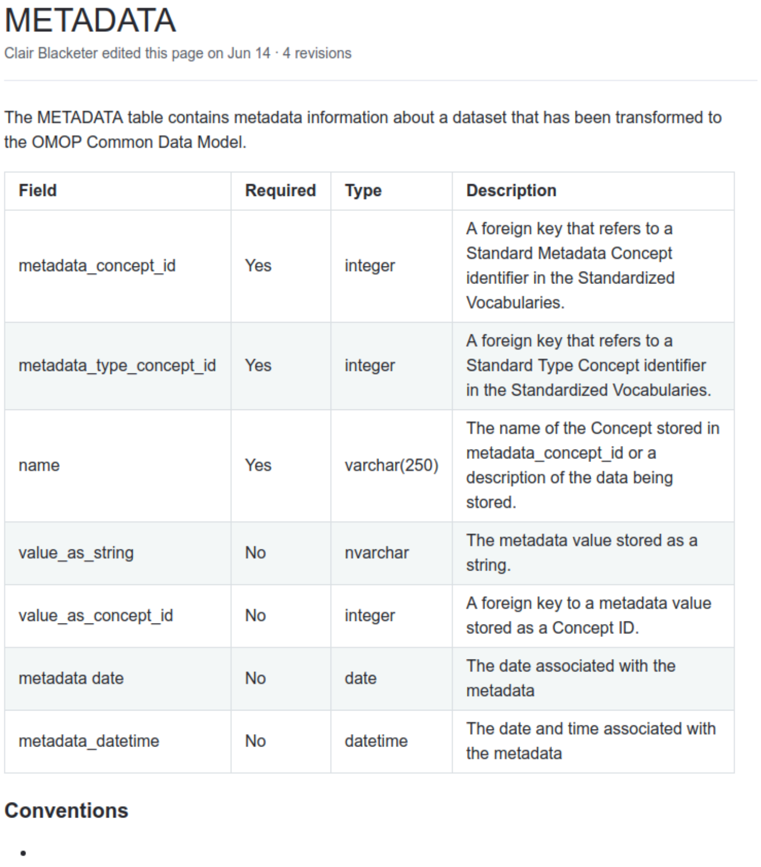
Update on Metadata and Annotations Work Group

Ajit Londhe

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Let’s Get Meta



METADATA

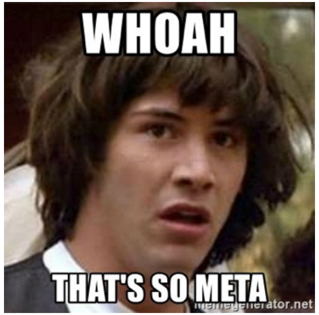
Clair Blacketer edited this page on Jun 14 · 4 revisions

The METADATA table contains metadata information about a dataset that has been transformed to the OMOP Common Data Model.

Field	Required	Type	Description
metadata_concept_id	Yes	integer	A foreign key that refers to a Standard Metadata Concept identifier in the Standardized Vocabularies.
metadata_type_concept_id	Yes	integer	A foreign key that refers to a Standard Type Concept identifier in the Standardized Vocabularies.
name	Yes	varchar(250)	The name of the Concept stored in metadata_concept_id or a description of the data being stored.
value_as_string	No	nvarchar	The metadata value stored as a string.
value_as_concept_id	No	integer	A foreign key to a metadata value stored as a Concept ID.
metadata date	No	date	The date associated with the metadata
metadata_datetime	No	datetime	The date and time associated with the metadata

