Explore in CDISC ARS by a Linked Data lover

Ippei Akiya

ICON Clinical Research GK, Biometrics

Disclaimer

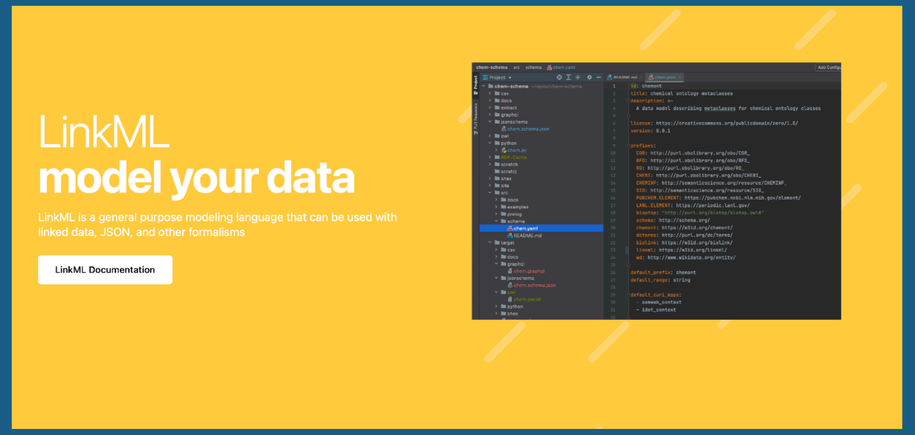
- The content of this presentation includes my personal research, reflections, and opinions, and has no connection whatsoever with the intellectual property rights or business activities of the company I belong to or its clients.

- Furthermore, it does not represent the views or opinions of any academic societies or organizations I am affiliated with.

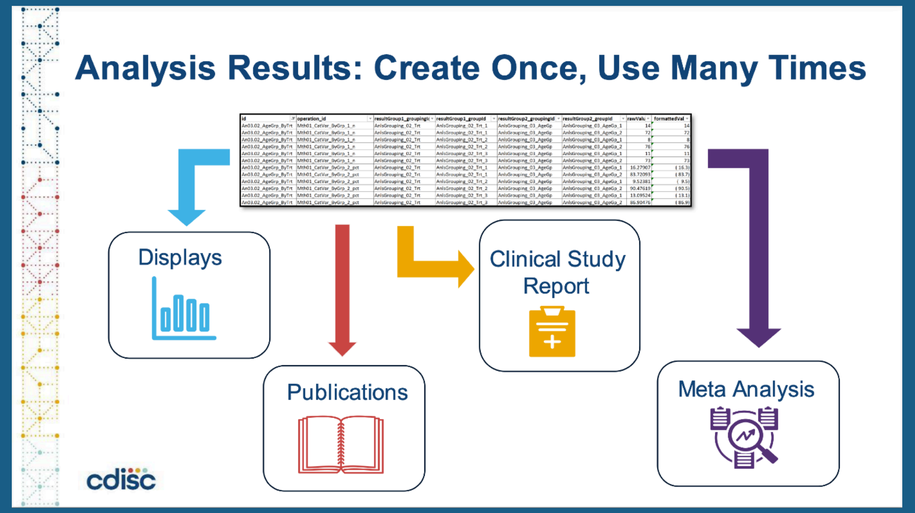
- The copyrights and intellectual property rights for the software and other products introduced in this presentation belong to their respective owners.

Introduction

- This is an explanation of CDISC Analysis Results Standard v1.0 (ARS) from the perspective of a Linked Data enthusiast.
- The reason for explaining it from a Linked Data enthusiast's perspective is that ARS uses LinkML, a Linked Data Modeling Language.
- A highly useful guidebook for understanding ARS is Getting Started with the New CDISC Analysis Results Standard by Bess LeRoy and Richard Marshall.



https://linkml.io/



https://www.cdisc.org/sites/default/files/2024-04/CDISC EU 2024 ARS 20240419.pdf

Overview of CDISC ARS

Analysis Results Key Objectives

Leverage analysis results metadata to drive the automation of results

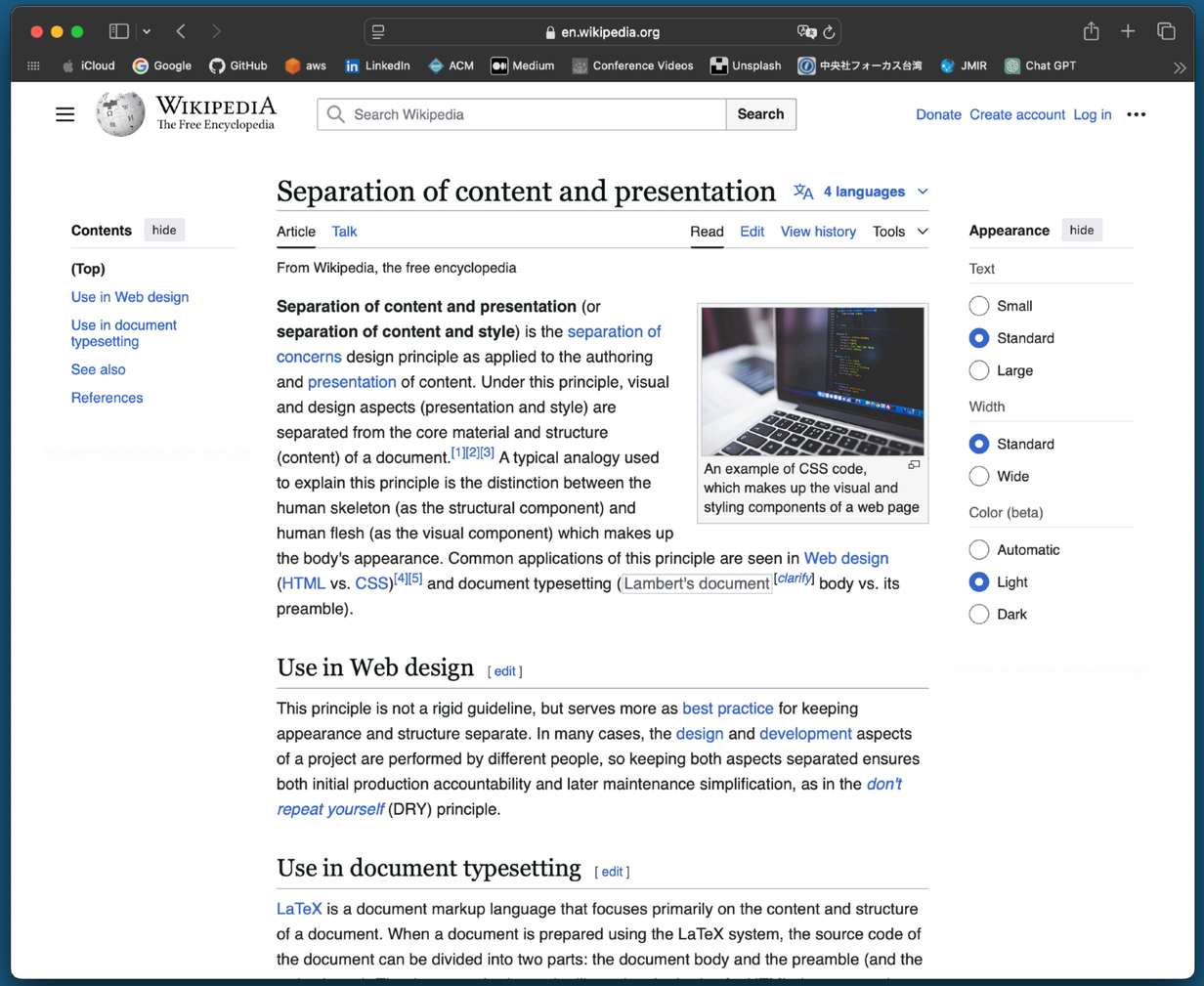
Support storage, access, processing, traceability and reproducibility of results

Getting Started with the New CDISC Analysis Results Standard
Bess LeRoy and Richard Marshall

https://www.cdisc.org/sites/default/files/2024-04/CDISC EU 2024 ARS 20240419.pdf

Why CDISC ARS is developed?

