Project documentation: Synonym improvement

❓Unmet needs

In the course of the search result relevance exercise, a comprehensive analysis of 20 queries was undertaken to evaluate the performance of the search algorithms. This examination revealed a critical disparity between user expectations and the actual outcomes of the search process. Furthermore, a subset of these queries clearly highlighted the root cause of this significant gap, some of which includes:

1. **Biological Cohorts Identification and Prioritization:**
   For the queries like "*response to exercise in young and old individuals,*" the algorithm fails to recognize and prioritize the crucial element of biological cohorts, thereby hindering the precision and relevance of search results.
2. **Acronyms Handling and Keyword Expansion:**
   Examining the query "*npm-1 mutated cell-lines in AML,*" we observe that the system struggles to handle acronyms effectively, leading to suboptimal results. Additionally, keyword expansion, at times, leads to associations with irrelevant subtypes.
3. **Prioritization of Search Results**:
   For the queries like "*differentially expressed miRNAs in lung cancer*" the search algorithm does not adequately prioritize miRNAs. Instead, it tends to expand and rank "lung cancer" higher, resulting into more partially relevant results.

In this document, we focus on a specific gap particularly in **handling acronyms and keyword expansion** as previously highlighted**.** The gap is underscored by an illustrative example derived from the exercise:

Query:

- npm-1 mutated cell-lines in AML

Expected outcome:

- The expected outcome was to retrieve information about "npm-1 mutated cell-lines"(nucleophosmin1 gene mutation) in the context of "AML" (acute myeloid leukemia).

What did not work well:

- The search result was mostly associated with “Acute promyelocytic leukemia (APML)," a subtype of AML.

Why it did not work well:

- The current data model failed to recognize "AML" as an acronym for "acute myeloid leukemia."
- Although "AML" was identified as a synonym for "acute promyelocytic leukemia (APML)," the system predominantly presented results pertaining to the more specific term "APML."
- This bias led to an outcome that failed to encompass the broader context of "AML."

The issue highlights the critical role of keyword expansion in ensuring contextually accurate results aligned with user intentions. In this instance, the challenge stemmed from a lack of comprehensive synonym enrichment, not accounting for diverse expressions of medical terms used by users.

🏹 Objectives

This project aims to:

- **Keyword Expansion:** Introduce a wider array of synonyms to disease terms, broadening search keywords for enhanced query results.
- **Improve Result Relevance:** Increase the relevancy of search outcomes by incorporating more synonyms, addressing user queries comprehensively.

Methodology

The steps for the exercise in synonym enrichment are shown in the flowchart below:

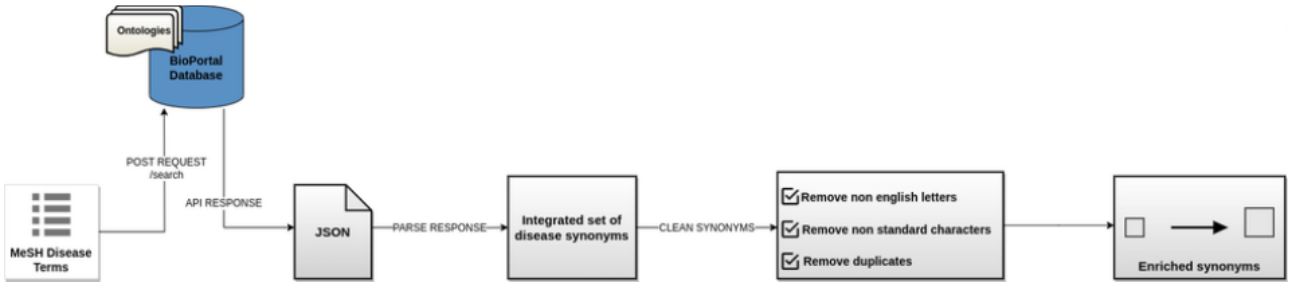
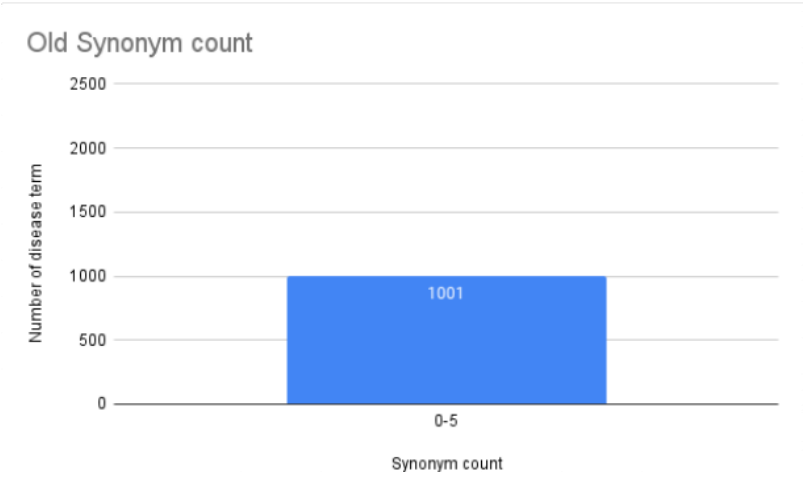


Figure1: Synonym enrichment methodology

Result analysis

The following histogram illustrates the distribution of synonyms before and after the intervention.



