##### Collaborative Enhancement of The Act Ontology

##### Project Report

##### Amy Y. Wang, MD, MBI, Associate Professor of Medicine

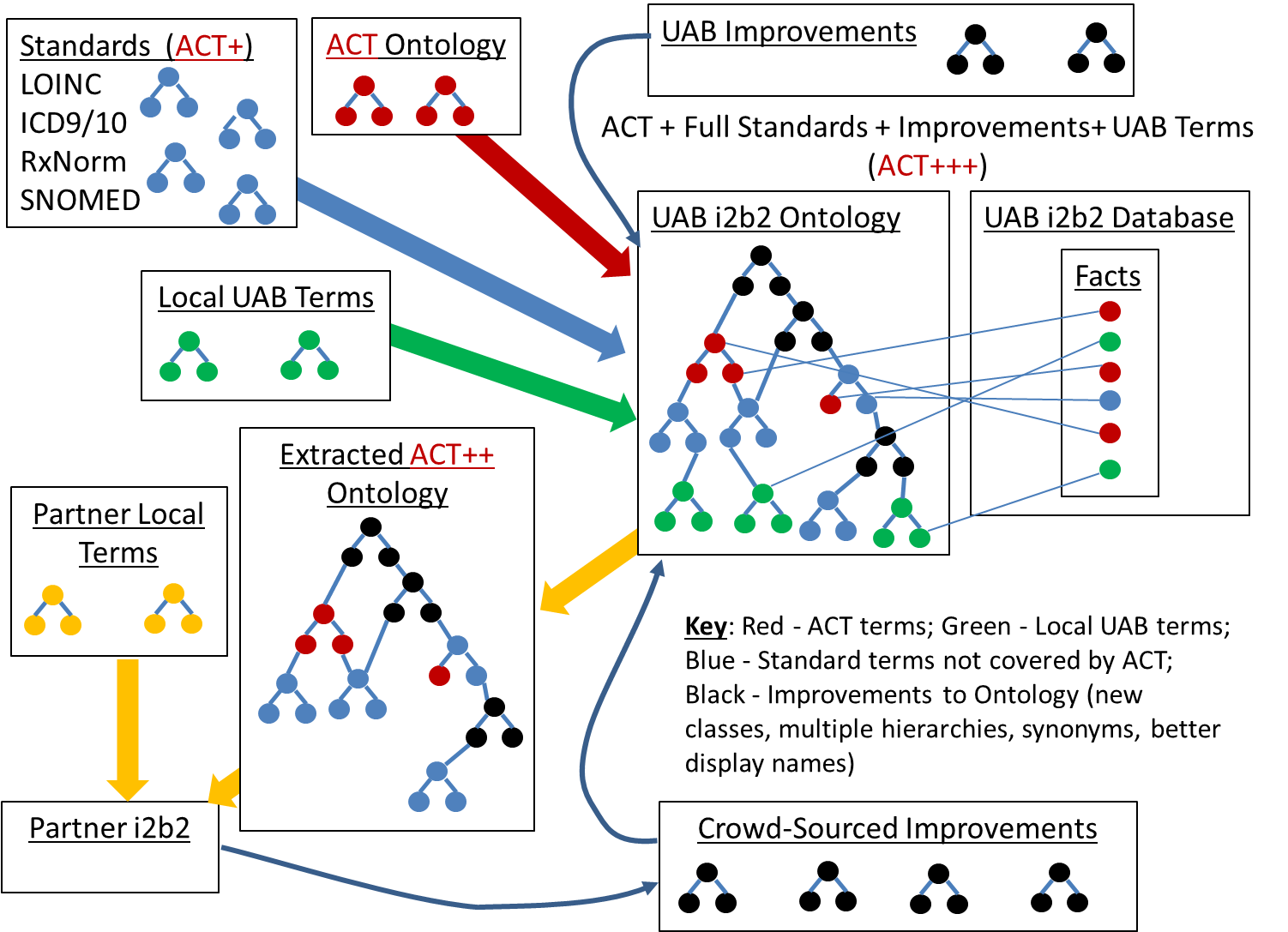
##### University of Alabama at Birmingham