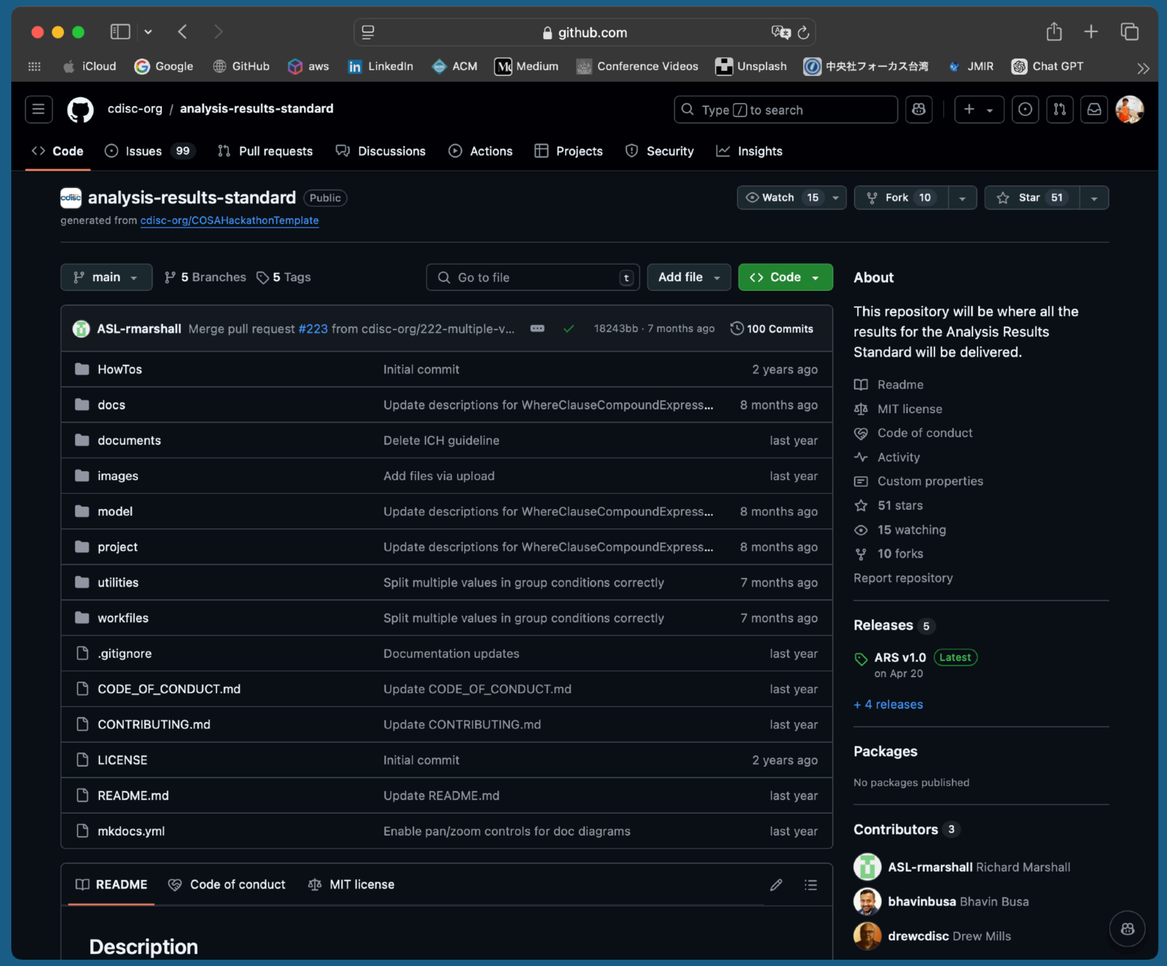
- Currently, the analysis results (TFLs) of clinical trials are output as files, such as RTF or PDF, through analysis programs.
- These files do not separate the results data or content from styles such as fonts and borders, making it difficult to extract the results data as literal values.
- As a result, it becomes challenging to reuse the result data.
- However, Define-XML serves as an example where content and style are completely separated.
  - define.xml <- Contents
  - define.xslt <- Style



https://en.wikipedia.org/wiki/Separation_of_content_and_presentation

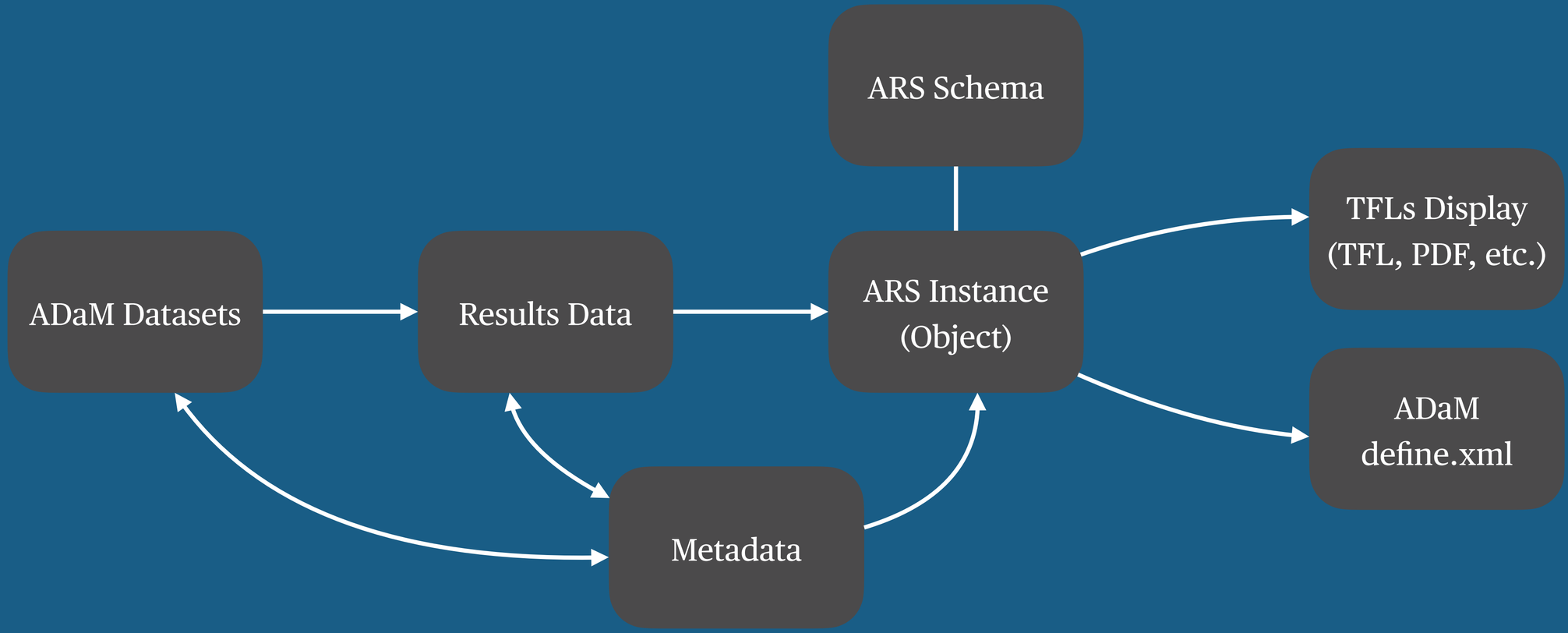
Why CDISC ARS is developed?

- The following are my assumptions.
  - It becomes easier to extract data or results data for conducting meta-analysis, enabling the accurate selection of target trials.
  - Utilization by healthcare authorities during approval reviews.
  - As an advanced topic, converting not only results data but also SDTM and ADaM data into Linked Data would make it easier to confirm data traceability. This would enhance the transparency of statistical analyses.



https://github.com/cdisc-org/analysis-results-standard

Workflow


ARS Schema
ADaM Datasets
Results Data
ARS Instance
(Object)
TFLs Display
(TFL, PDF, etc.)
ADaM
define.xml
Metadata

Mapping Results Data from ARS Instance to TFLs

```
- name: Summary of Age by Treatment
  label: Age
 - operationId: Mth02_ContVar_Summ_ByGrp_2_Mean
   resultGroups:
   - groupingId: AnlsGrouping_01_Trt
     groupId: AnlsGrouping_01_Trt_1
   rawValue: '75.2093023'
   formattedValue: '75.2'
 - operationId: Mth02_ContVar_Summ_ByGrp_2_Mean
   resultGroups:
   - groupingId: AnlsGrouping_01_Trt
     groupId: AnlsGrouping_01_Trt_2
   rawValue: '75.6666667'
   formattedValue: '75.7'
 - operationId: Mth02_ContVar_Summ_ByGrp_2_Mean
   resultGroups:
   - groupingId: AnlsGrouping_01_Trt
     groupId: AnlsGrouping_01_Trt_3
   rawValue: '74.3809524'
   formattedValue: '74.4'
 - operationId: Mth02_ContVar_Summ_ByGrp_3_SD
   resultGroups:
   - groupingId: AnlsGrouping_01_Trt
     groupId: AnlsGrouping_01_Trt_1
   rawValue: '8.5901671'
   formattedValue: ( 8.59)
```

