Addendum to Agenda

Sample Use Cases and Cross Node Queries to Inform Semantic Infrastructure Needs

This is a compilation of anticipated real use cases derived from discussions with HTAN, PDC, IDC, SEER and others, as well as some additional sample use cases/queries. The goal is to provide background to help workshop attendees better understand where semantic support may be needed for the Cancer Data Aggregator, described below, and CDRC in general. Note: The CDA will not actually do analysis but it will need to determine what data or pointers to retrieve. These should help inform the breakout session discussions.

Cancer Data Aggregator

Scope

The Cancer Data Aggregator (CDA) will allow users to find and access data stored on domain-specific nodes, or other appropriate sites, within the Cancer Research Data Commons (CRDC). The CDA will enable users to search via relevant criteria such as case, sample, study, disease, spatial location, or tissue type. It is envisioned that searches, requiring semantic translation, will be supported as new and required semantic resources are identified and/or developed. Query results will be returned as a listing of data resources meeting the search criteria, possibly as pointers to their cloud locations, to support subsequent analyses outside of CDA. Access control information will also be included. Analyses of identified data will be performed using tools and infrastructure enabled within the CRDC but outside the CDA tool. The CDA will not attempt to create a single, overarching data or semantic model. Instead, the CDA will serve as a basic integration point for more purpose-driven models, not as a replacement for those models. Functionally, the CDA will be a middleware layer interacting with the domain-specific nodes via their APIs and presenting results via the CDA API.

Search Use Case Approach:

A researcher is searching for colon cancer patients with proteomics, image and genomics data, each of which is stored in a domain-specific commons node on the CRDC. The researcher logs into the CDA, enters the search criteria, and is returned a list of data resources and data sets that meet the researcher's criteria. For data they are authorized to access, the cloud-

based data pointers are available to them through their Workspace. The user analyzes the aggregated data using the Cloud Resources or local infrastructure. Note that some applications might also wish to interact directly with CDA API to provide a search interface for their users.

Repository Use Case Approach:
A domain or study specific repository (such as a new domain-specific CRDC node or other approved site) wants to make their data available to discovery and search via the CDA. One possibility is they provide their data model and metadata concepts to the CDA in a manner that can be incorporated and provide a method for CDA to query their data. If the repository contains controlled data, there should be information on how a user can gain access. Once the repository information is incorporated into the CDA, the data from the repository is available to researchers.

More specific examples of Use Cases:

1. Basic Use Case - Retrieve all the data associated with a case / individual spread across various repositories (nodes) for a particular project/study such as APOLLO or CPTAC. The context of any sample within the study design must be retrieved. See Slide 1 from attached Reference Slides for the CPTAC distribution of samples as an example. (Query: Multiple nodes)

2. Query for patients across the nodes to investigate genomic, proteomic and morphologic differences of tumors from patients with X disease subtype, who survived 5-years and those who did not. Assume there could be some patients with genomic or proteomic data but not both. (Query: Multiple nodes)

3. See Slides 2-6 from attached Reference Slides for examples of Sequence-centric Proteogenomics and Analysis of Proteogenomic Relationships: (Query: PDC/GDC). Query for cases across the nodes that can help:

    a. Identify variant protein sequences corresponding to somatic mutations and to evaluate the relationship between mutation frequency and variant protein expression.
    b. Determine how copy number variation translates into protein expression differences.

    c. Evaluate the impact of genomic features on the status of signaling networks through direct analysis of phosphoprotein intermediates.
    d. Derive preliminary associations with clinical characteristics, such as platinum resistance in ovarian cancer.

4. Query for cases with pathology sample data/images from a cohort with X disease subtype, and pull in some Human Cell Atlas molecular & imaging data on normal cells to look at differences in tumor microenvironment from normal. (Query: Multiple nodes plus external data)

5. Human Tumor Atlas Network – Goal of HTAN is construction of Human Tumor Atlases that describe multidimensional cellular, morphological and molecular mappings of human cancers over time. HTAN will want to share some data with CRDC to make available to broader community. Complex and driving use case for CDA. (Query: Multiple nodes and/or External Data)
    a. Many complex technologies. See examples in HTAN slides 7-10 in attached Reference Slides. Users will need to compute across multiple dimensions at different levels of biological scale (molecular to whole organism)
    b. The technologies will change rapidly. This speaks to the importance of flexible, extensible data models.
    c. Key need for semantic infrastructure – how to capture a complex and rapidly evolving data model on which to hang the data. Graph models and RDF are infinitely flexible. However, for the data to have meaning the models must be well described. How?

6. SEER Use cases needing semantic support:
    a. Expand the clinical data collected through linkages to capture current and new data items
    b. Support real-time case ascertainment for clinical trials in SEER registries
    c. Submit SEER public use data of different types to node(s) in Cancer Research Data Commons (CRDC) (Query: Multiple nodes)

    Anticipated Semantic Requirements:

a. Semantic resources need to be accessible by public API, as well as need to be installed and run completely within secure enclaves. (Unique semantic infrastructure requirement for consideration)

b. For CRDC node, data mining, data science
    - Metadata for NAACCR elements (SEER data, CDC data)
    - Metadata for linkages (pharmacy, claims, radiation oncology, EMR, genomic tests, etc.)
    - Metadata for clinical trials inclusion/exclusion criteria
    - Concepts and relationships graph of mappings, views, transformations

c. For NLP and deep learning algorithms, the CRDC needs to provide access to:
    - Metadata for document annotation schemas
    - Concept definitions and synonyms
    - Cancer registry coding rules(e.g., site/histology, marital status)

7. Broadening Genomic Data Storage - >200 projects with wide variety of genomic data storage and access needs. Semantic requirements (to be discussed in breakouts) such as:
    a. Need to capture what scientist doing, e.g. study design
    b. Semantic description of relationships/associations
    c. Ensure models are independent of particular implementations such as forms.

8. Query across nodes to select cases to compare DNA and RNA expression patterns in clonal populations or single cell, before and after treatment and compare to the histology for patients with and without progression of subtype X breast cancer e.g. triple negative? (Query: GDC/PDC/Imaging nodes).