LLMDapCat: An LLM-based Data Catalogue System for Data Sharing and Exploration

Shang Ferheng Karim[1], Aisha Kelifa[1], Amanda Marie Holsæter Kjær[1], Shanshan Jiang[2,*], Sondre Sørbø[2] and Dumitru Roman[2]

[1]*OsloMet - Oslo Metropolitan University, Norway*
[2]*SINTEF AS, Norway*

Abstract

Good data catalogues are essential for effective data sharing and discovery to cope with the rapid expansion of datasets and scientific literature available on the Web. In this paper, we present LLMDapCAT, an LLM-based metadata and data catalogue system that exploits Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) for efficient data profiling, sharing, and exploration. We demonstrate how the system serves both data providers and consumers: on the one hand, it allows providers to automatically generate standardized and semantically accurate metadata from scientific papers using an LLM and RAG-based pipeline, and to publish the metadata in the catalogue system; on the other hand, it enables consumers to browse available datasets and explore them in chat-like Q&A sessions using an external LLM service. The system can be applied to curate custom domain-specific scientific databases that facilitate search, understanding, and exploration of domain-specific datasets.

Source: https://github.com/SINTEF-SE/LLMDap

Demo: https://github.com/SINTEF-SE/LLMDapCat_Demo

Keywords

Data Catalogue System, Large Language Model, Retrieval Augmented Generation, Data Discovery, Data Exploration

1. Introduction

A key challenge in the current era of information overload is the limited effectiveness of traditional keyword-based search over scientific literature and shared datasets. With the exponential growth of available publications and data, platforms for data discovery must go beyond simple keyword matching. There is an increasing need for data discovery on the Web that leverages semantic matching, which requires high-quality, semantically rich metadata to accurately describe published datasets [1]. This is inline with the Findability, Accessibility, Interoperability, and Reuse (FAIR) principles of data sharing.

Manual annotation of metadata is labor-intensive, often inconsistent, and varies significantly across domains. Automated approaches augmented with human-in-the-loop feedback mechanisms emerges as a promising approach to address these limitations. Recent advancements in generative AI—particularly the development of Large Language Models (LLMs), including both general-purpose models such as ChatGPT, LLaMA, and Mistral, and domain-specific models such as BioGPT [2] and BioMedLM [3]—offer robust technical capabilities to enhance the scalability, accuracy, and contextual relevance of metadata annotation. Techniques such as Retrieval Augmented Generation (RAG) [4] further contribute to this progress by integrating external knowledge sources into the generation process, thereby improving factual grounding and domain adaptation.

Exploiting these recent advances, this paper presents the LLMDapCat web application, including a Streamlit[1]-based interactive user interface and two LLM/RAG pipelines for metadata generation and dataset exploration.

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

*Corresponding author.

kakashang96@msn.com (S. F. Karim); aishakelifa@outlook.com (A. Kelifa); ahk_95@hotmail.com (A. M. H. Kjær); shanshan.jiang@sintef.no (S. Jiang); sondre.sorbo@sintef.no (S. Sørbø); dumitru.roman@sintef.no (D. Roman)

[1] https://streamlit.io

Our main contributions are as follows:

- We introduce a data catalogue system designed for intuitive data sharing, featuring an accessible user interface (UI) and automated metadata generation using LLMs. The generated metadata is aligned with domain ontologies to ensure consistency and improve findability.
- We present a method to curate customized, domain-specific scientific databases that support research exploration and analysis.

2. LLMDapCat: A Web Application for FAIR Data Sharing

LLMDapCat provides a ChatGPT-like interface for dataset profiling and discovery. While the system is demonstrated using biomedical datasets, it can be extended to other scientific domains.