Study – CDISC 360 Page x of y

Table 14.1.1
Summary of Demographics
Safety Population

Characteristics	Placebo (N=XX)	Xanomeline Low Dose (N=XX)	Xanomeline High Dose (N=XX)	p-value [1]
Age (years)				
n	XX	XX	XX	X.XXXX
Mean (SD)	XX.X (XX.XX)	XX.X (XX.XX)	XX.X (XX.XX)	
Median	XX.X	XX.X	XX.X	
Q1, Q3	XX.X, XX.X	XX.X, XX.X	XX.X, XX.X	
Min, Max	XX, XX	XX, XX	XX, XX	
Age Group, n (%)				
< 65 years	XX (XX.X)	XX (XX.X)	XX (XX.X)	X.XXXX
≥ 65 years	XX (XX.X)	XX (XX.X)	XX (XX.X)	
Gender, n (%)				
Male	XX (XX.X)	XX (XX.X)	XX (XX.X)	X.XXXX
Female	XX (XX.X)	XX (XX.X)	XX (XX.X)	
Ethnicity, n (%)				
Hispanic or Latino	XX (XX.X)	XX (XX.X)	XX (XX.X)	X.XXXX
Not Hispanic or Latino	XX (XX.X)	XX (XX.X)	XX (XX.X)	

https://github.com/cdisc-org/analysis-results-standard/blob/main/workfiles/examples/ARS v1/Demog_Table_Shell.rtf

TFL DESIGNER

- Defining the design of TFLs in text can be a tedious task, so a visual editor called TFL DESIGNER has been developed to alleviate this burden.

TFL DESIGNER

- A community tool to create study specific analysis displays
- Export analysis results metadata per the CDISC ARS model in JSON and Excel formats

Source: Bhavin Busa, Clymb Clinical; https://tfldesigner.org/

cdisc

Table 02

Baseline Demographic and Clinical Characteristics [FDA STF-IG]

Safety Population

Pooled Analyses (or Trial X)

Characteristics	Drug Name Dosage X N = XXX n (%)	Drug Name Dosage Y N = XXX n (%)	Placebo N = XXX n (%)	Active Control N = XXX n (%)	Total Population N = XXX n (%)
Ethnicity, n (%)	**n (%)**	**n (%)**	**n (%)**	**n (%)**	**n (%)**
Hispanic	n (%)	n (%)	n (%)	n (%)	n (%)
Not Hispanic or Latino	n (%)	n (%)	n (%)	n (%)	n (%)
Unknown	n (%)	n (%)	n (%)	n (%)	n (%)
Age, Years					
Mean (SD)	X (Y)	X (Y)	X (Y)	X (Y)	X (Y)
Median (Min, Max)	X (Y, Z)	X (Y, Z)	X (Y, Z)	X (Y, Z)	X (Y, Z)
Interquartile range -	X - Y	X - Y	X - Y	X - Y	X - Y
Total exposure (person years)	X (Y)	X (Y)	X (Y)	X (Y)	X (Y)

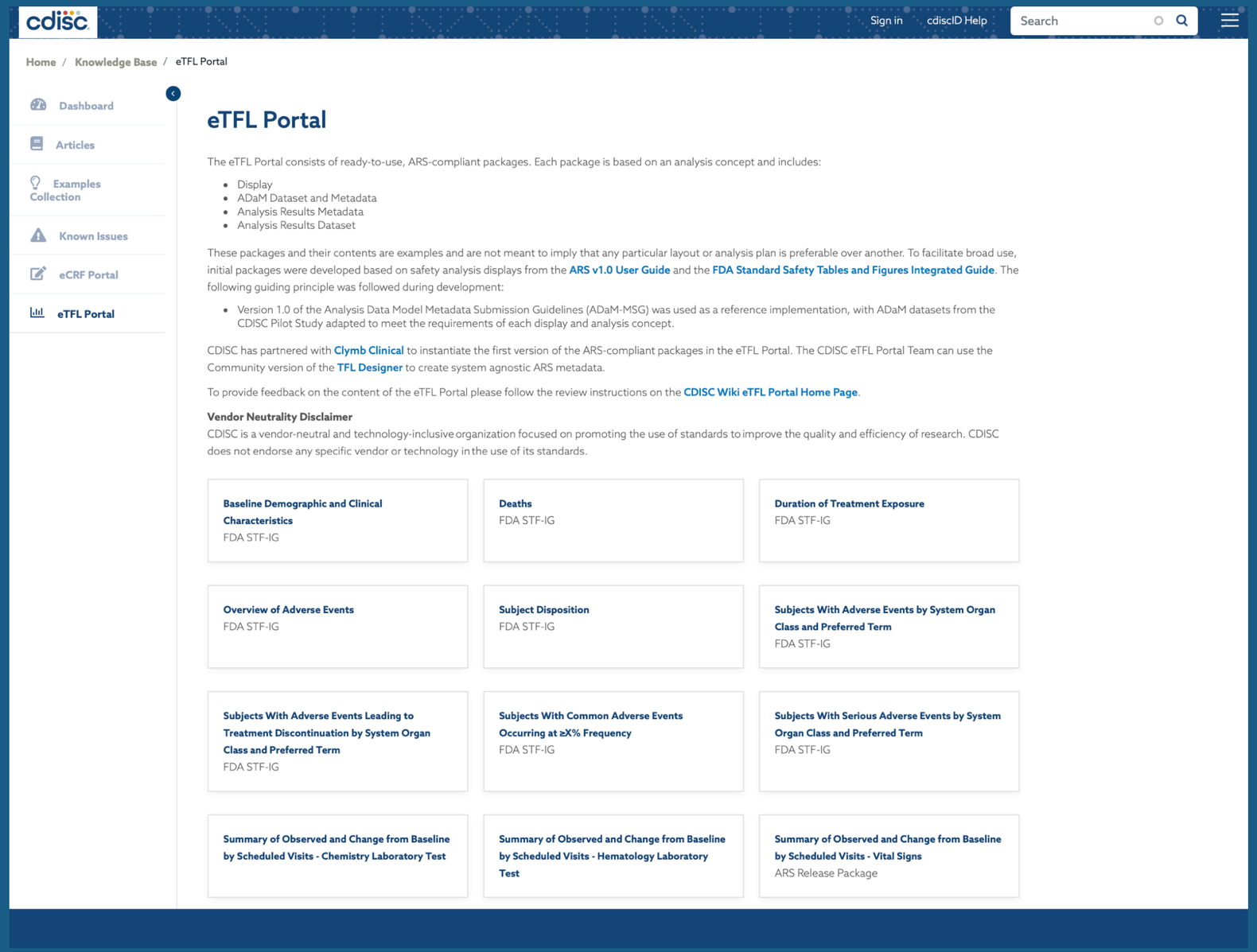

```
{
 "name": "FDA Standard Safety Tables and Figures – Integrated Guide, Table 2",
 "id": "FDA_STF_T2",
 "listOfPlannedAnalyses": {
  "listItems": [
   {
    "name": "Table 2. Baseline Demographic and Clinical Characteristics, Safety Population, Trial CDISCPILOT01",
    "level": 1,
    "order": 1,
    "outputId": "O_FDA_STF_T2",
    "sublist": {
     "listItems": [
      {
       "name": "Count of Subjects by Treatment",
       "level": 2,
       "order": 1,
       "analysisId": "A_SAF_CNT_USUBJID_TRT"
      },
      {
       "name": "Count of Subjects (Total Population)",
       "level": 2,
       "order": 2,
       "analysisId": "A_SAF_CNT_USUBJID"
      },
      {
       "name": "Sex, n (%)",
       "level": 2,
       "order": 3,
       "sublist": {
        "listItems": [
         {
          "name": "Summary of Subjects by Treatment",
          "level": 3,
          "order": 1,
          "analysisId": "A_SAF_SUM_USUBJID_TRT_SEX"
         },
         {
          "name": "Summary of Subjects (Total Population)",
          "level": 3,
          "order": 2,
          "analysisId": "A_SAF_SUM_USUBJID_SEX"
         }
        ]
       }
      },
```


https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/US/Bethesda/PRE_DS04.pdf

eTFL Portal

https://www.cdisc.org/kb/etfl

- Various samples of TFLs are available on the CDISC eTFL Portal.
- By referring to these samples, it should also be possible to define ARS-format TFLs in text.
- An AI trained on these samples might be able to generate ARS-format TFLs (Shells) with simple prompt inputs.

