Searching for Reliable Facts over a Medical Knowledge Base

Fabio Giachelle
fabio.giachelle@unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Stefano Marchesin
stefano.marchesin@unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Gianmaria Silvello
gianmaria.silvello@unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Omar Alonso*
omralon@amazon.com
Amazon
Santa Clara, California, USA

ABSTRACT

This work presents CoreKB, a Web platform for searching reliable facts over gene expression-cancer associations Knowledge Base (KB). It provides search capabilities over an RDF graph using natural language queries, structured facets, and autocomplete. CoreKB is designed to be intuitive and easy to use for healthcare professionals, medical researchers, and clinicians. The system offers the user a comprehensive overview of the scientific evidence supporting a medical fact. It provides a quantitative comparison between the possible gene-cancer associations a particular fact can reflect.

CCS CONCEPTS

• **Information systems → Web searching and information discovery**; **Search interfaces**.

KEYWORDS

Knowledge Discovery, Fact Search, Gene Cancer Associations

ACM Reference Format:
Fabio Giachelle, Stefano Marchesin, Gianmaria Silvello, and Omar Alonso. 2023. Searching for Reliable Facts over a Medical Knowledge Base. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591822

1 INTRODUCTION

According to the World Health Organization, cancer prevention has become one of the most critical public health challenges of the 21st century. In recent years, cancer research has increasingly relied on microarray and next-generation sequencing technologies, which produce a wealth of experimental data on gene expression-cancer interactions [4, 11]. These interactions contain critical information for diagnosing cancer, assessing prognosis, and predicting therapy response [6, 14]. However, the sheer volume and complexity of the data produced can make it difficult to analyze and derive meaningful insights. Peer-reviewed publications often describe the analysis and interpretation of this experimental data, making scientific literature a critical source to complement and validate them.

It appears clear that the need for efficient and reliable methods to store and organize knowledge from various sources has become increasingly important. To this end, KBs have emerged as essential resources for cancer researchers [10, 13]. They contain organized and structured information (i.e., scientific facts) that can be used to support data-driven research. KBs are often represented as graphs, where nodes represent concepts or entities, and edges represent relationships between them. To make the underlying data accessible, machine-readable, and interoperable, KBs can be represented as a Resource Description Framework (RDF) graph and queried with SPARQL [15]. However, searching within KBs can be daunting for several reasons. First, these resources can be schema-less, meaning that the data structure is not explicitly defined in a fixed schema. This lack of a fixed schema can make it challenging to understand the organization of the data within the KB and know how to formulate the SPARQL queries to retrieve the required information. It may also lead to ambiguity and confusion when trying to interpret search results. Second, searching a KB through SPARQL queries is difficult, even for expert users [16]. Third, the sheer volume of data within a KB can make it hard to identify and extract relevant information. The complexity of the queries required to retrieve that information can add an additional layer of difficulty [2, 16].

In addition, cancer research requires highly specialized KBs, comprising unique terminologies and ontologies. Within these resources, the large presence of complex terms and concepts and the use of acronyms and morphosyntactic variants pose a critical challenge for search [1]. Even with a thorough understanding of such specialized jargon, formulating effective search queries is difficult. As a result, researchers might struggle to effectively use the data within KBs, which could limit the potential for discoveries and insights. Therefore, it is necessary to have easy-to-use search applications that can quickly access and retrieve relevant information [16].

Contribution We present CoreKB, a Web-based search platform to discover gene expression-cancer associations, which is publicly available at https://gda.dei.unipd.it along with a demonstration video (https://gda.dei.unipd.it/static/videos/demo.mp4).

CoreKB builds upon one of the largest KBs containing fine-grained facts about gene expression-cancer associations [12], extracted semi-automatically from the vast scientific literature by

*Work done prior to joining Amazon

mining PubMed. The KB comprises more than 230,000 fine-grained facts, each categorized as reliable or unreliable based on the amount of supporting or conflicting evidence available. CoreKB provides a streamlined and intuitive interface for searching and exploring knowledge about these associations. The search platform is equipped with natural language and faceted search, enabling users to search for entities and facts easily. CoreKB includes several features, such as infometrics and entity cards. It is intended for specialized users, such as medical researchers and clinicians, who want to quickly and easily search for knowledge about gene expression-cancer associations without knowing the exact terminology or concept code of interest. Compared to other search applications that are sentence-oriented, CoreKB takes a fact-oriented approach by providing users with a comprehensive overview of the key information concerning each fact. This includes gene expression-cancer associations and the corresponding aggregated data, allowing users to assess the consensus supporting a fact. The aggregated data consists of: (i) the gene class (role) distribution among the sentences extracted from the scientific literature; (ii) the number of supporting and conflicting sentences; and (iii) the number of supporting publications per year. Hence, CoreKB offers a powerful platform for comprehensive knowledge discovery in precision medicine.