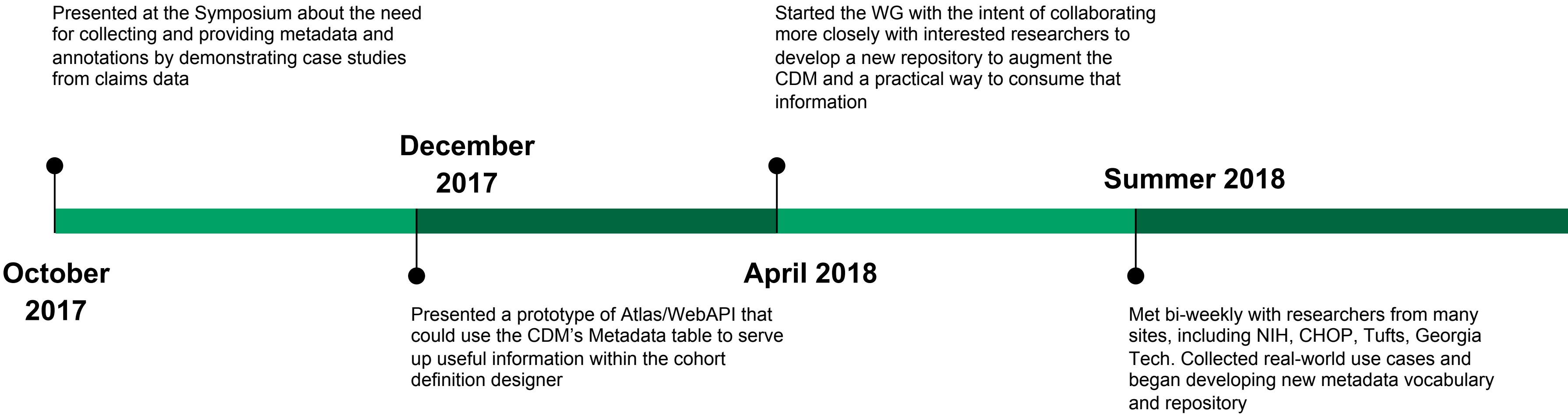
Conventions

-



1. **What is this table?**
   a. An early attempt at providing a space to land metadata about a CDM
2. **Where did it come from?**
   a. A proposal from Huser, Londhe, and Voss
3. **How should it get populated?**
   a. Manually by CDM data custodians
4. **When was it last changed?**
   a. June 2017
5. **How much utilization does it get?**
   a. Admittedly, not much. It’s probably missing a lot of useful information for most sites

The Journey since the Metadata table



October 2017

Presented at the Symposium about the need for collecting and providing metadata and annotations by demonstrating case studies from claims data

December 2017

Presented a prototype of Atlas/WebAPI that could use the CDM’s Metadata table to serve up useful information within the cohort definition designer

April 2018

Started the WG with the intent of collaborating more closely with interested researchers to develop a new repository to augment the CDM and a practical way to consume that information

Summer 2018

Met bi-weekly with researchers from many sites, including NIH, CHOP, Tufts, Georgia Tech. Collected real-world use cases and began developing new metadata vocabulary and repository

Goals and Deliverables

Goals

- Our goal is to define a standard process for storing human- and machine-authored metadata and annotations in the Common Data Model to ensure researchers can consume and create useful data artifacts about observational data sets.

Deliverables

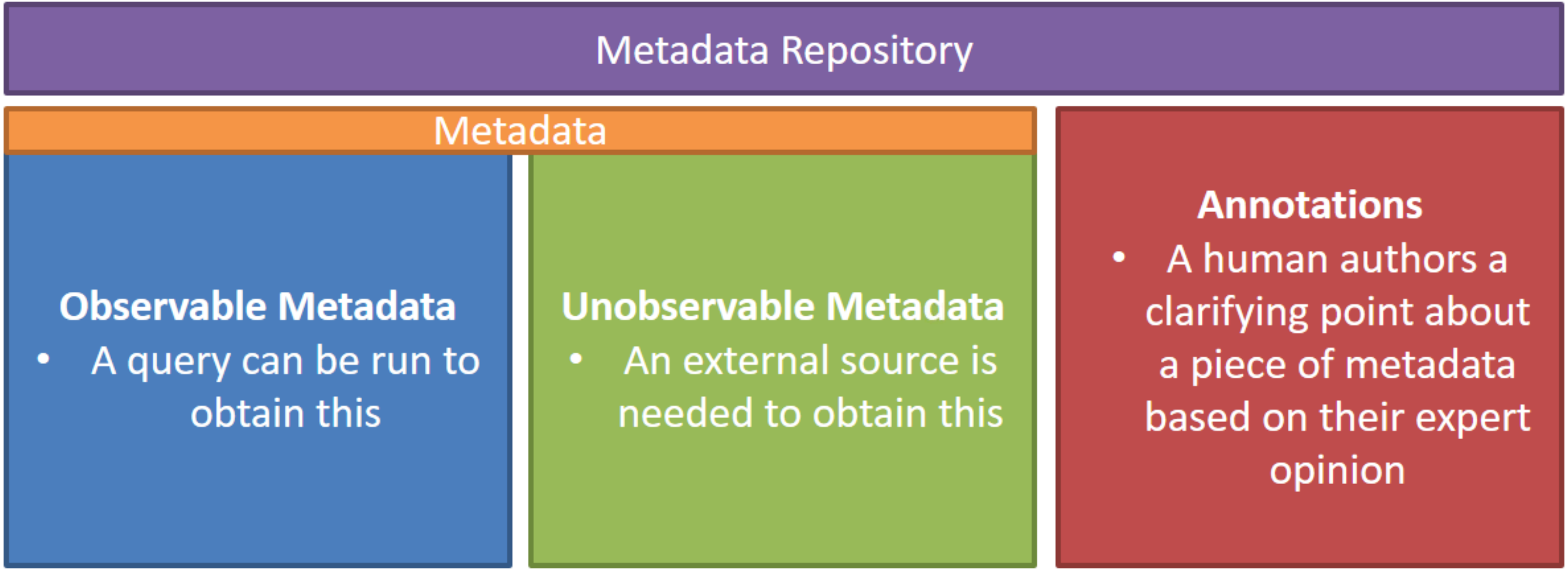
- We will design structures for metadata and annotations, construct algorithms for identifying potential metadata opportunities, and create requirements for new Atlas and WebAPI enhancements that can allow for consumption and maintenance of metadata and annotations.

