Use Case Exercise 4:

Fulfilling researcher data requests

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation. This technical data deliverable was developed using contract funds under Basic Contract No. W56KGU-18-D-0004.

Distribution Statement A. Approved for public release: distribution is unlimited
MITRE Public Release Case Number 24-2089

Table of Contents

Introduction to Use Case Exercise 4: Fulfilling researcher data requests

The purpose of exercise is to develop tools that enable biomedical researchers to easily and quickly find and assemble datasets for their research. These tools should allow researchers to interact via natural language and/or other low/no-code means, with minimal reliance on faceted search.

There are three component capabilities in this exercise:

A. Finding and retrieving data from publicly available data repositories
B. Finding and retrieving data in papers and their supplementary materials
C. Integrating and harmonizing data from multiple sources

Use Case Exercise 4 materials consist of:

1. **Data request use cases & exercises** corresponding to the capabilities above.
2. **Example tasks for tool development**, consisting of query and answer pairs for capabilities A and B above and datasets for capability C.
3. **Supporting materials characterizing challenges in searching for data in repositories.** These materials include a description of an exploratory experiment we did to understand the current baseline for searching for datasets with repository user interfaces and an example of a data search using an R library. These materials highlight how one of the major, time-consuming challenges is mapping a researcher's initial, high-level ideas for data to the various variables and values in repositories and illustrate the iterative nature of the search process.

Each of these are described below and accompanying materials are in the folder with this document.

1. Data Request Use Cases & Exercises

1.1 Data Request Use Cases

We have created a set of data request use cases based on real researcher uses of Cancer Research Data Commons (CRDC) repository datasets. These use cases are inspired by 1) statements in papers about the use of CRDC datasets for validation or data supplementation and 2) Research Technical Use Statements in dbGAP Authorized Data Access Requests, which are available on dbGAP project pages (e.g., see dbGaP Study phs000218).

For each data request use case, we have written a 'Research Need Statement', a 'Data Request Prompt', and a 'Results Request Prompt', which could be used with LLMs.

The **Research Need Statements** can be viewed as researchers' high-level ideas about intent and kinds of data they want, before refinement. Here's an example:

> *We plan to use machine learning to identify critical differentially expressed genes in colorectal cancer. We want whole-transcriptome profiling datasets from CRC samples. We would also like patient data if available to characterize our study sample.*

The **Data Request Prompts** are potential interpretations of the Research Need Statements and specify in greater detail the kinds of data that are sought. These prompts can be used to more directly formulate queries. Here's the Data Request Prompt for the example Research Needs Statement above:

> *Find gene expression data (RNA-seq, scRNA-seq, or RNA microarray data) from colorectal cancer samples. Download metadata about these patient cases, specimens, and sequencing data files.*

The **Results Request Prompts** elaborate on the research objectives in the Research Need Statements and ask for 'results data' that are highly relevant to the intended research. By 'results data', we mean statistical results from analyses of aggregated sample data. Examples include the frequency of mutations in genes in a particular kind of tumor, statistics about differences in the expression of genes between tumor and normal samples, and statistics about associations of gene transcript or protein abundances in tumor samples with patient outcomes. Results datasets can be found in scientific papers and their supplementary materials and can take the form of spreadsheets, tables, and figures. Here's the Results Request Prompt for the example above:

> *Find results data about differentially expressed genes in colorectal cancer samples compared to normal tissue samples.*

The data request use cases are in the excel file "Research Data Request Prompts.xlsx". The sheet "Res. Needs & Req Prompts" has the Research Need Statements, Data Request Prompts, and Results Request Prompts. Sheets "Paper sources info" and "dbGAP request metadata" have information about the papers and dbGAP Authorized Data Access Requests that formed the bases for these use cases. Sheet "Additional Info for Component A" includes mappings of the Data Request Prompts to repository queries; see the exercise for Capability A below for more information.