Related Work. DEXTER [8], a text mining approach to extract associations from scientific literature, provides a basic search endpoint that requires users to search for gene-cancer pairs and does not support natural language search. OncoMX [5] offers a unified and easy-to-use interface for different datasets, but its literature mining capabilities are limited. OncoSearch [9] allows users to search for sentences that mention gene expression changes in cancer and offers advanced faceted search functionalities, but lacks natural language search and does not provide entity cards or infometrics for supporting evidence. DisGeNET [7] provides a well-designed faceted navigation system, but it only returns gene-disease associations at the sentence level, making it difficult to establish reliable associations for gene-disease pairs (fact level). Finally, BioKB [3] provides access to the semantic content of biomedical literature, but it does not support natural language search and focuses on the exploration and visualization of the structure of the underlying KB.

2 COREKB

Figure 1 depicts the Knowledge Base Construction (KBC) system used to build the large-scale KB on gene expression-cancer associations (left side), together with the CoreKB architectural components (right side).

2.1 Knowledge Base Construction

To build the KB that underpins CoreKB, we used a system that acquires text from literature and processes it to obtain sentences. The sentences pass through a Named Entity Recognition and Disambiguation (NERD) component that annotates them with gene-cancer pairs and then undergoes bootstrapping and deployment processes. In the bootstrapping process, fine-grained relationships between entities are manually annotated, and these annotations are used to train Relation Extraction (RE) methods and populate the KB. In the deployment process, automatic annotations are provided by the RE methods, and facts are generated to populate the KB. Each

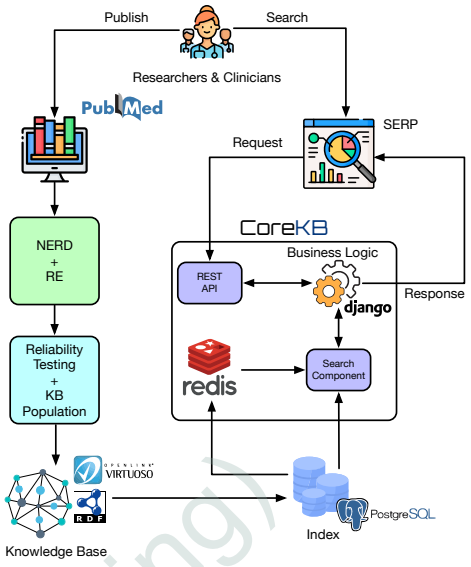

Figure 1: Overview of the KBC system (left side) and the CoreKB architectural components (right side).

fact within the KB is associated with a probability distribution that reflects the likelihood of assuming one of the possible gene-cancer associations. The probabilities are used to perform reliability tests that tag the facts as reliable or unreliable based on the collected evidence's quantity and level of consensus. This probabilistic approach allows the system to capture the uncertainty inherent in the scientific discourse and helps users to understand the strength of the evidence supporting a particular gene-cancer association.

The KB contains 23,879 genes and 11,530 cancer diseases for a total of 230,000 fine-grained facts, supported by 1,037,845 sentences from 251,038 research articles [12].

2.2 Search Platform Architecture

The architecture of CoreKB consists of (i) a web-based front-end interface built with *React.js*; (ii) a back-end for the business logic, REST APIs, and services built with the Python web framework *Django*; (iii) a PostgreSQL database coupled with the *Virtuoso* RDF triple store for saving the knowledge base data; (iv) *Redis* for efficient in-memory data store and access; and (v) a search/retrieval component implemented in Python which performs NERD on the entity mentions present in the user-provided query, by leveraging a Redis in-memory dictionary of entities, and then a structured search in the database. So, given a user query, we assign a maximum score for each entity in case of an exact match or proportional to the number of matching terms in case of a partial match (minus a discount proportional to the entity's length). When a single entity is recognized, all the related facts retrieved are ordered by scientific evidence support. Similarly, when multiple entities have been recognized, the facts concerning the most matching gene-cancer pairs are promoted in the results ranking. The KB contains facts about any living organism, not just humans; we rank human-specific genes on top.