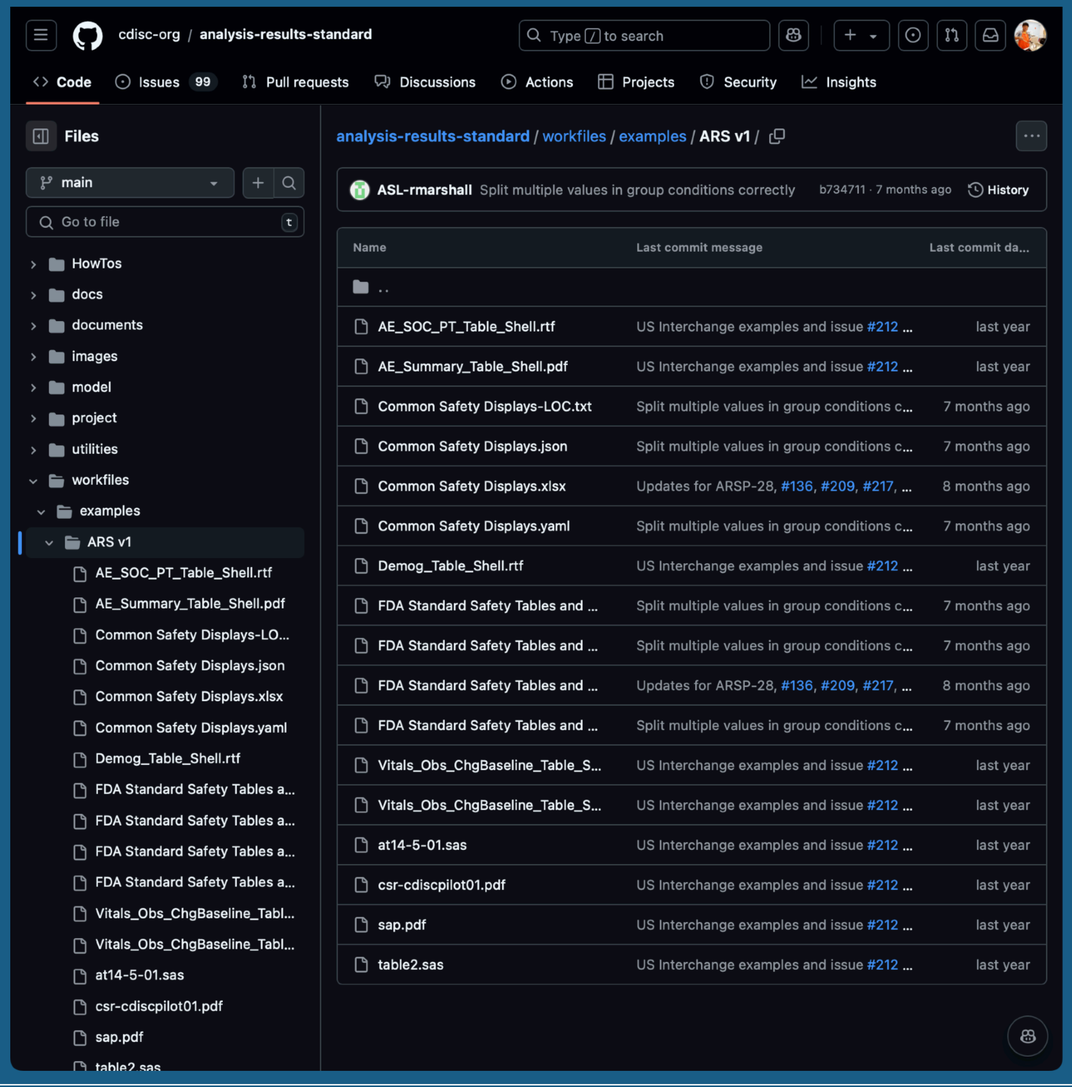


https://www.cdisc.org/kb/etfl

How to make TFLs with CDISC ARS?

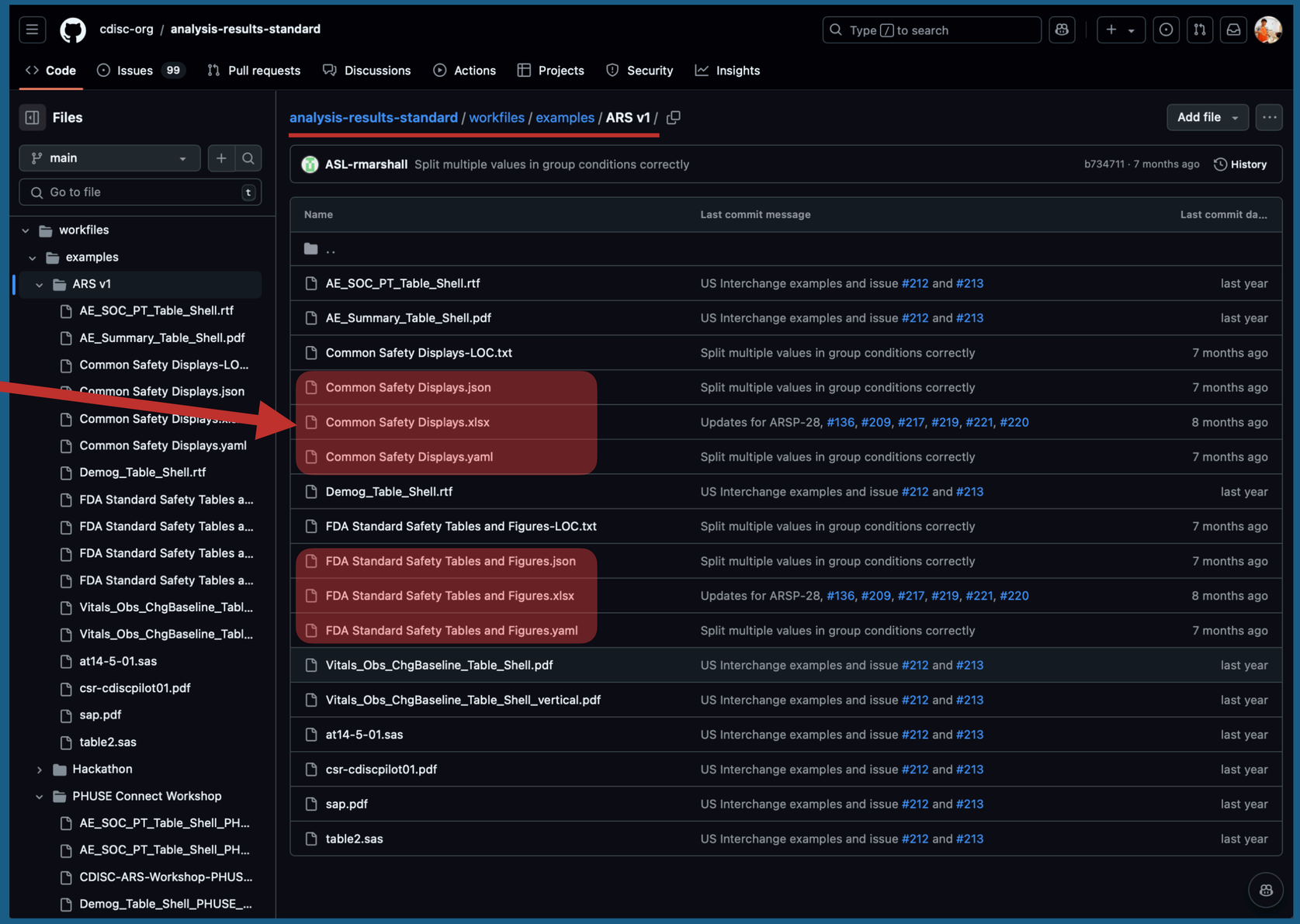
CDISC ARS Github Repository

- Examples are stored in the CDISC ARS GitHub Repository. https://github.com/cdisc-org/analysis-results-standard/tree/main/workfiles/examples/ARS v1



Work Process in CDISC ARS

- The repository contains materials from workshops held by CDISC and PHUSE.
- One of these files can be created, and TFLs (RTF, PDF, etc.) can be generated from it.
- Programs or software are required to create ARS data files.
- Additionally, programs or software are needed to generate TFLs from ARS data files.

