AI-READi Data Standards Workstream

Data Core Team

(Tools team + Data Standards team)

Data Standards Group

11/16/2023
Stephanie Hong, Yvette Chen, James Cavallon, Monique Bangudi, Jessica Mitchell
Dr. Chute's Lab

11/17/2023
Updated to include OMOP vocabulary extension process for AI-READI

3/13/2024
WBS, Workflow, EHR DI&H OMOP pipeline overview revisited

Overall Work breakdown Structure (WBS)

- Overall Data Core Tasks (JHU involvement in RED)

OMOPfying AI-READi datasets

REDCap
Survey
Survey/Vision/Labs/
biospecimens/
Environmental exposure

mappable

Not mappable
(20%)
Handle Via OMOP vocab extension if possible (and not use the local Vocab Extension

MOCA Score Data

How to Map?
Need to create OMOP Extension to handle MOCA Extension via OMOP Vocab Issue PR?

EHR - OMOPfied

EHR Data from the Sites merge into OMOP (UAB, UCSD, UW) – year 2 / JHU begin pipeline work Aug 2024

REDCap

MOCA

OMOP Merge of EHR data

LDS (de-id)

Is this required ? SafeHarbor (with date shifting/ 3digit -zip

Data Quality/ Unit Testing

REDCap Survey Data
EHR Data
MOCA Data
Demographics Age range data metrics

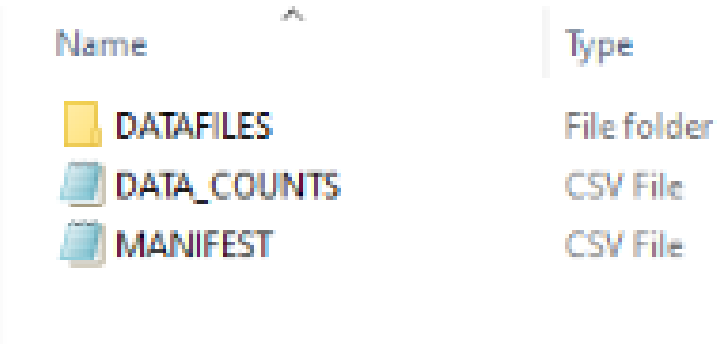
** DICOM IMAGING DATA INTO OMOP IMAGING EXTENSION CDM TBD **

EHR Data zip file submission zip file format

- The data file name should have the following format: <abbreviated_sitename>_OMOP_<mmddyyyy>.zip

  The content should include the Manifest table along with the datafiles and data counts.

- For example: JHU_OMOP_11162023.zip parent directory structure



Manifest table

- With each data payload manifest table containing one row of data is submitted. The manifest table contains metadata about the payload.

- Manifest table is described here ->

Field name	Definition	Sample value	Comment
SITE_ABBREV	Unique abbreviation for your site; will be provided by AI-READi	“UCSD”	Static
SITE_NAME	Full name of your site	“University of Cal, San Diego”	Static
CONTACT_NAME	Full name of N3C technical contact at your site	“Jane Doe”	Static
CONTACT_EMAIL	Email address of technical contact at your site	“jane_doe@ohdsi.org”	Static
CDM_NAME	CDM model – only OMOP CDM is used	“OMOP”	Static
CDM_VERSION	Numbered version of CDM	“5.3.1”	Static
VOCABULARY_VERSION	Version of OMOP vocabulary in use for this data pull.	"v5.0 19-AUG-23"	Will change if you update your vocabulary tables at your site
SHIFT_DATE_YN	Enter Y if your site is shifting dates prior to submission, otherwise enter N. Note: Date shifting is **not** required and, indeed, it is preferred for sites **not** to date shift prior to submission.	“Y”	Static
MAX_NUM_SHIFT_DAYS	The maximum number of days that you are shifting dates. Write Unknown if you do not know and NA if you do not shift dates.	“30”	Static
RUN_DATE	Date the current extract was run.	“2020-05-05”	Changing (use SYSDATE)
DATA_UPDATE_DATE	Date for which the data in this extract is current (i.e., the maximum date present in your dataset)	“2020-05-04”	Changing (use SYSDATE - # days latency at your site)
NEXT_SUBMISSION_DATE	Date on which you will submit your next extract	“2020-05-07”	Changing (use SYSDATE + # days between submissions)

Datafiles subdirectory structure – omop domain data

- For example, UCSD_OMOP_08162024.zip >
  >> DATAFILES subdirectory should contain all of the OMOP domain files

i.e.