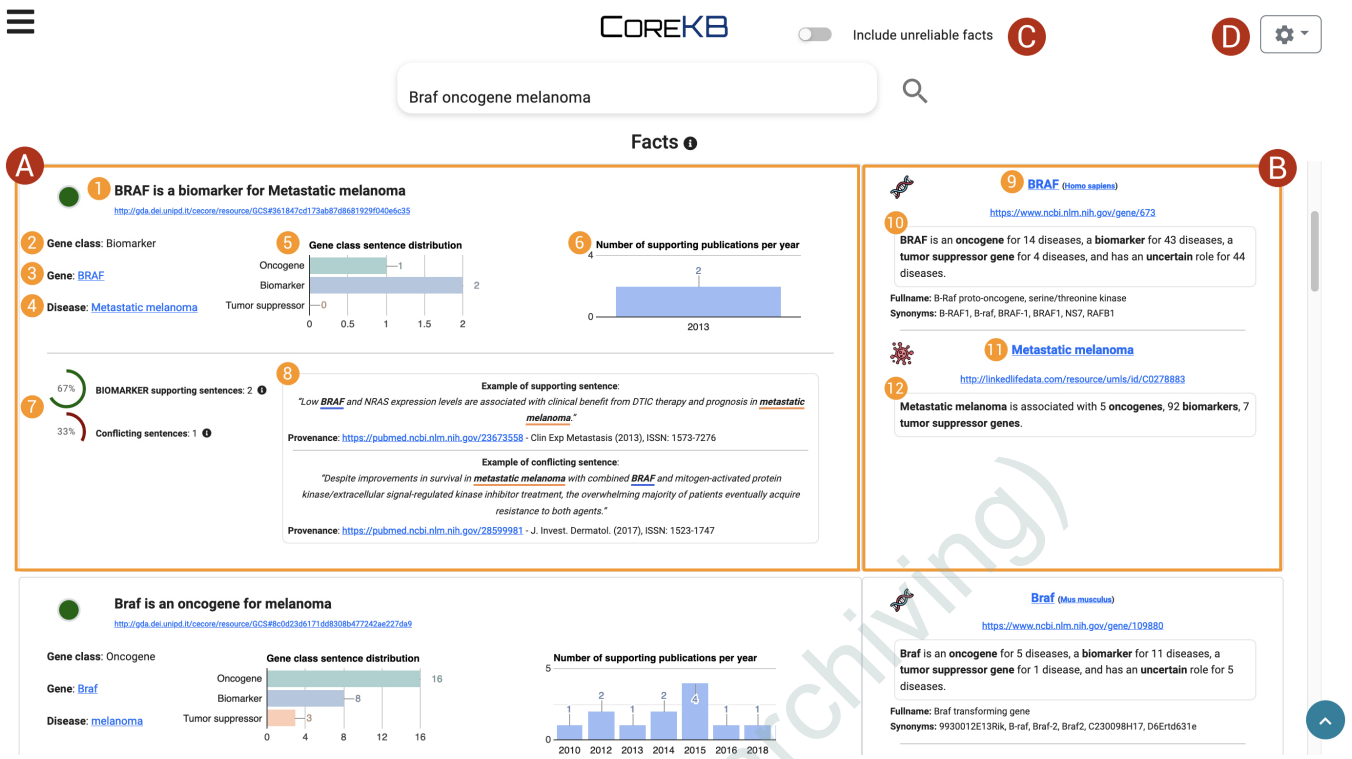

Figure 2: CoreKB Search Engine Results Page (SERP) for the query "*Braf oncogene melanoma*". The retrieved facts are organized as cards (A) providing information about the gene (3), cancer (4), and their relationship (2). The first fact depicted in the figure is "*BRAF is a biomarker for Metastatic melanoma*" (1). Two sentences support this fact (7). In contrast, a third conflicting sentence proposes the *oncogene* gene class instead of *biomarker*. Since the majority (two over three) of the scientific evidence available propose *BRAF* as a *biomarker*, the fact is considered reliable. Note that the gene *BRAF* is different from the *Braf* of the second result; the first is a human gene (9), while the second one is a gene for the *Mus musculus* organism.