What are “Metadata” and “Annotations”

Metadata is information that can be directly observed, indirectly inferred, or externally obtained about an observational dataset that provides us with a more complete understanding of the dataset.

Annotations are notes about metadata authored by those with relevant experience or expertise that are intended to improve study design for other researchers.

How do we delineate between Metadata and Annotation?



Metadata Repository

Metadata

Observable Metadata

- A query can be run to obtain this

Unobservable Metadata

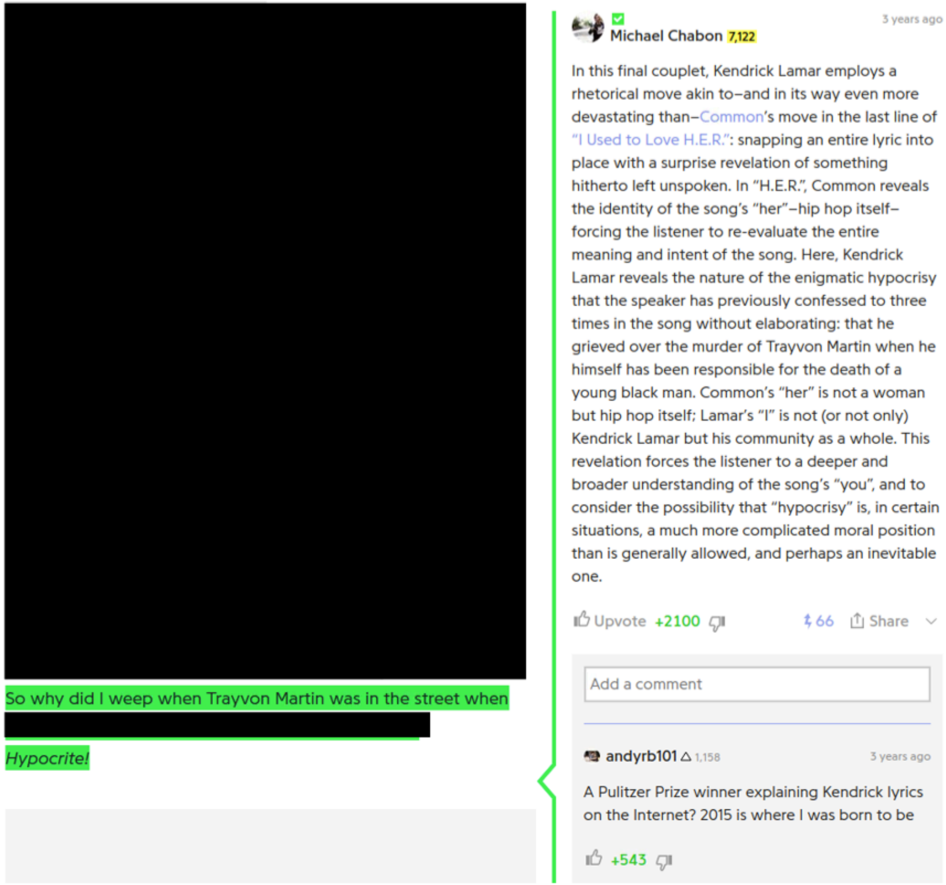
- An external source is needed to obtain this

Annotations

- A human authors a clarifying point about a piece of metadata based on their expert opinion

A fun way to think about Annotations

Genius.com, a site where song lyrics are annotated by the community….and sometimes the artists themselves or…..Pulitzer Prize winners?



[redacted]

So why did I weep when Trayvon Martin was in the street when

[redacted]

Hypocrite!

Michael Chabon 7,122

3 years ago

In this final couplet, Kendrick Lamar employs a rhetorical move akin to–and in its way even more devastating than–Common's move in the last line of "I Used to Love H.E.R.": snapping an entire lyric into place with a surprise revelation of something hitherto left unspoken. In "H.E.R.", Common reveals the identity of the song's "her"–hip hop itself–forcing the listener to re-evaluate the entire meaning and intent of the song. Here, Kendrick Lamar reveals the nature of the enigmatic hypocrisy that the speaker has previously confessed to three times in the song without elaborating: that he grieved over the murder of Trayvon Martin when he himself has been responsible for the death of a young black man. Common's "her" is not a woman but hip hop itself; Lamar's "I" is not (or not only) Kendrick Lamar but his community as a whole. This revelation forces the listener to a deeper and broader understanding of the song's "you", and to consider the possibility that "hypocrisy" is, in certain situations, a much more complicated moral position than is generally allowed, and perhaps an inevitable one.

