Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
2018

OXFORD

Semantic integration of disease-specific knowledge

Anastasios Nentidis [1,*], Konstantinos Bougiatiotis [1], Anastasia Krithara [1] and Georgios Paliouras [1]

[1] Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", Athens, Greece.

[*] To whom correspondence should be addressed.

Abstract

Motivation: Biomedical researchers working on a specific disease need up-to-date and unified access to knowledge relevant to the disease of their interest. Knowledge is continuously accumulated in scientific literature and other resources such as biomedical ontologies. Identifying the specific information needed is a challenging task and computational tools can be valuable. In this study, we propose a pipeline to automatically retrieve and integrate relevant knowledge based on a semantic graph representation, the *iASiS Open Data Graph*.
Results: The disease-specific semantic graph can provide easy access to resources relevant to specific concepts and individual aspects of these concepts, in the form of concept relations and attributes. The proposed approach is applied to three different case studies: Two prevalent diseases, Lung Cancer and Dementia, for which a lot of knowledge is available, and one rare disease, Duchenne Muscular Dystrophy, for which knowledge is less abundant and difficult to locate. Results from exemplary queries are presented, investigating the potential of this approach in integrating and accessing knowledge as an automatically generated semantic graph.
Availability: The source code for the platform developed in Java and Python is available in GitHub.*
Contact: tasosnent@iit.demokritos.gr

1 Introduction

A lot of biomedical knowledge is published every day in the literature and structured forms like biomedical ontologies. It is a challenge for biomedical experts to identify and process all available knowledge. For example, 1.8 million citations were added to PubMed during 2017[1], which corresponds to more than three citations per minute. Not all published literature is relevant to the work of every researcher and identifying the relevant articles can be challenging. Efficient access to relevant knowledge is crucial and simple term-based search can retrieve irrelevant documents, e.g. due to homonyms, or miss relevant documents because some synonyms, abbreviations or mismatch of terms between the query and the relevant documents.

Much effort has been made to address this issue, including semantic search approaches that use predefined concepts which can have several associated synonyms and relations with other concepts, expanding the query terms. PubMed is the main knowledge resource considered for biomedical literature and a variety of search tools have been proposed to search in PubMed as described in Lu, 2011. PubMed supports semantic search based on the Medical Subject Headings (MeSH) hierarchy[2]. A team of curators in the U.S. National Library of Medicine (NLM) continuously annotates articles added in PubMed with the appropriate MeSH terms that represent the topics of the article. Information systems can exploit these topic annotations for information retrieval and extraction of knowledge in the form of relations between MeSH terms. An overview of systems following this approach is available in Zhang *et al.*, 2014.

Gathering a set of articles relevant to a topic of interest is often not sufficient. An article may contain different pieces of knowledge, which are more or less relevant to the interests of a researcher. Additionally, the value of these knowledge items can change, if they are combined with information from other articles or resources. A knowledge base supports this process of organizing and storing domain knowledge, in order to be easily accessible. A biomedical knowledge base, like DrugBank (Wishart *et al.*, 2008), often consists of knowledge which has been manually reviewed by domain experts. Such manually curated knowledge bases require human effort for their creation and maintenance, which is

[*] https://github.com/tasosnent/Biomedical-Knowledge-Integration

[1] Statistical Reports on MEDLINE/PubMed Baseline Data, available at https://www.nlm.nih.gov/bsd/licensee/baselinestats.html

[2] https://www.nlm.nih.gov/mesh/

usually supported by computational tools. Hence, automated extraction and integration of knowledge plays a key role in efficiently managing the continuously growing biomedical knowledge.

Biomedical knowledge is often extracted from literature in the form of relations between biomedical entities. For example protein-protein interactions or gene-disease associations. A variety of text-mining approaches has been proposed for this task, including term co-occurrence in the literature and approaches based on syntactic analysis of articles. Rebholz-Schuhmann *et al.*, 2012 presents a review of text-mining techniques that have been applied to the biomedical domain so far.

One of the most systematic approaches to integrating biomedical knowledge automatically extracted from literature is provided by the Semantic MEDLINE web application, in the context of the Semantic Knowledge Representation (SKR) project[3] of the NLM. Utilizing this tool, biomedical researchers have uniform access to knowledge from MEDLINE abstracts matching their queries, in the form of a graph. The extracted knowledge is retrieved from the latest version of the SemMedDB database and is visualized as a network of concepts linked by a range of relations. The user can browse this network to discover the information needed as described by Rindflesch *et al.*, 2017. Another recent effort to biomedical knowledge integration is the Hetionet[4] which exploits the power of graph databases to integrate knowledge from a variety of resources as a heterogeneous network. This work focuses on structured manually-curated resources. Hetionet incorporates literature mining, focusing on a limited number of relations extracted through co-occurrence analysis from the MEDLINE corpus. The knowledge in Hetionet can be accessed through graph queries which can effectively target combinations of multiple relations between entities, in order to support complex tasks such as computational predictions for drug repurposing (Himmelstein *et al.*, 2017).