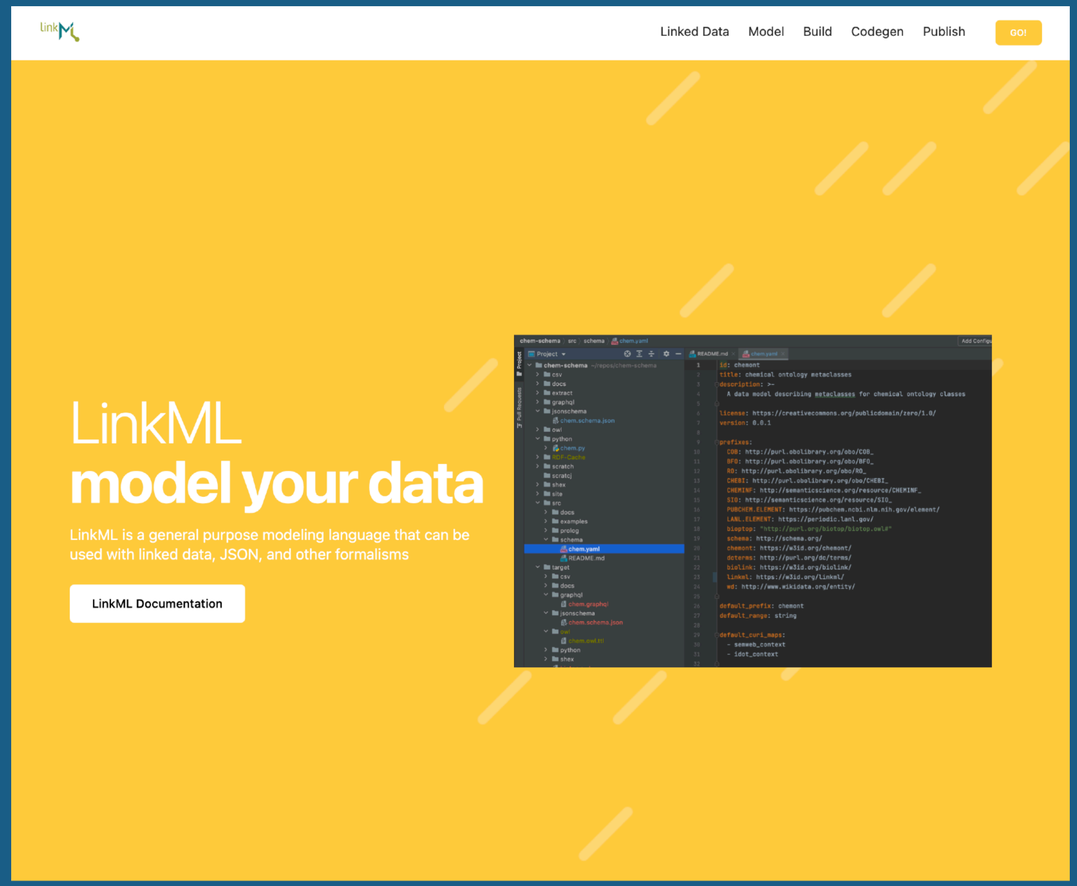


https://github.com/cdisc-org/analysis-results-standard/tree/main/workfiles/examples/ARS v1

LinkML

LinkML

- Meta Schema
  - Schema generator for ...
    - JSON-Schema
    - RDFS
    - OWL
    - GraphQL
    - SQL DDL
    - Python dataclass



https://linkml.io/

Explanation of Specific Terms

Term	Definition	Synonym	Used in
Schema	A definition of the data structure for a specific domain.	Class	Database, JSON, and etc.
Class	A definition of the data structure for a specific domain. However, in the context of programming languages, it includes not only data but also the definition of behaviors.	Schema	JSON, Python, Property Graph, and etc.
Entity	An entity that includes the data itself.	Object	Property Graph, and etc.
Object	An entity that includes the data itself.	Entity	JSON, Python, and etc.
Literal	Just charactors or numbers.		

A simple example to understand LinkML.

An example of generating a Python dataclass from a LinkML schema

LinkML Schema
(Meta-Schema)

YAML File: personinfor.yaml

```yaml
id: https://w3id.org/linkml/examples/personinfo
name: personinfo
prefixes:
  linkml: https://w3id.org/linkml/
  personinfo: https://w3id.org/linkml/examples/personinfo
imports:
  - linkml:types
default_range: string
default_prefix: personinfo

classes:
  Person:
    attributes:
      id:
      full_name:
      aliases:
      phone:
      age:
```

Python Dataclass

Python Class: personinfor.py

```python
@dataclass(repr=False)
class Person(YAMLRoot):
    _inherited_slots: ClassVar[List[str]] = []

    class_class_uri: ClassVar[URIRef] = PERSONINFO["Person"]
    class_class_curie: ClassVar[str] = "personinfo:Person"
    class_name: ClassVar[str] = "Person"
    class_model_uri: ClassVar[URIRef] = PERSONINFO.Person

    id: Optional[str] = None
    full_name: Optional[str] = None
    aliases: Optional[str] = None
    phone: Optional[str] = None
```

A simple example to understand LinkML.

An example of generating a JSON schema from a LinkML schema

LinkML Schema
(Meta-Schema)

YAML File: personinfor.yaml

```yaml
id: https://w3id.org/linkml/examples/personinfo
name: personinfo
prefixes:
  linkml: https://w3id.org/linkml/
  personinfo: https://w3id.org/linkml/examples/personinfo
imports:
  - linkml:types
default_range: string
default_prefix: personinfo

classes:
  Person:
    attributes:
      id:
      full_name:
      aliases:
      phone:
      age:
```

JSON Schema

JSON Schema File: personinfo.schema.json

```json
{
    "$defs": {
        "Person": {
            "additionalProperties": false,
            "description": "",
            "properties": {
                "aliases": {
                    "type": [
                        "string",
                        "null"
                    ]
                },
                "full_name": {
                    "type": [
                        "string",
                        "null"
                    ]
                },
```

A simple example to understand LinkML.

An example of generating a YAML data file from a CSV data file

CSV

CSV File: personinfor_data.csv

```
id,full_name,phone,age
ORCID:1234,Clark Kent,555-555-5555,32
ORCID:4567,Lois Lane,,33
```

YAML

JSON Schema File: personinfo_data.yaml

```
persons:
  - id: ORCID:1234
    full_name: Clark Kent
    phone: 555-555-5555
    age: 32
  - id: ORCID:4567
    full_name: Lois Lane
    age: 33
```

A simple example to understand LinkML.

An example of generating a JSON-LD file from a CSV data file

CSV

CSV File: personinfor_data.csv

```
id,full_name,phone,age
ORCID:1234,Clark Kent,555-555-5555,32
ORCID:4567,Lois Lane,,33
```