andyrb101 △ 1,158

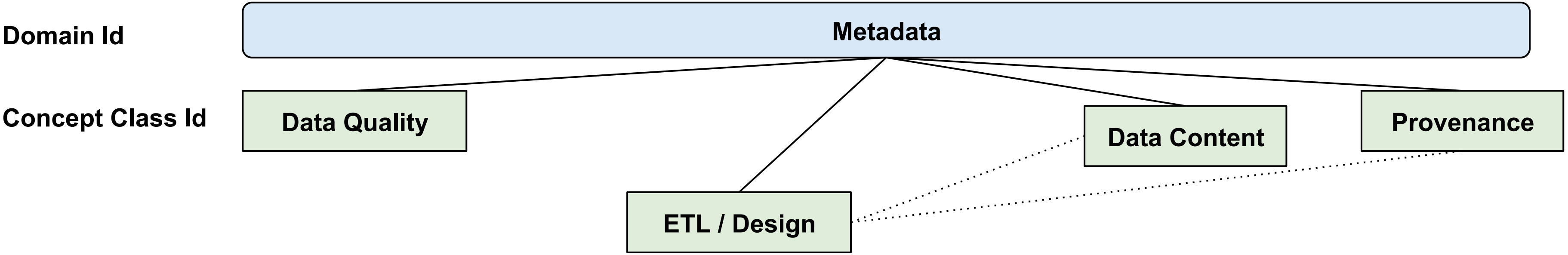
3 years ago

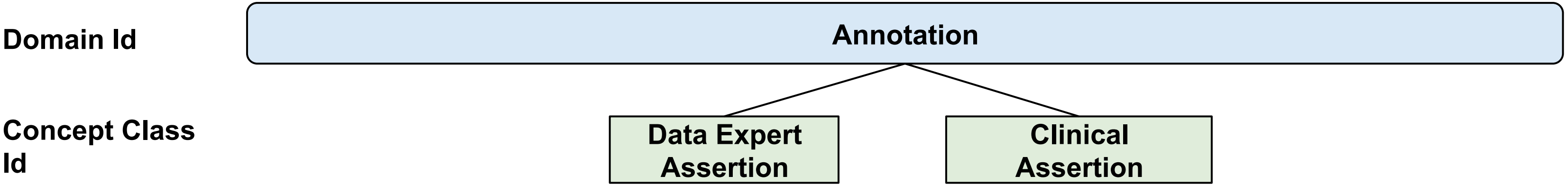
A Pulitzer Prize winner explaining Kendrick lyrics on the Internet? 2015 is where I was born to be

