" • "SUPINE" = "Supine" • "STANDING" = "Standing"	RF Annotated Case Report Form [51 97 97]
VSORRES VLM		Result or Finding in Original Units	text	5		
VSTESTCD = "DIABP" (Diastolic Blood Pressure)		Diastolic Blood Pressure	integer	3		CRF Annotated Case Report Form [51 97 97]
VSTESTCD = "HEIGHT" (Height)		Height	float	4		CRF Annotated Case Report Form [51 97 97]
VSTESTCD = "PULSE" (Pulse Rate)		Pulse Rate	integer	3		CRF Annotated Case Report Form [51 97 97]
VSTESTCD = "SYSBP" (Systolic Blood Pressure)		Systolic Blood Pressure	integer	3		CRF Annotated Case Report Form [51 97 97]

Define.xml



Output Structure

The parser returns a comprehensive dictionary:

```
{
  "metadata": {
    "study": {...},
    "standard": {"name": "SDTM-IG", "version": "3.4"},
    "domains": {
      "VS": {
        "name": "VS",
        "label": "Vital Signs",
        "variables": [...],
        "repeating": "Yes"
      }
    },
    "codelists": {...}
  },
  "datasets": {...},
  "key_sequences": {"VS": ["STUDYID", "USUBJID", ...]},
  "attributes": {...}
}
```

Define.xml

Metadata Extraction Details

A. Study-Level Information

XML Path: `ODM/Study/GlobalVariables`

Extracts:

- Study Name
- Study Description
- Protocol Name
- Standard Name (SDTM-IG)
- Standard Version (e.g., 3.4)

```
namespaces = {'odm': 'http://www.cdisc.org/ns/odm/v1.3'}
study_name = root.find('..//odm:GlobalVariables/odm:StudyName', namespaces)
```

C. Variable-Level Details

XML Path: `ItemDef` + `ItemRef` elements

Per variable:

- Name, Label, Data Type
- Length, Significant Digits
- **Mandatory flag** (Yes/No)
- **Key Sequence** (for sorting keys)
- **Codelist Reference** (OID linkage)
- **Origin** (Collected, Derived, Assigned)

Smart Key Handling:

```
# Extracts and sorts key variables by sequence
key_sequences[dataset_name].sort(key=lambda x: x[1])
key_sequences[dataset_name] = [x[0] for x in key_sequences[dataset_name]]
```

B. Dataset (Domain) Metadata

XML Path: `ItemGroupDef` elements

For each domain, captures:

- **Basic:** Name, Label, OID
- **Structure:** Repeating, IsReferenceData, Purpose
- **Variables:** Complete variable list with attributes

Key Technical Detail: Uses lookup dictionary for performance

```
item_lookup = {item.get('OID'): item for item in items}
```

D. Codelist Parsing

XML Path: `Codelist` elements

Handles **two formats**:

1. `EnumeratedItem` (simple lists)
2. `CodeListItem` (with external dictionary refs)

Each codelist contains:

- **OID** (unique identifier)
- **Name** (human-readable)
- **DataType** (text, integer, float)
- **Items:** Array of {value, decode} pairs

```
codelist_dict[codelist_oid] = {
    'name': codelist_name,
    'dataType': codelist_data_type,
    'items': [{'value': 'Y', 'decode': 'Yes'}, ...]
}
```

Define.xml

Projects > DEMO STUDY 0002 > Synthetic Data

Synthetic Data Generation

Generate CDISC-compliant synthetic SDTM data for testing and validation

CDISC Dataset-JSON v1.1

Define-XML Files

1 file(s) uploaded

define.xml
63 domains • Invalid Date



Upload Another File

Configurations

1 configuration(s) saved

PHUSE DEMO
10 subjects • 3 domains

Generate

Create New Configuration

Generation Complete

Restored from last session

Synthetic data has been generated successfully

Generation ID: gen-20251007-214401

3

Domains

1186

Total Records

2

Formats

Domains Generated:

DM

10 records
Click to preview data

AE

56 records
Click to preview data

VS

1120 records
Click to preview data

Validate with CDISC CORE

Download Data

Dismiss



Synthetic Data Dependencies

- DM is the seed for all subject level domains
- Trial needs to be anchored around an arbitrary date
- Combine with TV to map out subject visits
- Combine with TA for trial arms –ratio controlled
- RACE, ETHNIC, CONTRY, SEX – from codelists
- AGE – from slider
- DTHFL=Y in < 10% of subjects - tweak