2.1. System Architecture

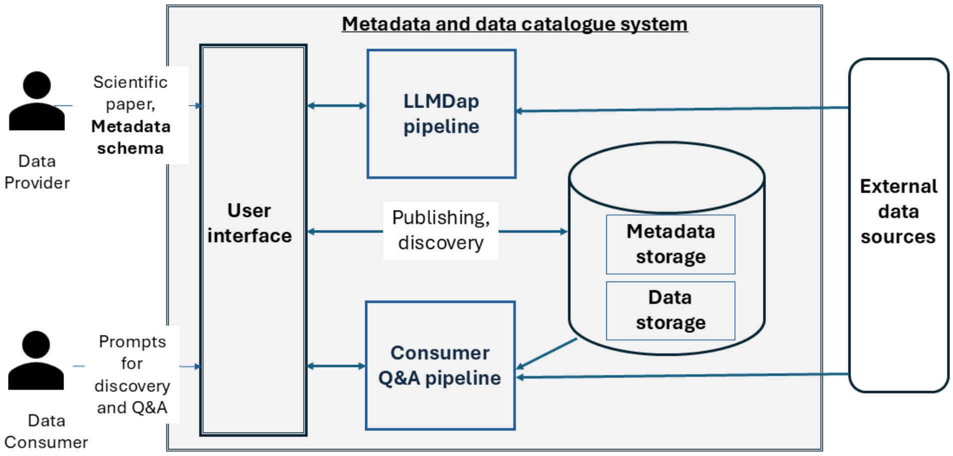


Figure 1: Architecture for LLMDapCat.

Figure 1 illustrates the architecture of the LLMDapCat application, which consists of the following components:

- **Streamlit UI:** A user interface with multiple pages, including Provider View, Dataset Browser, Consumer Q&A, and Configuration.
- **LLMDap Pipeline:** An LLM and RAG-based backend pipeline for automatic extraction of dataset metadata as described in [5]. The pipeline consists of the following steps:
  - *Data processing:* Extract and preprocess text from scientific papers.
  - *Document chunking:* Split documents into semantically coherent segments.
  - *Context retrieval:* Retrieve relevant chunks using vector-based search.
  - *LLM generation:* Generate metadata or answers using LLM prompts informed by retrieved context.
- **Consumer Q&A Pipeline:** A front-end pipeline that interacts with an external LLM service for answering user queries.
- **Metadata and Data Storage:** Metadata is stored in an SQLite database, while user-uploaded files (e.g., PDFs) are stored locally.
- **External Data Sources:** External APIs such as NCBI E-utilities[2] [6], BioStudies[3] [7], ArrayExpress[4] [8], and EuropePMC[5] [9] are used to access external data sources for metadata enrichment.

[2] https://www.ncbi.nlm.nih.gov/home/develop/api
[3] https://www.ebi.ac.uk/biostudies
[4] https://www.ebi.ac.uk/biostudies/arrayexpress
[5] https://europepmc.org/RestfulWebService

LLMDapCat provides two main user interfaces:

- **Provider View:** Allows users to submit textual documents and automatically generate metadata using the LLMDap pipeline.
- **Consumer View:** Enables users to search, explore, and query datasets using a ChatGPT-style Q&A interface.

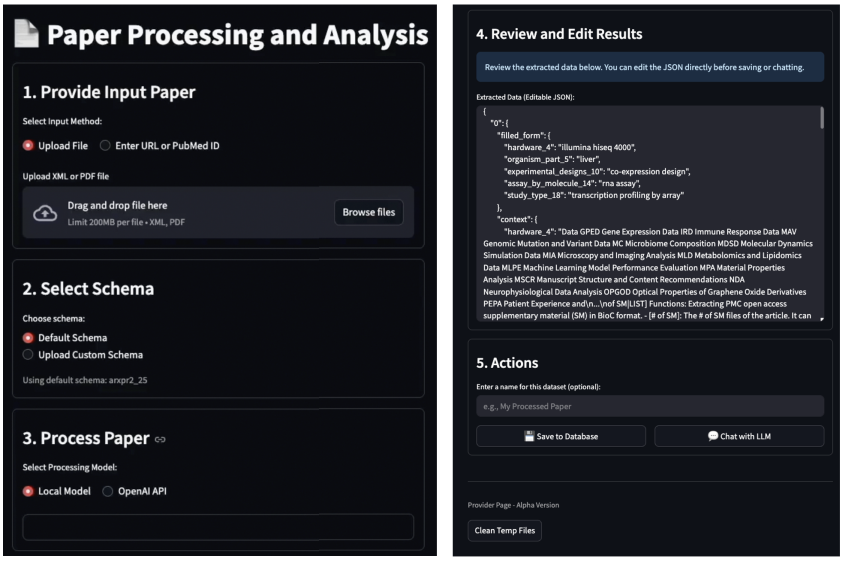
Metadata Schema: One important input to the system is the metadata schema that includes metadata fields defined according to domain ontologies to ensure semantic alignment and interoperability. The LLMDap will extract metadata from the scientific papers based on the information contained in the schema, consisting of metadata name, description, and value ranges.