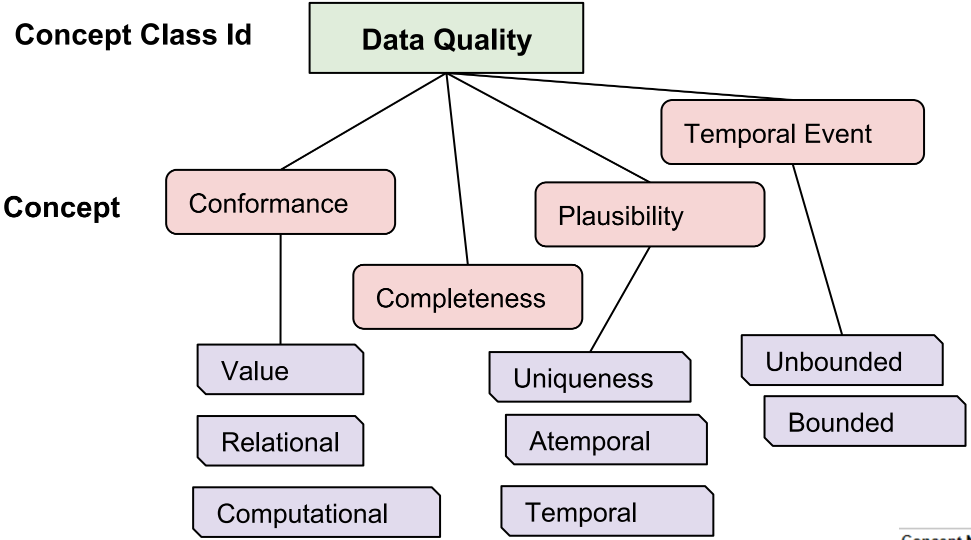
Examples from the WG

- **Data Quality**
  - Achilles Heel: ERROR: 101-Number of persons by age, with age at first observation period; should not have age < 0
  - In November 2011, the Social Security Administration stopped including death information whose source was solely state-level records.
  - In October 2015, US Claims records transitioned from ICD9CM to ICD10CM and ICD9Proc to ICD10PCS
- **Source Provenance**
  - Data come from observational trial, hence there are not life time data. They span only 2 years.
  - Dataset is derived from patients in clinical trials, patients with claims only, and patients with claims/EHR/cancer registry
- **ETL/Design**
  - Visit dates are inferred. (imputed)
  - Data after age 90 were deleted (due to policy)
  - Data was shifted by -+7 days and date-shift revealing events were redacted (fully deleted)
  - The Ambulatory and Other Ambulatory visits are difficult to disambiguate. We have standardized definitions for each type of visit. The 9202 visit is a face-to-face visit while the Other Ambulatory visit are administrative. Transfusion and radiology visits are still 9202 but lab visits are Other Ambulatory.
  - In order to standardize data more efficiently, we made a decision to not follow OHDSI mappings for concepts that are mapped to measurement but do not have an actual result or value associated with it. An example would be something like concept_id = 45553744, with the concept_name = 'Elevated blood glucose level'. In designing the database, these concepts that appear to be metadata about a lab and not the actual lab, should be rerouted to either Observation or Condition.
- **Data Content**
  - PAD phenotype from Mayo Clinic identified patient to have confirmed case of PAD, however, clinician disagreed based on patient profile case adjudication

Concept Hierarchies







Concept Class Id

Data Quality

Concept

Conformance

Completeness

Plausibility

Temporal Event

Value

Relational

Computational

Uniqueness

Atemporal

Temporal

Unbounded

Bounded

Type Concept

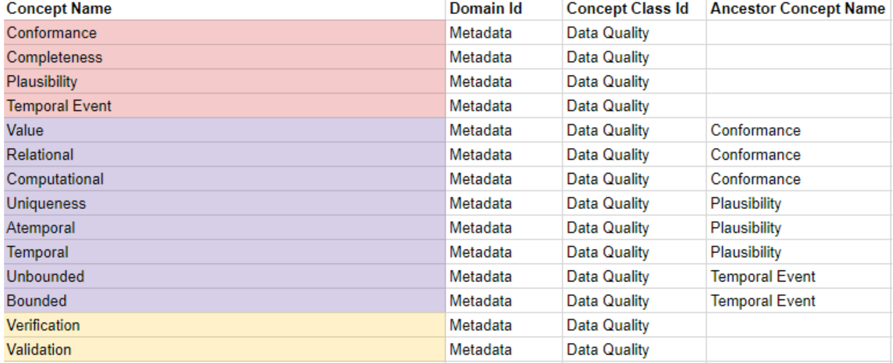
Verification

Validation

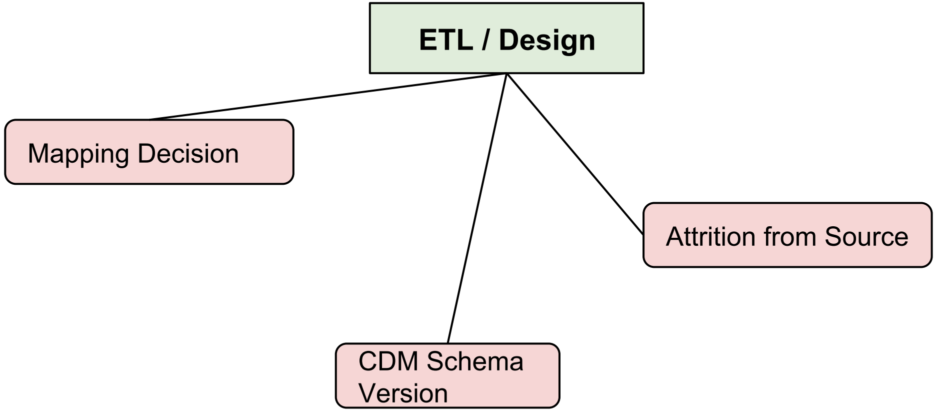
Data Quality concept hierarchy:

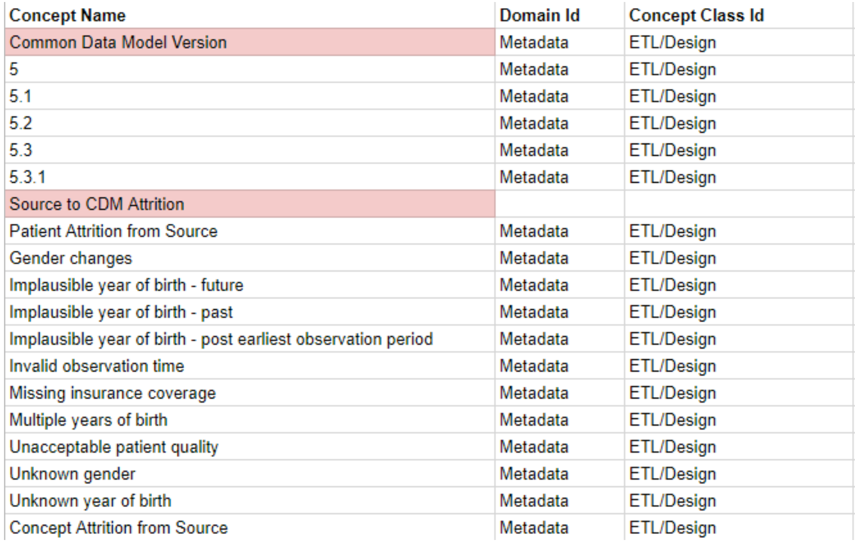
Based on Kahn paper in order to use a standard vision of DQ that has been adopted by OHDSI sites already.