2.3 User Interaction and Features

CoreKB allows the users to search for facts (i.e., gene-cancer associations supported by the scientific literature) and entities, namely, genes and cancers. It provides free-text and structured search facilities so that users can search using natural language queries or facets, providing autocomplete features. Users can switch from free-text to structured search interface and vice versa using the settings menu placed on the top-right corner of the interface, as shown in Figure 2.D. In addition, the interface provides three clickable sample queries to let the users try the system and easily realize its capabilities. Figure 2 shows the CoreKB SERP, where the user has a list of scientific facts matching the given query. In this regard, users can choose whether to include unreliable facts in the results or to view only the facts considered reliable, that is, with proper literature support and consensus on a specific gene class (i.e., *oncogene*, *biomarker*, and *tumor suppressor gene*) for a gene-cancer association. By default, users can view only *reliable* facts; they need to click on the dedicated switch (see Figure 2.C) to include the unreliable facts in the SERP. The facts are presented within cards (A), including gene class (2) and symbol (3); cancer label (4); statistics concerning the number of supporting and conflicting sentences (7); key examples of supporting and conflicting sentences as well as information about the associated publications (8); a horizontal bar chart presenting the gene class distribution among the related sentences extracted from the literature (5); and bibliometrics about the number of supporting publications per year (6). The fact claim is reported on the topside of the card (1) and is emphasized using boldface. On the left side of the fact claim, there is a circle that is colored differently according to the informativeness and reliability of the fact - i.e., green for reliable facts, gray for facts without proper support, and red for unreliable facts, that is, without proper consensus on a specific gene class. In addition, on the right side of each fact, card (B) reports specific information concerning the gene and cancer involved. Genes (9) and cancer diseases (11) are provided with links and references to the corresponding entries in the National Center for Biotechnology Information (NCBI)[1] and Linked Life Data[2] platforms. Besides, aggregated information (10, 12) is reported for each gene and disease, indicating, for instance, the number of cancer diseases a gene is associated with the role of oncogene, biomarker, or tumor suppressor. These numbers and all the other quantitative information are clickable so that users can easily consult them on dedicated landing pages. Similarly, a