🔗 Detailed Date Calculation			
Variable	Formula	Example Result	Purpose
RFICDTC	<code>2023-08-01 + random(-30, 30)</code>	2023-07-15	Informed Consent
RFSTDTC	<code>RFICDTC + 30 + random(-5, 5)</code>	2023-08-14	Study Start
RFXSTDTC	<code>= RFSTDTC</code>	2023-08-14	Treatment Start (ANCHOR)
RFXENDTC	<code>= RFENDTC</code>	2024-06-28	Treatment End
RFENDTC	<code>RFICDTC + 350 + random(-10, 10)</code>	2024-06-28	Study End
RFPENDTC	<code>RFICDTC + 370 + random(-10, 10)</code>	2024-07-18	End of Participation

Synthetic Data - Findings Domains

- Need a master list of test information per domain
- Labs dictionary be used for this purpose but adhoc elsewhere
- Ranges used to give realistic values, outlier % controlled
- Combine with DM + TV – assign results
- Assumed tests collected in SI units
- EPOCH / baseline flags assigned

VSTESTCD	VSTEST	VSORRESU	VSCAT	Lower	Upper	DCP
SYSBP	Systolic Blood Pressure	mmHg	PULSE AND BLOOD PRESSURE	90	120	0
DIABP	Diastolic Blood Pressure	mmHg	PULSE AND BLOOD PRESSURE	60	80	0
HR	Heart Rate	beats/min	PULSE AND BLOOD PRESSURE	60	100	0
RESP	Respiratory Rate	breaths/min	RESPIRATORY	12	20	0
TEMP	Temperature	C	BODY TEMPERATURE	36.5	37.5	1
HEIGHT	Height	cm	BODY MEASUREMENT	170	185	0
WEIGHT	Weight	kg	BODY MEASUREMENT	60	90	1
BMI	Body Mass Index	kg/m2	BODY MEASUREMENT	18.5	24.9	1

Synthetic Data - Adverse Events

- Similar to findings, need to assign MedDRA variables as a group
- Subjects given random no. of AEs between 0 and X - controlled
- If DM.DTHFL=Y then AESDTH=Y, AEOUT=FATAL, AESER=Y, dates must match across AE / DM
- If AESER=Y at least one of AESDTH, AESLIFE, AESDISAB, AESHOSP, AESCONG, AESMIE = Y
- If AEACN=DRUG WITHDRAWN impact on exposure



Dataset JSON - Resources

- CDISC published Data Exchange Standard JSON v1.1 in Dec 2024
- CDISC Wiki User Guides: <https://wiki.cdisc.org/display/PUB/Dataset-JSON+v1.1+User%27s+Guide>
- CDISC Dataset JSON Github: <https://github.com/cdisc-org/DataExchange-DatasetJson>

Dataset-JSON v1.1 User's Guide

Created by Omar Garcia Calderon, last modified by Alana St. Clair on Dec 05, 2024

- 1 ALIGNMENT OF DATASET-JSON WITH DEFINE-XML
- 2 CHARACTER ENCODING AND ESCAPING
- 3 DATASET-JSON EXTENSIONS
- 4 PRECISION AND ROUNDING
- 5 REPRESENTING DATASET-JSON AS NDJSON
- 6 REPRESENTING NUMERIC DATES
- 7 USE OF SOURCE SYSTEM

README License

CDISC Dataset-JSON repository

This repository contains the schema, specification, and examples of the Dataset-JSON standard for Clinical Datasets.

The Dataset-JSON v1.1 release can be downloaded as a [zip](#) file.

Dataset-JSON Specification

The [doc](#) folder contains the specifications of:

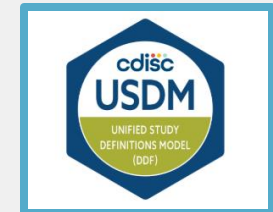
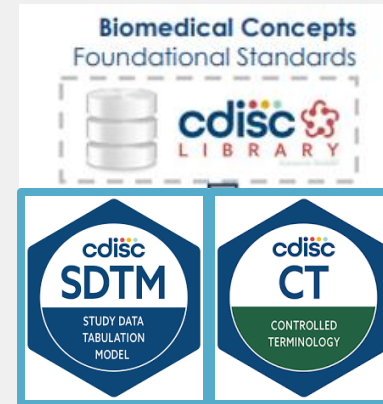
- [Dataset-JSON v1.1](#).
- The [NDJSON](#) representation of Dataset-JSON v1.1

```
{
  "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": [ ... ]
}
```