2.2. Demonstration

Provider View (Figure 2)

This page allows users (data providers) to submit documents (e.g., research papers) for automated LLM-based metadata extraction with the following steps:

1. *Provide input paper:* Users can upload a PDF/XML file or input a URL/PubMed ID.
2. *Select schema:* Users can use a default or custom JSON schema to guide metadata generation.
3. *Process paper:* Clicking "Process Input" button sends the paper and schema to the backend LLM pipeline for metadata generation.
4. *Review metadata:* The generated metadata is displayed for user validation or edits.
5. *Save results:* When confirmed, the metadata is saved to the database and linked with the input document.



Paper Processing and Analysis

1. Provide Input Paper

Select Input Method:

● Upload File ○ Enter URL or PubMed ID

Upload XML or PDF file

2. Select Schema

Choose schema:

● Default Schema

○ Upload Custom Schema

Using default schema: arxpr2_25

3. Process Paper

Select Processing Model:

● Local Model ○ OpenAI API

4. Review and Edit Results

Review the extracted data below. You can edit the JSON directly before saving or chatting.

Extracted Data (Editable JSON):

```
{
  "0": {
    "filled_form": {
      "hardware_4": "illumina hiseq 4000",
      "organism_part_5": "liver",
      "experimental_designs_10": "co-expression design",
      "assay_by_molecule_14": "rna assay",
      "study_type_18": "transcription profiling by array"
    },
    "context": {
      "hardware_4": "Data GPED Gene Expression Data IRD Immune Response Data MAV
Genomic Mutation and Variant Data MC Microbiome Composition MDSD Molecular Dynamics
Simulation Data MIA Microscopy and Imaging Analysis MLD Metabolomics and Lipidomics
Data MLPE Machine Learning Model Performance Evaluation MPA Material Properties
Analysis MSCR Manuscript Structure and Content Recommendations NDA
Neurophysiological Data Analysis OPGOD Optical Properties of Graphene Oxide Derivatives
PEPA Patient Experience and\n...\nof SM|LIST] Functions: Extracting PMC open access
supplementary material (SM) in BioC format. - [# of SM]: The # of SM files of the article. It can
```

5. Actions

Enter a name for this dataset (optional):

e.g., My Processed Paper

Provider Page - Alpha Version

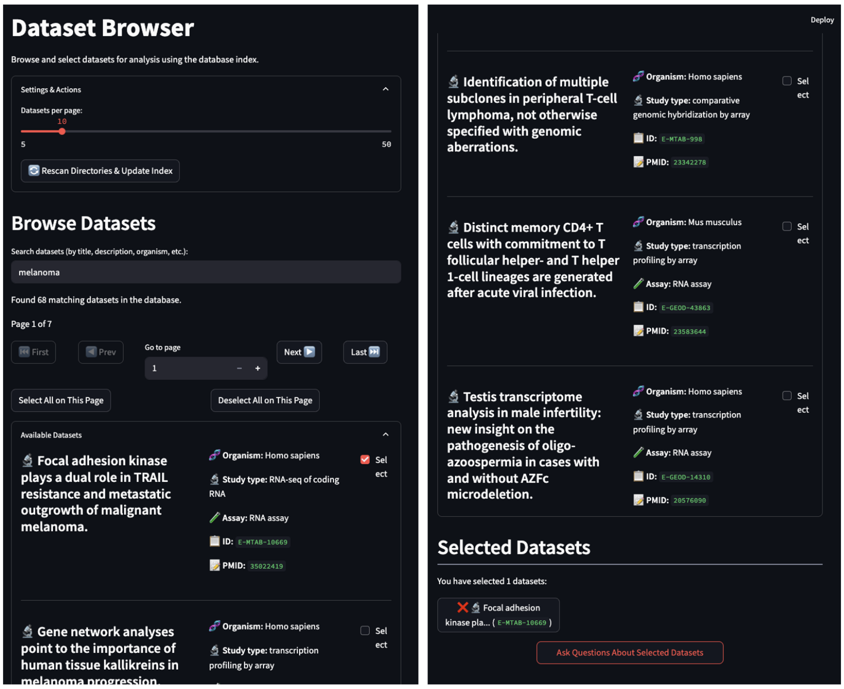
Figure 2: The provider view.

Dataset Browser (Figure 3)

This page offers dataset browsing and management functionality. It allows users (data providers or consumers) to browse, search and select datasets processed and indexed in the catalogue. After a dataset is selected for Q&A (for data consumers), it will initiate the Consumer View page. In addition, this page offers an optional database utility function to allow users (data providers or system administrators) to initiate index-update in addition to the automatic index-update associated with uploading of datasets and their metadata.

