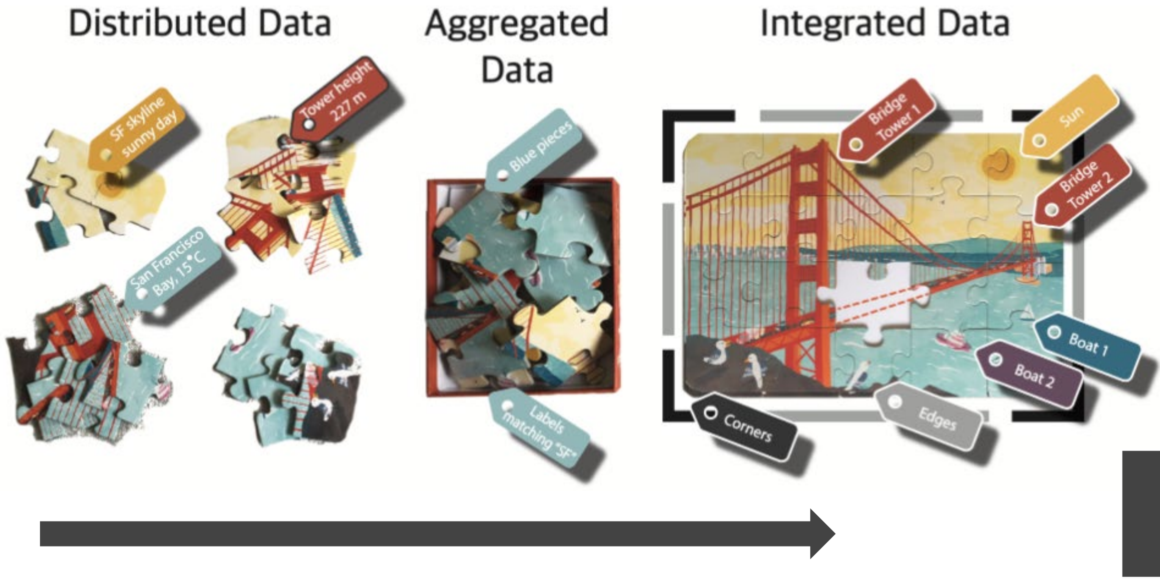
Making the “Common” in CDEs more Common

Anne E Thessen, Christopher Mungall, Sierra Moxon, Melissa Haendel
University of Colorado Anschutz Medical Campus

March 7, 2024
Advancing the Use and Development of Common Data Elements in Research Workshop

We’ve Come a Long Way, But…



>1800 DATABASES

>1500 STANDARDS

>900 ONTOLOGIES

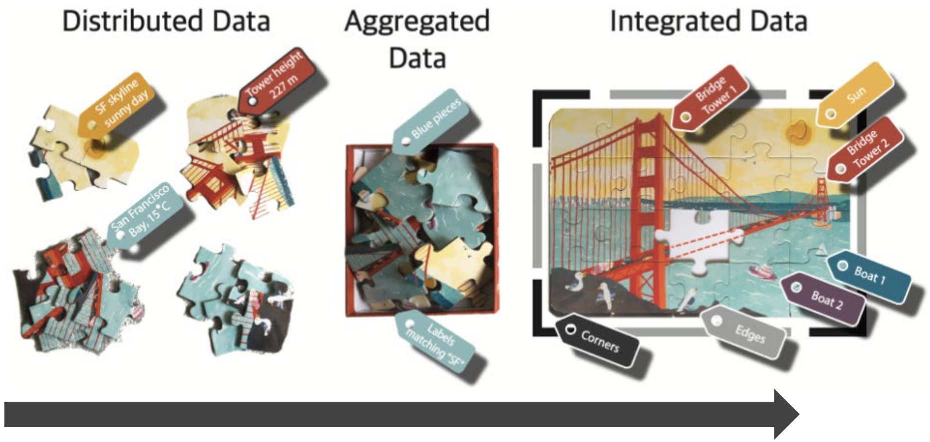
>23K CDES

Search Term	CDEs in NLM	Ontologies in BioPortal
Date of Birth	28	49
Blood Pressure	104	33
Smoking Status	22	48

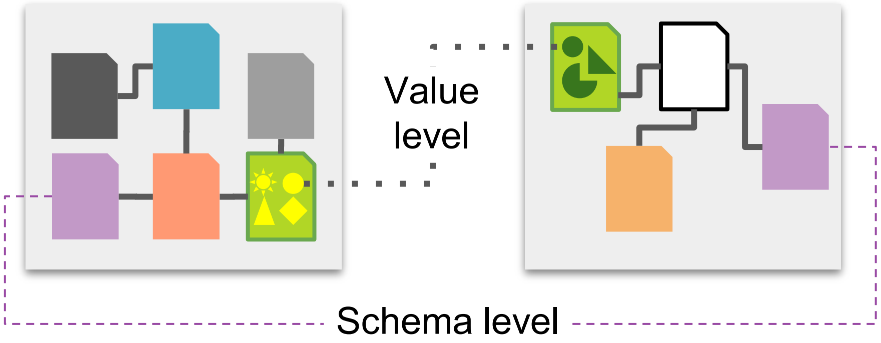
..our data is still not integrated nor interoperable

CDEs Help, But They Are Not Yet Computable

- Good governance and validation approaches
- Working with CDEs is still largely manual - not scalable
- Many overlapping CDE repositories (NLM, caDSR, PhenX, Heal, and more)
- Context and metadata are implied - including mappings, provenance, etc
- Limits data interoperability at scale - this is what we want!