[1]https://www.ncbi.nlm.nih.gov/gene/
[2]http://linkedlifedata.com/

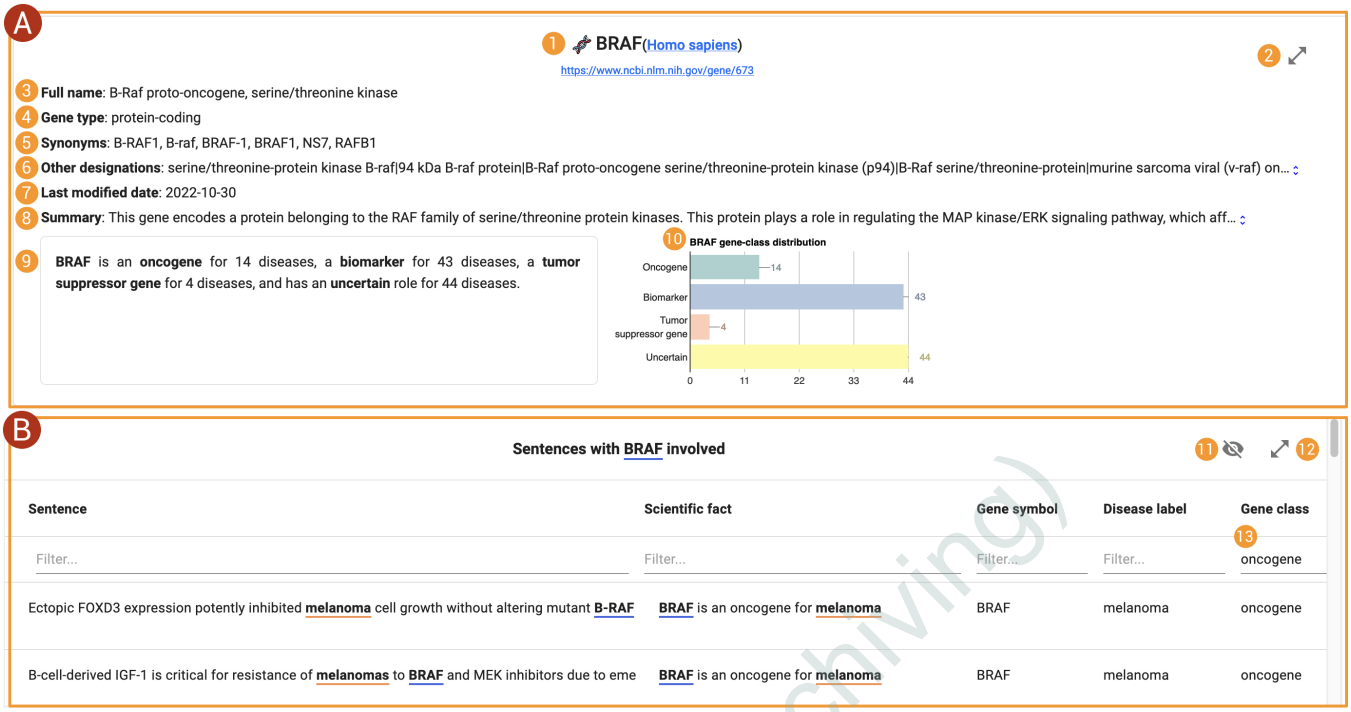

Figure 3: Landing page for the human gene *BRAF*. The interface consists of two cards; the first (A) reports gene-specific information, while the second (B) shows the sentences where *BRAF* is involved, according to the user-specified column filters and sorting. In the figure, only the sentences where *BRAF* is involved as an oncogene are shown due to the filtering on the *Gene class* column (13). Gene/cancer mentions are highlighted in the sentence text respectively in blue and orange.

dedicated landing page reports all available information for each entity, including quantitative and aggregated information, related facts, and supporting/conflicting sentences.

Figure 3 shows the landing page for the human gene *BRAF*. The interface is organized into two major cards (A, B); the first (A) presents all the information about the gene, such as its symbol (1), full name (3), type (4), synonyms (5), designations (6), last modified date (7), summary (8), and the gene class distribution for different cancer diseases both as textual content (9) and with a horizontal bar chart (10). For what concerns long textual information like the gene summary, by default, they are truncated to save space in the interface. Still, they can be expanded/collapsed by clicking the dedicated button after the ellipsis. Instead, the second card (B) reports the sentences involving the gene. The sentences are presented in tabular form, so users can narrow the sentences arbitrarily using filtering and sorting features. For instance, in Figure 3, we can notice that under the column *Gene class* (13), the value *oncogene* is specified, thus only the sentences where *BRAF* acts as an oncogene are shown. In case of long sentences to view, users can resize columns or hover with the mouse on a sentence to visualize a tool-tip with the full sentence's text or even click on the sentence to better visualize it on a separate pop-up. On the top-right corner of each card, two action buttons enable the users to (i) expand/collapse each card to fit the full-screen size (2, 12); (ii) choose the columns to be shown in the sentence table (11).

3 CONCLUSIONS

We presented CoreKB, a Web-based search platform for scientific facts concerning gene expression-cancer associations. CoreKB enables users to search using natural language queries or structured facets, providing autocomplete facilities. The platform provides a unique set of features unavailable in similar systems. CoreKB's fact-oriented approach distinguishes it from its sentence-oriented counterparts, offering users a comprehensive overview of the scientific evidence supporting or conflicting with a fact. Additionally, CoreKB provides users with a quantitative comparison between the possible gene-cancer associations a fact can reflect, making it easier to determine if there is a consensus on a specific gene role. It is worth noting that even though CoreKB focuses on gene expression-cancer associations, the same model can be adapted for other kinds of relationships with the necessary adjustments.

Through its set of advanced features, CoreKB aims to support clinicians and researchers in searching for reliable scientific findings and corroborating evidence, thus promoting knowledge discovery from a serendipity perspective. In future work, we plan to conduct a user study with clinicians to improve the search interface and integrate additional functionalities.

