DeepLife: An Entity-aware Search, Analytics and Exploration Platform for Health and Life Sciences

Patrick Ernst, Amy Siu, Dragan Milchevski, Johannes Hoffart, Gerhard Weikum
Max-Planck Institute for Informatics
Campus E1 4
66123 Saarbrücken, Germany
{pernst, siu, dmilchev, jhoffart, weikum}@mpi-inf.mpg.de

Abstract

Despite the abundance of biomedical literature and health discussions in online communities, it is often tedious to retrieve informative contents for health-centric information needs. Users can query scholarly work in PubMed by keywords and MeSH terms, and resort to Google for everything else. This demo paper presents the DeepLife system, to overcome the limitations of existing search engines for life science and health topics. DeepLife integrates large knowledge bases and harnesses entity linking methods, to support search and exploration of scientific literature, newspaper feeds, and social media, in terms of keywords and phrases, biomedical entities, and taxonomic categories. It also provides functionality for entity-aware text analytics over health-centric contents.

1 Introduction

There is an ever-growing abundance of biomedical information and health-related contents on the Internet: scientific publications in PubMed, ontologies and knowledge bases on genes, proteins, drugs, etc., health portals like the one by the Mayo Clinic, online communities where patients and doctors discuss diseases, therapies, drug side effects, etc., and more. However, this wealth of information is in contrast to the limited support of finding relevant contents, especially when laymen search for specific topics off the mainstream or when experts want high recall on advanced topics from many sources. A typical user approach is to combine keywords with Medical Subject Headings (MeSH) terms when searching PubMed, and to use Google for everything else.

As an example, consider a user who takes asthma medication and plans to go for a 3-month trip to China including rural areas. Which vaccinations are needed, which asthma drugs are not compatible with these vaccines or other drugs that may be needed and purchased during the trip (e.g. diarrhea, sinusitis, influenza)? What is the experience of other travelers? As an example for an expert user's needs, consider a medical student who is investigating the conditions and risk factors under which Zika spreads and causes health problems.

State of the Art and its Limitations: These kinds of users face the following shortcomings of available search engines:

- *Restricted search functionality:* The search engines for PubMed or health portals like uptodate.com or mayoclinic.org support only keyword queries with some support for MeSH-like annotations, but lack query functionality that can incorporate hierarchical taxonomies and linkage with knowledge bases. Search over social media sites is even more limited.
- *Limited coverage and diversity:* Other than Google, all search engines can tap only into one kind of content: either scholarly publications or user-provided social media, but never both. The same holds for intermediate-style contents like health portals.
- *Restriction to molecular entities:* For contents about genes, proteins, pathways, etc., there are structured-data sites that come with richer query and exploration functionality. However, for entities at the level of diseases, therapies,

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 19–24,
Berlin, Germany, August 7-12, 2016.

symptoms, risk factors, etc., there are no services of this kind.

- *Lack of support for interactive exploration:* The only user-friendly support for interactive sessions is auto-completion suggestions for user queries. However, these are solely based on the query-and-click history of previous users. This has no awareness of emerging topics in the underlying contents and entity-level background knowledge.

This state of the art for health-related search is in sharp contrast with the state of the art for general-purpose search, say over daily news or general-purpose social media (e.g., discussing celebrities). Advances in recognizing and disambiguating textual mentions of named entities and the linkage to comprehensive knowledge bases like DBpedia, Freebase, Wikidata and Yago have enabled powerful and user-friendly retrieval systems. Google supports entity-centric search through transparent interlinkage with the Google Knowledge Graph; Microsoft, Facebook, etc. have similar functionalities. Academic systems such as Broccoli (Bast and Buchhold, 2013), STICS (Hoffart et al., 2014) and Semantic Scholar (Valenzuela et al., 2015) are highly expressive in their capabilities for querying and exploration, with entity-centric auto-completion suggestions and other advanced features. However, none of these covers biomedical or health contents.

Our Approach and Contribution: This paper presents a novel system, called *DeepLife*, which provides this kind of user-friendly and expressive support for health-related contents from a wide variety of sources, including scholarly publications, newspaper articles and online communities. Our approach is inspired by the STICS system (Hoffart et al., 2014). However, our content is completely different, and coping with textual mentions of biomedical entities is much harder than recognizing and disambiguating prominent people or places in news articles. This paper presents the system architecture of DeepLife, demonstrates its usefulness for various use cases, and discusses how we overcame the aforementioned limitations of prior work and the challenges regarding coverage, scale and usability.