Dataset JSON - Purpose



- JSON is the de facto data exchange standard for RESTful APIs
- Lightweight / human readable - HTTP
- Language independence
- Automation / validation
- Avoid encoding limitations

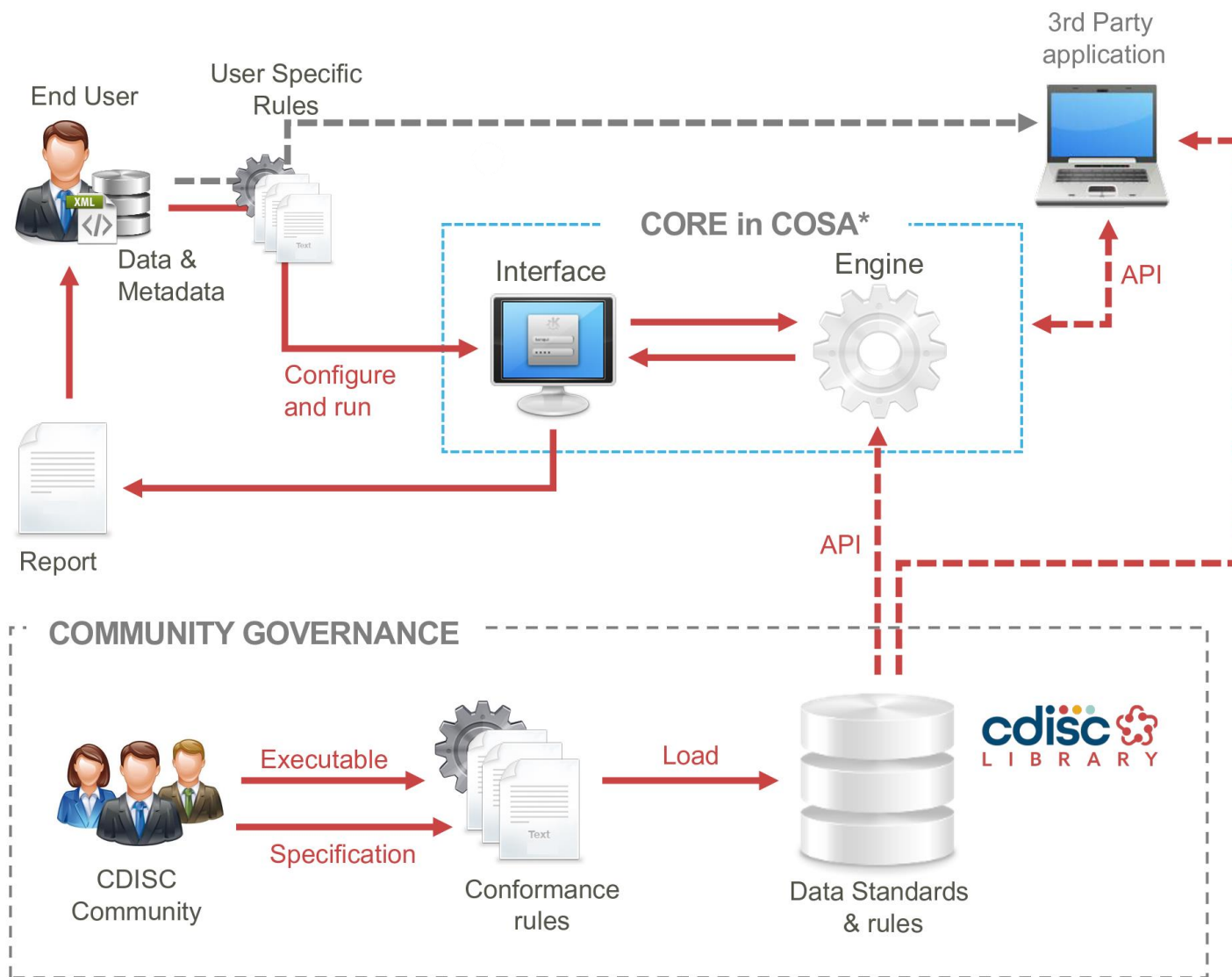


- NDJSON – Newline delimited JSON
- Helps with streaming of larger datasets

```
{
  "datasetJSONCreationDateTime": "2024-08-01T16:35:31",
  "datasetJSONVersion": "1.1.0"
}
[1, "CDISCPIL0T01", "DM", "CDISC001", 84, ... ]
[2, "CDISCPIL0T01", "DM", "CDISC002", 76, ... ]
[3, "CDISCPIL0T01", "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

CDISC CORE



* CDISC Open-Source Alliance

CDISC CORE Validation

CDISC CORE Validation

Validate SDTM datasets against CDISC Conformance Rules for DEMO STUDY 0002 (1 or more datasets, XPT or JSON format)

Validation Configuration

Select the CDISC standard, version, and controlled terminology package

Standard

SDTM ▾

Configuration

SDTM-IG 3.4 (FDA) ▾

SDTM CT

2025-03-28 ▾

Source Data

Upload SDTM datasets in XPT or JSON format (1 or more files)



Drop your files here to select them

Select XPT or JSON files (ae.xpt, cm.json, dm.xpt, etc.)

or

 Browse...

Define.xml (Optional)

Upload Define-XML to validate metadata and cross-check with datasets

 Browse...

Optional

Settings

QC trackers

SAS Papers

Programming Other

PhUSE

Phastar groups

GoogleSheet raw2s...

Sharepoint QC Trac...

All Bookmarks

SM

SDTM Mapper

Dashboard

Projects

SDTM Standards

Administration

User Management

Client Management

Project Management

Role Management

Standards Management

🔍

Search projects, standards...

Projects > DEMO STUDY 0002

📁 Training Area

draft

DEMO STUDY 0002

For different ALS

📄

Project Information

Client:

Status:

Created:

Last Updated:

Training Area

draft

18/08/2025

02/10/2025

📍

Mapping Progress

Current progress on SDTM mapping

Total Entries

Mapped Entries

Completion

0

0

0%

Go to Mapping

→

👤

Team Members

Project team and collaborators

No team members assigned

🔍

Show Training Config

🔍

Show Protocol Config

📖

Mapping

View and edit SDTM mappings

📄

Specifications

SDTM specifications and metadata

📄

Define.xml

Data definition metadata

📊