One tweak: addition of temporal events that are either unbounded (point in time) or bounded (have a start and end).



Concept Name	Domain Id	Concept Class Id	Ancestor Concept Name
Conformance	Metadata	Data Quality	
Completeness	Metadata	Data Quality	
Plausibility	Metadata	Data Quality	
Temporal Event	Metadata	Data Quality	
Value	Metadata	Data Quality	Conformance
Relational	Metadata	Data Quality	Conformance
Computational	Metadata	Data Quality	Conformance
Uniqueness	Metadata	Data Quality	Plausibility
Atemporal	Metadata	Data Quality	Plausibility
Temporal	Metadata	Data Quality	Plausibility
Unbounded	Metadata	Data Quality	Temporal Event
Bounded	Metadata	Data Quality	Temporal Event
Verification	Metadata	Data Quality	
Validation	Metadata	Data Quality	

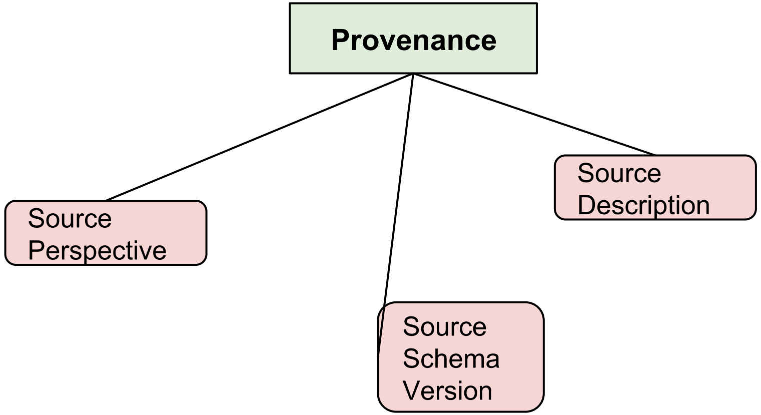




Concept Name	Domain Id	Concept Class Id
Common Data Model Version	Metadata	ETL/Design
5	Metadata	ETL/Design
5.1	Metadata	ETL/Design
5.2	Metadata	ETL/Design
5.3	Metadata	ETL/Design
5.3.1	Metadata	ETL/Design
Source to CDM Attrition		
Patient Attrition from Source	Metadata	ETL/Design
Gender changes	Metadata	ETL/Design
Implausible year of birth - future	Metadata	ETL/Design
Implausible year of birth - past	Metadata	ETL/Design
Implausible year of birth - post earliest observation period	Metadata	ETL/Design
Invalid observation time	Metadata	ETL/Design
Missing insurance coverage	Metadata	ETL/Design
Multiple years of birth	Metadata	ETL/Design
Unacceptable patient quality	Metadata	ETL/Design
Unknown gender	Metadata	ETL/Design
Unknown year of birth	Metadata	ETL/Design
Concept Attrition from Source	Metadata	ETL/Design

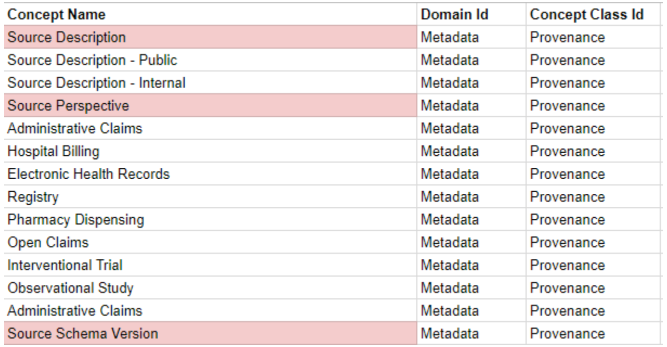
ETL/Design:

1. Decisions made by the data custodian in order to map the native data into the CDM
2. Information about the CDM schema itself (version number, deviations from the spec)
3. Quantifying the ways in which we drop patients or events from the native data

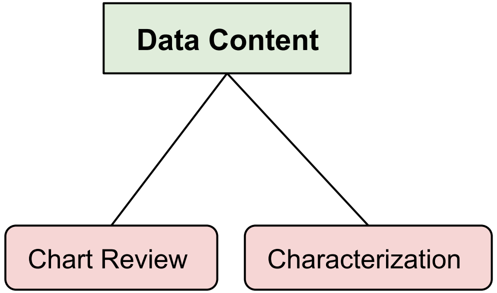


Provenance:
Information about where the native data comes from, its versioning, what kinds of system(s) provided the data.