- care_site.csv
- condition_era.csv
- Condition_occurrence.csv
- Death.csv
- Device_exposure
- Dose_era.csv
- Drug_era.csv
- Drug_exposure.csv
- Location.csv
- Measurement.csv
- Observation.csv
- Observation_period.csv
- Person.csv
- Procedure_occurrence.csv
- Provider.csv
- Visit_occurrence.csv

Example of Datacounts.csv

TABLE_NAME String	ROW_COUNT String
OBSERVATION_PERIOD	71409
VISIT_DETAIL	288600
DRUG_EXPOSURE	3635705
PROCEDURE_OCCURRENCE	1879836
OBSERVATION	12947361
LOCATION	66386
PROVIDER	16108
NOTE	0
PERSON	71409
VISIT_OCCURRENCE	1905204
CONDITION_OCCURRENCE	12696818
DEVICE_EXPOSURE	175786
MEASUREMENT	54958997
DEATH	2070
CARE_SITE	401
DRUG_ERA	999815
NOTE_NLP	0
CONDITION_ERA	5115801

Some more details about the data file format

- Column heading should be in the first row of the csv file
- The data should be in quotes and delimited by pipe|
- Place quotes around the “data” and use the | delimiter to accommodate those data that can contain delimiters like “|” in the text fields.
- For example:

  Condition_occurrence_id|person_id|condition_concept_id|condition_start_date|condition_end_date

  “100”|”1”|”4332245”|”2023-10-04”|”2023-10-30”

OMOP VOCABULARY EXTENSION FOR AI-READI DATA ELEMENTS

- AI-READi vocabulary – Custom extension to support AI-READi data elements is used for the survey data ingestion

vocabulary_id	vocabulary_name	vocabulary_reference	vocabulary_version	vocabulary_concept_id
AI-READi	AI-READi	AI-READi generated	v5.0 31-AUG-23	1
MOCA Extensions	MOCA extensions	AI-READi generated or OMOP generated	v5.0 31-AUG-23	2

AI-READi extension created to support AI-READi survey instruments data elements and MOCA concepts

- AI-READi concepts in the concept table:
- Concept_id – 2 billion range ids
- Concept_name
- Domain_id
- Vocabulary_id
- Concept_class_id
- Standard_concept
- Concept_code
- Valid_start_date
- Valid_end_date
- Invalid_reason

AI-READi Vocabulary Extension

- Example

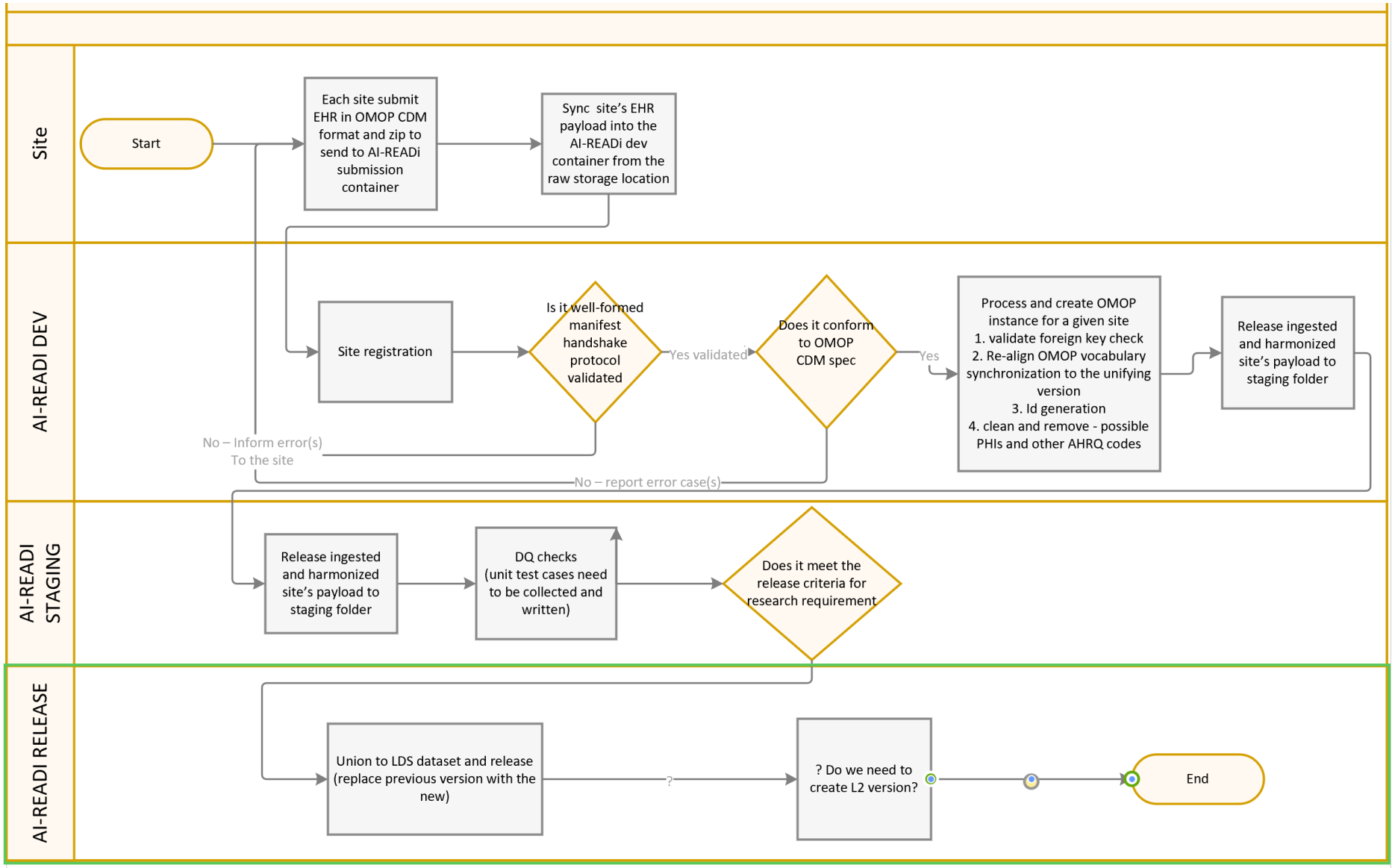
concept_id	concept_name	domain_id	vocabulary_id	subvocabulary_id	concept_code	predicate_id	predicate_label
2005200000	Recruitment Survey Started Timestamp (from REDCap)	Observation	AIREADI	recruitment_survey	99423-6	skos:exactMatch	Maps to
2005200001	Recruitment Survey Completed Timestamp (from REDCap)	Observation	AIREADI	recruitment_survey	99423-6	skos:exactMatch	Maps to
2005200007	Do you use lifestyle changes to control your A1C and blood glucose levels? Examples: regular exercise, avoiding sugary foods and beverages, eating a balanced diet with lots of vegetables, sticking to a consistent eating schedule	Observation	AIREADI	screening_survey		skos:exactMatch	Maps to
2005200008	What is your ancestry or ethnic origin?(Examples: Italian, Jamaican, African American, Cambodian, Cape Verdean, Cherokee, Navajo, Nez Pearce, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)	Observation	AIREADI	screening_survey		skos:exactMatch	Maps to

