Observations on challenges of finding data in repositories

September 2024

This is part of the Use Case Exercise 4 materials.

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

I’ll start with my take-aways:

What does searching data repositories involve?

And why is it kind of hard?

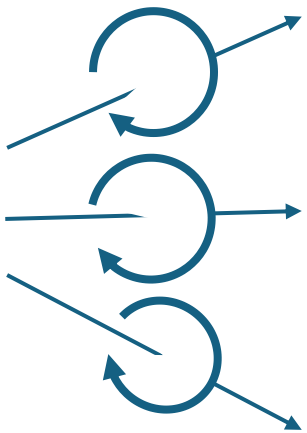
Processes involved in finding and retrieving data of interest

"Mapping" can be hard and time-consuming!

Mapping

Retrieval Implementation

Data Request



Query for Repository A → **For Repository A**

Query for Repository B → **For Repository B**

Query for Repository C → **For Repository C**

- Involves harmonizing a user's concepts & vocabulary w/ those in repositories
- Requires knowledge of a repository's variables & values
- Involves human decision making
- Is often iterative and time-consuming
- Would probably benefit from AI suggestions

Can take the form of any of the following, depending on repository infrastructure:

- A sequence of UI button selections
- An API call
- A GraphQL query
- Code using R/python software packages

All have learning curves

An illustration of the mapping challenge:

Mapping a data request to multiple repository queries

Data Request

Find gene expression data (single-cell RNA-seq or bulk RNA-seq) from glioma samples from patients.

Same data request, but the queries are quite different and sometimes complex!

Query for Genomic Data Commons

PRIMARY DIAGNOSIS in ('astrocytoma, anaplastic', 'astrocytoma, nos', 'glioblastoma', 'glioblastoma multiforme', 'glioma, malignant', 'mixed glioma', 'oligodendroglioma, anaplastic', 'oligodendroglioma, nos') and **EXPERIMENTAL STRATEGY** in ('RNA-Seq', 'scRNA-Seq') and **TISSUE TYPE** in ['tumor']

Query for Cancer Data Service

PRIMARY DIAGNOSIS IN ('Glioblastoma, NOS', 'Astrocytoma, Anaplastic, NOS', 'Glioma, Malignant', 'Diffuse Midline Glioma, H3 K27M-Mutant', 'Astrocytoma, NOS', 'Pilocytic Astrocytoma', 'Angiocentric Glioma', 'Glioblastoma, NOS, Medulloblastoma, NOS', 'Pleomorphic Xanthoastrocytoma, NOS', 'Giant Cell Glioblastoma', 'Oligodendroglioma, Anaplastic, NOS', 'Glioblastoma,Glioblastoma Multiforme', 'Oligodendroglioma, NOS', 'Pilomyxoid Astrocytoma', 'Glioblastoma', 'Astrocytoma, anaplastic', 'Glioma', 'Astrocytoma, Benign,Glioma', 'Oligodendroglioma', 'Astrocytoma', 'Astrocytoma, NOS, Glioma', 'Ganglioglioma,Glioma', 'Dysplasia,Glioma', 'Diffuse Midline Glioma, H3 K27M-Mutant,Glioma', 'Astrocytoma,Subependymal Giant Cell Astrocytoma', 'Glioma,Sarcoma', 'Glioma,Neurofibroma,Schwannoma', 'Gliomatosis cerebri', 'Astrocytoma, NOS,Embryonal Tumor,Glioma', 'Malignant glioma', 'Glioma, malignant, Malignant peripheral nerve sheath tumor, NOS', 'Adenocarcinoma,NOS, Glioblastoma, NOS') AND **'EXPERIMENTAL STRATEGY** IN ('RNA-Seq') AND **SAMPLE TUMOR STATUS** in ('tumor')

Query for Human Tumor Atlas Network

Disease in ('Glioblastoma' OR 'Glioma malignant' OR 'Astrocytoma anaplastic') AND **Assay** in ('Bulk RNA-seq' OR 'scRNA-seq')

An Exploratory Exercise: Performing Data Searches in NCI Repositories with their User Interfaces

Exercise: Perform data searches in NCI Repositories

Purpose: Get a ballpark estimate for how long searches currently take and identify pain points

Methods: Carried out two searches from Proposed Use Case 4, Task A (starting with data requests I had not previously written queries for.) Using NCI repository UIs, identified variables and values to select, and downloaded resulting metadata files. Recorded the query used. Timed how long this took for each repository and afterwards took notes on the process and challenging elements.

Data Request Prompts Used in Exercise

Find lung cancer specimens with processed gene copy number and gene expression quantification data. Download metadata about the patient cases, specimens, gene copy number, and gene expression data files.

Find gene expression data (single-cell RNA-seq, bulk RNA-seq, or RNA microarray data) from glioma patient samples. Download metadata about these patient cases, specimens, and sequencing data files.