##### June 30, 2020

##### Introduction

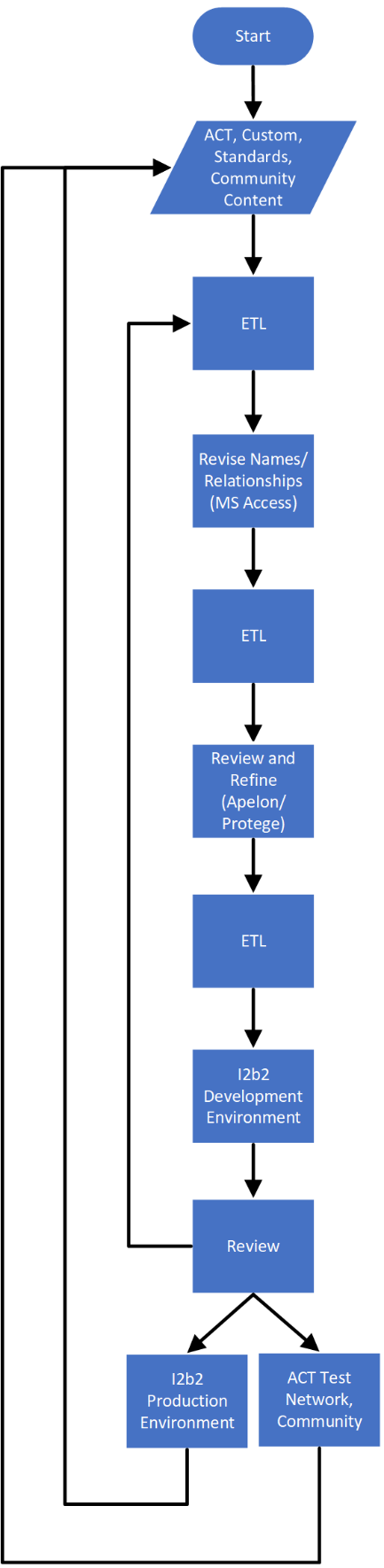
The initial phase of work on the Collaborative Enhancement of the ACT Ontology (CEAO) project was to perform a workflow analysis with the aim of improving workflow and efficiency. The CEAO is based on the UAB Foundational Ontology (UFO) at the University of Alabama at Birmingham (UAB, Figure 1) and is designed to facilitate self-service queries for cohorts for clinical research. The UFO contains concepts and terms from standard, local, and custom terminologies, including those used for representing patient data, user-friendly terms, new classes for grouping, and multiple hierarchies for supporting searching. Significant challenges in data models and formats have affected UFO development. The goal of this phase was to 1) perform a workflow analysis was to identify challenges in UFO development, 2) identify opportunities for addressing these challenges and streamlining and optimizing workflow, and 3) start a pipeline of content development, including rationale for prioritizing this content.

**Figure 1.** The UFO contains standard terms (blue), research ontologies (red), local UAB terms (green), and improvements (black).