Mapping CDEs (And More) Is Necessary For Interoperability



Data model alignment : Each source models things differently

For example, no direct link from Sample-to-Diagnosis in one model

Would need to “remodel” Sample-to-Case, and Diagnosis-to-Case to align with Sample-to-Diagnosis

Value Set alignment : Each source uses different values

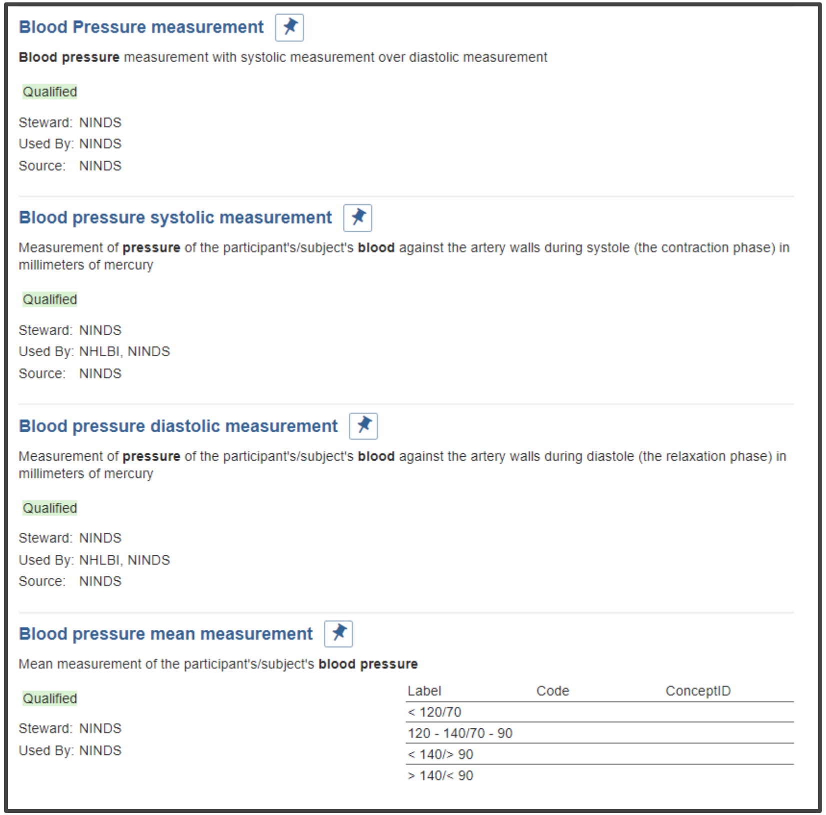
For example, one node encodes race like this:

- not reported
- white
- american indian or alaska native
- black or african american

While another does it like this:

- not allowed to collect
- unknown
- white
- native hawaiian or other pacific islander
- american indian or alaska native
- asian
- other
- black or african american

Example: Many CDEs for Blood Pressure



Blood Pressure measurement

Blood pressure measurement with systolic measurement over diastolic measurement

Qualified

Steward: NINDS
Used By: NINDS
Source: NINDS

Blood pressure systolic measurement

Measurement of **pressure** of the participant's/subject's **blood** against the artery walls during systole (the contraction phase) in millimeters of mercury

Qualified

Steward: NINDS
Used By: NHLBI, NINDS
Source: NINDS

Blood pressure diastolic measurement

Measurement of **pressure** of the participant's/subject's **blood** against the artery walls during diastole (the relaxation phase) in millimeters of mercury

Qualified

Steward: NINDS
Used By: NHLBI, NINDS
Source: NINDS

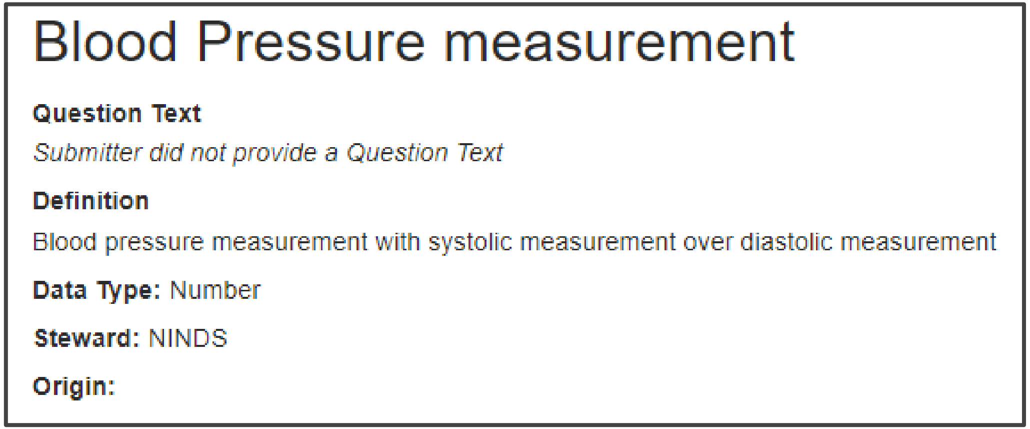
Blood pressure mean measurement

Mean measurement of the participant's/subject's **blood pressure**

Qualified

Steward: NINDS
Used By: NINDS

Label	Code	ConceptID
< 120/70		
120 - 140/70 - 90		
< 140/> 90		
> 140/< 90		



Blood Pressure measurement

Question Text

Submitter did not provide a Question Text

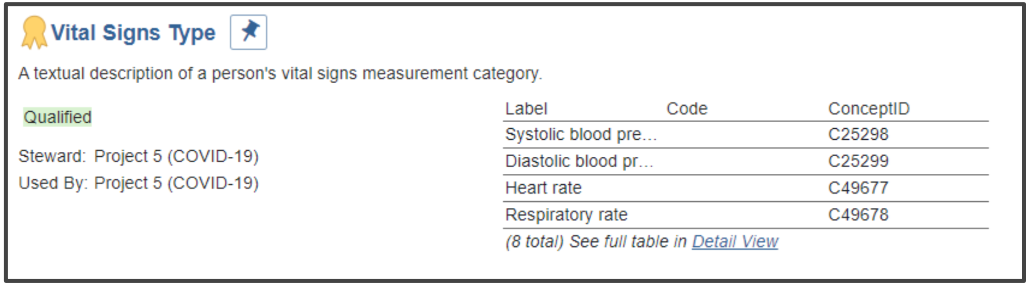
Definition

Blood pressure measurement with systolic measurement over diastolic measurement

Data Type: Number

Steward: NINDS

Origin:



Vital Signs Type

A textual description of a person's vital signs measurement category.