Could replace CDM_SOURCE.



Concept Name	Domain Id	Concept Class Id
Source Description	Metadata	Provenance
Source Description - Public	Metadata	Provenance
Source Description - Internal	Metadata	Provenance
Source Perspective	Metadata	Provenance
Administrative Claims	Metadata	Provenance
Hospital Billing	Metadata	Provenance
Electronic Health Records	Metadata	Provenance
Registry	Metadata	Provenance
Pharmacy Dispensing	Metadata	Provenance
Open Claims	Metadata	Provenance
Interventional Trial	Metadata	Provenance
Observational Study	Metadata	Provenance
Administrative Claims	Metadata	Provenance
Source Schema Version	Metadata	Provenance



Data Content:

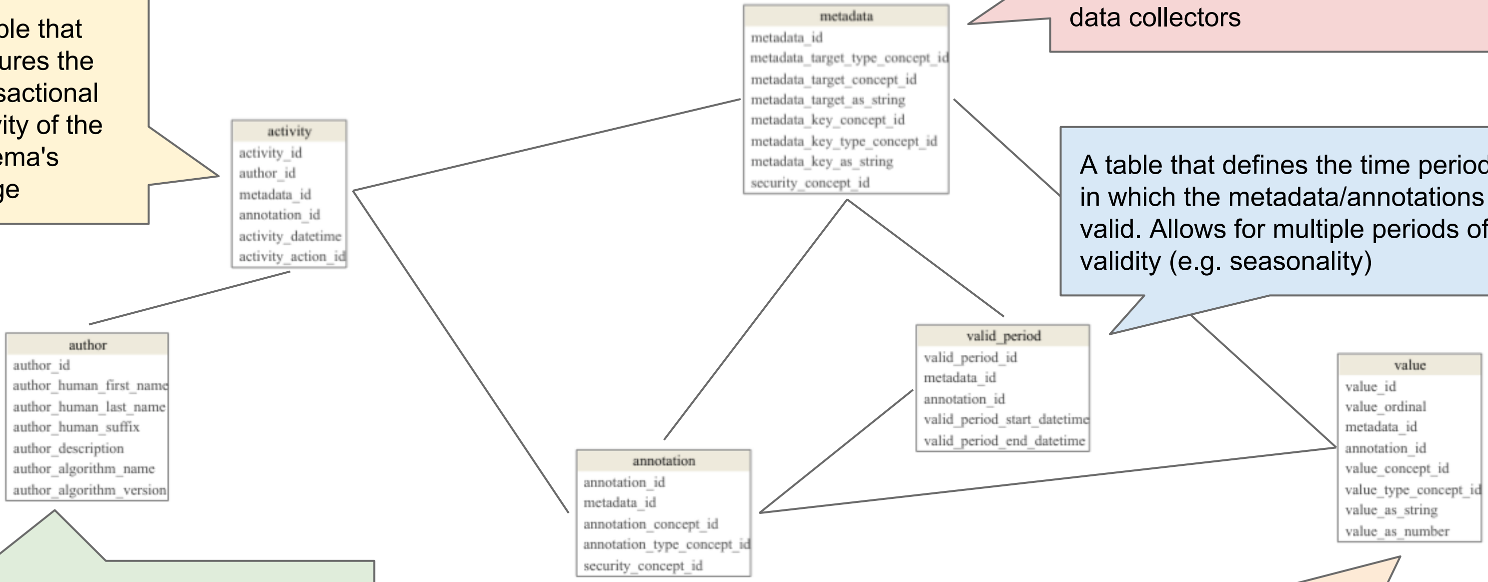
Specific pieces of information about data within the CDM schema. Patient chart review, phenotype performance, characterization of a cohort.

Isn’t there a Chart Review WG?

Collaboration with Chart Review WG

- As the Chart Review WG is further along with their deliverables, they will be creating their own application tables to be stored within the WebAPI repository and, for now, storing their data in a custom set of tables
- However, we have been reviewing the application and the draft Metadata schema and we feel confident that the Chart Review application can be refactored to store its questions and answers in the CDM Metadata schema
- One key need from the Chart Review WG: tracking authorship
  - *Elena MD, PhD, Regulator at FDA;* Elena has a background in internal medicine and has been working at the FDA for 20 years. She is supportive of advancing the quality of real-world evidence-based analytics to improve health safety. She must ensure an extremely high level of rigor in the studies that she uses as evidence in her regulatory work. Elena is interested in the potential of research networks like OHDSI.