2 Approach

In this work, we develop a platform hosting the *iASiS Open Data Graph*, that combines state-of-the-art approaches and tools to automatically integrate knowledge from both literature and structured resources into a disease-specific knowledge base which can be incrementally updated. The architecture of the platform is shown in Figure 1. In particular, all literature currently available for the disease of interest is retrieved online, considering both the PubMed abstracts and full-text documents from PubMed Central[5] (PMC) when available. Literature analysis tools are used to automatically extract knowledge from the text, which is then integrated in a knowledge base as a graph of inter-related concepts. In addition, knowledge from biomedical ontologies and databases is integrated into the same knowledge graph. Therefore, uniform access is provided to up-to-date knowledge, automatically extracted from literature, and high-quality manually reviewed knowledge from structured resources. Access to the above integrated knowledge is provided through graph queries, which can be used either for the development of advanced user interfaces or for computational graph analysis of the graph. The pipeline developed in this study is tested on the automated creation of knowledge bases for three distinct diseases. Example graph queries are used to illustrate the value of the knowledge bases.

[3] https://skr3.nlm.nih.gov/

[4] http://neo4j.het.io/browser/

[5] https://www.ncbi.nlm.nih.gov/pmc/

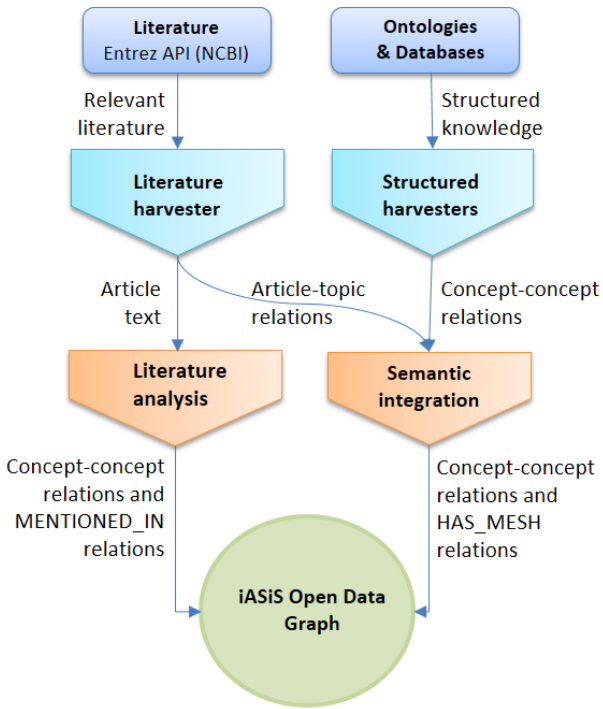

Fig. 1. The architecture of the the iASiS Open Data Graph platform for the automated integration of biomedical knowledge from different resources into a common semantic graph.

3 Methods

3.1 Platform architecture

The proposed platform for the semantic retrieval and integration of disease-specific knowledge has been designed and developed as a pipeline of distinct modules that perform well-defined tasks and can be reused independently. For the online retrieval of literature, a harvesting module has been developed that interacts with the REST API of the Entrez Programming Utilities in NCBI[6]. Then, an analysis module has been developed that employs state-of-the-art tools to extract knowledge from biomedical natural language text and extract structured knowledge for the integrated graph. Concerning the structured resources, harvesting modules have been developed to transform the data in their latest version into a format appropriate for integration. Additionally, an integration module has been developed to map all entities of structured resources in the coding system of the Unified Medical Language System (UMLS) Metathesaurus. All knowledge is integrated in a graph database, under a simple but powerful representation, where graph queries can be employed to serve the information needs of biomedical researchers.

3.2 Data harvesting

The basic source of available knowledge is the biomedical scientific literature. Yet different resources for biomedical literature exist, as reviewed by Masic, 2012, PubMed[7] is the most established resource offering more than 28 million citations. Most of these citations are accompanied by an abstract which summarizes the content of the article. As already mentioned in Section 1, PubMed also supports a powerful

[6] https://www.ncbi.nlm.nih.gov/books/NBK25497/

[7] https://www.ncbi.nlm.nih.gov/pubmed/

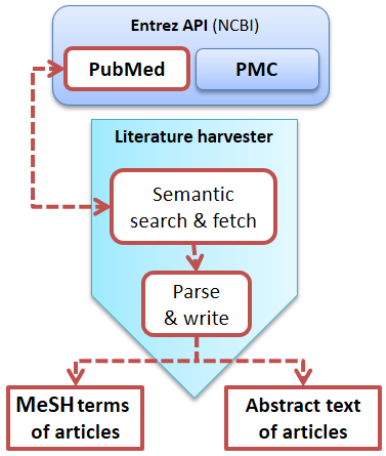

Fig. 2. The literature harvester software module retrieves relevant literature from PubMed and extracts only the abstract and the topic areas for each article.

semantic retrieval functionality based on MeSH indexing. In addition, it is uniformly accessible, through the Entrez REST API, with PMC which offers the full-text of about 4.7 million articles. In this work we exploit these facilities to retrieve online all relevant article abstracts and their full-text when available. In particular, a software module was developed to interact with the Entrez REST API and harvest the relevant literature.