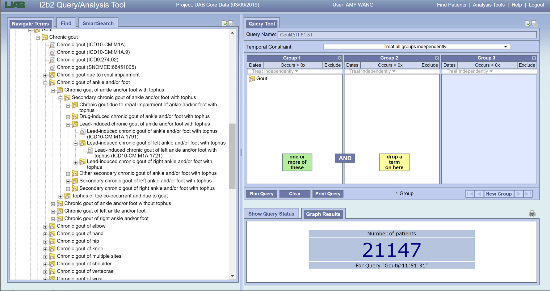
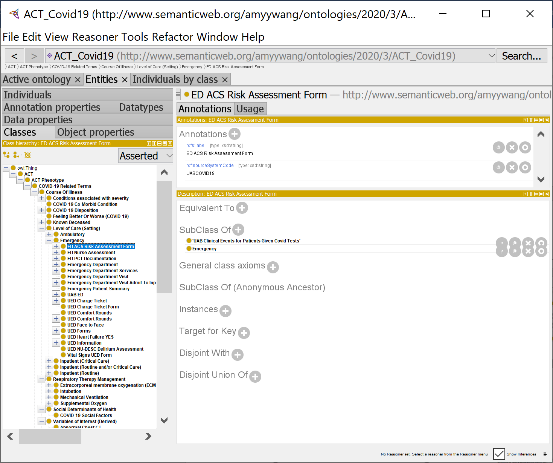
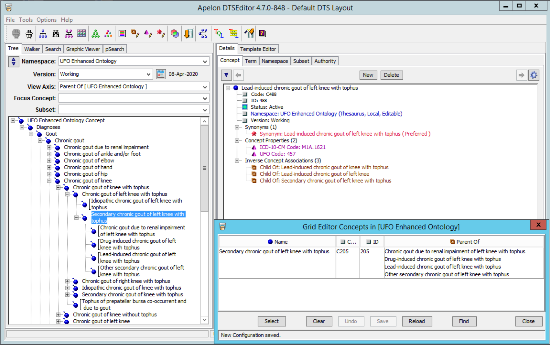
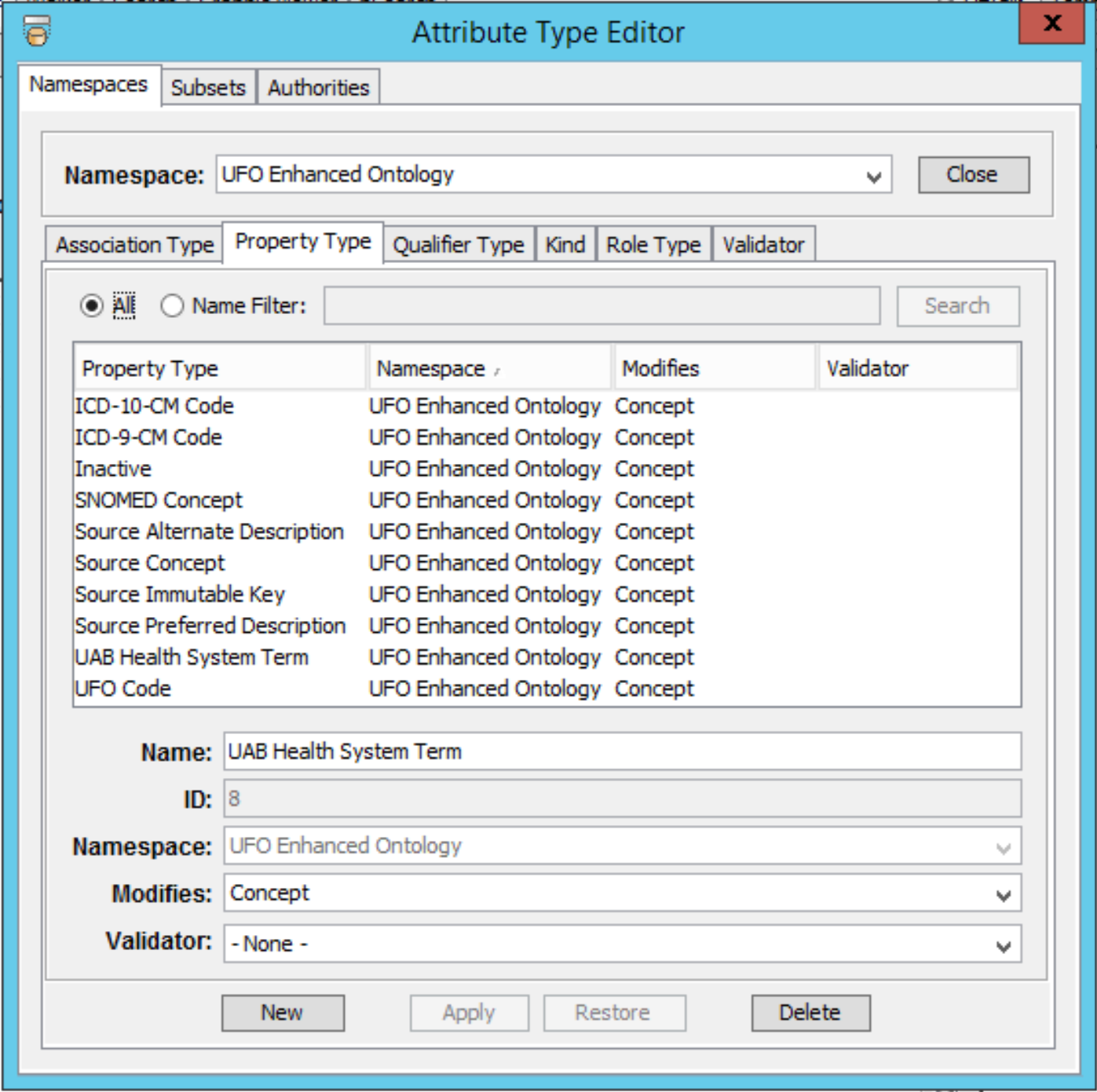
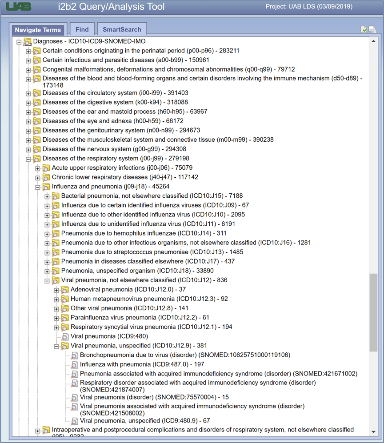