Qualified

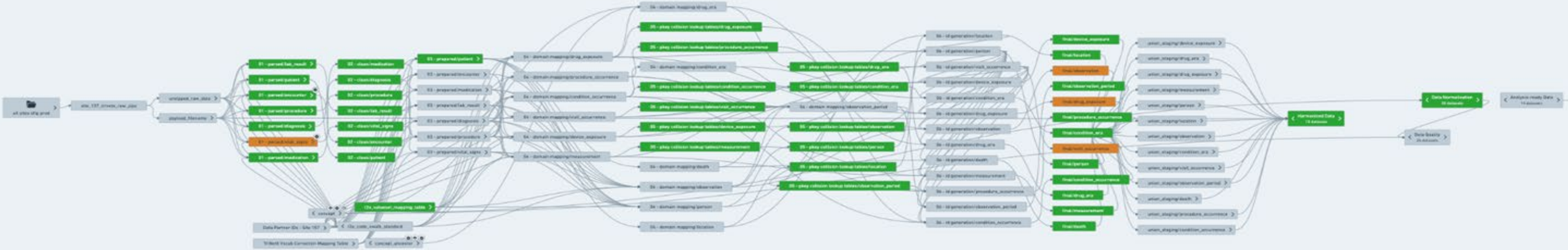
Steward: Project 5 (COVID-19)
Used By: Project 5 (COVID-19)

Label	Code	ConceptID
Systolic blood pre…		C25298
Diastolic blood pr…		C25299
Heart rate		C49677
Respiratory rate		C49678

(8 total) See full table in Detail View

Making CDEs Computable Will Scale Up Data Interoperability

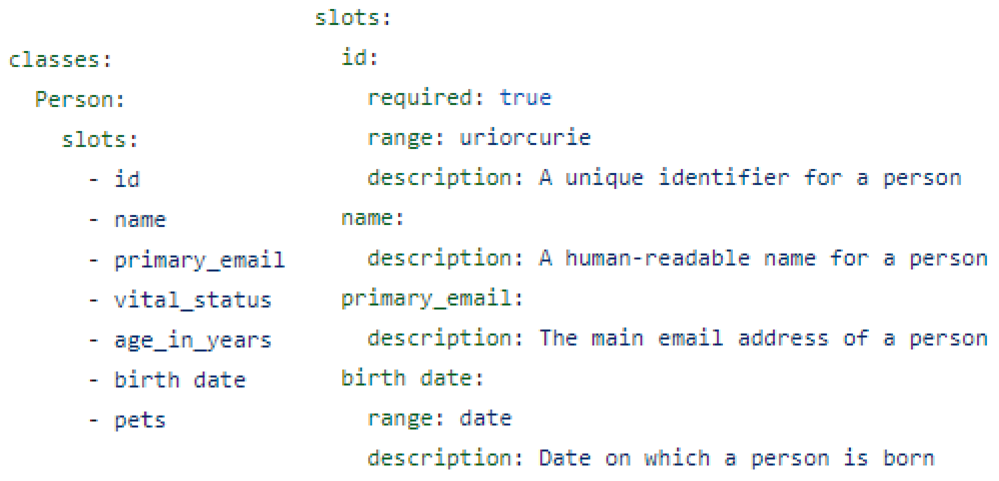
- Take advantage of knowledge modeling languages, tools, and services
- Capture detail and nuance in data that you couldn't before
- Leverage power of reasoners, transformers, and compliance checkers to automate data QA/QC, inference, search, versioning, and format changes downstream
- **If we make CDEs semantically interoperable we can make data interoperable at scale…**



Our Proposal: LinkML

- Simple, flexible, agnostic - YAML
- Suite of supportive tooling to create, manage, export models and data
- Allows for the capture of EVERYTHING needed in a CDE
- Not just about the CDE - supports the data to which you apply the CDE

Example YAML



```
classes:
  Person:
    slots:
      - id
      - name
      - primary_email
      - vital_status
      - age_in_years
      - birth date
      - pets
```

```
slots:
  id:
    required: true
    range: uriorcurie
    description: A unique identifier for a person
  name:
    description: A human-readable name for a person
  primary_email:
    description: The main email address of a person
  birth date:
    range: date
    description: Date on which a person is born
```