To date, we have developed eight data request use cases. We have nine additional Research Need Statements for further tool development and/or validation if program performer teams are interested.

1.2 Exercises

The exercises below make use of the Data Request Prompts and Results Request Prompts.

Capability A. *Finding and retrieving data from publicly available data repositories*

- Identify and download data matching each Data Request Prompt in publicly available data repositories. Each Data Request Prompt has at least one matching dataset in at least one of the following NCI repositories: Genomic Data Commons (GDC), Proteomic Data Commons (PDC), Cancer Data Service (CDS), and the Human Tumor Atlas Network (HTAN). Other repositories that are also likely to contain relevant datasets include cBioPortal for Cancer Genomics, National Center for Biotechnology Information (NCBI) repositories such as GEO and SRA, and the European Genome-Phenome Archive (EGA).
- With results, return information about each repository query or search used, including any variables and terminology terms used.

There is additional information for this component in “Research Data Request Prompts.xlsx” on the sheet “Additional Info for Component A”. For each Data Request Prompt, we provide queries we constructed to search the NCI repositories and the reasoning behind these (see the columns under “Example searches”, in peach). These can be used to assess the query selections a tool makes. We have not provided full answers, which would consist of the actual downloaded files with matching data or metadata. However, for each Data Request Prompt, we provide the name of at least one project or study in a CRDC repository with matching data (see columns D-F, “Repository”, “Studies”, and “Program”, in green). These come from dbGAP project IDs associated with studies listed in CRDC repositories and from mentions of CRDC programs in papers, and can be used for very light-weight benchmarking by checking that answer datasets include some data from these studies. Note: there will sometimes be additional projects or studies that have matching data within the repository listed, as well as additional repositories with matching data.

Capability B. *Finding and retrieving data in papers and their supplementary materials*

- Find information about datasets matching the Data Request Prompts and the Results Request Prompts in published papers and their supplementary materials. PubMed Central has full-text freely available for many relevant papers; use the open access filter to include only papers with licenses consistent with data processing.
- If matching data is found, provide identifiers and details about the dataset, including whether it is in a data repository, the body of the paper, or supplementary file, and the repository name, dataset identifier, URL to obtain the

dataset, the name of the dataset, and/or name of the sheet in a file that the dataset is in, depending on what information is available in the paper.

- Also include a brief description of the dataset including the methods used to generate the data or results.

For datasets in the paper body or supplementary materials, we recommend downloading the table, figure, or supplementary file; processing this may be necessary to fully identify the dataset, e.g., the sheet the data is on in an excel file. For datasets that are in repositories, this exercise only involves providing information from the paper about the dataset and does not involve obtaining the dataset from the repository. It is very rare that there is a direct link in a paper to the appropriate subset of a study's data in a repository, and consequently obtaining this data is a Capability A task and involves generating and executing a query for a repository.

For this exercise, we do not provide answers here, but example answers to some of these and additional similar prompts are available in the *Example Tasks for Tool Development* materials. Note, LLMs such as GPT-4o do well at finding datasets within a single given paper—the challenge here lies in doing this across many papers.

Capability C. *Integrating and harmonizing data from multiple sources*

For Data Request Prompts where there are patient case or specimen metadata from multiple sources, harmonize and integrate these into a single dataset. There is no specified format for this; options include choosing the format from one of the sources, using a standard format like one from the GDC, or creating a new integrated format.

If you want to work on Capability C without finding datasets first, you can use the source papers that are associated with some of the research requests (see sheet "Paper sources info" in the "Research Data Request Prompts.xlsx" file). All these papers utilize datasets from multiple sources, and you can collect and integrate metadata from the same sources. We do not provide example harmonized results in these materials.