As graphically shown in Figure 2, the first step of the harvester performs the appropriate semantic queries to PubMed, based on MeSH terms in order to identify the relevant articles. The next step is to fetch the relevant article citations in MEDLINE/Pubmed XML format. This format includes a variety of information for each article using more than 100 different types of hierarchically related elements. The harvester extracts only the abstract text and the topic areas for each article and provides them in a simple JSON format. The basic functionality of the PMC harvester is similar. The main difference is the extraction of the full-text from the PMC XML format for each article. In addition, no MeSH topics are retrieved from PMC since they are offered from PubMed. The inclusion of full-text in this study was a key decision. The target was to perform an integration of knowledge as deep as possible. Querying the integrated graph with a concept should return, apart from articles where this concept is important and possibly mentioned in the abstract, also articles that may contain specific details related to the concept, available only in the full text.

Biomedical ontologies are another important source of domain knowledge in the field of biomedicine. A harvester was developed to support integration of biomedical ontologies available in the Open Biomedical Ontologies (OBO) format (Smith *et al.*, 2007). Currently, more than 140 ontologies are available in the OBO Foundry[8]. Three basic ontologies have been selected for the creation of the case study datasets:

- The Disease Ontology (DO) (Schriml *et al.*, 2012) semantically integrates disease terms and identifiers from different resources including NCI's thesaurus, SNOMED Clinical Terms and OMIM. It includes more than 10,000 concepts with more than 6,000 mappings to the UMLS. In this work only concepts mapped to the UMLS are integrated graph.
- The Gene Ontology (GO) (Ashburner *et al.*, 2000) provides more than 24,000 concepts to represent three basic categories of genomic knowledge, namely, biological processes, molecular functions and cellular components. GO is integrated into the UMLS, providing mappings for all the GO concepts.
- MeSH is a hierarchical controlled vocabulary developed by NLM to semantically index biomedical articles primarily based on their topics. It provides more than 28,000 descriptors organized in sixteen tree structures that cover a broad spectrum of knowledge, from chemicals and organisms to humanities and geographical locations. All MesH descriptors are linked to specific concepts which are integrated in the UMLS.

The OBO harvester currently extracts only hierarchical (is-a) relations from OBO ontologies. This restricts the exploitation of the available knowledge in the ontologies, but since hierarchical relations are the most important, we decided to focus these first. Especially for MeSH which was not available in the OBO format, a preprocessing script was developed to parse the MeSH XML format and produce a simpler version of MeSH in the OBO format, that could be further integrated in the graph using the OBO parser.

Apart from biomedical ontologies, other structured databases can also be integrated, as long as the knowledge they provide can be expressed as relations between concepts that can be mapped to UMLS. For example, in this framework another harvester was developed to extract relations expressing drug-to-drug interactions from DrugBank that is available in XML format. DrugBank is a comprehensive, manually maintained resource of information for drugs, containing more than 10,000 drug entries with more than 200 fields of information per drug. Drug interactions is the most abundant type of information in DrugBank, exceeding 300.000 which makes it a rich resource of knowledge (Wishart *et al.*, 2017).

All harvesters for structured data resources described in this section, including topic relations for articles, produce datasets of relations in the same simple JSON format. As a result, all harvested data are uniformly integrated in the semantic graph which is easily extensible with additional knowledge. Coverage can be extended both towards more types of relations for the supported resources and towards more resources, aiming at a knowledge graph as comprehensive as possible.

3.3 Literature analysis

Discovering biological relations that include drugs, proteins, diseases and other entities, through the extraction of knowledge from the scientific literature is a challenging task. Despite multiple knowledge acquisition efforts to catalog biological events in databases, a considerable amount of unstructured knowledge is still buried in the scientific literature. Text mining offers the potential to tap into the knowledge hidden in the ever-increasing body of biomedical articles. Such automated extraction would provide scientists with insights into underlying interactions and hidden patterns with regard to co-occurrences of diseases, drug re-purposing and a plethora of other tasks.