Validators

Data Converters

Code Generation

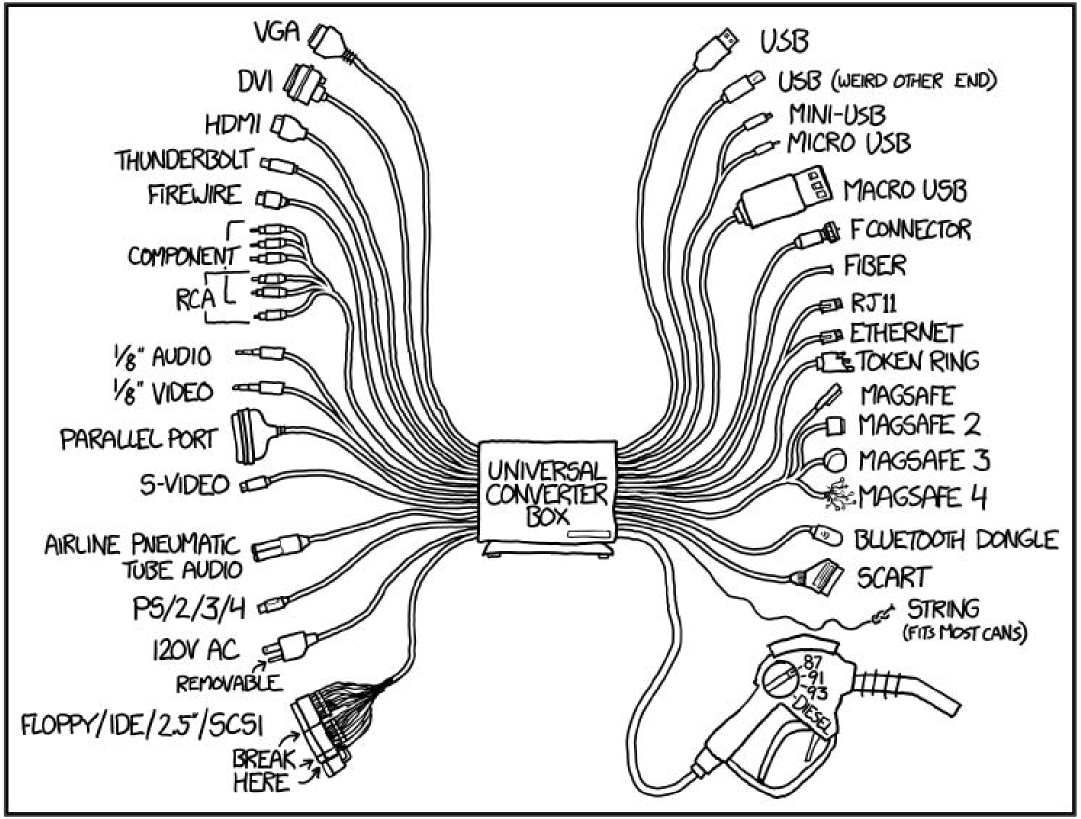
Data entry tooling

Schema inference

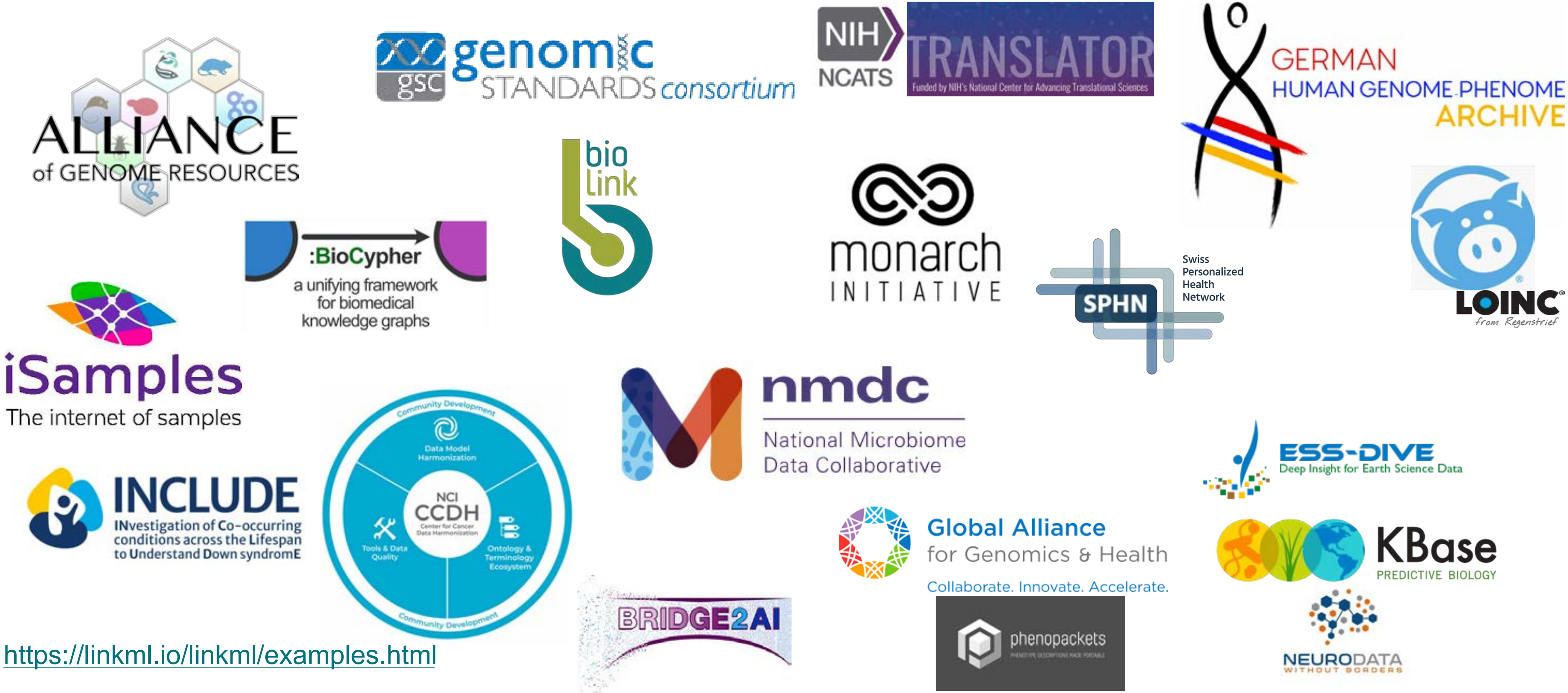
https://linkml.io
https://github.com/linkml/linkml

https://github.com/linkml/linkml-tutorial
https://linkml.io/linkml/intro/tutorial.html