Salient features of DeepLife include the following novelties:

- integrating large knowledge bases like the Unified Medical Language System (UMLS) and KnowLife (Ernst et al., 2015) into a search engine over a variety of health-related sources and document feeds,
- providing capabilities for search and exploration based on flexible combinations of keywords (and phrases), biomedical entities, facts, and taxonomic categories,
- supporting users by powerful auto-completion suggestions, interactive query sessions, and basic forms of entity-aware text analytics.

DeepLife is available for interactive use at `https://gate.d5.mpi-inf.mpg.de/deeplife/en-health/`.

2 Related Work

In the biomedical domain, the majority of information retrieval systems limit their scopes to PubMed scientific publications. Kim et al. (2008) only uses molecular entities for query expansion. The scopes of Textpresso (Müller et al., 2004), GoPubMed (Doms and Schroeder, 2005), FACTA+ (Tsuruoka et al., 2011), EVEX (Van Landeghem et al., 2012), BioTextQuest+ (Papanikolaou et al., 2014) and CRAB (Guo et al., 2014) are restricted to genes, proteins, or chemicals. MEDIE (Miyao et al., 2006) and GeneView (Thomas et al., 2012) annotate PubMed articles with various kinds of biomedical entities and events, but both systems do not offer interactive real-time exploration and analytics. Contrary to the systems aforementioned, PolySearch2 (Liu et al., 2015) goes beyond scientific publications, but its search and exploration interface is not entity-aware.

Besides scientific publications, bio-surveillance systems aggregate and analyze news articles to identify health threats, such as disease outbreaks and food hazards. HealthMap (Freifeld et al., 2008) and EpiSpider (Keller et al., 2009) rely on user created ProMED reports and do not process documents automatically. Contrary to these user-based approaches, Global Health Monitor (Don et al., 2008) and the Medical Information System (MedISys) (Rortais et al., 2010) in combination with PULS (Steinberger et al., 2008) automatically extract entities and events from relevant medical news. However, the amount of entities both systems can distinguish is limited.

Pang et al. (2015) emphasize the need for better exploratory search capabilities for health content,

Genre	Sources	Documents	Entity Occurrences	Distinct Entities
Clinical Trials	2	16,476	49,170	8,962
Encyclopedic Articles	44	11,139	405,795	16,505
News	121	76,534	3,058,111	38,295
Scientific Publications	15	19,884,225	214,531,153	453,647
Social Media	1	9,473	117,421	4,433
Total	**182**	**19,997,847**	**218,161,650**	**454,620**

Table 1: Input corpus snapshot on June 1st, 2016

but they do not consider semantic assets, like entities or a knowledge base.

3 DeepLife's Knowledge Base

Knowledge bases (KBs) store facts about entities, their properties, and the relationships between entities. A fact is a triple consisting of two entities e_1, e_2, which serve as left- and right-hand arguments of a relation R, denoted by $R(e_1, e_2)$. We augment and integrate two large knowledge bases to generate DeepLife's KB covering the entire spectrum of biomedical entities, together with an extensive type system featuring salient facts.

UMLS: As entity dictionary, we rely on the Unified Medical Language System (UMLS). UMLS is the largest collection of biomedical entities and covers 3,221,702 entities with 12,842,558 entity names. It integrates source vocabularies from different biomedical domains into a coherent structure. Due to its broad coverage, we are able to detect all kinds of entities in text, i.e. entities about diseases, anatomy, genes, treatments, etc. However, the UMLS semantic type system is shallow, i.e. it only assigns 127 types to more than 3 million entities. Therefore, we generate our own type system by automatically augmenting UMLS with type hierarchies from its source vocabularies. For each vocabulary, we compute its entity coverage in our text corpus depending on the entities' semantic types. The hierarchy of the vocabulary with the highest coverage for a particular semantic type is then used, i.e. for genes the Gene Ontology (GO) is used, for anatomical entities the Foundational Model of Anatomy (FMA) and for drugs and diseases the Medical Subject Headings (MeSH).