Our goal is not to create a new text mining tool, but rather to create a framework where any such tool can be incorporated in the pipeline of extracting biomedical relations from text and consequently enriching a unified knowledge graph. Working towards this goal, the first non-trivial obstacle is the variation of biomedical terms due to a variety of reasons, such as orthographic (e.g. Hodgkins disease - Hodgkins Disease) or synonym (e.g. headache - head-pain) variation. Fortunately, due to the existence of many terminological resources, like curated ontologies (Ashburner *et al.*, 2000) and lexicons (Liu *et al.*, 2005), we can overcome this problem. In our work, we use the UMLS (Bodenreider, 2004) as the reference basis for all entities and relations.

Specifically, we utilize two of the main components of UMLS, the Metathesaurus (Schuyler *et al.*, 1993) and the Semantic Network (McCray, 2003). The Metathesaurus is essentially a vocabulary database containing

[8] http://www.obofoundry.org/

more than 100 vocabularies with more than one million concepts. The key aspect of this thesaurus is that differing names for a biomedical meaning are linked and aggregated under a single *concept*. Therefore, the Metathesaurus deals with term variation and at the same time creates links between different vocabularies, accumulating knowledge for a concept from different domains (e.g. chemical, physical, gene associations). On the other hand, the Semantic Network enriches the concepts of the Metathesaurus with meaning, equipping them with semantic types and relations between those types. There are 133 semantic types expressing a high-order grouping of the concepts into categories, such as diseases, body parts, genes or genomes etc., which are also hierarchically structured. Each concept is mapped to at least one semantic type and always at the deepest possible level of the hierarchy (e.g. "trout" is a fish, which is a more specific type of animal). Together with 55 semantic relations between these types, a rich network spanning the whole biomedical domain is created.

In order to harness the power of this representation, we use SemRep (Rindflesch and Fiszman, 2003), a UMLS-based tool that extracts biomedical predications, i.e. semantic triples in the form of subject-predicate-object, from unstructured text. The subject and object arguments in these predications are concepts from the UMLS and the predicate is one of the semantic relations of the semantic network, connecting the semantic types of the subject and object in the context of the specific sentence. To achieve this, it also relies on MetaMap (Aronson, 2001) which is a tool that uses symbolic natural-language processing (NLP) and computational-linguistic techniques to map biomedical text to Metathesaurus concepts.

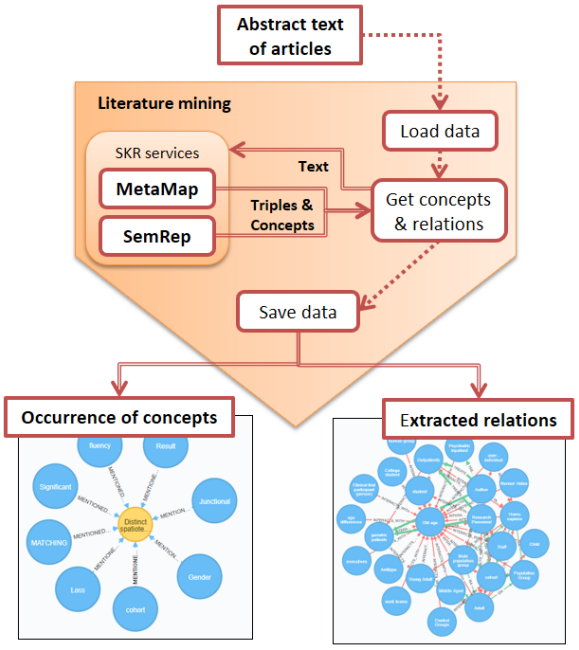

Fig. 3. The literature analysis module.

For our purposes, we incorporated the use of these tools in our pipeline for analyzing the harvested literature as shown in Figure 3. Specifically, after some preprocessing to remove artifacts not needed in the text (e.g. tables, LaTeX code etc.) the body of the abstracts, or the full-text of biomedical articles when available, is processed by SemRep with the help of MetaMap, to produce concepts and predications found in the text. We then transform the recognized entities and triples in a format suitable to be inserted in the knowledge graph. Alongside the extracted predications stemming from the Semantic Network, we also create a new type of relation, which we call "*MENTIONED IN*". This relation expresses the occurrence of a concept in an article. The motivation behind this approach is that biomedical literature contains relationships between medical concepts that have been "distilled" through research and one way to leverage this knowledge is to study the co-occurrence of concepts in the documents. Co-occurence provides a way of measuring the association between two terms and potentially helps in targeting interesting and clinically important associations (for example between medications and disorders). Such associations may have not been examined extensively before and may prove interesting signals, e.g. for adverse drug reactions or lead to novel off-label uses of existing drugs (Liu *et al.*, 2012).