LinkML Is A Converter Box

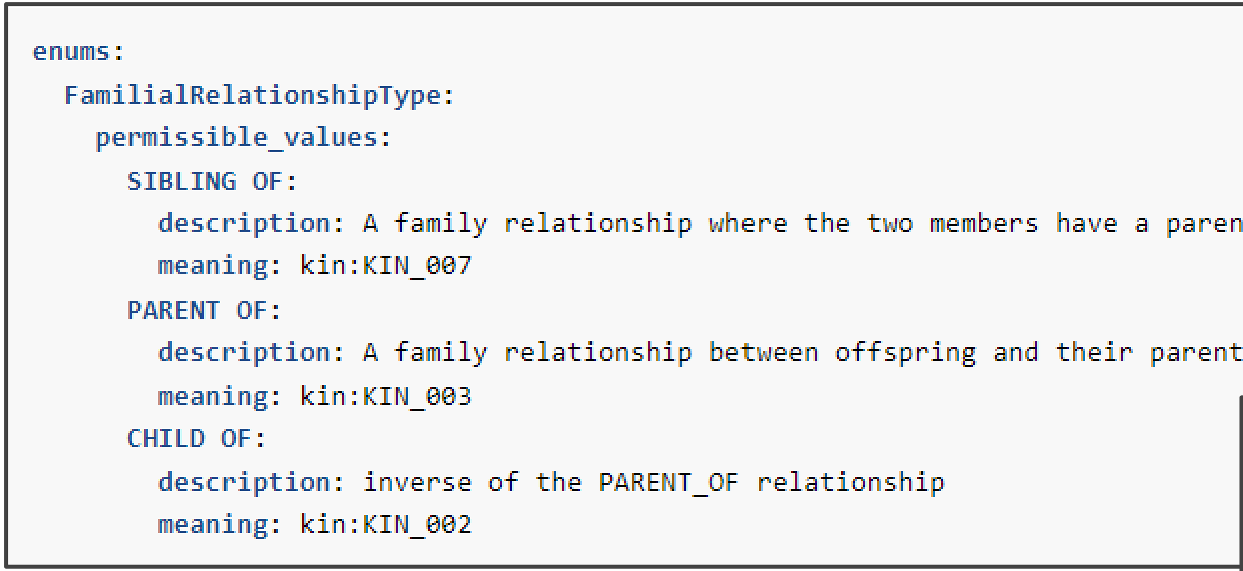


Adoption: Who Is Using LinkML?

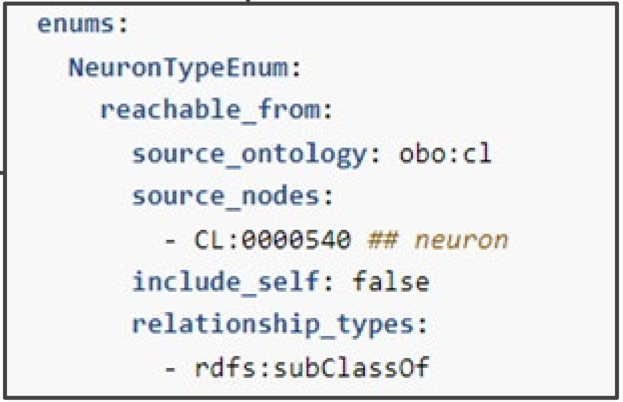

ALLIANCE of GENOME RESOURCES
gsc genomic STANDARDS consortium
NIH NCATS
TRANSLATOR Funded by NIH's National Center for Advancing Translational Sciences
GERMAN HUMAN GENOME-PHENOME ARCHIVE
bio link
:BioCypher a unifying framework for biomedical knowledge graphs
monarch INITIATIVE
SPHN Swiss Personalized Health Network
LOINC from Regenstrief
iSamples The internet of samples
nmdc National Microbiome Data Collaborative
ESS-DIVE Deep Insight for Earth Science Data
INCLUDE INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE
Community Development
Data Model Harmonization
NCI CCDH Center for Cancer Data Harmonization
Tools & Data Quality
Ontology & Terminology Ecosystem
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.
KBase PREDICTIVE BIOLOGY
BRIDGE2AI
phenopackets
NEURODATA WITHOUT BORDERS

https://linkml.io/linkml/examples.html

Ontologies Provide Enumerated Values and Logical Structure



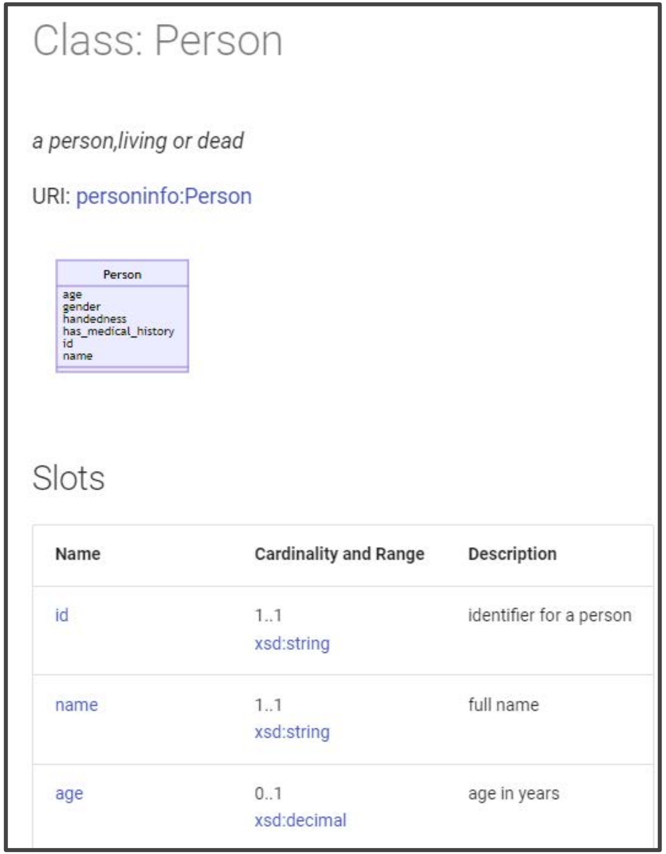
```
enums:
  FamilialRelationshipType:
    permissible_values:
      SIBLING OF:
        description: A family relationship where the two members have a parent on common
        meaning: kin:KIN_007
      PARENT OF:
        description: A family relationship between offspring and their parent
        meaning: kin:KIN_003
      CHILD OF:
        description: inverse of the PARENT_OF relationship
        meaning: kin:KIN_002
```



```
enums:
  NeuronTypeEnum:
    reachable_from:
      source_ontology: obo:cl
      source_nodes:
        - CL:0000540 ## neuron
      include_self: false
      relationship_types:
        - rdfs:subClassOf
```

Build On Foundation: Make The Implicit, Explicit

- Humans know that blood pressure is systolic over diastolic - make computable - in the context of previous work
- Create computable data models
- Create mappings
- Use an open, community driven approach (OBO Foundry good example)
- Make documentation easy