Metadata Schema



A table that captures the transactional activity of the schema's usage

activity
activity_id
author_id
metadata_id
annotation_id
activity_datetime
activity_action_id

author
author_id
author_human_first_name
author_human_last_name
author_human_suffix
author_description
author_algorithm_name
author_algorithm_version

A table that captures the author of the metadata/annotation records. Used only when (1) Shiro is not enabled or (2) Shiro is enabled, but algorithms are being used to populate the metadata table

metadata
metadata_id
metadata_target_type_concept_id
metadata_target_concept_id
metadata_target_as_string
metadata_key_concept_id
metadata_key_type_concept_id
metadata_key_as_string
security_concept_id

A table for capturing metadata, which we define as objective facts about the CDM database or its usage that can be observed through query or obtained from data collectors

valid_period
valid_period_id
metadata_id
annotation_id
valid_period_start_datetime
valid_period_end_datetime

A table that defines the time period(s) in which the metadata/annotations are valid. Allows for multiple periods of validity (e.g. seasonality)

annotation
annotation_id
metadata_id
annotation_concept_id
annotation_type_concept_id
security_concept_id

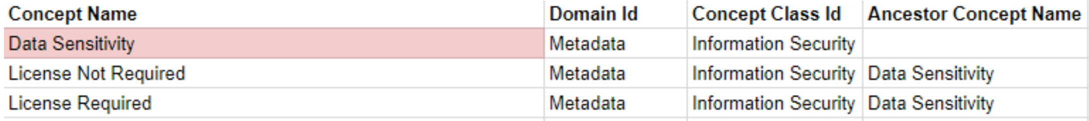
A table for capturing annotations, which we define as subjective assertions about record(s) in Metadata from subject matter experts

value
value_id
value_ordinal
metadata_id
annotation_id
value_concept_id
value_type_concept_id
value_as_string
value_as_number

A table that is used to capture values associated with metadata/annotation record(s). Values can be represented in various formats and can be ordered using the value_ordinal field

A Note about Data Sensitivity

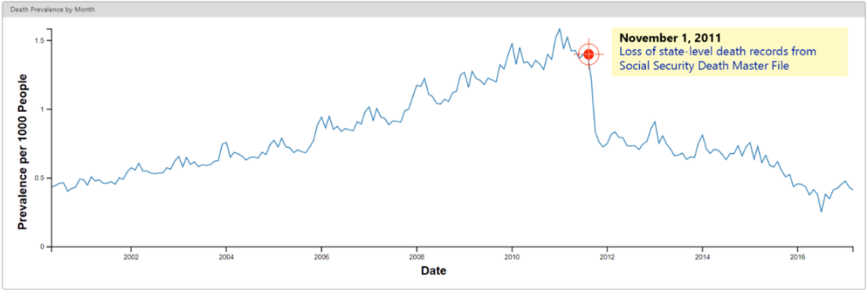
Each piece of metadata or annotation should be tagged with a security concept that indicates whether it can be shared with those without a license and whether it can be kept even after the license expires.



Concept Name	Domain Id	Concept Class Id	Ancestor Concept Name
Data Sensitivity	Metadata	Information Security	
License Not Required	Metadata	Information Security	Data Sensitivity
License Required	Metadata	Information Security	Data Sensitivity

Future Considerations

- Kronos integration
  - Store results on time series analyses and allow data custodians to provide annotations on each finding
- Migration of Achilles results into the CDM Metadata schema
  - Achilles is classic metadata, why keep it separate?



November 1, 2011
Loss of state-level death records from Social Security Death Master File

Kronos could identify this structural break, Metadata schema could hold this DQ record and a suggestion in the annotations table

- Metadata repositories that reside at a site and network level
  - Each site could collect metadata that is stored within their WebAPI repository
  - Each site could submit metadata about their dataset that is allowed to be shared into an OHDSI Community repository (e.g. Truven CCAE is known to have ICD9CM to ICD10CM concept instability starting in October 2015)

What’s Next?

- Finish development of new concepts to submit to Vocabulary team
- Lee Evans has provided us with a public Postgres instance, WG members will use this to test their Metadata use cases
- WebAPI development to support SQL operations to the CDM Metadata schema (volunteers welcome)
- Atlas development to provide a User Interface (Atlas UI wizards welcome)
- Development of a SQL library for non-Atlas users to be able to execute standard Metadata workflows

Thanks to the WG team

- Andrew Williams
- Vojtech Huser
- Yurang Park
- Michael Gurley
- Hanieh Razzaghi
- Michael Kahn
- Jon Duke
- Robert Miller