1.3 Variation of Capability A and B Exercises

The process of searching for data often involves iteratively clarifying, refining, and specifying what a researcher wants. For example, a researcher may want gene expression data, but not initially specify whether they want raw or processed data, or certain types of files. For Capabilities A and B, you could start with the Research Need Statements instead and create tools that assist researchers in refining their data searches. Your tools could be interactive, and respond with an interpretation of the request for confirmation, and ask researchers follow-up questions, e.g., after a request for RNA-seq data, a tool could respond with a follow-up question like "Are you looking for raw sequence data, aligned reads, and processed data, such as gene counts?" On the sheet "Additional Info for Component A" in the excel file, there are descriptions of our interpretations in the column “The interpretation used for formulating the Data Request Prompt” and example follow-up questions in the column “Potential follow-up questions” (these columns are in blue).

2. Example Tasks for Tool Development

We created example tasks with answers to illustrate what useful responses could look like and for research teams to assess their tools during development. The example tasks and associated materials are in the folder “Example_tasks”. These sets of example tasks can be expanded for benchmarking.

2.1 Capability A, Repository Search Tasks

These example tasks were requested by the Jataware team, who designed most of the task format. Each example task includes the following elements:

- *repository_search_task_id*, a task identifier
- *web*, a url for a repository to search within
- *ques*, short for question; a description of the kind of data or metadata to find.
- *query_information*, information about how the question maps to the content of the repository of the question. This uses variables, values and logic for the repository being searched. The format varies depending on the repository and sometimes includes information about corresponding buttons in a repository user interface.
- *rubric*, the name of a file that is an answer dataset (from the repository in question) that can be compared against. The rubric files were downloaded from the repositories and provided along with the example tasks.
- comment lines, which may include 1) information about any former task ids and relationships to older task versions and 2) notes about the query information selections and appropriate alternative variable and values selections

Here is an example task:

```
repository_search_task_id: GDC0003

# former_task_id: example B3

web: https://portal.gdc.cancer.gov/analysis_page?app=

ques: From the Genomic Data Commons, download a metadata file with information about the samples with WGS or WXS aligned reads data from normal tissue samples from patients with neuroblastomas.

query_information: Under Cohort Builder: 'Tumor Code' in ['neuroblastoma (nbl)'] and 'Tissue Type' in ['normal'] and ‘Experimental Strategy’ in [‘WGS’, 'WXS'] and 'Data Type' in ['Aligned Reads']. Under Repository: ‘Experimental Strategy’ in [‘WGS’, 'WXS'] and 'Data Type' in ['Aligned Reads'] and 'Tissue Type' in ['normal']. Then under Download Associated Metadata, select 'Sample Sheet' to download the file.

# note: 'Primary Diagnosis' in ['neuroblastoma, NOS] is a reasonable alternative to using 'Tumor Code' here. They do show different numbers of cases though.

rubric: ‘GDC0003_gdc_sample_sheet.2024-12-04.tsv’ is provided with this task as the reference.
```

We provide 28 example repository search tasks with answers, which are under the ‘Example_tasks/Capability_A_Data_repository_search_tasks’ folder in the file ‘repository_search_tasks.txt’. The tasks span the Proteomic Data Commons (PDC; 8 tasks), Genomic Data Commons (GDC; 6 tasks), the Human Tumor Atlas Network (HTAN; 3 tasks), and cBioPortal (2 sets;11 tasks). Some of the tasks are based on the Data Request Use Cases, some are based on other real-world examples from papers, and others were developed by exploring the repositories. The second set of cBioPortal tasks are based on steps within a larger search task that is illustrated in a file in *Supporting Materials Characterizing Challenges in Searching for Data in Repositories*; for more information, also see the comments within ‘repository_search_tasks.txt’.

The rubric answer sets to compare against are in the folder “rubric_answer_sets”. For the PDC, GDC, and HTAN tasks, we used the repository user interfaces to locate and download these datasets; all these datasets were downloaded on Dec 4 and Dec 5, 2024. See the document ‘How results metadata files were obtained.docx’ in the folder “info_about_generation_of_answer_sets” for more information. For the cBioPortal tasks, we used the R library cbioportalR; the folder contains scripts used to generate the answer sets for this data portal. Note, because the content of repositories changes over time, it is possible that in the future correct answers will contain more and/or different content than these rubric answer sets. Furthermore, for some tasks there are alternative query variables and values that could also be considered correct and consequently some manual review of non-matching tool answers is warranted.