Class: Person

a person,living or dead

URI: personinfo:Person

Person

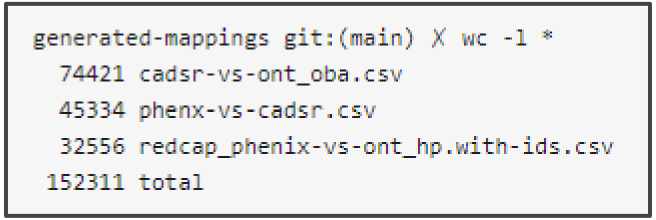
age
gender
handedness
has_medical_history
id
name

Slots

Name	Cardinality and Range	Description
id	1..1 xsd:string	identifier for a person
name	1..1 xsd:string	full name
age	0..1 xsd:decimal	age in years

Proposed Workflow: Making CDEs Computable At Scale

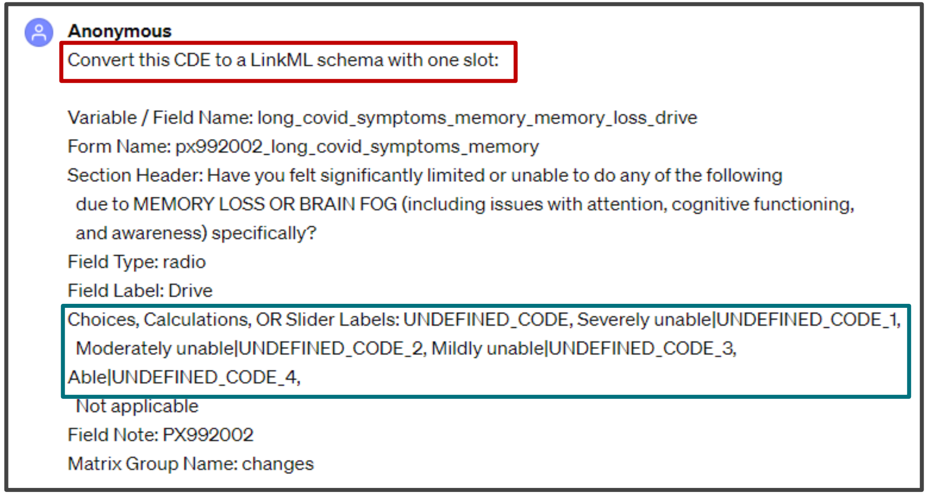
- LinkML Schema Helper to generate YAML
- CurateGPT to generate mappings
- SSSOM to express mappings
- Will need human review
- Use these automated results to design a strategy for curation
- With these tools, this process is tractable in years, not decades
- Let’s look at some examples…



```
generated-mappings git:(main) X wc -l *
  74421 cadsr-vs-ont_oba.csv
  45334 phenx-vs-cadsr.csv
  32556 redcap_phenix-vs-ont_hp.with-ids.csv
 152311 total
```

PhenX	HPO	HPO Label	Similarity
px020101_phx_arm_span	HP:0012771	Increased arm span	0.8588712
px020501_phenx_child_head_circumference	HP:0040194	Increased head circumference	0.8714960

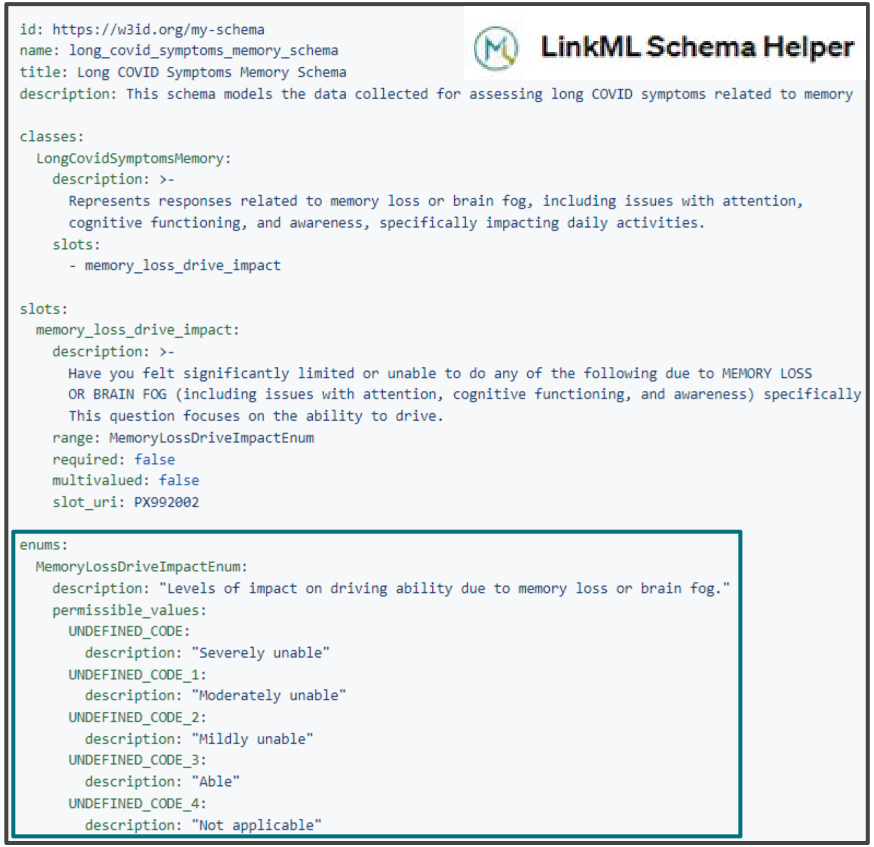
Example: Using LinkML Schema Helper



Anonymous

Convert this CDE to a LinkML schema with one slot:

Variable / Field Name: long_covid_symptoms_memory_memory_loss_drive
Form Name: px992002_long_covid_symptoms_memory
Section Header: Have you felt significantly limited or unable to do any of the following
 due to MEMORY LOSS OR BRAIN FOG (including issues with attention, cognitive functioning,
 and awareness) specifically?
Field Type: radio
Field Label: Drive
Choices, Calculations, OR Slider Labels: UNDEFINED_CODE, Severely unable|UNDEFINED_CODE_1,
 Moderately unable|UNDEFINED_CODE_2, Mildly unable|UNDEFINED_CODE_3,
Able|UNDEFINED_CODE_4,
 Not applicable
Field Note: PX992002
Matrix Group Name: changes