3.4 Knowledge Graph

The main motivation behind our work was to create a semantic knowledge graph where concepts are related to other concepts with different types of edge. The *iASiS Open Data Graph* integrates both structured and unstructured knowledge in the same semantic graph. To accomplish that, we use a single node per concept, based on the concepts of the UMLS Metathesaurus. Using a single node per concept leads to a highly integrated graph, where all available knowledge is linked to the node, in the form of interactions with other concepts, regardless of the source of this knowledge. Articles are also integrated as nodes in the graph connected with concepts extracted from their text through a "*MENTIONED IN*" edge. In order to incorporate the knowledge from biomedical ontologies and other structured data sources in the graph, we took advantage of the available UMLS REST API[9], which helped to map terms from different vocabularies (e.g. DrugBank, MeSH, GO) to UMLS concepts. This allowed the integration of concepts and relations to the unified schema that was already in place for biomedical literature.

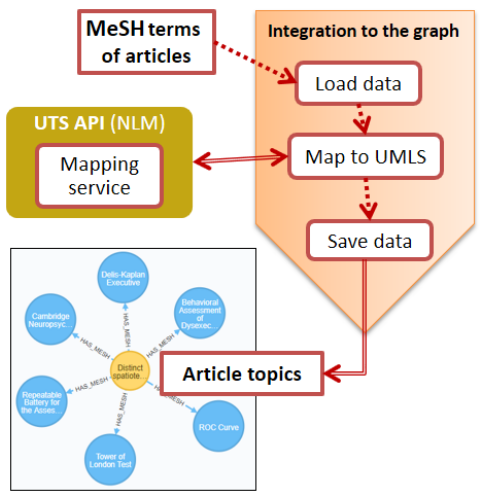

Fig. 4. Integration of "HAS MESH" edges in the knowledge graph.

Alongside the entities and relations extracted from text and structured data sources, we added another type of relation in the knowledge graph, which we call "*HAS MESH*". Each article in PubMed is associated with a set of MeSH tags. All MeSH tags are first mapped to UMLS concepts using the UMLS REST API and a new kind of triple, in the form of "article-HAS MESH-concept" is created for each MeSH tag associated with the article. These triples express the topics to which each article is

[9] https://documentation.uts.nlm.nih.gov/rest/home.html

connected. This is very important, as it captures knowledge that humans have deducted, making the links rather robust and semantically rich. The procedure followed is shown in Figure 4.

Regarding the technical implementation of the knowledge graph we used the capabilities of Neo4j[10]. Traditional databases based on SQL are suitable for storing structured data, but are somewhat rigid in expressing relationships between the data of different tables[11]. On the other hand, graph databases explicitly focus on the connectivity between the nodes. Taking into account their superior performance when traversing many levels of connectivity (Jaiswal and Agrawal, 2013), they become a natural choice for storing biomedical concepts and relations. This is the reason why Neo4j is increasingly adopted for bioinformatics projects that model biological connectivity(Yoon *et al.*, 2017, Mungall *et al.*, 2016, Lysenko *et al.*, 2016).

4 Case studies and discussion

The methodology described above was applied to three distinct cases of disease, namely Lung Cancer (LC), Dementia including Alzheimer Disease (ADD) and Duchenne Muscular Dystrophy (DMD). A different knowledge graph was developed for each case and appropriate graph queries were used to investigate different properties of the proposed approaches to effectively identify information of interest. In the following sections, details are presented regarding the development of the three datasets, as well as the results of queries that aim to investigate the potential use of the semantic graph.

4.1 Dataset creation

The platform described above was employed three times to create three corresponding semantic graphs configured with appropriate MeSH terms for each case. In particular, the MeSH term used for LC was "Lung Neoplasms" (D008175), for ADD was "Dementia" (D003704) and for DMD was "Muscular Dystrophy, Duchenne" (D020388). Based on these semantic topics, the pipeline shown in Figure 1 retrieved all available relevant literature[12] and created three graphs. As expected, the volume of available knowledge was different for the three cases resulting in different size. For ADD and LC which are highly prevalent diseases, more than 100,000 articles were available complicating the identification of specific information needed for detailed scientific questions. Even for DMD, which is a rare disease, the number of directly relevant articles exceeded 4,000, which is still difficult to manage. Details about the composition of these datasets are presented in Table 1. For the majority of the articles only the abstract was available. DMD had the higher percentage of full text availability exceeding 24%. In terms of quantity, knowledge about concept occurrence in the articles ("*MENTIONED IN*") is dominant in the knowledge graphs. Relations indicating the relevance of articles to specific concepts ("*HAS MESH*") are also quite abundant and domain knowledge, in the form of extracted relations between concepts is the least frequent. This could be attributed to limitations in relation extraction since this is the most complex and less mature part of the processing.