1. *Browse datasets:* A paginated view displays key metadata for available datasets.

2. *Search datasets:* A search bar allows keyword-based filtering.
3. *Select for Q&A:* Users select datasets via checkboxes and initiate Q&A by clicking on the "Ask Questions About Selected Datasets" button shown on the right image of Figure 3.
4. *Update index:* Users can press the "Rescan Directories & Update Index" button (above the *Browse Datasets* header in the left image of Figure 3) and start a background job to rescan and reindex datasets.



Dataset Browser

Browse and select datasets for analysis using the database index.

Settings & Actions

Datasets per page: 10

Browse Datasets

Search datasets (by title, description, organism, etc.): melanoma

Found 68 matching datasets in the database.

Page 1 of 7

Go to page: 1

Available Datasets

Focal adhesion kinase plays a dual role in TRAIL resistance and metastatic outgrowth of malignant melanoma.
Organism: Homo sapiens
Study type: RNA-seq of coding RNA
Assay: RNA assay
ID: E-MTAB-10669
PMID: 35022419
☑ Select

Gene network analyses point to the importance of human tissue kallikreins in melanoma progression.
Organism: Homo sapiens
Study type: transcription profiling by array
□ Select

Identification of multiple subclones in peripheral T-cell lymphoma, not otherwise specified with genomic aberrations.
Organism: Homo sapiens
Study type: comparative genomic hybridization by array
ID: E-MTAB-998
PMID: 23342278
□ Select

Distinct memory CD4+ T cells with commitment to T follicular helper- and T helper 1-cell lineages are generated after acute viral infection.
Organism: Mus musculus
Study type: transcription profiling by array
Assay: RNA assay
ID: E-GEOD-43863
PMID: 23583644
□ Select

Testis transcriptome analysis in male infertility: new insight on the pathogenesis of oligo-azoospermia in cases with and without AZFc microdeletion.
Organism: Homo sapiens
Study type: transcription profiling by array
Assay: RNA assay
ID: E-GEOD-14310
PMID: 20576090
□ Select

Selected Datasets

You have selected 1 datasets:

Focal adhesion kinase pla... (E-MTAB-10669)

5
50
Rescan Directories & Update Index
First
Prev
Next
Last
Select All on This Page
Deselect All on This Page
Deploy
Ask Questions About Selected Datasets

Figure 3: The dataset browser.

Consumer View (Figure 4)

After users (data consumers) selected datasets for Q&A, the Consumer View page is displayed for the datasets selected:

1. *Display context:* A summary of selected datasets is shown as Q&A context.
2. *Ask question:* Users type a question and submit it to the LLM system.
3. *Display answer:* The system provides a generated answer with chat history maintained.

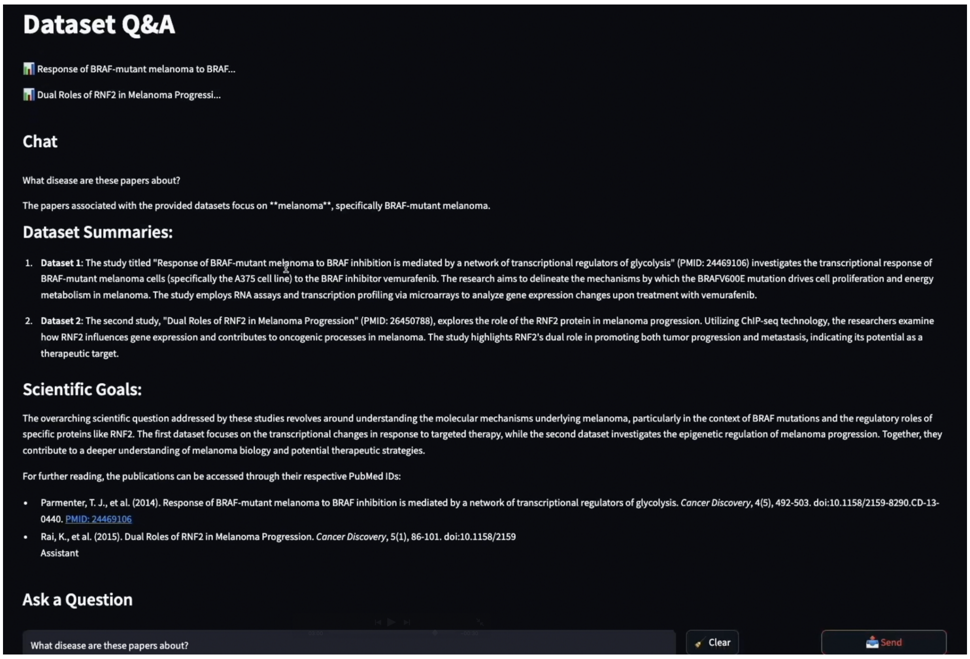
The **configuration** page (Figure 5) allows selection of LLM models, tuning of parameters (e.g., temperature, max tokens), and prompt template customization.