##### Workflow Analysis

##### Challenges in the development of UFO have included different data models and formats for source terminologies, editing environments, and implementation. In addition, there is no single editing tool that supports end-to-end development and is optimized for editing individual concepts, organizing hierarchical relationships between concepts, and creating mappings between different terminologies. As a result, development has involved the use of multiple tools and environments. These include Microsoft Access, Apelon Distributed Terminology Services (DTS), and Protégé, and i2b2. Much time and effort have been spent on extract, transform, load (ETL) processes throughout the development workflow (Figure 2). The use of multiple tools increases complexity, decreases consistency, and hampers distributed development by multiple editors. The multiple ETLs require considerable time and resources, leaving less time for developing UFO content and affecting throughput, Many of the development and ETL steps require highly manual processes, affecting consistency and increasing the risk of error.



**Figure 2.** Existing UFO development workflow.



##### During and before this phase of the project, there has been an ongoing search for other development environments. These efforts have included discussions with informaticians, ontologists, vendors, terminology development organizations, and researchers. While many terminology and ontology editing tools exist, including open source and proprietary systems, they are designed for different or specific use cases, and few adequately meet the needs of this project. Required functionality includes editing concepts, developing hierarchical structures, inferred views, cross-mappings, subsets, distributed development, versioning, publishing, and subsets. Graceful evolution requires identification of ambiguous or duplicate concepts and splitting, merging, and retiring concepts. Many systems do not support these functions, or they are not user friendly for terminology or domain experts.

##### Costs for proprietary systems can be significant and prohibitive for research centers and institutes that are developing freely available solutions for the research community. These systems are often intended for use cases that generate direct revenue, such as improving health system quality or developing applications for market. Open source and freely available solutions can be challenging or require extensive technical support (either in-house or consulting). Open-source or freely available systems may lack required functionality and even be no longer actively updated or supported.

##### After the main development phases have been completed, the content is imported into UAB’s main i2b2 environment. As this i2b2 instance uses the same database as the production environment, this ETL must be performed by database analysts who are also maintaining and updating i2b2, supporting i2b2 users, and helping researchers query and obtain patient data. These demands have only increased during the current COVID-19 pandemic due to the urgent need for data to support COVID-19 research. Since UFO is still in development and there may be changes to the data model, ETL processes and SQL scripts often need to be updated.

##### There are also frequent iterations of UFO content, such as when source terminologies are updated, users request new content, and issues are found during review in i2b2. may necessitate querying the enterprise data warehouse (EDW) for researchers Since UFO is in development, there may be changes to the data model and content, requiring frequent ETLs, rework in the various tools, and repeat loads into i2b2.

##### Opportunities for Tooling and Process Improvements

##### *Separate i2b2 Instance for UFO*

##### A major challenge in reviewing and testing UFO content in i2b2 locally is the need for UAB database analysts to perform the ETL into i2b2. The current proposal is to install a separate instance of i2b2 used only for reviewing and testing UFO. The UFO team would have full administrative rights to this instance and be able to load content into it. Individuals in the CD2H HOT community have been identified who can provide support on the installation and administration of a local i2b2 instance. Having a separate and local i2b2 instance would provide greater flexibility to perform ETL, supporting more rapid iteration.