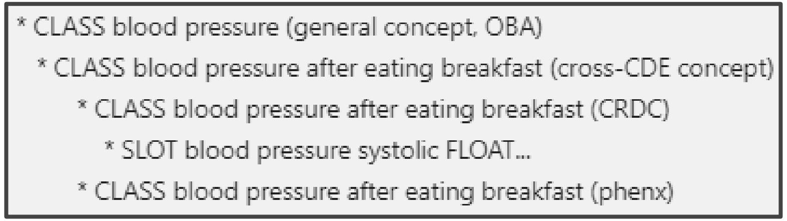


LinkML Schema Helper

```yaml
id: https://w3id.org/my-schema
name: long_covid_symptoms_memory_schema
title: Long COVID Symptoms Memory Schema
description: This schema models the data collected for assessing long COVID symptoms related to memory

classes:
  LongCovidSymptomsMemory:
    description: >-
      Represents responses related to memory loss or brain fog, including issues with attention,
      cognitive functioning, and awareness, specifically impacting daily activities.
    slots:
      - memory_loss_drive_impact

slots:
  memory_loss_drive_impact:
    description: >-
      Have you felt significantly limited or unable to do any of the following due to MEMORY LOSS
      OR BRAIN FOG (including issues with attention, cognitive functioning, and awareness) specifically
      This question focuses on the ability to drive.
    range: MemoryLossDriveImpactEnum
    required: false
    multivalued: false
    slot_uri: PX992002

enums:
  MemoryLossDriveImpactEnum:
    description: "Levels of impact on driving ability due to memory loss or brain fog."
    permissible_values:
      UNDEFINED_CODE:
        description: "Severely unable"
      UNDEFINED_CODE_1:
        description: "Moderately unable"
      UNDEFINED_CODE_2:
        description: "Mildly unable"
      UNDEFINED_CODE_3:
        description: "Able"
      UNDEFINED_CODE_4:
        description: "Not applicable"
```