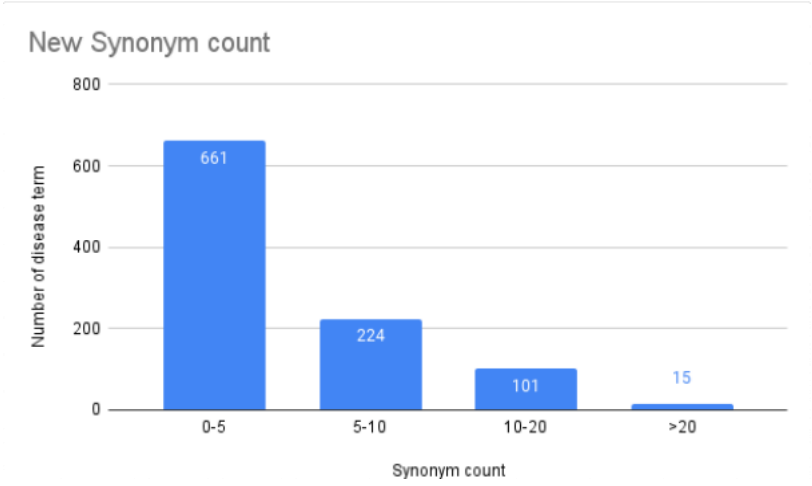


Figure 2: Histogram depicting the distribution of synonyms before and after enrichment.

Impact and Benefits

Here are a few illustrative examples demonstrating how synonym addition worked well vs terms where it did not:

Worked well:

Disease term	Original synonyms	New synonyms	Example improvement
NGLY1 deficiency	1. Congenital disorder of deglycosylation 2. Alacrimia-choreoathetosis-liver dysfunction syndrome	1. CDDG 2. deficiency of N-glycanase 1 3. CDG IV 4. congenital disorder of glycosylation, type IV	Discover relevant information even if they use alternative terms or abbreviations like “CDDG“ or “CDG IV“
Growth hormone excess	_	1. somatotroph adenoma 2. pituitary giant	Enhanced coverage of related concepts and specificity
Amoebiasis due to Entamoeba histolytica	_	1. amebic dysentery 2. Amebiases, intestinal 3. Colitides, amoebic 4. entamoebiasis, intestinal 5. colitis, amoebic	Enhanced coverage of related concepts and specificity
Alexanders leukodystrophy	_	1. AxD 2. Alexander's disease 3. megalencephaly in infancy accompanied by progressive spasticity and dementia 4. ALXDRD	Discover relevant information even if they use alternative terms or abbreviations
Muscular dystrophy, limb-girdle, type 1A	_	1. myofibrillar myopathy type 3 2. limb-girdle muscular dystrophy due to myotilin deficiency 3. distal myotilinopathy	Increased coverage of disease subtypes

		4. spheroid body myopathy 5. LGMD1A 6. MFM3 7. MYOT autosomal dominant distal myopathy	
Pancreatic carcinoma, familial	_	1. cancer of pancreas 2. exocrine cancer 3. pancreatic cancer 4. pancreatic cancer not islets 5. carcinoma of exocrine pancreas	Increased coverage of alternative terms
Familial primary gastric lymphoma	1. Gastric Lymphoma	1. MALT lymphoma 2. lymphoma of mucosa-associated lymphoid tissue 3. Extranodal marginal zone B-cell lymphoma 4. Immunocytoma 5. MALToma 6. gastric lymphoma primary	Enhanced representation of specific cancer or its primary site
Malignant mesenchymal tumor		1. soft tissue sarcoma 2. non-Rhabdo soft tissue sarcoma 3. connective tissue sarcoma 4. soft part sarcoma 5. malignant soft tissue tumor	Enhanced coverage of specificity
Leukemia, Myeloid, Acute		1. leukemia, myelogenous, acute 2. leukemia, acute nonlymphocytic 3. acute myeloblastic leukemia 4. anll 5. acute myeloid leukemia with maturation 6. acute myeloid leukemia without maturation 7. Acute leukemic myelosis 8. Acute granulocytic leukemia 9. AML 10. myeloid leukemia, acute, **m1** 11. myeloid leukemia, acute, **m2**	Enhanced representation of the disease

Did not work well:

Disease term	Original synonyms	New synonyms	Limitations
Mucor infections	1. Mucor infections	1. infection by mucor 2. mucor infection	Limited scope for expanding keyword variations
Diffuse alopecia	1. Diffuse alopecia 2. Patchy alopecia 3. Vitiligo 4. Alopecia Celsi	1. alopecia diffuse 2. diffuse alopecia	Limited scope for expanding keyword variations
Acute malaria	_	1. Chronic malaria	Minimal addition of synonyms
Cerebral astrocytoma, adult	1. Adult cerebral astrocytoma	1. Adult cerebral astrocytoma	Limited scope for expanding keyword variations

👀 Constraints faced

1. **Iterative API Requests:**

- The process of sending individual API requests for each disease term led to significant time consumption.
- The inability to batch multiple terms together hindered the efficiency of the synonym enrichment process.

2. **Pagination Impact:**

- The API responses were structured with multiple pages, necessitating additional requests for each page of every disease term.
- This approach further prolonged the time required for data retrieval and processing.

Potential Solutions:

1. **Batch Processing:** Combine multiple disease terms in a single API request to save time
2. **Efficient Pagination :** Streamline the process by retrieving all term pages in one request.