One-pager for PCDC Project

Cancer research suffers from 'a sea of standards' -- multiple data models exist to bind concepts to standardized definitions. Therefore, one of the goals for the PCDC (Pediatric Cancer Data Commons) project is to connect shared concepts across different cancer models and generate a searchable graph, which is of great significance for future cancer research. For example, it allows researchers to link and compare datasets from different data models.

The obstacles to achieve this goal comes from the diversity of data structures of different cancer models. Given this diversity, it is currently impossible to search for shared concepts across different models. For example, data from mCode (a cancer data model) is constructed based on the FHIR (Fast Healthcare Interoperability Resources) standard, while data from PCDC (another cancer data model) is manually recorded with its own unique format. In order to allow for linking and comparing concepts between different models, the data structures from models need to be rewritten in a common format.

The task for the Data Science Clinic team was to reconstruct the mCODE data model, over the course of the quarter, the team completed the entire process of reconstructing the mCode data structure. (See Figure 1 for the original data structure of mCode, Figure 2 for the code for reconstructing the same structure) This work was completed using linkML, which allows for the creation of schemas in YAML to describe the structure of data and build a searchable graph. During this process, a set of rules were identified for simplifying the data structure while preserving the main information related to the clinical concepts. Based on the reconstructed schema, more human-readable representations of the data were generated and recorded in markdown files. All code related to the project was uploaded and documented on a github repository. The schemas and markdown files will be necessary for the PCDC team to connect shared concepts between different cancer research models and build a searchable graph in the future.

Name	Flags	Card.	Type	Description & Constraints
ContactPoint	Σ I N		Element	Details of a Technology mediated contact point (phone, fax, email, etc.) *+ Rule: A system is required if a value is provided.* Elements defined in Ancestors: id, extension
system	Σ I	0..1	code	phone | fax | email | pager | url | sms | other ContactPointSystem (Required)
value	Σ	0..1	string	The actual contact point details
use	?! Σ	0..1	code	home | work | temp | old | mobile - purpose of this contact point ContactPointUse (Required)
rank	Σ	0..1	positiveInt	Specify preferred order of use (1 = highest)
period	Σ	0..1	Period	Time period when the contact point was/is in use

Figure1: The original data format of mCode

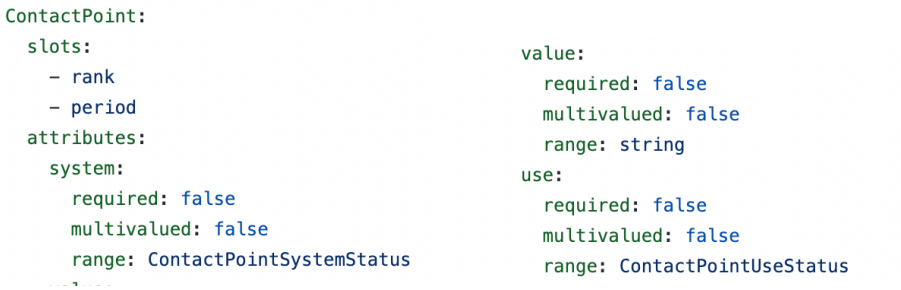

Figure 2: The rewritten format of mCode