##### *NCI Protégé*

##### For a terminology editing environment, one clear winner has emerged in the search for a highly functional, fully integrated solution. The National Cancer Institute (NCI) Enterprise Vocabulary Services (EVS) uses [NCI Protégé](https://wiki.nci.nih.gov/display/Protege/NCI+Protege) to develop the NCI Thesaurus (NCIt). [Gilberto Fragoso, PhD](https://datascience.cancer.gov/about/staff-directory/gilberto-fragoso), Biomedical Informatics Program Manager at NCI Center for Biomedical Informatics & Information Technology (CBIIT), along with others, developed NCI Protégé (T able 1). He gave a presentation about Protégé NCI at the CD2H Healthcare Open Terminology (HOT) Ecosystem meeting on June 3, 2020.

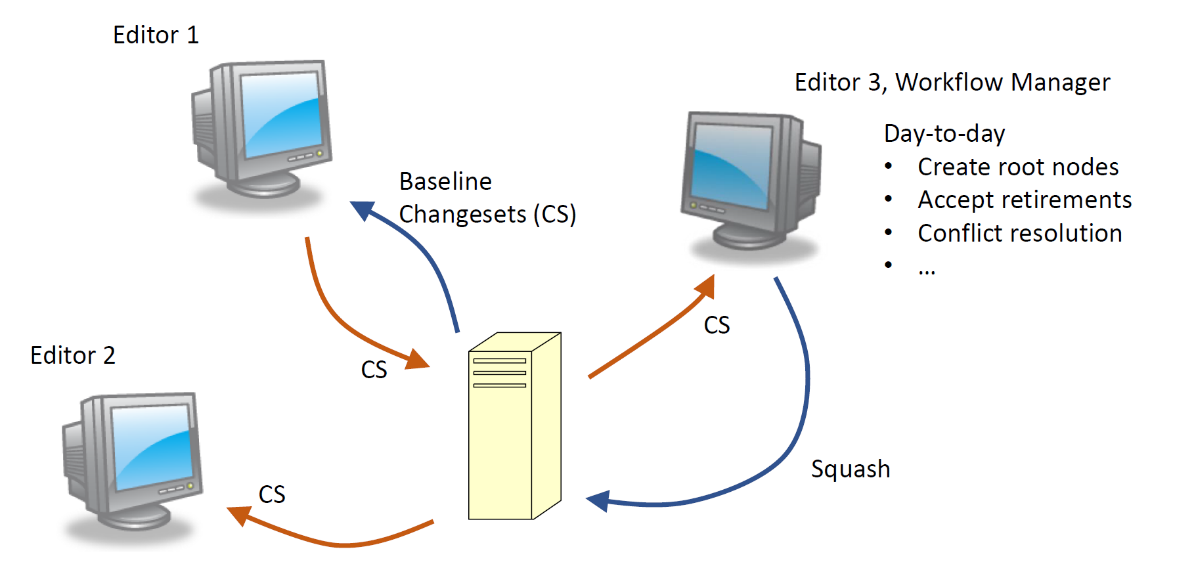
##### NCI Protégé is a set of custom Web Ontology Language (OWL) plug-ins developed for [Protégé](https://wiki.nci.nih.gov/display/Protege/Protege) which are designed for developing NCIt. [Protégé](https://protegewiki.stanford.edu/wiki/Main_Page) is an open-source ontology editing environment developed by Stanford University. While NCIEVS has used other solutions, none of these fully met their needs, prompting them to develop their own solution. NCI Protégé operates in a client-server model and is designed to be user-friendly for subject matter experts (SMEs). About 20 SMEs and domain experts in geographically disparate locations use NCI Protégé to develop NCIt. NCIt development involves frequent collaboration in real-time on editing decisions.



**Table 1.** Plugins and tabs in NCI Protégé.

##### NCI Protege has SME-friendly views, such as annotations, and it supports complex edits such as merging, splitting, and retiring concepts and maintaining the history of such changes. There is also a history of all revisions and published versions. Other features not available in native Protégé include batch loading and editing, support for complex searches (e.g., Lucene, SPARQL, Boolean, property), workflow management, distributed development, administration, access control, assignments, and resolution of editing conflicts (Figure 3).