JSON-LD

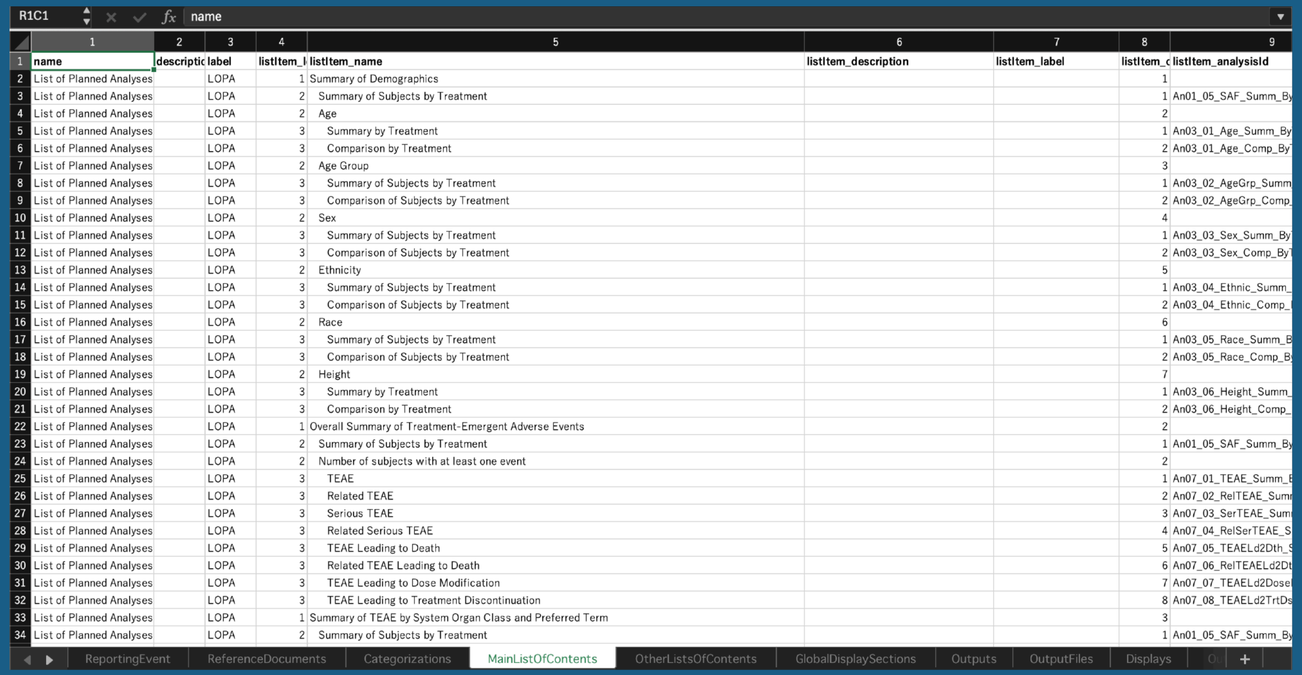
JSON Schema File: personinfo_data.jsonld

```
{
  "persons": [
    {
      "id": "ORCID:1234",
      "full_name": "Clark Kent",
      "phone": "555-555-5555",
      "age": 32
    },
    {
      "id": "ORCID:4567",
      "full_name": "Lois Lane",
      "age": 33
    }
  ],
  "@type": "Container",
  "@context": {
    "xsd": "http://www.w3.org/2001/XMLSchema#",
    "linkml": "https://w3id.org/linkml/",
```

The advantages of using LinkML

- One of the major advantages of using LinkML for data modeling is the ease with which you can convert schemas and data between various file formats.

- The LinkML Schema can also output an Excel template for data input. By entering data into this template, it is possible to convert the data into other formats and handle it as Linked Data.


R1C1 name

	1	2	3	4	5	6	7	8	9
1	name	descriptic	label	listItem_l	listItem_name	listItem_description	listItem_label	listItem_c	listItem_analysisId
2	List of Planned Analyses		LOPA	1	Summary of Demographics			1	
3	List of Planned Analyses		LOPA	2	Summary of Subjects by Treatment			1	An01_05_SAF_Summ_By
4	List of Planned Analyses		LOPA	2	Age			2	
5	List of Planned Analyses		LOPA	3	Summary by Treatment			1	An03_01_Age_Summ_By
6	List of Planned Analyses		LOPA	3	Comparison by Treatment			2	An03_01_Age_Comp_By
7	List of Planned Analyses		LOPA	2	Age Group			3	
8	List of Planned Analyses		LOPA	3	Summary of Subjects by Treatment			1	An03_02_AgeGrp_Summ
9	List of Planned Analyses		LOPA	3	Comparison of Subjects by Treatment			2	An03_02_AgeGrp_Comp
10	List of Planned Analyses		LOPA	2	Sex			4	
11	List of Planned Analyses		LOPA	3	Summary of Subjects by Treatment			1	An03_03_Sex_Summ_By
12	List of Planned Analyses		LOPA	3	Comparison of Subjects by Treatment			2	An03_03_Sex_Comp_By
13	List of Planned Analyses		LOPA	2	Ethnicity			5	
14	List of Planned Analyses		LOPA	3	Summary of Subjects by Treatment			1	An03_04_Ethnic_Summ_
15	List of Planned Analyses		LOPA	3	Comparison of Subjects by Treatment			2	An03_04_Ethnic_Comp_
16	List of Planned Analyses		LOPA	2	Race			6	
17	List of Planned Analyses		LOPA	3	Summary of Subjects by Treatment			1	An03_05_Race_Summ_B
18	List of Planned Analyses		LOPA	3	Comparison of Subjects by Treatment			2	An03_05_Race_Comp_B
19	List of Planned Analyses		LOPA	2	Height			7	
20	List of Planned Analyses		LOPA	3	Summary by Treatment			1	An03_06_Height_Summ_
21	List of Planned Analyses		LOPA	3	Comparison by Treatment			2	An03_06_Height_Comp_
22	List of Planned Analyses		LOPA	1	Overall Summary of Treatment-Emergent Adverse Events			2	
23	List of Planned Analyses		LOPA	2	Summary of Subjects by Treatment			1	An01_05_SAF_Summ_By
24	List of Planned Analyses		LOPA	2	Number of subjects with at least one event			2	
25	List of Planned Analyses		LOPA	3	TEAE			1	An07_01_TEAE_Summ_B
26	List of Planned Analyses		LOPA	3	Related TEAE			2	An07_02_RelTEAE_Sum
27	List of Planned Analyses		LOPA	3	Serious TEAE			3	An07_03_SerTEAE_Sum
28	List of Planned Analyses		LOPA	3	Related Serious TEAE			4	An07_04_RelSerTEAE_S
29	List of Planned Analyses		LOPA	3	TEAE Leading to Death			5	An07_05_TEAELd2Dth_S
30	List of Planned Analyses		LOPA	3	Related TEAE Leading to Death			6	An07_06_RelTEAELd2Dt
31	List of Planned Analyses		LOPA	3	TEAE Leading to Dose Modification			7	An07_07_TEAELd2Dose
32	List of Planned Analyses		LOPA	3	TEAE Leading to Treatment Discontinuation			8	An07_08_TEAELd2TrtDs
33	List of Planned Analyses		LOPA	1	Summary of TEAE by System Organ Class and Preferred Term			3	
34	List of Planned Analyses		LOPA	2	Summary of Subjects by Treatment			1	An01_05_SAF_Summ_By

ReportingEvent | ReferenceDocuments | Categorizations | MainListOfContents | OtherListsOfContents | GlobalDisplaySections | Outputs | OutputFiles | Displays

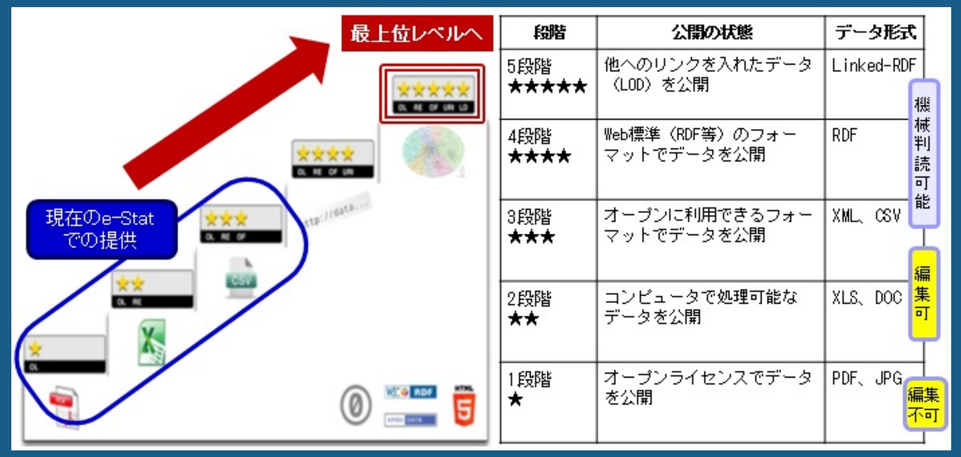
Power of Linked Data

Machine Readable

- For example, the COVID-19 infection status was published on various countries' websites, but aggregating this data was not an easy task.
- Simply being able to read an Excel file on a computer does not necessarily make it "machine-readable" in the context of structured data.
- For a data value (literal) to be truly "machine-readable," the computer must be able to interpret what type of data it is. This ensures that the data is in a sufficiently machine-readable state.
- Therefore, it is necessary to add extensive metadata to ensure that the computer can interpret and understand the data correctly.

Linked Data

- To solve the various issues mentioned above and achieve the goal, Linked Data was conceived. However, it has not become widely popular…
- A system designed to connect data across the web worldwide. https://en.wikipedia.org/wiki/Semantic_Web
- It is the foundational technology for W3C's Web 3.0. (Note: This is different from Web3, also known as Web Three!)