These tasks were designed for formative tool development more than for summative evaluation, and there is flexibility in how teams choose to use the tasks elements. *ques* and *web* are meant to be “input” material, and *rubric* has information about the “output” answers. In contrast, *query_information* can be treated as either input or output depending on the capabilities a team is working on and stage of development. *query_information* can be used as input when developing capabilities to retrieve data from a repository with known query values, and can be used as output answers to compare against when developing capabilities to formulate repository-specific queries based on natural language requests.

Because the task questions were designed to facilitate assessment, they have some important differences compared to the Data Request Prompts. First, each task specifies a particular repository to search rather than being open-ended like the Data Request Prompts; this enabled us to provide answers for evaluation, but removes the real-world challenge of selecting appropriate repositories to search in. Second, the task questions tend to be more specific than the Data Request Prompts, so that each task question has a correct answer that can be used for evaluation. This makes the task questions less representative of how a researcher would typically search for data but enables semi-automated assessment.

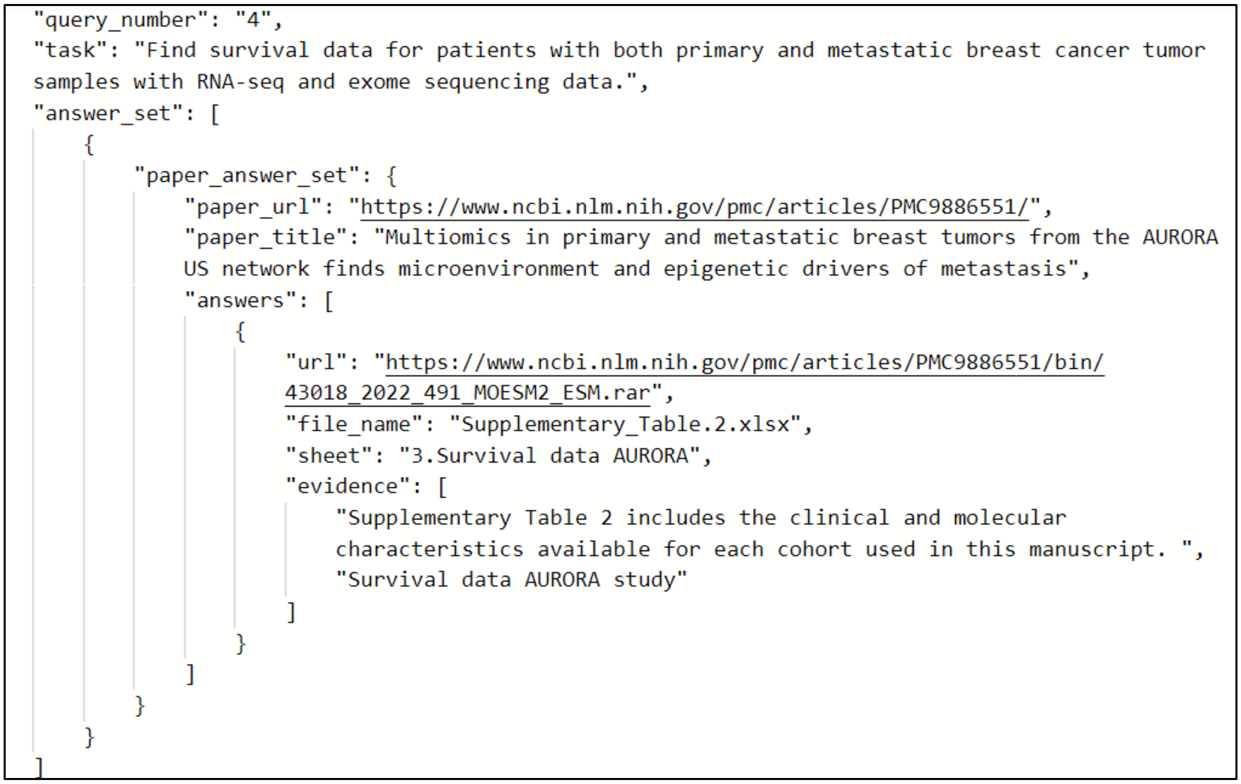
2.2 Capability B, Search Tasks for Finding Datasets in Papers