##### NCIEVS has been using NCI Protégé since 2007 and released the most current version, NCI Protégé 5, in 2016, with periodic updates to scripts since then. While NCI Protégé was designed specifically for developing NCIt, it has functionality that is applicable to other terminology development efforts. While NCI Protégé was designed for developing NCIt and NCIEVS is not aware of anyone else using the current version, NCI Protégé’s features are applicable to the development of other terminologies, including health terminologies. Dr. Fragoso has offered to assist CD2H in installing and using NCI Protégé for UFO development. Having an open source and actively developed solution that has required functionality through multiple phases of development provides greater integration, reducing the need for multiple ETLs. Given its features, use of NCI Protégé, which is open source, and its use in production and ongoing updates, and current support NCI Protégé shows great promise for many terminology and ontology efforts in health and biomedicine.



**Figure 3.** Distributed development in NCI Protégé.

##### Content Development Pipeline

##### The original plan was to develop concepts, descriptions, and hierarchical relationships in infectious, respiratory and cardiovascular diseases to support COVID-19 research. ACT has created the ACT COVID-19 Ontology (see [wiki](https://dbmi-pitt.github.io/ACT-Network/covid.html) and [GitHub](https://github.com/shyamvis/ACT-COVID-Ontology)) to help researchers perform focused searches based on concepts that are highly relevant to COVID-19 and SARS-CoV-2. These concepts include course of illness, severity of illness, need for intubation and mechanical ventilation, diagnosis, laboratory testing, and treatments. The current ontology development for this project has focused on mapping local UAB concepts to the ACT COVID-19 Ontology to facilitate searching. These mappings can be found [here](https://github.com/amywangmd/ACT-UAB-COVID-19). Version 3 of the ACT COVID-19 Ontology was released June 1, 2020. Work is in progress to update the mappings to this newest of the ACT COVID-19 Ontology.

##### Future Work

##### The next phase will focus on improving workflow and developing content.

##### Priorities for workflow improvement are to install and migrate workflow to NCI Protégé in collaboration with NCIEVS and collaborate with others in the CD2H/HOT community to install a local instance of i2b2 devoted to UFO development along with scripts for more automated ETL. These changes will reduce ETL, provide a more integrated development environment, allow the UFO team to readily import, export, and update content, and increase efficiency and throughput. Improving workflow will support rapid prototyping and iterative development. In addition to improving efficiency and throughput for UFO, this work will provide improved terminology development processes throughout the CD2H community.

##### Priorities for content development are to update maps from UAB content to Version 3 and subsequent releases of the ACT COVID-19 Ontology. Another focus will be to map the local UAB content to standard terminologies when possible. Work will continue on developing concepts and mappings to respiratory, cardiovascular, and infectious diseases to integrate content from different standards and local content. This content will be mapped to the ACT COVID-19 Ontology as appropriate.

##### Conclusion

##### This project addresses longstanding challenges well known to research institutions and aims to improve the use of health information in clinical and translational research by enhancing the use of local and standard ontologies. The COVID-19 pandemic highlights the urgent need for continuing to enhance ontological tools, processes, content, and expertise for connecting health information with researchers, analysis, and knowledge discovery.

##### Acknowledgements

We would like to acknowledge organizational support from [the National Center for Data to Health (CD2H)](https://cd2h.org/), [the Johns Hopkins Institute for Clinical & Translational Research (ICTR)](https://ictr.johnshopkins.edu/), [Oregon Clinical and Translational Research Institute (OCTRI)](https://www.ohsu.edu/octri) at Oregon Health & Science University (OHSU), [the UAB Informatics Institute](https://www.uab.edu/medicine/informatics/), and [the UAB Center for Clinical and Translational Sciences (CCTS)](https://www.uab.edu/ccts/). This project was funded in part by NCATS/JHU (U24 TR002306), UAB CCTS (UL1 TR003096), and the UAB Informatics Institute. We would like to thank Dr. Chris Chute, Dr. Melissa Haendel, Davera Gabriel, Dr. Harold Solbrig, Dazhi Jiao, Dr. Gilberto Fragoso, Dr. Jim Cimino, Matt Wyatt, Robert Johnson, and Ayme Miles for their ongoing support and assistance on this project.