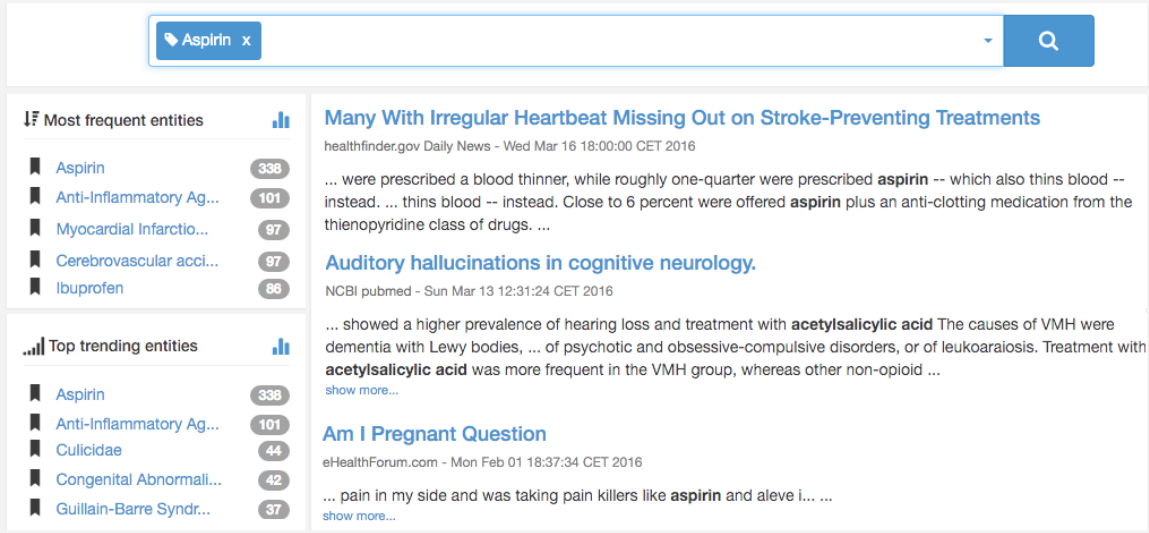
KnowLife: Although UMLS is rich on entities and types, it lacks cross-domain facts, i.e. relationships connecting different biomedical domains. To bridge this gap, we integrate KnowLife, a large knowledge base for health and life sciences, automatically constructed from Web sources (Ernst et al., 2015). KnowLife contains more than 500,000 of such cross-domain facts at a precision of 96% connecting different biomedical areas such as genes, diseases, anatomic parts, symptoms, treatments, as well as environmental and lifestyle risk factors for diseases. To integrate the facts into our system, we represent them as types. For all facts, $R(e_1, e_2)$, we create a new type by using the relation R and the right-hand argument e_2 as type name. For example, for all left-hand arguments e_1 appearing in facts such as $isRiskFactor(e_1, Asthma)$ we create the type $RiskFactorsforAsthma$.

Altogether, DeepLife's knowledge base covers 3.2 million entities with around 12.8 million entity names and synonyms, 64,568 custom types from source vocabularies and 136,437 fact types.