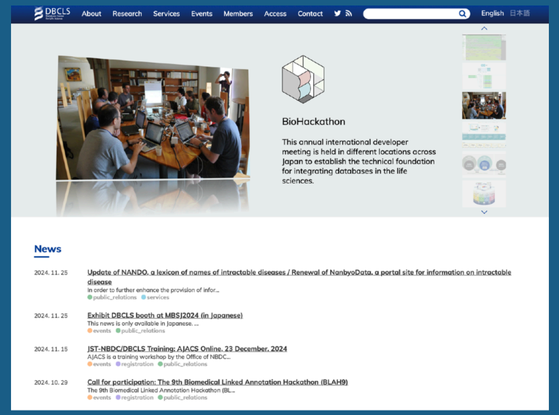


段階	公開の状態	データ形式	
5段階 ★★★★★	他へのリンクを入れたデータ（LOD）を公開	Linked-RDF	機械判読可能
4段階 ★★★★	Web標準（RDF等）のフォーマットでデータを公開	RDF	
3段階 ★★★	オープンに利用できるフォーマットでデータを公開	XML、CSV	
2段階 ★★	コンピュータで処理可能なデータを公開	XLS、DOC	編集可
1段階 ★	オープンライセンスでデータを公開	PDF、JPG	編集不可

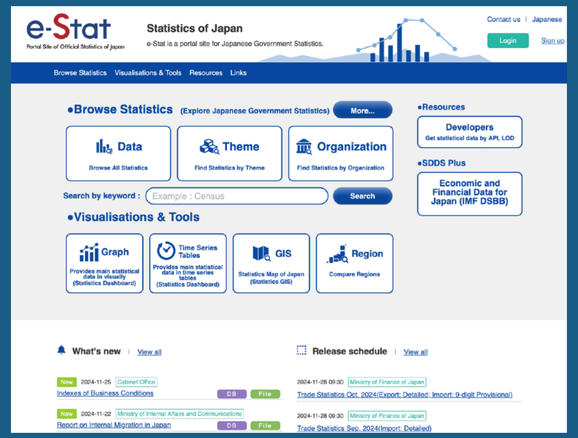
https://www.stat.go.jp/info/today/095.htm

Examples of Linked Data Utilization in Japan

- Database Center for Life Science
  https://dbcls.rois.ac.jp/



- e-Stat(Statistics of Japan)
  https://www.e-stat.go.jp/



- National Diet Library, Japan
  https://www.ndl.go.jp/jp/dlib/standards/lod/index.html

The fundamental of Linked Data is a property graph

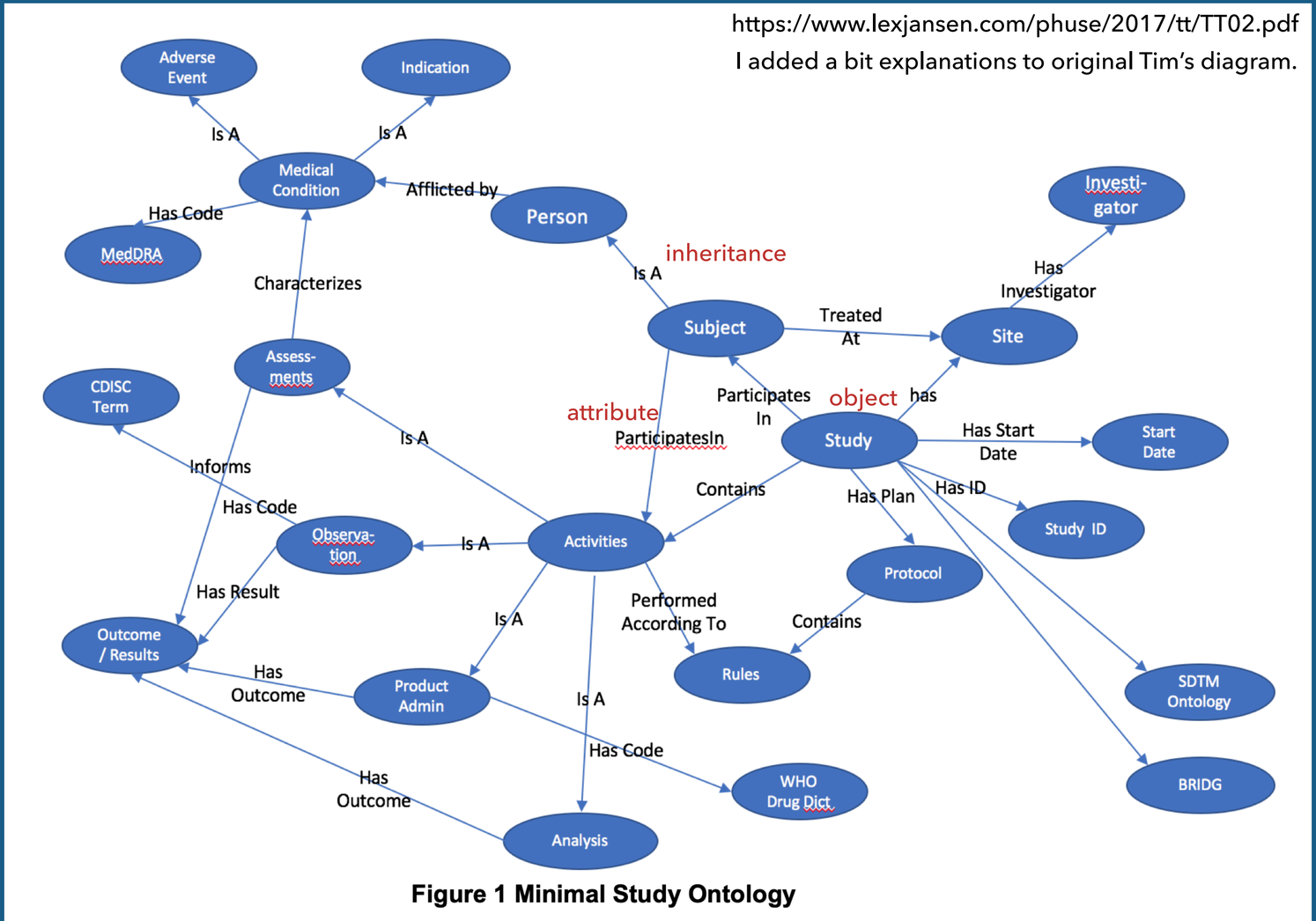


Figure 1 Minimal Study Ontology

- By linking the Results Data from clinical trials conducted worldwide, you can execute a query once to extract all the necessary Results Data from all trials that match the query conditions.

Advantages and Disadvantages of Using Linked Data

- Advantages
  - It becomes easier to extract the Data and Results Data needed for performing a meta-analysis, allowing for correct selection of the trials to be included.
  - By converting not only Result Data but also SDTM and ADaM data into Linked Data, it becomes easier to verify data traceability. This enhances the transparency of statistical analysis!
  - You can extract the required data from trial data and trial results worldwide with a single query.
  - Utilization by regulatory authorities, such as during approval reviews.
- Disadvantages
  - Conceptually, it's difficult. -> However, AI will eventually handle this, so we may not need to be consciously aware of Linked Data. (AI will create the graph, and the graph will form the AI's RAG).
  - It is necessary to add a lot of metadata.
  - It is necessary to create a Linked Data application.
  - Even without using Linked Data, things are manageable for now, and since AI seems to be handling it well, motivation tends to be low.

I like Linked Data because methods in graph theory and Graph Neural Network are available .

Is ARS Linked Data?