As described in Section 3, knowledge from structured resources is also integrated in the graph. These structured resources are not disease-specific and details about the corresponding datasets are presented separately in Table 2. In this setup, only hierarchical relations were harvested from DO, GO and the MeSH hierarchy and only interactions between drugs from the DrugBank. These types of relation are among the most important ones and constitute a proof of concept for the integration of any kind of relation from structured resources. In practice, different setups would be more suitable, depending on the domain areas and the specific interests of the users. These setups can be implemented combining different structured resources or parts of them per case. Apart from selecting resources relevant for each use case other important issues, like the potential overlap or conflict in the content of different resources, should also be carefully considered.

[10] https://neo4j.com/

[11] https://neo4j.com/blog/demining-the-join-bomb-with-graph-queries/

[12] until October 11 2017

Table 1. Number of articles with abstract only and articles with full-text are reported for each case study, along with the corresponding number of distinct concepts and relations between concepts extracted from literature. The number of "HAS MESH" and "MENTIONED IN" relations is also reported.

Case study	article abstracts	article full-texts	concepts extracted	relations extracted	relations has MeSH	relations mentioned in
DMD	4,403	1,075	21,982	27,954	113,136	335,071
ADD	108,458	6,000	75,985	392,421	3,651,698	7,587,772
LC	141,712	10,000	92,846	608,759	4,228,785	9,940,847

Table 2. Types of knowledge and integration details about structured resource datasets. Only selected relation types are considered for each dataset. For DO only concepts with mappings to the UMLS are integrated.

Resource	relation type	concepts integrated	relations integrated
DO	is_a	5,307	5,129
GO	is_a	64,751	125,629
MeSH	is_a	55,400	123,287
DrugBank	drug interactions	581,055	1,628,077

Table 3. Top 10 semantic types for each disease ranked by the number of distinct concepts extracted from corresponding literature.

Semantic type	DMD rank	ADD rank	LC rank
Gene or Genome	2	1	1
Amino Acid, Peptide, or Protein	1	3	3
Organic Chemical	4	2	2
Pharmacologic Substance	5	4	4
Finding	3	5	5
Disease or Syndrome	7	6	6
Biologically Active Substance	6	7	7
Body Part, Organ, or Organ Component	11	8	10
Therapeutic or Preventive Procedure	12	10	8
Intellectual Product	10	9	14
Quantitative Concept	8	13	12
Qualitative Concept	9	15	18
Neoplastic Process	59	36	9

4.2 Query examples

The amount of accumulated knowledge and the variability observed among different diseases highlight the importance of two distinct but complementary needs regarding knowledge access. It is crucial to have precise access to highly detailed information and at the same time it is important to have a broad overview of all knowledge available to select specific areas of focus. Given the integrated disease-specific semantic graph, both these needs can be addressed through corresponding queries. For example, ranking the semantic types of distinct concepts extracted from literature for each case study can indicate directions for further study. In Table 3 the top 10 semantic types are reported with their corresponding rank in each knowledge graph. Some semantic types, like *Gene or Genome*

Table 4. Top 10 semantic types with highest differentiation among the three diseases, based on the standard deviation (STDV) of their ranks in the corresponding graphs.

Semantic type	LC rank	ADD rank	DMD rank	STDV
Plant	23	22	79	32.62
Bacterium	53	72	104	25.78
Neoplastic Process	9	36	59	25.03
Fungus	72	94	120	24.03
Mental or Behavioral Dysfunction	65	28	53	18.88
Eukaryote	39	43	70	16.86
Activity	81	76	51	16.07
Mental Process	62	39	34	14.93
Hazardous or Poisonous Substance	32	45	61	14.53
Organism Attribute	77	71	50	14.18

and *Amino Acid, Peptide, or Protein* are ranked highly for all three diseases. However, some interesting differences can be observed. For example, the semantic type *Neoplastic Process* is ranked higher for LC than for the other diseases, as expected. Less clear differences may also have some explanation. For example concepts for *Amino Acid, Peptide, or Protein* are ranked first only in DMD. The fact that DMD is caused due to the lack of the protein dystrophin could be a potential explanation. Potential explanations could also be investigated for other less clear differences like concepts for *Organic Chemical* being ranked higher in ADD and LC.

In order to emphasize the differences among disease, we re-ranked all semantic types according to the standard deviation of the three ranks of each semantic type for the three use cases. Table 4 presents the ten semantic types that differ the most in the three knowledge graphs. The most extreme case for the three diseases is the semantic type *Plant* which seems to be more frequent in LC and ADD than for DMD. This observation lead to further investigations regarding the role of plants in research for LC and ADD. Other similar observations can also provide directions for further study, constructing a profile for each case study. Such observations are the higher ranking of *Mental Process*, *activity* and *organism attributes* for DMD.

Table 5. Five most frequently occurring concepts of semantic type Plant in the literature for LC and ADD.