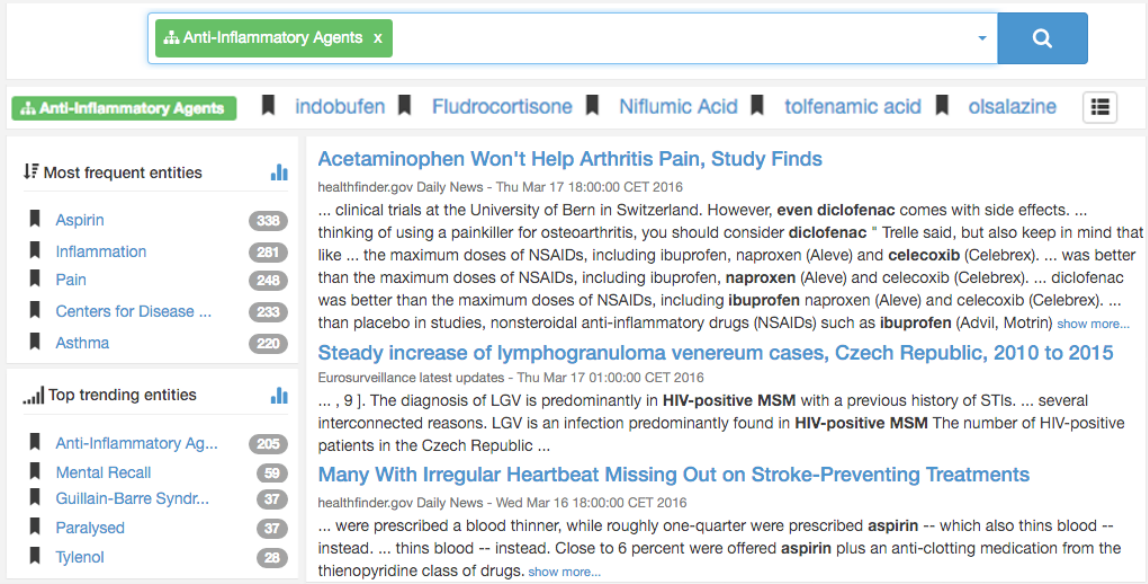
4 Entity Extraction

DeepLife has currently indexed 19,997,847 documents and extracted 218,161,650 entities from a continuous stream of 182 RSS feeds spanning five text genres. As Table 1 shows, this constantly growing and diverse corpus covers the full spectrum of biomedical information on the web. Clinical trials and scientific publications describe research findings and target professionals. DeepLife thereby includes the entire Pubmed MEDLINE collection. Encyclopaedic articles are educational resources providing insights to laymen. Social media, such as patient discussion forums, are mainly used to share experiences and to receive advice. By including news articles, our system is always up-to-date on the latest health topics, such as disease outbreaks or lifestyle information.

Entity Recognition: To process incoming articles in real-time and to stay up-to-date, our system applies an agile entity recognition method. StanfordCoreNLP is used to split sentences, tokenize words and determine part-of-speech tags. OpenNLP Chunker is used to generate an initial set of noun chunk candidates. We extend this set by applying a rule-based approach, e.g. splitting or merging prepositional phrases, conjunctions, as well as proper and common nouns. Candidates are then matched against the entity names in UMLS using string-similarity, giving preference to the longest possible matching chunk. To efficiently handle the large dictionary and volume of candidates, we use our own method which is based on



(a) Entity search for aspirin also includes its synonym acetylsalicylic acid



(b) Searching for anti-inflammatory agents expands to all agents in this category

