2020-07-13 CPTAC DCC Data Migration Meeting notes

Date

13 Jul 2020

Attendees

- Beyers, Matthew (NIH/NCI) [C]
- Karen Ketchum, Anand, Rajesh, John

Goals

- Discuss next steps for CPTAC DCC Data Migration.

Discussion items

<table>
<tr><th>Time</th><th>Item</th><th>Who</th><th>Notes</th></tr>
<tr><td></td><td></td><td></td><td>Collections (Phase refers to CPTAC phases, not clinical phases):
Phase 1 studies (2006-2011) (6-8 studies, 2-3 publications) - Not in PDC
Phase 2 studies (2012-2016) - 1/2-2/3 proof of principle studies not in PDC
Phase 3 studies (2017-Ongoing) - ongoing different cancer cohorts (discovery/confirmatory); Most/All of them are in PDC
External Studies: (Friends of CPTAC - Various Dates)
Types of Data:
RAW, mzML, Peptide Spectral Matches, Protein Reports, QC reports, etc.
User Generated processed data
Two types of views:
Study specific pages
Publication specific pages</td></tr>
<tr><td></td><td></td><td></td><td>Considerations in Migration of Data
preserve data provenance, analysis methods, parameters, IDENTIFIERS
Preserve publication related data includuing user generated processed data (links in publication to CPTAC data portal)
Allow download of data from new home in PDC
Preserve clinical metadata, currently stored as excel/text files in CPTAC</td></tr>
<tr><td></td><td></td><td></td><td>Where we are today:<table>
<tr><th>Collections/Phase</th><th>In CPTAC</th><th>In PDC</th></tr>
<tr><td>CPTAC 1</td><td>9</td><td>-</td></tr>
<tr><td>CPTAC 2</td><td>23</td><td>6</td></tr>
<tr><td>CPTAC 3</td><td>11</td><td>8</td></tr>
<tr><td>External Studies</td><td>9</td><td>3</td></tr>
</table></td></tr>
</table>

			Challenges and Possible Solutions:
			
			• Phase I studies have minimal metadata, making it almost unusable in a curated portal like PDC
			  ◦ Possible solution: provide packaged file sets for anyone who might still be interested separately (new area of the PDC portal)
			  ◦ Archive separately and not have any linkages on PDC
			• Phase II & III - several additional studies that are not primary cancer cohorts from CPTAC program
			  ◦ Possible solution: absorb all or a select few
			• Publication related and user generated processed data
			  ◦ Possible solution: absorb in the original study
			• Additional metadata
			  ◦ Possible solution: extend PDC portal to capture some if not all
			• Download Datasets
			  ◦ Possible solution: non-AWS based data center that does not have egress charges to make it cost effective.
			
			Looking at where we are today slide (John): of those projects in CPTAC portal vs. in PDC - can we get a sense of how used those projects are in CPTAC? Has anyone accessed the 9 Phase I studies in the last year? For CPTAC 3, the three studies that are not in PDC - are they planned to move to PDC or is there a reason not to do so? CPTAC 3 should move forward. CPTAC 1 is low usage at this point. CPTAC 2, sometimes people are looking for a simple data set and might want this, but more likely if they are looking for a full cancer cohort, then not interested. Make a list of all the studies in CPTAC and include the download metrics to make individual decisions.
			
			Henry has concern that PDC doesn't get a lot of use relative to the DCC portal. Question regarding studies in both portal and PDC: which site gets more use? Since there aren't downloads from PDC, that metric may not be easy to get. If they find it on PDC, they probably go to the portal to download it. May include a work habit adjustment since PDC is still new and people haven't adapted their workflow. Also difficult to download the data from PDC since users need signed URL. Easy to download from CPTAC data portal.
			
			Can we estimate the cost if everything is transferred to PDC and allow downloads from AWS? Anand thinks this should be easy to calculate.

			ROM (Rough Order of Magnitude) estimates for migration: • Each CPTAC 2/3 study ~ 1 month of data curation, prep, load and test into PDC • Map/maintain all publications links - 2 months • Supplemental Studies - Unknown at this time (each study is different), might require new portal development to support data types
			Given Erika isn't available on Wednesday and we want to talk about CPTAC data portal migration, do we want to hold off that discussion or should we carryon without Erika? Does Henry want to hold the meeting without Erika? Matt to touch base with Erika. Default: wouldn't want to hold meeting without her - still give progress update to Henry. Question is about this particular agenda item.

Action items

- [ ]