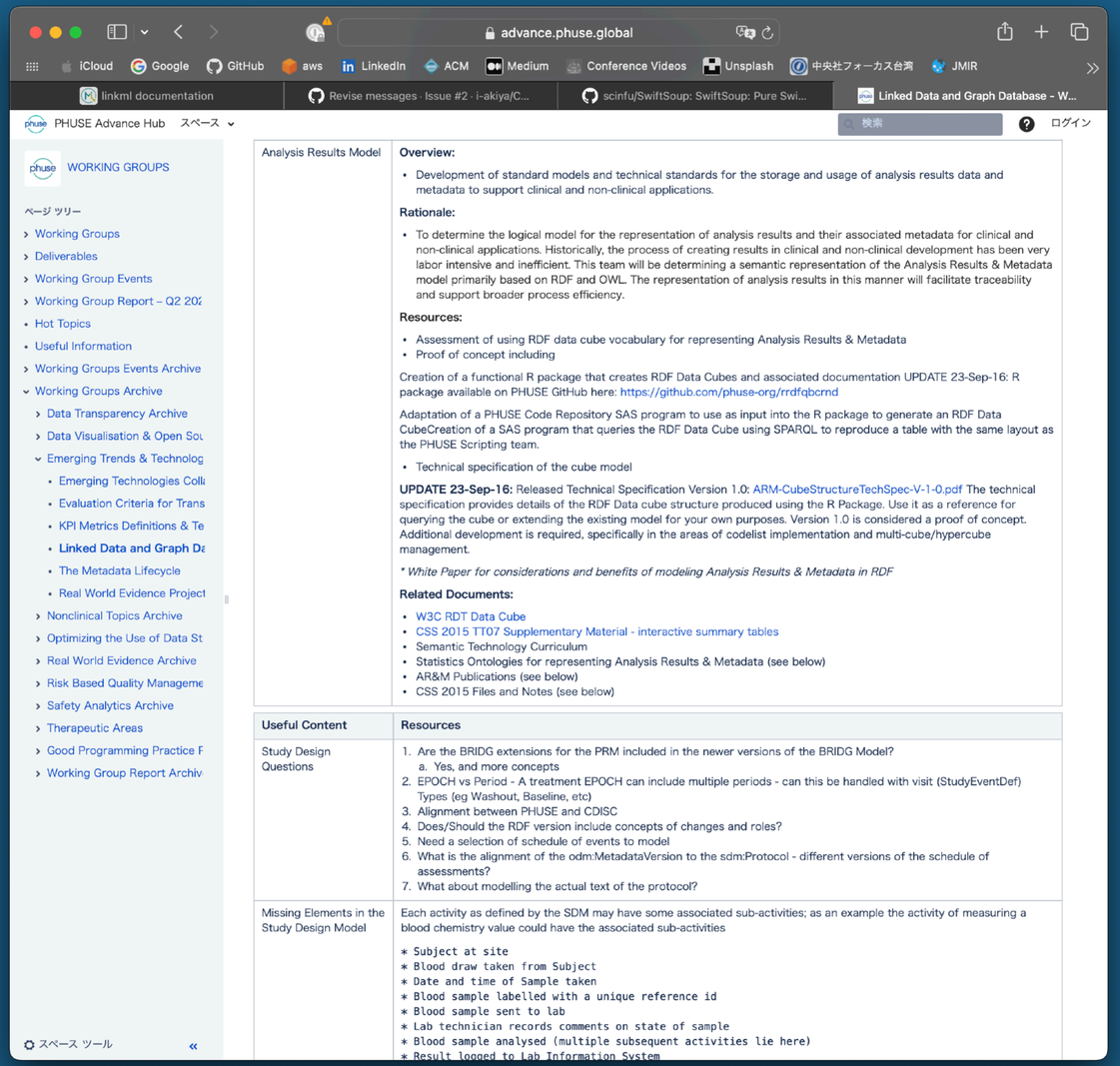
- Yes, but not yet enough to be perfect Linked Data.
  - CDISC ARS is developed with LinkML that is Linked Data modeling language.
  - Currently, the focus seems to be on automating the output of TFLs and define.xml.
  - It seems that there has been little mention of utilizing Linked Data.
- It seems there has been no mention of query languages such as SPARQL or GraphQL for extracting Results Data across multiple items and trials.

Expectations for CDISC ARS

- I would like to see more diverse examples for eTFL, such as survival analysis.
- If examples of creating Summary Tables for CTD (Common Technical Document) from TFLs of multiple trials were provided, it would help people better understand the benefits of CDISC ARS.
- Currently, there are examples of Safety Tables, but there is a need for data implementation examples for creating various charts.
- By incorporating a standardized set of statistical terms, the statistical analysis methods themselves would also become Machine Readable, making ARS even more useful. By the way, the standardization of statistical terminology (Ontology) is addressed by STATO. https://stato-ontology.org/
- I hope that software vendors and open-source developers will become more active in developing software related to ARS.
- There is a desire for the development of software that allows regulatory authorities to review TFLs more easily when using ARS.

Past work on Linked Data at PHUSE

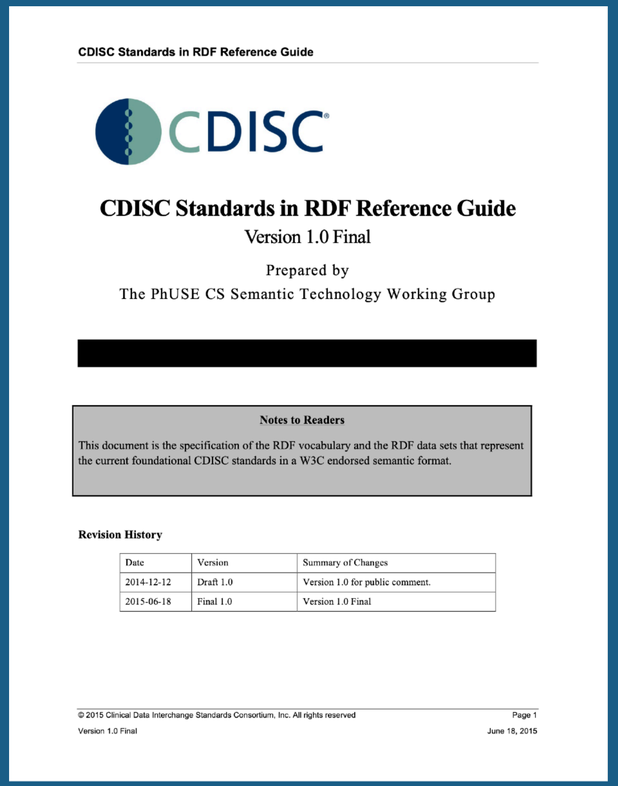
- Linked Data and Graph Database https://advance.phuse.global/display/WEL/Linked+Data+and+Graph+Database



CDISC's Guide related to Linked Data

- CDISC Standards in RDF Reference Guide
- CDISC Standards in RDF User Guide

https://www.cdisc.org/standards/data-exchange/rdf



CDISC Standards in RDF Reference Guide

CDISC

CDISC Standards in RDF Reference Guide

Version 1.0 Final

Prepared by

The PhUSE CS Semantic Technology Working Group

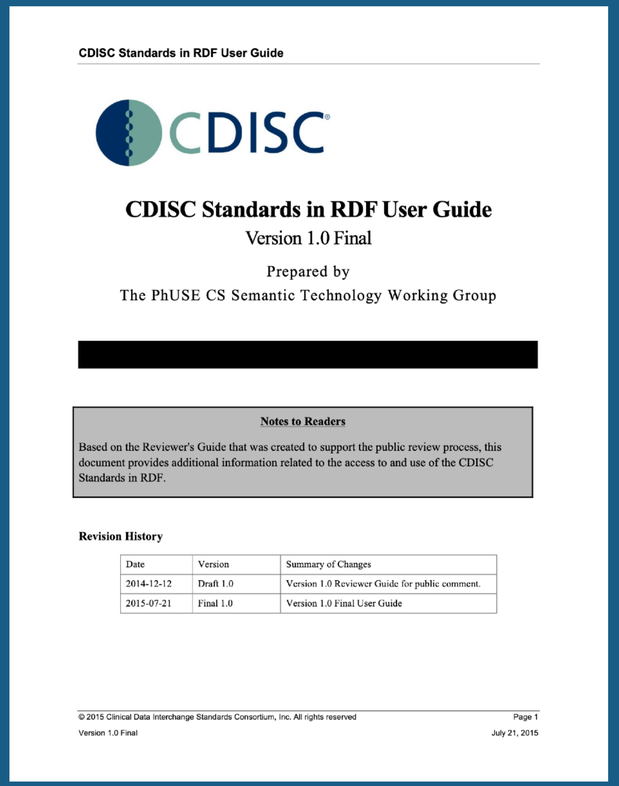
[redacted]

Notes to Readers

This document is the specification of the RDF vocabulary and the RDF data sets that represent the current foundational CDISC standards in a W3C endorsed semantic format.

Revision History

Date	Version	Summary of Changes
2014-12-12	Draft 1.0	Version 1.0 for public comment.
2015-06-18	Final 1.0	Version 1.0 Final



CDISC Standards in RDF User Guide

CDISC

CDISC Standards in RDF User Guide

Version 1.0 Final

Prepared by

The PhUSE CS Semantic Technology Working Group

[redacted]

Notes to Readers

Based on the Reviewer's Guide that was created to support the public review process, this document provides additional information related to the access to and use of the CDISC Standards in RDF.

Revision History

Date	Version	Summary of Changes
2014-12-12	Draft 1.0	Version 1.0 Reviewer Guide for public comment.
2015-07-21	Final 1.0	Version 1.0 Final User Guide

Thank you for your Attention