Rank	LC	ADD
1	Plants	Bark - plant part
2	Nicotiana	Parkinsonia
3	Gossypium	Bikinia le-testui
4	Bikinia le-testui	Plants
5	Rosa	Ginkgo biloba

Next, we focus on the first observation about semantic type *Plant*, in order to illustrate the power of the semantic graph. In particular, we query the three knowledge graphs for the five most frequently occurring concepts of this type. The results presented in Table 5 show that *Plants* is the most frequent concept with semantic type *Plant* in the LC knowledge graph. A quick overview of the knowledge available about each concept can be retrieved with corresponding graph queries. In the LC knowledge graph there are 795 articles, where the concept *Plants* occurs 1463 times and 13 articles having a topic relevant to *Plants*. In addition, there are 195 distinct relations of 5 different types between *Plants* and 194 distinct concepts in the LC literature. Table 6 summarizes all relations involving the concept *Plants* extracted from the LC literature. The majority of extracted relations involving *Plants* provide general taxonomic knowledge and structural information about *Plants*. The top ten concepts more frequently related with *Plants* in LC are also available in Table 7. Links to specific sentences in the articles where these relations have been extracted from are also available in the graph as relation properties.

Table 6. Types of relation involving the concept Plants in the LC literature. The number of relations among distinct concepts and the corresponding extracted instances are also reported.

Relation type	distinct relations	relation occurrences
LOCATION_OF	120	202
ISA	59	84
PROCESS_OF	9	10
PART_OF	6	7
INTERACTS_WITH	1	1

Table 7. Top 10 concepts more frequently related to Plants in LC literature, the frequency of the relations and the number of corresponding articles.

Concept label	relation occurrences	distinct articles
Curcuma longa	13	13
Chrysotile	11	8
polyphenols	10	9
3-hydroxyflavone	8	8
Asbestos	7	7
Antineoplastic Agents	7	7
Oils, Volatile	7	7
Chlorophyll	4	4
Magnolia	4	4

Examination of the source articles confirms that most of the related concepts are plant species (e.g. *Curcuma longa*, *Magnolia*) or chemicals found in plants (e.g. *polyphenols*, *3-hydroxyflavone*, *Chlorophyll*) that have been studied for their potential effect on lung cancer. The fact that articles are organized per related concept occurrence allows a selective examination of them. For example, looking at the article (Saha *et al.*, 2010) for *Curcuma longa* confirms that this is indeed a plant studied for LC effects. Following the same approach and using similar queries for other concepts in the tables, we observe that three concepts of semantic type *Plant* are of interest in LC research (*Plants*, *Nicotiana* and *Gossypium*), and two in the context of ADD (*Plants* and *Ginkgo biloba*). On the other hand, as indicated by the initial observation in Table 4 the role of plants in DMD studies is limited.

In an alternative scenario, a researcher may be interested in the effect of drug combinations in long surviving LC patients. Contrary to the previous example, this information need is highly targeted and involves distinct inter-related concepts. A central concept in this question is long-term survival of patients which can be expressed by the concept *long Term Survivorship* in the UMLS. This concept is mentioned in 2303 articles in LC dataset, but none of them is annotated with it as a topic. We further segment this set of articles, based on drug concepts co-occurring with *long Term Survivorship*. To restrict the search in drug concepts only, we can add a clause in the graph query requiring that the co-occurring concept should be a child of the *Pharmaceutical Preparations* concept (i.e. related with the “is a” relation). This query returns more than 300 concepts with supporting articles.

In order to identify the most relevant concepts out of them, we can search for concepts that are related to *long Term Survivorship* instead of co-occurring. This query retrieves 14 distinct concepts related with *Long Term Survivorship* with five distinct types of relation. Each occurrence provides a reference to the corresponding sentence of the article which can facilitate consulting the initial resource text. In addition, each of these concepts can also be enriched with other supplementary information from the knowledge graph. For example, the number of other entities directly interacting with each concept can be retrieved using an appropriate graph query. The 14 related concepts with the accompanying information are presented in Table 8.

Table 8. Drug concepts related to Long Term Survivorship in LC literature with the frequency of the relation and the number of other interacting drug concepts.

Concept label	frequency	interacting concepts
Antineoplastic Agents	3	210
Cisplatin	3	991
Aim	2	243
Melphalan	1	58
everolimus	1	640
cetuximab	1	71
complement C3a, des-Arg-(77)-	1	6
Interferons	1	12
animal allergen extracts	1	87
gefitinib	1	985
Altretamine	1	110
Topotecan	1	561
Paclitaxel	1	1518
Carboplatin	1	147

In this example relations automatically extracted from the literature are not distinguished from relations coming from structured resources. In particular, "is a" relations from the integrated ontologies were used to retrieve drug concepts while drug interactions from DrugBank have been taken into account in interacting concepts column of Table 8. Since structured resources are in general more reliable than automatic extraction, we could also restrict to it, to increase precision to the detriment of recall. For example, restricting the queries for drug concepts to use ontological "is a" only, results in no drug concepts found related to *Long Term Survivorship* and only the seven drug concepts presented in Table 9 co-occurring with it in 117 distinct articles.