Figure 1: Entity and Category Search

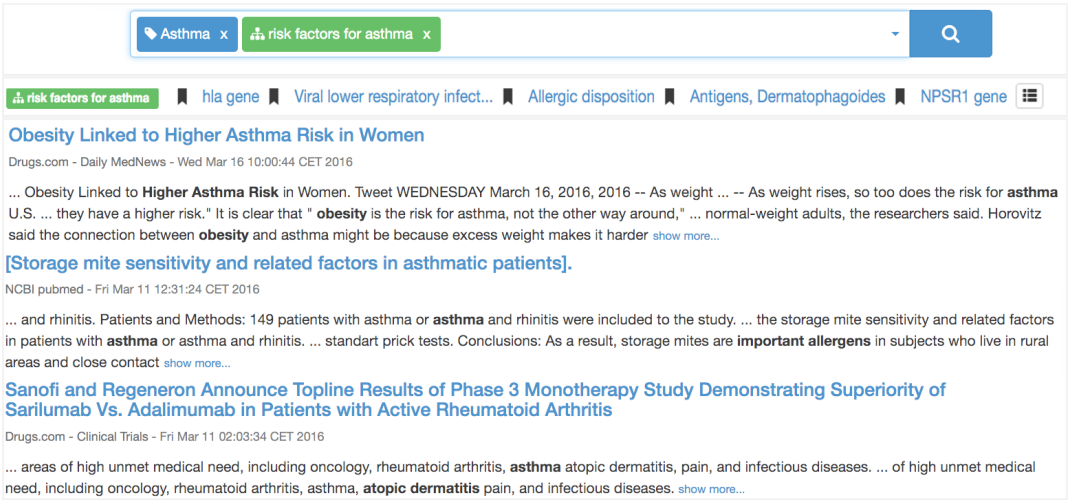


Figure 2: Combined Category and Entity Search for Asthma Risk Factors

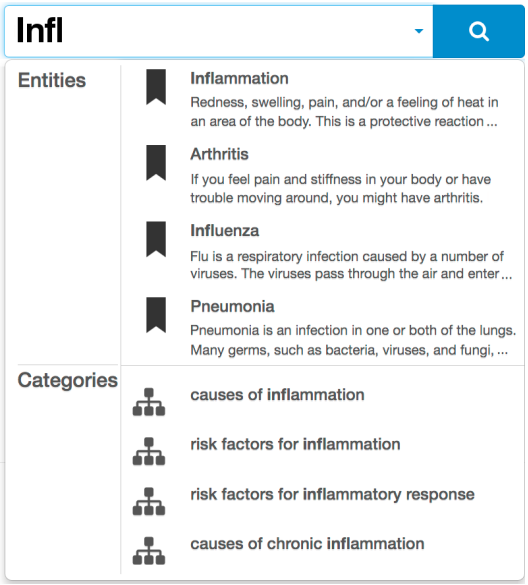


Figure 3: Entity/Category Auto-completion

locality sensitive hashing (LSH) with min-wise independent permutations (MinHash) to quickly find matching candidates (Siu et al., 2013). LSH probabilistically reduces the high-dimensional space of all character-level 3-grams, while MinHash quickly estimates the similarity between two sets of 3-grams. A successful match provides us also with the entity's semantic type.

Entity Disambiguation: The entity type information is used to disambiguate between multiple entity candidates matched to the same noun chunk in the input text. In the first filtering step, we reduce the number of entity candidates by only retaining the most specifically typed entities according to the UMLS semantic type system. Taking into account that UMLS provides a ranked list of entities for every possible name, we further disambiguate between the remaining candidates by determining the highest ranked entities. In case two entities share the same rank, we determine their popularities by the number of occurrences in different UMLS source vocabularies and prefer the more popular entity. As shown in Table 1, our system has currently extracted 218,161,650 mentions of 454,620 distinct entities.

5 Demo Scenarios

Entities are at the core of our system. Combining them with facts and types from DeepLife's knowledge base enables us to realize different use cases showcasing novel features of our system.

Entity-aware Auto-completion: Formulating queries with DeepLife is user-friendly and responsive. Providing an entity auto-completion which combines prefix matching with entity popularity, the system lets users easily explore and navigate through an extensive amount of entities and categories. For a user-provided prefix, the system retrieves entity and category candidates, where any token of their name or synonyms matches the prefix. These candidates are then ranked by corpus statistics which the system constantly updates. For example, Figure 3 depicts *Arthritis* as the second suggestion, because its synonym *Joint Inflammation* matches the prefix, and because of its high prevalence in the corpus.

Entity and Category Search: The entity-based search of our system excels over traditional keyword-based search. It increases recall, since the system automatically includes all synonyms of an entity, as well as precision, since the disambiguation removes unwanted occurrences. For example, as depicted in Figure 1a, if users search for *Aspirin*, documents mentioning its synonym *Acetylsalicylic Acid* are also retrieved. An important feature of our system is the possibility to search for categories of entities. This allows users to broaden their search request to all entities of the same type, i.e. entities which share common attributes or features. For example, to search for all "aspirin like drugs" which share therapeutic properties, one can search for the category *Anti-inflammatory Agents* (see Figure 1b). The system automatically determines all entities within the category via DeepLife's type system to retrieve relevant documents. Figure 1b also highlights DeepLife's diverse set of sources. The search results cover news, publications, as well as discussions. To tap into specific sources, users can easily customize queries with search filters.

Cross-domain Combined Search: DeepLife's knowledge base empowers our system to provide an intuitive method for searching facts by combining category and entity search. This is especially useful for layman users. Consider a user who is suffering from asthma and is interested in finding all risk factors triggering the disease. In this case, using the category *Risk Factors for Asthma* generated from facts together with the entity *Asthma* as search query, the system retrieves all documents mentioning *Asthma* with its risk factors (see Figure 2). Displaying the individual risk factors (e.g. *HLA Gene, Viral Lower Respiratory Infection, etc.*) as an expansion of the category provides immediate insights and facilitates further

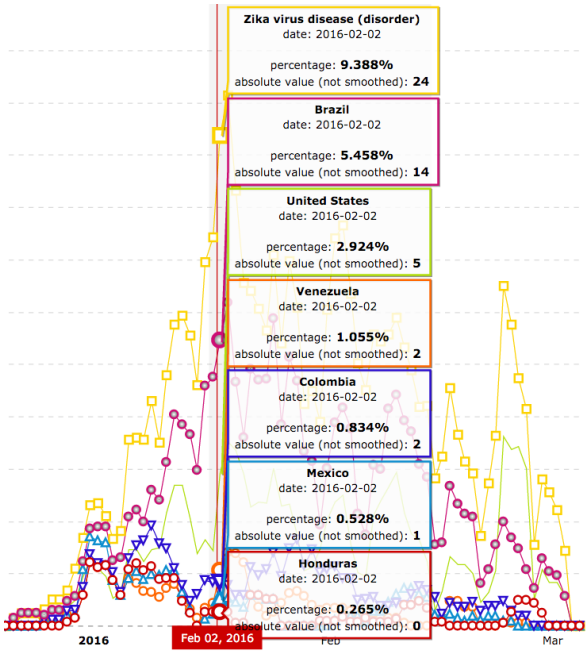


Figure 4: Countries co-occurring with Zika

exploration.