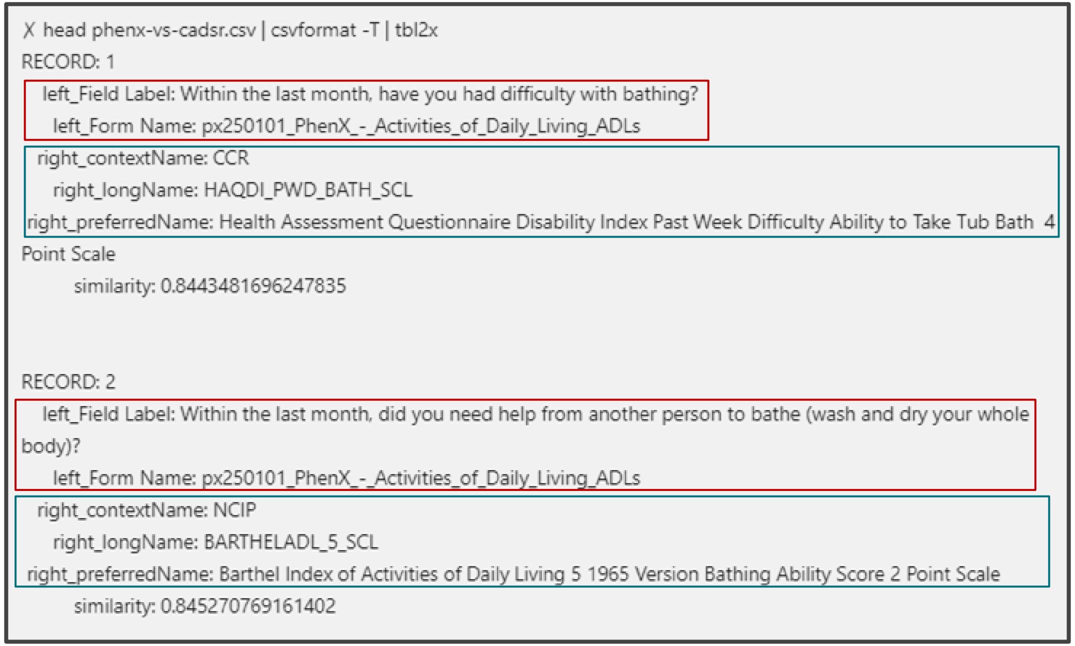
Example: CurateGPT

- Semantic similarity
- Mapping terms to ontologies
- Mapping CDEs



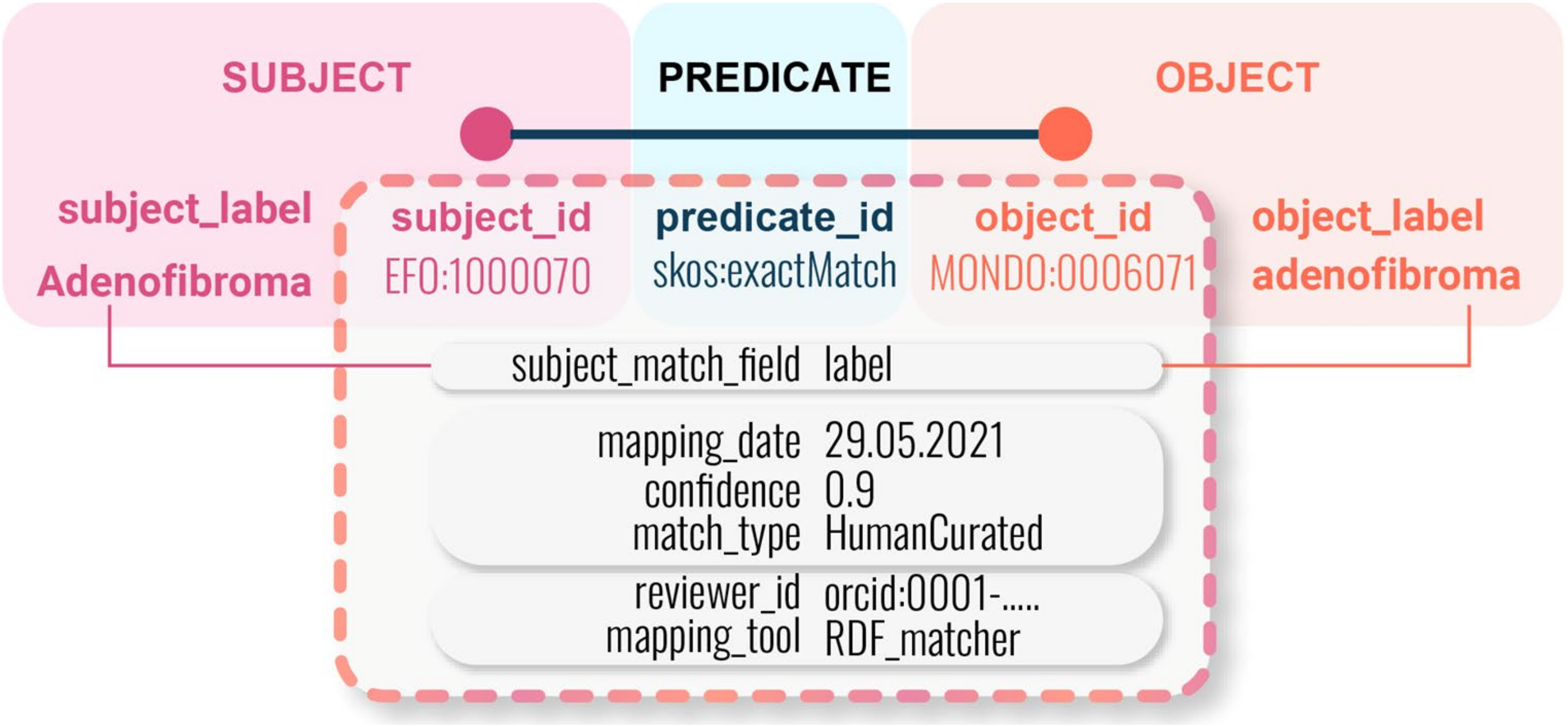
```
* CLASS blood pressure (general concept, OBA)
  * CLASS blood pressure after eating breakfast (cross-CDE concept)
     * CLASS blood pressure after eating breakfast (CRDC)
        * SLOT blood pressure systolic FLOAT...
     * CLASS blood pressure after eating breakfast (phenx)
```

Left: PhenX Right: caDSR



```
X head phenx-vs-cadsr.csv | csvformat -T | tbl2x
RECORD: 1
  left_Field Label: Within the last month, have you had difficulty with bathing?
   left_Form Name: px250101_PhenX_-_Activities_of_Daily_Living_ADLs
 right_contextName: CCR
   right_longName: HAQDI_PWD_BATH_SCL
right_preferredName: Health Assessment Questionnaire Disability Index Past Week Difficulty Ability to Take Tub Bath  4
Point Scale
      similarity: 0.8443481696247835


RECORD: 2
  left_Field Label: Within the last month, did you need help from another person to bathe (wash and dry your whole
body)?
   left_Form Name: px250101_PhenX_-_Activities_of_Daily_Living_ADLs
 right_contextName: NCIP
   right_longName: BARTHELADL_5_SCL
right_preferredName: Barthel Index of Activities of Daily Living 5 1965 Version Bathing Ability Score 2 Point Scale
      similarity: 0.845270769161402
```

Example: SSSOM Mapping Model



SUBJECT

PREDICATE

OBJECT

subject_label
Adenofibroma

subject_id
EFO:1000070

predicate_id
skos:exactMatch

object_id
MONDO:0006071

object_label
adenofibroma

subject_match_field label

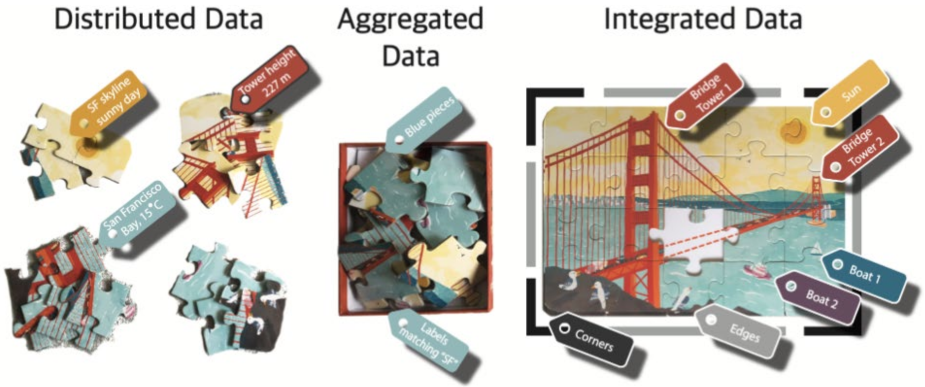
mapping_date 29.05.2021
confidence 0.9
match_type HumanCurated

reviewer_id orcid:0001-.....
mapping_tool RDF_matcher

A Simple Standard for Sharing Ontological Mappings https://doi.org/10.1093/database/baac035

Conclusions

- CDEs are not computable and that reduces interoperability
- We can update CDEs to make them more computable and interoperable
- Recently developed mapping standards and LLM-based tools now make this work tractable in years instead of decades
- Preliminary output can be used to develop a curation strategy
- The benefits in terms of increased data interoperability will be enormous



Acknowledgements

- Berkeley Bioinformatics Open-source Projects (BBOP)
- Translational and Integrative Science Lab (TISLab)
- Monarch Initiative
- LinkML developer community
- National COVID Cohort Collaborative
- NIH for funding so much of the work in this field

Funding:
NHGRI CEGS RM1 HG010860
NHLBI 1OT2HL167310