mapping table explained

- Adopted from N3C

columns	description
DATA_SOURCE	It name of the data source refers to the source data Common Data Model abbreviated name, i.e. OMOP, ACT, TriNetX, or PCORnet.
TBL_NAME	CDM table name where specifid value is found in the data source
TABLE_COLUMN_NAME	column name within the table list in the CDM_TBL_NAME column where the data values is found.
SRC_CD	permissible data values, it is often an enumerated list of permissble values found in the source data.
SRC_CD_DESCRIPTION	Description of the value listed in the SRC_CD column.
TARGET_CONCEPT_ID	This is the standardized concept identifier that represents a clinical event, measurement, observation, drug or procedure in OMOP CDM. The target_concept_id is a field used to represent the concept that is the focus or target of a particular record in the source database.
TARGET_CONCEPT_NAME	The text description associated with the target_concept_id
TARGET_DOMAIN_ID	The domain id represents the various tables in the OMOP CDM. It is used to categorize and classify data into different domains based on the type of healthcare information it represents. In essence, this field referes to the table where the target concept should be inserted into. It serves as a way to codify and standardize the clinical content of the data in a consistent manner.
TARGET_VOCABULARY_ID	The vocabulary id represents the standardized vocabulary or the terminology from which the concept code is derived. The source of the concept code in the source data in the SRC_CD column is often terminology based coded value. Some commonly used vocabularies in the source data include: SNOMED, RxNorm, LOINC, ICD10CM, ICD0PCS, CPT4, HCPCS and NDC. The vocabulary_id and the concept_code field uniquely identifies a concept within a specific vocabulary.
TARGET_CONCEPT_CLASS_ID	This field is used to categorize concepts based on their class or category. It helps to organize concepts into groups that share similar characteristics or properties.
TARGET_STANDARD_CONCEPT	The target_standard_concept column is used to identify the standard representation of a concept within a vocabulary. This column helps to distinguish between standard and non-standard. The possible values in this column can be S(standard), C(classification), D(Deprecated), or null.
TARGET_CONCEPT_CODE	This column represents the codified data element or the equivalent code found in the source data that is associated with the concept within a specific vocabulary. For example, SARS-CoV-2 (COVID-19) Ag [Presence] in Upper respiratory specimen by Rapid immunoassay laboratory test is equivalent to LOINC(Logical Observation Identifiers Names and Codes) 97097-0. The corresponding codified value of the concept in source_cd column LOINC code 97097-0 is represented in this field.
TARGET_TBL_COLUMN_NAME	used as a reference to specify the location of data element in the source data.
PREDICATE_ID	It defaults to omoprel:maps_to. Currently, not generated. However, it is used to describe the mapping relationship between the source data and the OMOP target concept. The OMOP relationship of "Maps to" is used, which can be either skos:exactMatch or skos:broadMatch. The common mapping predicates are explained here.
PREDICATE_label	It defaults to "Maps to".

EHR OMOP CDM payload ingestion & harmonization (ai-readi-ehr-DI&H)



High-level Pipeline steps

1. Validate manifest handshake protocol
2. Perform CDM conformance check – i. e. foreign key checks, person_id check, and encounter_id check
3. Clean – Remove PHI fields and “Never” events if found
4. Re-align OMOP concept_ids using the latest unifying version of the OMOP vocabulary table downloaded from the Athena
5. Regenerate domain_ids using registered data_partner_id and mark all rows with data_partner_id such that we can replace site’s data with subsequent updates

AI-READi EHR OMOP pipeline DI&H staging area

- Databricks EHR-OMOP pipeline dev/ prod environment