Semantic Technologies: Metadata fields are aligned with established ontologies to ensure that LLM outputs are both accurate and interoperable.

3. Conclusion and Future Work

We have introduced LLMDapCat, a web-based application and LLM-backed pipelines for enabling FAIR data sharing and exploration. Our approach uses RAG and LLMs to improve the quality and trustworthiness of generated metadata and Q&A interactions. Quantitative evaluation has been conducted on the proposed system to validate the performance of the pipeline and showed systemic improvement in the annotation task [10].

In future work, we plan to integrate domain ontologies more tightly into the profiling process to extend and refine metadata schemas. This will enhance semantic discovery, accuracy, and coverage of



Dataset Q&A

📊 Response of BRAF-mutant melanoma to BRAF...

📊 Dual Roles of RNF2 in Melanoma Progressi...

Chat

What disease are these papers about?

The papers associated with the provided datasets focus on **melanoma**, specifically BRAF-mutant melanoma.

Dataset Summaries:

1. **Dataset 1**: The study titled "Response of BRAF-mutant melanoma to BRAF inhibition is mediated by a network of transcriptional regulators of glycolysis" (PMID: 24469106) investigates the transcriptional response of BRAF-mutant melanoma cells (specifically the A375 cell line) to the BRAF inhibitor vemurafenib. The research aims to delineate the mechanisms by which the BRAFV600E mutation drives cell proliferation and energy metabolism in melanoma. The study employs RNA assays and transcription profiling via microarrays to analyze gene expression changes upon treatment with vemurafenib.
2. **Dataset 2**: The second study, "Dual Roles of RNF2 in Melanoma Progression" (PMID: 26450788), explores the role of the RNF2 protein in melanoma progression. Utilizing ChIP-seq technology, the researchers examine how RNF2 influences gene expression and contributes to oncogenic processes in melanoma. The study highlights RNF2's dual role in promoting both tumor progression and metastasis, indicating its potential as a therapeutic target.

Scientific Goals:

The overarching scientific question addressed by these studies revolves around understanding the molecular mechanisms underlying melanoma, particularly in the context of BRAF mutations and the regulatory roles of specific proteins like RNF2. The first dataset focuses on the transcriptional changes in response to targeted therapy, while the second dataset investigates the epigenetic regulation of melanoma progression. Together, they contribute to a deeper understanding of melanoma biology and potential therapeutic strategies.

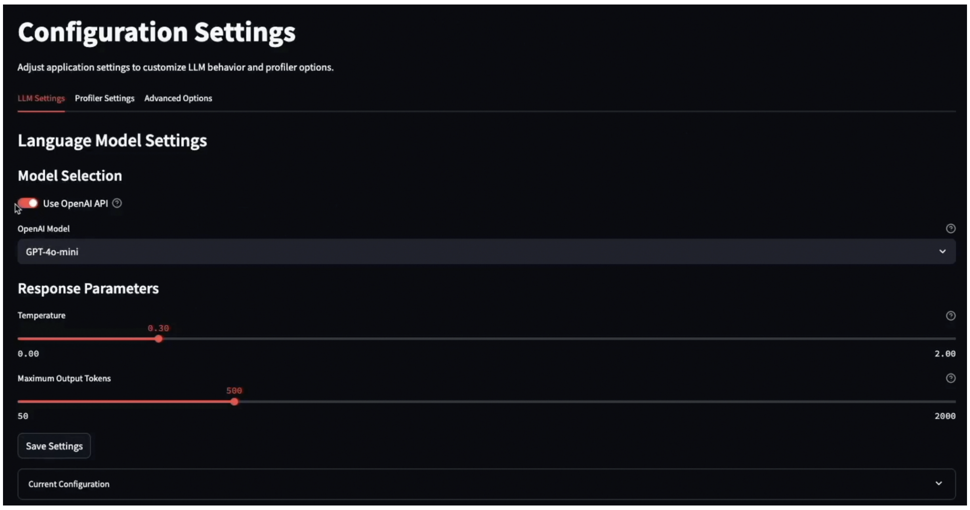
For further reading, the publications can be accessed through their respective PubMed IDs:

- Parmenter, T. J., et al. (2014). Response of BRAF-mutant melanoma to BRAF inhibition is mediated by a network of transcriptional regulators of glycolysis. *Cancer Discovery*, 4(5), 492-503. doi:10.1158/2159-8290.CD-13-0440. PMID: 24469106
- Rai, K., et al. (2015). Dual Roles of RNF2 in Melanoma Progression. *Cancer Discovery*, 5(1), 86-101. doi:10.1158/2159
  Assistant

Ask a Question

What disease are these papers about?

Figure 4: The consumer view.



Configuration Settings

Adjust application settings to customize LLM behavior and profiler options.

Language Model Settings

Model Selection

Use OpenAI API

OpenAI Model: GPT-4o-mini

Response Parameters

Temperature: 0.30 (0.00 – 2.00)

Maximum Output Tokens: 500 (50 – 2000)

Figure 5: The configuration page.

domain-specific metadata. In addition, qualitative user evaluation is planned to validate the usefulness of the system.

Potential Impact: The proposed system facilitates semantic data discovery and exploration for researchers. LLMDapCat can also be used to build customized scientific metadata catalogues in any domain by tailoring the metadata schema and integrating with domain-specific APIs and ontologies.

Acknowledgments

The work is funded through the projects UPCAST (HE 101093216), enRichMyData (HE 101070284), and DataPACT (HE 101189771).

Declaration on Generative AI

The author(s) used GPT-4 for grammar and spelling checks. The author(s) have reviewed and edited the content and take full responsibility for the publication's content.

References

[1] S. Jiang, T. F. Hagelien, M. Natvig, J. Li, Ontology-based semantic search for open government data, in: 2019 IEEE 13th International Conference on Semantic Computing (ICSC), 2019, pp. 7–15. doi:10.1109/ICOSC.2019.8665522.

[2] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics 23 (2022) bbac409.

[3] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin, et al., Biomedlm: A 2.7 b parameter language model trained on biomedical text, arXiv preprint arXiv:2403.18421 (2024).

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[5] S. Jiang, S. Sørbø, P. Tinn, S. F. Karim, D. Roman, Llmdap: Llm-based data profiling and sharing, in: VLDB 2025 Workshop: 3rd Data EConomy Workshop (DEC), 2025.

[6] E. Sayers, A general introduction to the e-utilities, Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US) (2010).

[7] U. Sarkans, M. Gostev, A. Athar, E. Behrangi, O. Melnichuk, A. Ali, J. Minguet, J. C. Rada, C. Snow, A. Tikhonov, et al., The biostudies database—one stop shop for all data supporting a life sciences study, Nucleic acids research 46 (2018) D1266–D1270.

[8] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, A. Brazma, Arrayexpress—a public database of microarray experiments and gene expression profiles, Nucleic Acids Research 35 (2006) D747–D750. URL: https://doi.org/10.1093/nar/gkl995. doi:10.1093/nar/gkl995. arXiv:https://academic.oup.com/nar/article-pdf/35/suppl_1/D747/3893619/gkl995.pdf.

[9] S. Rosonovski, M. Levchenko, R. Bhatnagar, U. Chandrasekaran, L. Faulk, I. Hassan, M. Jeffryes, S. I. Mubashar, M. Nassar, M. Jayaprabha Palanisamy, M. Parkin, J. Poluru, F. Rogers, S. Saha, M. Selim, Z. Shafique, M. Ide-Smith, D. Stephenson, S. Tirunagari, A. Venkatesan, L. Xing, M. Harrison, Europe pmc in 2023, Nucleic Acids Research 52 (2023) D1668–D1676. URL: https://doi.org/10.1093/nar/gkad1085. doi:10.1093/nar/gkad1085. arXiv:https://academic.oup.com/nar/article-pdf/52/D1/D1668/55040834/gkad1085.pdf.

[10] P. Tinn, S. Sørbø, S. Jiang, K. Voutetakis, S. M. Giounis, E. Pilalis, O. Papadodima, D. Roman, Pre-meta: Priors-augmented retrieval for llm-based metadata generation, Bioinformatics (2025). Accepted for publication.