Repositories Searched: Genomic Data Commons (GDC), Cancer Data Service (CDS), Human Tumor Atlas Network (HTAN) (Proteomic Data Commons was not relevant for these Data Requests)

Description of My Searching Process (slide 1 of 2)

For each data request prompt, I first thought about what to look for. I then used the UIs to look at the variables available in each resource and selected variables to match what was in the prompt. Sometimes this involved choosing among multiple variables, e.g., there are multiple variables that can capture cancer types, such as Primary Diagnosis, Tissue or Organ or Origin, and Tumor Code, and data types, such as Experimental Strategy and Data Type.

I also had to identify values that matched what I was looking for. There were almost always multiple values that matched what I was looking for. As I don't have expertise on the many cancer diagnosis terms, this involved internet searches to identify cancer terms that matched those in the data request and to assess whether some terms in the UI lists were matches. To include values, I had to click on all the matching values. For the CDS, I had to search through and click on many values for Primary Diagnosis.

I navigated to the download buttons and downloaded the three types of metadata files and put them in a folder. As I have experience with these UIs, this wasn't hard but if I had not used these recently, I would have had to use the tutorial materials for the GDC, like I did the first few times I used this site. I also manually copied the query/filter selections to record the queries I used.

Description of My Searching Process (slide 2 of 2)

I did this for each repository that might contain data types matching the request. Sometimes this involved assessing whether a repository had the kind of data I was looking for.

Sometimes in doing one search or reviewing a query, I realized and learned new things that made me rethink the variables and/or values I selected in another repository search. For example, I learned that additional primary diagnoses and primary site terms were relevant. Once I also learned that I should have added selections from another variable, e.g., I should use Tissue or Organ of Origin in addition to Primary Site. When this happened, I went back and redid the repository searches.

Results

Time in Minutes to Search and Obtain the Metadata Files		
	First Data Request	Second Data Request
GDC*	29	28
CDS	6	38
HTAN	2	7
Total	37	73

* I searched and downloaded data from the GDC twice for each data request, as I learned things in searching the other sites. GDC times are totals from the two searches.

Final queries used for the First Data Request

GDC:

In Cohort Builder:
PRIMARY SITE in ['bronchus and lung'] and
TISSUE OR ORGAN OF ORIGIN in ['lower lobe, lung', 'lung, nos', 'main bronchus', 'middle lobe, lung', 'overlapping lesion of lung', 'upper lobe, lung', 'pleura, nos'] and
DATA TYPE in ['Gene expression quantification', 'Gene level copy number'] and
TISSUE TYPE in ['tumor']
Under Repository:
Data type in ['Gene expression quantification', 'Gene level copy number'] and Tissue type in ['tumor'].

CDS:

SAMPLE TUMOR STATUS IS Tumor AND PRIMARY DIAGNOSIS IN ('Lung Adenocarcinoma, NOS'' Lung Squamous Cell Carcinoma, NOS'' Bronchio-alveolar carcinoma, mucinous', 'Bronchiolo-alveolar carcinoma, non-mucinous')

Then found that there was no data of the kind matching the request

HTAN:

<None>. Investigated site and determined it did not contain the type of data matching the request

Final queries used for the Second Data Request are on slide 4

Observations: Time-consuming steps

- Exploring potentially relevant repository variables and values
- Understanding what the values mean and which are relevant
- Making decisions about which variables and values to use, especially when variables have overlapping information content (e.g., primary diagnosis, tumor types, and sites of origin)
- Having to go back and redo things in the light of new information
- Going through the many overlapping primary diagnosis values in the CDS. The CDS—the catch-all repository for data that don't fit in the other repositories--lacks term standards for primary diagnoses. e.g., for the second data request, I ended up selecting ~37 primary diagnosis terms, which I found by searching the screen for segments of relevant cancer terms. Some term differences are trivial, e.g., in capitalization or the presence/absence of "NOS"
- Some slowness in the UIs: In the CDS, waiting for the selection clicks to process. Some slow file download times in GDC

Additional Challenges and Questions

- Some variables, particularly in the GDC, appear to have missing data. This makes how to filter unclear and could result in someone getting much less matching data than is actually in the repository (e.g., case numbers for some tumor code values are substantially less than numbers for related primary diagnosis terms.) Having a good mental model of the data is important but hard to get from the UI.
- The lack of provenance is a problem for data reuse. The repositories are missing information about associated papers. There is very little information about methods. E.g., how do we know what methods were used to select the samples?
- Do people really use the CRDC interfaces to find data, or are they finding out about data from reading papers and then locating that data in the CRDC repositories to download it?
- Developing benchmarking sets for realistic repository search tasks is likely to be challenging. For most realistic requests, there are likely to be multiple ‘good’ answers rather than single right answers because there is interpretation involved about what to include. We would have to make the statements very specific, with good alignment with the variables and values in the repositories, if we want there to be single right answers; that may rarely reflect reality though and would take away what I think is the most challenging part of searching—figuring out how to map one’s ideas to the data structures in repositories.