Table 9. Drug concepts* co-occurring with Long Term Survivorship in LC literature.

Concept label	distinct articles of co-occurrence
Solutions	91
Drug Combinations	18
Prodrugs	5
Investigational New Drugs	2
Xenobiotics	2
Drugs, Non-Prescription	1
Controlled substance	1

*Retrieved based on ontological "is a" relations only.

5 Conclusion

Computational tools can support biomedical experts in identifying and organizing knowledge relevant to their work, based on an ever-increasing biomedical literature and a proliferation of disparate resources. In this paper we propose a framework for the retrieval and semantic integration of disease-specific knowledge we focusing on automated and incremental update of the knowledge. Knowledge coming from relevant publications is integrated with biomedical ontologies and databases into a common semantic graph representation. This graph provides both uniform way for biomedical experts to query and access domain knowledge while also facilitates knowledge discovery using computation approaches.

The *iASiS Open Data Graph* has been used to create semantic graphs for three case studies and example queries have been employed to investigate the potential of the platform and its limitations. Directions and ideas for improvement have been identified and discussed based on preliminary experimentation with the resulting datasets. A detailed evaluation is planned to estimate the importance of observed caveats, reveal new ones and set priorities for future work. A specialized user interface for easy graphical interaction with the knowledge graph is being developed towards this direction.

Funding

This work was supported by the EU H2020 programme, under grant agreement No 727658 (project iASiS).

References

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, **32**(suppl_1), D267–D270.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**.

Jaiswal, G. and Agrawal, A. P. (2013). Comparative analysis of relational and graph databases. *IOSR Journal of Engineering (IOSRJEN)*.

Liu, H., Hu, Z.-Z., Zhang, J., and Wu, C. (2005). Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**(1), 103–105.

Liu, Y., Bill, R., Fiszman, M., Rindflesch, T., Pedersen, T., Melton, G. B., and Pakhomov, S. V. (2012). Using semrep to label semantic relations extracted from clinical text. In *AMIA annual symposium proceedings*, volume 2012, page 587. American Medical Informatics Association.

Lu, Z. (2011). PubMed and beyond: A survey of web tools for searching biomedical literature. *Database*, **2011**(February), 1–13.

Lysenko, A., Roznovăţ, I. A., Saqi, M., Mazein, A., Rawlings, C. J., and Auffray, C. (2016). Representing and querying disease networks using graph databases. *BioData mining*, **9**(1), 23.

Masic, I. (2012). Review of most important biomedical databases for searching of biomedical scientific literature. *Donald School Journal of Ultrasound in Obstetrics and Gynecology*, **6**(4), 343–361.

McCray, A. T. (2003). An upper-level ontology for the biomedical domain. *Comparative and Functional Genomics*, **4**(1), 80–84.

Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., *et al.* (2016). The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, **45**(D1), D712–D722.

Rebholz-Schuhmann, D., Oellrich, A., and Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, **13**(12), 829–839.

Rindflesch, T. C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic

propositions in biomedical text. *Journal of biomedical informatics*, **36**(6), 462–477.

Rindflesch, T. C., Blake, C. L., Fiszman, M., Kilicoglu, H., Rosemblat, G., Schneider, J., and Zeiss, C. J. (2017). Informatics support for basic research in biomedicine. *ILAR Journal*, **58**(1), 80–89.

Saha, A., Kuzuhara, T., Echigo, N., Fujii, A., Suganuma, M., and Fujiki, H. (2010). Apoptosis of human lung cancer cells by curcumin mediated through up-regulation of "growth arrest and dna damage inducible genes 45 and 153". *Biological and pharmaceutical bulletin*, **33**(8), 1291–1299.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, **40**(Database issue), D940–6.

Schuyler, P. L., Hole, W. T., Tuttle, M. S., and Sherertz, D. D. (1993). The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, **81**(2), 217.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25**(11), 1251–5.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, **36**(Database issue), D901–6.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.* (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, **46**(D1), D1074–D1082.

Yoon, B.-H., Kim, S.-K., and Kim, S.-Y. (2017). Use of graph database for the integration of heterogeneous biological data. *Genomics & informatics*, **15**(1), 19–27.

Zhang, Y., Sarkar, I. N., and Chen, E. S. (2014). PubMedMiner: Mining and Visualizing MeSH-based Associations in PubMed. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, **2014**, 1990–9.