ACKNOWLEDGMENTS

The work was supported by the ExaMode project as part of the EU H2020 program under Grant Agreement no. 825292

REFERENCES

[1] M. Agosti, S. Marchesin, and G. Silvello. 2020. Learning Unsupervised Knowledge-Enhanced Representations to Reduce the Semantic Gap in Information Retrieval. *ACM Trans. Inf. Syst.* 38, 4 (2020), 38:1–38:48. https://doi.org/10.1145/3417996

[2] H. Bast, B. Buchhold, and E. Haussmann. 2016. Semantic Search on Text and Knowledge Bases. *Found. Trends Inf. Retr.* 10, 2-3 (2016), 119–271. https://doi.org/10.1561/1500000032

[3] M. Biryukov, V. Grouès, and V. P. Satagopam. 2017. BioKB - Text mining and semantic technologies for the biomedical content discovery. In *Proc. of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4LS 2017), Rome, Italy, December 4-7, 2017 (CEUR Workshop Proceedings, Vol. 2042)*. CEUR-WS.org. http://ceur-ws.org/Vol-2042/paper5.pdf

[4] P. Borry, H. B. Bentzen, I. Budin-Ljøsne, M. C. Cornel, H. C. Howard, O. Feeney, L. Jackson, D. Mascalzoni, Á. Mendes, B. Peterlin, B. Riso, M. Shabani, H. Skirton, S. Sterckx, D. Vears, M. Wjst, and H. Felzmann. 2018. The Challenges of the Expanded Availability of Genomic Information: an Agenda-Setting Paper. *J. Community Genet.* 9, 2 (2018), 103–116.

[5] H. M. Dingerdissen, F. Bastian, K. Vijay-Shanker, M. Robinson-Rechavi, A. Bell, N. Gogate, S. Gupta, E. Holmes, R. Kahsay, J. Keeney, H. Kincaid, C. H. King, D. Liu, D. J. Crichton, and R. Mazumder. 2020. OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data. *JCO Clin. Cancer Inform.* 4 (2020), 210–220.

[6] S. Dugger, A. Platt, and D. Goldstein. 2018. Drug development in the era of precision medicine. *Nat. Rev. Drug. Discov.* 17 (2018), 183–196.

[7] J. Piñero González, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong. 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, Database-Issue (2020), D845–D855.

[8] S. Gupta, H. Dingerdissen, K. E. Ross, Y. Hu, C. H. Wu, R. Mazumder, and K. Vijay-Shanker. 2018. DEXTER: Disease-Expression Relation Extraction from Text. *Database J. Biol. Databases Curation* 2018 (2018), bay045.

[9] H. J. Lee, T. Cuong Dang, H. Lee, and J. C. Park. 2014. OncoSearch: cancer gene search engine with literature evidence. *Nucleic Acids Res.* 42, Webserver-Issue (2014), 416–421.

[10] X. Li and J. L. Warner. 2020. A Review of Precision Oncology Knowledgebases for Determining the Clinical Actionability of Genetic Variants. *Front. Cell Dev. Biol.* 8 (2020). https://doi.org/10.3389/fcell.2020.00048

[11] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, and R. Ferrari. 2016. Genome, Transcriptome and Proteome: the Rise of Omics Data and Their Integration in Biomedical Sciences. *Briefings in Bioinformatics* 19, 2 (2016), 286–302.

[12] S. Marchesin, L. Menotti, G. Silvello, and O. Alonso. 2023. *CORE: Gene Expression-Cancer Knowledge Base.* https://doi.org/10.5281/zenodo.7577127

[13] S. Marchesin and G. Silvello. 2022. TBGA: a large-scale Gene-Disease Association dataset for Biomedical Relation Extraction. *BMC Bioinform.* 23, 1 (2022), 111.

[14] B. Neary, J. Zhou, and P. Qiu. 2021. Identifying Gene Expression Patterns Associated with Drug-Specific Survival in Cancer Patients. *Scientific Reports* 11, 1 (2021), 1–12.

[15] G. Weikum, X. L. Dong, S. Razniewski, and F. M. Suchanek. 2021. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Found. Trends Databases* 10, 2-4 (2021), 108–490.

[16] W. Wu. 2013. Proactive natural language search engine: tapping into structured data on the web. In *Proc. of the Joint 2013 EDBT/ICDT Conferences, EDBT '13, Genoa, Italy, March 18-22, 2013*. ACM, 143–148. https://doi.org/10.1145/2452376.2452394