Reviewer's Guide

Study data reviewer guide

📄

aCRF

Annotated case report forms

⚙️

SDTM Programs

Data transformation programs

📄

Synthetic Data

Generate test data

⚙️

Validation Reports

Pinnacle Z1, CDISC CORE

Configuration

SDTM standards and protocol configuration

SDTM IG Version

CT Version

3.4

2025-03-28

admin@example.com

admin@example.com

MAPIT v1.0.0

© 2025 Phastar. All rights reserved.

SDTM Mapper

Dashboard

Projects

SDTM Standards

Administration

User Management

Client Management

Project Management

Role Management

Standards Management

Search projects, standards...

Browse...Optional

Issues by Domain

Click a domain card to view detailed issues

AE

44

44 Errors

Rules Triggered: 3

Records Affected: 43

DM

36

36 Errors

Rules Triggered: 3

Records Affected: 18

VS

No Issues Found

New Validation

Export to Excel

Validation Rules Fired

Select domains and rules to show in Issue Summary and Issue Details

SDTM Domains

AE 44

DM 36

Validation Rules

CORE-000207 1

CORE-000726 12

CORE-000767 42

CORE-000841 1

CORE-000844 3

CT_PHUSE-001 21

Select All

Clear All

Select All

Clear All

Conformance

Datasets

Issue Summary

Issue Details

Rules Report

Conformance Details

Validation configuration and summary statistics

Standard

SDTMIG

Version

3-4

CT Packages

sdtmct-2025-03-28

Elapsed Time

Total Datasets

3

Total Records

2,302

Rules Checked

140

Total Issues

admin@example.com

admin@example.com

MAPIT v1.0.0

© 2025 Phastar. All rights reserved.

Summary

- High quality synthetic data multiple uses
 - Control timelines
 - Tool development
- Dataset JSON is a powerful file format
 - Interoperability
 - Easy visualisation
- CDISC CORE is a very flexible validation engine for any workflow
 - Various deployment options
 - Easy to enhance reporting – custom rules / filtering / RG automation
 - Real-time validation
- CDISC 360i enablement

Acknowledgements

Sincere thanks to Rajwanur Rahman, introduced me to everything AI and web app development.

I would also like to thank the Dataset-JSON and CDISC CORE working groups, the opportunity to implement this project is based on their hard work.

Thank You

John.mcdade@phastar.com