These example tasks were requested by the MIT team, who designed some of the task structure and format. Each example task includes:

- *query_number*, an identifier for the task example
- *task,* a statement with a description of data to find
- *answer_set*, which includes
    - *paper_answer_set*, each of which includes the information below. While we include just one *paper_answer_set* in our examples, when the tasks are applied to a large set of papers, often there will be multiple papers with matching datasets.
        - *paper_url* and *paper_title*, a url and title for a paper with information about a matching dataset. Note, the paper is part of the answer—the task exercise includes identifying papers with relevant datasets from a publication corpus. However, the specified paper can be used by tools as a steppingstone in tool development.
        - *answers*, information about the matching datasets within the paper, including the information below. There can be more than one ‘answer’ matching dataset per paper.
            - Information about the dataset available in the paper and its supplementary materials, which includes one or more of the following: *dataset_name* (e.g. “Table S1”); *sheet*, an excel file sheet name; *file_name*; *zip_file_name*; *url*, a url at which a dataset can be obtained; *source* (e.g. ‘GEO’); d*ataset_id* (e.g. “GSE269746”). Usually only a subset of these is available from a paper for a dataset.

- *evidence*, one or more sentences, phrases, and/or table titles verbatim from the paper that provide information about the dataset and suggest it matches the task

Here is an example task in json format:



```
"query_number": "4",
"task": "Find survival data for patients with both primary and metastatic breast cancer tumor
samples with RNA-seq and exome sequencing data.",
"answer_set": [
    {
        "paper_answer_set": {
            "paper_url": "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9886551/",
            "paper_title": "Multiomics in primary and metastatic breast tumors from the AURORA
            US network finds microenvironment and epigenetic drivers of metastasis",
            "answers": [
                {
                    "url": "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9886551/bin/
                    43018_2022_491_MOESM2_ESM.rar",
                    "file_name": "Supplementary_Table.2.xlsx",
                    "sheet": "3.Survival data AURORA",
                    "evidence": [
                        "Supplementary Table 2 includes the clinical and molecular
                        characteristics available for each cohort used in this manuscript. ",
                        "Survival data AURORA study"
                    ]
                }
            ]
        }
    }
]
```

We provide 13 example tasks with answers. These are in the format above in the ‘Example_tasks/Capability_B_Paper_search_tasks’ folder in the file ‘paper_search_tasks_in_json.txt’. We also provide the tasks in an excel file, ‘paper_search_tasks_with_additional_info.xls’, with includes information about how these tasks were selected and their scope. Briefly, all the papers in the answer sets are in the PubMed Central open access set, are about cancer 'omics or biomarkers research, and were published within the last few years. The tasks involve three kinds of data types: patient case or sample metadata; molecular data from individual samples; and results data aggregated across samples. All the answer datasets are either located in repositories or in supplementary files.

We experimented with GPT-4o to assess how well this LLM does in finding the dataset information for these tasks when given the answer paper. We found that it did very well -- it generated completely correct answers with just one prompt for 11 of 12 example tasks we assessed; the prompt we used and the GPT-4o responses are in the file ‘paper_search_tasks_with_additional_info.xls’. We also wrote a prompt for GPT-4o to extract information about all datasets in a paper and generate queries for each. The results were very good and suggest that we could use GPT-4o with light manual review to efficiently create a substantially sized benchmarking set for evaluation. This also highlights that LLMs like GPT-4o already do very well at identifying dataset information in individual papers, and that the current technical challenges probably lie with efficiently finding datasets over a large set of papers.