Analytics: Our system offers interactive entity-based analytics to spot trends and topic shifts. Such analyses benefit from the improved recall and precision aspects aforementioned. Statistics, based on entity occurrences in documents over time, are computed and visualized. For example in Figure 4, entity occurrences of *Zika* and related countries in our corpus (Y-Axis) are visualized over time (X-Axis). Users can zoom into specific time frames and explore documents the statistics are based on. Not only can entities be tracked individually, the analytics can also be constrained on one entity of main interest, i.e. only those documents in which this entity appears. In the same example in Figure 4, to gather insights about countries affected by the virus, the user set *Zika virus disease* as the main entity to compute analytics based on documents where Zika and a particular country were mentioned.

References

Hannah Bast and Björn Buchhold. 2013. An index for efficient semantic full-text search. In *Proc. of CIKM*. pages 369–378.

Andreas Doms and Michael Schroeder. 2005. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Res* 33:W783–6.

Son Don, Ai Kawazoe, and Nigel Collier. 2008. Global health monitor - a web-based system for detecting and mapping infectious diseases. In *Proc. of IJCNLP*. pages 951–956.

Patrick Ernst, Amy Siu, and Gerhard Weikum. 2015. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* 16(1):1–13.

Clark Freifeld, Kenneth Mandl, Ben Reis, and John Brownstein. 2008. Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *JAMIA* 15(2):150–157.

Yufan Guo, Diarmuid Ó Séaghdha, Ilona Silins, Lin Sun, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2014. Crab 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment. In *Proc. of COLING*. pages 76–80.

Johannes Hoffart, Dragan Milchevski, and Gerhard Weikum. 2014. Stics: Searching with strings, things, and cats. In *Proc. of SIGIR*. pages 1247–1248.

Mikaela Keller, Michael Blench, Herman Tolentino, Clark C. Freifeld, Kenneth D. Mandl, and Abla Mawudeku et al. 2009. Use of unstructured event-based reports for global infectious disease surveillance. *Emerging Infectious Disease Journal* 15(5):689.

Jung-jae Kim, Piotr Pezik, and Dietrich Rebholz-Schuhmann. 2008. Medevi: Retrieving textual evidence of relations between biomedical concepts from medline. *Bioinformatics* 24(11):1410–1412.

Yifeng Liu, Yongjie Liang, and David Wishart. 2015. Polysearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Research* 43(W1):W535–W542.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proc. of ACL*. pages 1017–1024.

Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2(11).

Patrick CI Pang, Karin Verspoor, Jon Pearce, and Shanton Chang. 2015. Better health explorer: Designing for health information seekers. In *Proc. of OZCHI*. pages 588–597.

Nikolas Papanikolaou, Georgios A. Pavlopoulos, Evangelos Pafilis, Theodosios Theodosiou, Reinhard Schneider, and Venkata P. et al. Satagopam. 2014. Biotextquest+: a knowledge integration platform for literature mining and concept discovery. *Bioinformatics* 30(22):3249–3256.

Agnès Rortais, Jenya Belyaeva, Monica Gemo, Erik van der Goot, and Jens Linge. 2010. Medisys: An early-warning system for the detection of (re-)emerging food- and feedborne hazards. *Food Research Internat.* 43(5):1553–1556.

Amy Siu, Dat Ba Nguyen, and Gerhard Weikum. 2013. Fast entity recognition in biomedical text. In *Proc. of Workshop on Data Mining for Healthcare at KDD*.

Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. 2008. *Text Mining from the Web for Medical Intelligence*, IOS Press, volume 19.

Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. 2012. Geneview: a comprehensive semantic search engine for pubmed. *Nucleic Acids Research* 40(W1):W585–W591.

Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* 27(13):i111–i119.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Proc. of the Workshop on Scholarly Big Data at AAAI*.

Sofie Van Landeghem, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2012. Exploring biomolecular literature with evex: Connecting genes through events, homology, and indirect associations. *Advances in Bioinformatics* 2012:12.