CCDH Internal Monthly Report - 2021-03-17

2021-03-17
Data Model Harmonization Workstream Update

1. **Progress over the last month:**
    a. Continued work on the remaining Administrative and Biospecimen subdomains entities including
        i. Consider the abstract real-world entities that need to be represented and then refactor ADM entities into CDM entities that align to that real-world model (e.g. Program, Project, Case); an attempt at a more top-down as opposed to a bottom-up approach
        ii. Pulling instance data from PDC and GDC to examine how administrative entity data on cases that appear in both DCs align or misalign
    b. Began work on Clinical subdomain with initial focus on diagnostic attributes (e.g. staging, grading, risk classification of disease) and ways to collapse duplicative attributes into a more parsimonious structure
        i. Multiple staging systems (FIGO, INSS, INRG, COG renal, COG liver, etc.) currently represented as individual variables might be refactored into a simpler complex data type for staging that includes Stage and Staging System
    c. Continued review of experiment-related entities in GDC to identify common data structures underpinning a lot of the surface variability (e.g. an abundance of ‘workflow’ entities) and how those align conceptually with structures in PDC
    d. Brian met with Gilberto and Denise Warzel on 3/11 to discuss model change processes and best practices in caDSR
        i. model updates are received pretty stochastically dependent upon the model owner; some provide regular updates, others more ad hoc.
        ii. changes to caDSR are noted in a template which records changes to classes and properties
        iii. eventually these get loaded into the caDSR and versioned
        iv. rather than rely upon nodes to supply us with manually curated lists of changes, it would be preferable to pull in the latest models from the node repositories and represent these models in a common modeling language like LinkML
            1. perhaps an area where new tooling could help
    e. Smita and her team from Samvit came to the data model harmonization meeting on 3/15 to discuss their FHIR work
        i. Previous work with CCDH was in mapping to FHIR v4, but the new project is specifically targeting the US Core IG and identifying research gaps in that

            ii. CCDH will supply ADM and CDM artifacts to Samvit along with a prioritized list of entities to map (from the CCDH perspective)

2. **Timeline check**: what's the status of your work compared to our project milestones?
    a. Need to review and prioritize efforts
        i. Deliverables listed in the contract are under-specified and in some cases may no longer be priority. For example, "2b1h: Develop and extend CRDC-H for use in the primary CRDC nodes" needs a lot of unpacking and has dependencies on node effort. What is the best forum in which to review these?

3. **Work for the upcoming months**:
    a. Cross-workstream collaboration with terminologies and tools
        i. Meeting scheduled for 3/19 with data model and terminologies to discuss a number of cross-cutting topics including representation of terminological bindings in LinkML given the CodeableConcept class we've chosen to use thus far
        ii. Future meeting to discuss LinkML instance data and validation of instance data
    b. Finalize work on the remaining Administrative and Biospecimen subdomains entities including
    c. Incorporate existing CDM entities into the CRDC-H MVP (e.g. SpecimenCreation, SpecimenProcessing, etc.); requires some further specification of these entities to make sure they conform with our existing modeling choices in the MVP
    d. Continued expansion of the CDM to include diagnostic and treatment entities
    e. Concierge work
        i. meeting with CDS to review minimum metadata standard scheduled for 3/17

2021-03-10

Terminology & Ontology Ecosystem Workstream Update.

1. **Progress over the last month**
    a. LinkML Model
        i. Implemented features to represent enumerated values, code sets, and code systems
        ii. Gave weekly talk/workshop on LinkML development and functions
        iii. Created examples on the Specimen analyte_type property using LinkML enum and code set definitions
            1. Who is the audience at these workshops and can you provide details on content? - They have been going on for 3 weeks and are recorded (contact Harold for links to recordings). The content is a

deep dive into the core functionality of LinkML, occurring on Friday afternoons.

2. Is there a recommended resource for an intro to LinkML? - Recent talk for Monarch Initiative given by Deepak, Moni can share if needed.
3. These weekly talks are aimed at developers.

b. TCCM (Terminology Core Common Model)
   i. Created a model (using LinkML) to capture the minimum information of concept references
   ii. Developed a REST API endpoint for retrieving the concept references based on the CURIEs and URIs
   iii. Created loaders to load ontologies into the service
   iv. Loaded ontologies such as NCIT and ICD-O Morphology into the service

c. Data Exploration
   i. Downloaded all the biospecimen and clinical instance data from the GDC portal. Analyzed certain fields such as the analyte type, analyte type id and their correlations in the analyte and aliquot data
   ii. Using the discoveries based on data to assist the modeling and terminology work.
      1. Should GDC model be used as the standard? Has not been optimized for all. - Alison (helped develop GDC model) says that model can be used as baseline, but it is not a standard to be used for all cases.
      2. Not intending to follow GDC model exactly, but want to understand data to create a model that makes sense. Looking at what is in the data to come up with a more accurate representation of our model in the valueset.

Notes:

- - CDA has approved content, format, formalization of our model. They are using the components from CRDC-H that are covered for their needs. They are also using elements from clinical setting that do not currently make part of CRDC-H model.
- Our terminological work is part of the CRDC-H model, and included in LinkML. CDA will be adopting the model.
- For MVP release they are taking the union of distinct values from those sets, we will be harmonizing to that. They won't be harmonizing to NCIt terms. So it will be post MVP.

2. **Timeline check**: what's the status of your work compared to our project milestones?
   a. Aiming to make more progress on model definitions and enumerated values/code sets in the rest of Phase II

3. **Work for the upcoming months**
   a. Continue to work on the LinkML model and TCCM

b. Starting to create LinkML-based enumerated values / code sets for CRDC-H models
c. Continue to collaborate with the tools and harmonization workstreams

---

2021-03-03

Meeting was cancelled in favour of GA4GH Connect Meeting.