To use these tasks in an exercise that involves finding datasets within many papers, a corpus of papers needs to be specified. This could be a set of say 100 papers that includes the papers in which we identified answers, or it could be as broad as the whole PubMed Central open access set. The corpus size can be left up to performers during tool development; we can also provide paper lists of various sizes with domain-relevant content. We also have ideas for metrics for measuring success over a corpus of papers, which include accounting for correct answers in papers that were not evaluated for answers in generating the tasks.

Like the Capability A tasks, these paper search tasks were written to facilitate semi-automated evaluation and tend to be more specific than the Data Request and Results Request Prompts so that they have clear, correct answers in papers. In addition, this specificity makes it unlikely that the tasks will have many correct answers in many papers; this will reduce the need for manual review and facilitate semi-automated benchmarking. However, we think this does make the tasks less representative of the way many researchers will typically search for data, which is often an iterative process involving adding specifications as researchers see results, as described in the subsection 1.3 above, Variation of Capabilities A and B Exercises. Consequently, benchmarking tasks like this, while useful for assessing some aspects of tool performance, cannot substitute for user testing.

2.3 Capability C, Datasets for Harmonizing Metadata

We have gathered data sets for a harmonization exercise that is similar to Use Case Exercise 1, which was based on the paper Li et al. 2023 and involved harmonizing patient demographic and clinical data from tables with varying content and format from ten papers. This new exercise involves harmonizing sample metadata from ten source datasets and is based on a supplementary material table from Zethoven et al. 2022; this paper was the basis for one of the Data Request Use Cases. These materials are in the “Category_C_Harmonization_exercise” folder within the “Example_tasks” folder, and include a description of the exercise in the file “Description of Data for Harmonization Based on Zethoven.pdf”, pointers to associated datasets, and a script to obtain some of the datasets. We also have a version of this exercise written as a prompt to be given to an LLM-based tool, which is in “Prompt for an LLM tool to Harmonize GEO datasets.pdf”. Unlike in Use Case Exercise 1 where the challenge was to create harmonized data in a GDC format, for this exercise we do not specify a particular format. Instead, we view part of the challenge to involve selecting a format or creating a new format, as this is something researchers often do. Consequently, manual review of output will be required but as this is small in scope, this review would not be a lot of effort.

3. Supporting Materials Characterizing Challenges in Searching for Data in Repositories

To better understand the challenges and opportunities for improving search for data in repositories, we worked through a few use cases with currently available tools. This is captured in two documents in the folder “Materials_characterizing_search_challenges”.

First, we carried out an exploratory exercise to understand the current baseline for searching for datasets with repository user interfaces. We searched for data matching two of the Data Request Prompts in three National Cancer Institute data repositories using the repository user interfaces; we recorded the time it took and observations about what was involved and what was difficult. The file “Observations on challenges in finding data in repositories.pdf” describes this exploratory exercise and our findings. One key conclusion was that a particularly challenging and time-consuming component is the process of mapping an idea about data to the various variables and values in repositories.

Second, we performed a search for data based on a real-world example from a paper in cBioPortal using R. The search task involved multiple elements, and the search process involved many steps. The file “cbioprotal_search_example.html” has step-by-step code with comments about the thought processes involved. Notably, many steps involved asking about and observing what datasets, variables, and values were present in the repository. The search also involved leveraging domain knowledge, such as about different types of copy number alteration data formats and reference genomes, and involved combining data from the repository with supplementary data from a paper. In addition, multiple judgement calls were made about which types of data were acceptable, and how to reasonably combine data from different studies. This illustrates how searching repositories involves learning about the content of a resource, leveraging user domain knowledge, bringing in information from other sources, and human decision making.