CCDH Internal Monthly Report - 2021-05-19

2021-05-19

Tools & Data Quality Workstream Update

1. **Progress over the last month**:
    a. Lots of work on CRDC-H model workflows
        i. Migrated code for Google sheets -> LinkML conversion to the model repo: https://github.com/cancerDHC/ccdhmodel/pull/9
        ii. Update to latest LinkML code (BiolinkML -> LinkML); regenerated artifacts and docs: https://github.com/cancerDHC/ccdhmodel/pull/6
    b. Example data transformation workflow: https://github.com/cancerDHC/example-data
        i. example notebook: https://nbviewer.jupyter.org/github/cancerDHC/example-data/blob/0a983991cbc274a7fbf3121aa8ae10047549fa1a/cptac2-subject-09CO022/CDA%20example%20for%20subject%2009CO022.ipynb
    c. Have deployed local instance of OLS containing NCIt OBO Edition & NCIt Plus
2. **Timeline check**: what's the status of your work compared to our project milestones?
    a. On track to provide demonstration data transformation and validation workflows for GDC/PDC data.
3. ***Work for the upcoming months***:
    a. Expand example validation to incorporate code set validation as that infrastructured becomes completed in LinkML.
    b. Plan and develop a tighter integration between ontology browser/triplestore & terminology server.

- Q: Does OLS have the capability to show the CodeSets that we are defining to support the data? Or is there a way to support look up and retrieval of specific value sets?
    - A: Not yet. The Terminology Services we are creating will have that. This will give programmatic access. OLS could provide a UI version - in case it is useful to stakeholders.
    - Could use OWL transformation and use for OLS service. Have not yet decided how the enums would work out. -- let's talk about this next!
- Migration from BL to LinkML - maybe present that as not specific to biology. ;)
- Q: moving the sample data in Jupyter notebooks - please clarify.
    - A: There are 2 different repos. There is the model repo that has a generator to build the model from the Google sheets. The second one, example-data, takes the model and takes data in some other model and uses the Python API that our model provides to create instances of specific data objects. These workflows can take GDC data and create a new data set that is CRDC-H conformant. This is the

part we are calling “ETL” - you could use this Python code for any kind of model. Any transformation to CRDC-H is going to be different depending on where the data are coming from.

- FU-q: Please clarify “ETL pipeline” -
  - Google Sheets representation is used to develop the model, and represent it in LinkML. There will not be a connection ‘back’ to the Google sheets. The LinkML computable model will be the product, and we can use that model to demonstrate how ETL can be done. CDA can use this for their transformation purposes. In the model repo, the /python/ folder contains Python DataClasses that can be used to do that transformation step. Offers perspective on the different ways that you can use the model.
  - Our idea is to be able to demonstrate how terminology services, the tools, and the model could demonstrate how we can harmonize data coming from node resources. Since it is not our remit to do so, this can help CDA to further their work. For our next contract, do you want us to put the harmonized data somewhere for the nodes? It doesn’t preclude the nodes using their own implementation of our tools. We could consider harmonizing the data and putting it somewhere, but it is not our remit - we are doing it because we had to demonstrate that our tools work. Food for thought for the future conversation.

2021-05-05

Community Development Workstream Update:

1. **Progress over the last month**:
   a. Presentations
      i. NCPI spring workshop (Melissa and Sam)
      ii. HioH inaugural talk (Melissa and Sam)
   b. Website updates (Debra, Jen, Brian)
      i. Tickets can be found here.
   c. Working on the Support and Engagement plan deliverable.
   d. April newsletter went out
      i. Sent to **186** contacts
      ii. Open rate of **39.7%** (this is a very good open rate for a newsletter)
      iii. WIll be added to the website soon
2. **Timeline check**:
   a. On track for May deliverables
3. **Work for the upcoming months**:
   a. Continue website updates as needed
   b. Finalize support and engagement plan by end of month
   c. July newsletter will go out at the end of this quarter
   d. Next HioH session?

i. **June (bi-monthly)** or July (quarterly)?
    1. Kat will send out invite for bi-monthly HioH today

ii. Topic
    1. Part 2 Harmonization talk - Sam and Melissa

2021-04-28

Data Model Harmonization Workstream Update

1. **Progress over the last month**:
    a. Ongoing modeling work
        i. Draft data structures for cancer staging and cancer grading to replace numerous individual variables for different classification systems. Basic pattern is to capture the classification system in one variable and the value in another (e.g. classification_system = 'FIGO Stage' and value = 'Stage 1', rather than figo_stage = 'Stage 1')
            1. ADM contains 23 different variables that capture staging using different classification systems
            2. ADM contains 8 different variables that capture grade using different classification systems

Notes: This allows us to accommodate various staging systems.
But does CCDH prefer a particular staging system?
Should CCDH suggest? (e.g., INRG vs. **AJCC**)
MarkM: Has done stage/grade modeling for OB.
How would this be serialized?
Link to FIGO
OLS Stage browser:
https://www.ebi.ac.uk/ols/ontologies/obi/individuals?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FOBI_0002327
Brian: Start creating JSON representations of specific values that we see in a node. This is inspired partly by mCODE's method.
Can we get Mark's prior work in this area.
Careful not to conflate stage and grade.
Chris: AJCC is the most widely used and "appropriate" as a generalized and canonical standard for harmonization
Be aware of lossy transform from source info - but you would preserve both - harmonized to support search but preserve source detail.
Shahim - Harmonizing is about harmonizing the shape/structure/fields. We're not really talking about harmonizing the original encoding.
Chris - Syntactic and semantic harmonized forms - need a canonical rendering of the dataset. Harmonized model is the target of the process for cross-domain query with

preservation of the original. But need to harmonize semantics as well....
Jim - Isn't this what we are working on?
Brain - Can bind staging value attribute to any number of different codesets for different staging systems. But are we asserting that this is the primary staging system?
Chris - AJCC is more about severity and extent and not about diagnosis.
BillD - Need to capture ICD-O histology - which are captured by PDC/GDC
Gilberto - wondering about structure for staging and grading. While some of the original annotations need to be maintained - there may also need to be mapping to another structure. Can this be "noted" but left up to CDA to perform the mapping for query purposes.

ii. Draft data structure for exposure (environmental, tobacco, alcohol)
    Real world evidence being included in the model (environment)
        E.g., PPM2.5 pollution data

iii. Draft generic observation class(es) that can take atomic measures or composite groupings of measures

iv. Review of outstanding modeling questions from initial MVP v0 release

v. Increased cross-workstream engagement
    1. Refining our model development output format to: 1) be more easily computable as input to downstream tooling that generates LinkML-compliant YAML, and 2) specify terminology bindings for enumerative types
    2. Learning about latest developments in LinkML and how these might be used in our model

vi. Continued engagement with CDA technical team
    1. As the primary consumer of the model to date, they continue to be an important collaborator
    2. Meetings on **4/1 and 4/15** focused on aligning efforts between groups by providing status updates on work in progress and planned effort
    3. On 4/15, CDA (Donovan) demonstrated a YAML-based mapping syntax that they were making use of in their transformations from cached DC data into their CRDC-H-aligned model
    4. CCDH discussed draft modeling approach for staging and grade; Dazhi demonstrated TCCM services that could be useful for CDA's transformation needs

vii. Meeting with Denise Warzel, and Samvit personnel to review initial work on CDM to FHIR mapping and research gap analysis
    1. CCDH review will be paused until the end of May when our deliverables are due

b. **Timeline check**: what's the status of your work compared to our project milestones?
    i. Model development pace is picking up and we're **making progress** towards our stated scope for Phase II

c. **Work for the upcoming months**:
    i. The focus in the coming month will be primarily in ongoing data model development leading up to our **Phase II deliverables**
    ii. After May 31, we should *evaluate priority areas* for ongoing model development to align with the needs of DCs and CDA
    Could do this in our Internal Meeting the next time the DMH group presents:
        1. Next group of entities
        2. Get feedback from federal oversight team