Intelligent Information Management Targeted Competition Framework ICT-2011.4.4(d)

Project **FP7-318652 / BioASQ**

Deliverable **D3.2**

Distribution **Public**

http://www.bioasq.org

Report on existing and selected datasets

Authors: George Tsatsaronis, Matthias Zschunke, Michael R. Alvers and Christian Plonka

Status: Final (Version 3.0)

January 2013

Project

Project ref.no.	FP7-318652
Project acronym	BioASQ
Project full title	A challenge on large-scale biomedical semantic indexing and question answering
Porject site	http://www.bioasq.org
Project start	October 2012
Project duration	2 years
EC Project Officer	Martina Eydner

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D3.2
Deliverable title	Report on existing and selected datasets
Contractual date of delivery	M3 (December 2012)
Actual date of delivery	January 2013
Relevant Task(s)	WP3/Task 3.2
Partner Responsible	TI
Other contributors	-
Number of pages	21
Author(s)	Authors: George Tsatsaronis, Matthias Zschunke, Michael R. Alvers and Christian Plonka
Internal Reviewers	Eric Gaussier, Ioannis Partalas
Status & version	Final
Keywords	Biomedical Resources, BioASQ Indexed Resources

Executive Summary

In this deliverable, as part of task 3.2 of the *BioASQ* project, we give a detailed description of the resources that are indexed for the purposes of the *BioASQ* challenges.

Chapter 1 gives a general overview of the biomedical landscape and the role of ontologies in the domain. In addition, it associates the needs of the *BioASQ* challenges with the wider biomedical domain.

Chapter 2 describes the most important publicly available resources in the biomedical domain.

Chapter 3 lists the resources that are selected and the criteria based on which the selection was conducted, and describes also the services that have been implemented for the access of the resources in the framework of the *BioASQ* project.

Chapter 4 summarizes the contents of this deliverable and concludes the activities of task 3.2 of the *BioASQ* project.

Contents

List of Figures

1 Introduction

1.1 Overview

In this deliverable, the selected data sources (documents, databases, ontologies) that will be used in the *BioASQ* challenges are described. The document also gives a wider presentation of the available biomedical resources, presents the basic philosophy, upon which the selection was made, and gives details of the indexed resources. In principle, challenge tasks 1a and 2a will use *PubMed* abstracts and *MeSH* concepts, while for challenge tasks 1b and 2b, we established a wider set of ontologies that will be used to annotate questions, as well as resources from which relevant snippets and triples will be retrieved. The choices described in this deliverable were made by consulting the team of biomedical experts of the *BioASQ* project. Local copies of the selected sources are already created on the *BioASQ* infrastructure, to construct preprocessed versions and allow data to be annotated. The rest of this deliverable is organized as follows; the remaining of this chapter gives a general introduction to the landscape of the available biomedical resources, and associates it with the principles and the aims of the *BioASQ* project. Chapter 2 provides a detailed view of the most important and widely used resources in the biomedical domain. Chapter 3 describes the selected resources for each of the *BioASQ* tasks, and lists the services that have been developed for their access. Finally, Chapter 4 summarizes the contents of this deliverable, and concludes the work conducted within *BioASQ* task 3.2.

1.2 The landscape in the biomedical domain

Ontologies such as the *Medical Subject Headings* (*MeSH*) and the *Gene Ontology* (*GO*) play a major role in biology and medicine since they facilitate data integration and the consistent exchange of information between different entities. They can also be used to index and annotate data and literature, thus enabling efficient search and analysis. In the past few years, the volume of the biomedical literature has been growing exponentially, expanding by almost 1 million new scientific papers per year, indexed by *Medline*. This fact makes the task of monitoring the knowledge and the changes in the biomedical domain extremely difficult. This in turn affects the maintenance of the existing biomedical ontologies, but in parallel motivates the creation of new, larger and more detailed thesauri in the domain, that may cover very different information needs.

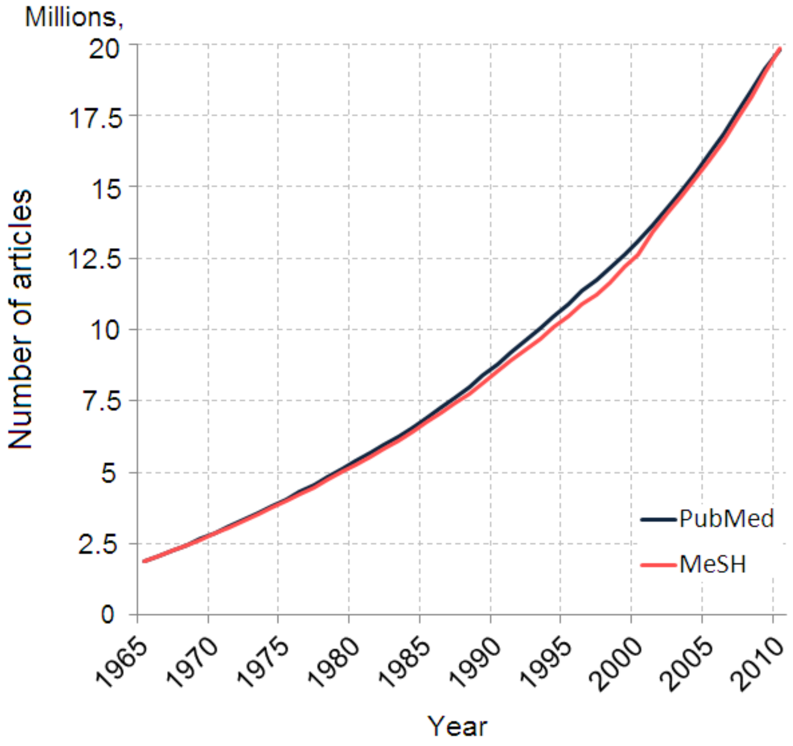


Figure 1.1: Growth of the biomedical scientific literature (indexed by *Medline*) in absolute number of articles over the past 45 years.

As the availability of biomedical resources, such as ontologies and thesauri, is expanding and covers more aspects, their use in biomedical retrieval systems and question answering systems is constantly gaining ground as well. For example, in the past decade we have witnessed a boom in the development of semantic enabled technologies and technologies that rely on advanced natural language processing techniques, that promise to deliver efficient search and provide the basis for advanced question answering approaches (Athenikos and Han (2010); Cairns et al. (2011); Cao et al. (2011); Doms and Schroeder (2005)).

Within the process pipeline of semantic search engines for the biomedical domain there are two major issues that need to be addressed efficiently. The first issue pertains to the ability of the engine to annotate timely and accurately the new scientific articles using concepts of the underlying ontologies. The latest advances in the field of text classification and text alignment have provided the respective research communities of semantic search with novel methodologies which can address efficiently this task. For example, Figure 1.2 shows how the *PubMed* engine is able to follow the exponential increase of published research articles (blue line) by annotating almost all of the new articles (red line) with *MeSH* concepts in a timely fashion. In addition, recent methodologies, e.g., Tsatsaronis et al. (2012a), seem to annotate efficiently and accurately the scientific literature text with *MeSH* concepts using machine learning techniques, despite the ambiguity of *MeSH* terms. Thus, maintaining the pace of the annotations in a level that can follow the increasing amount of newly published articles is an issue that has been sufficiently addressed in the bibliography and can be conducted in a satisfactory manner with automated methods.

The second problem faced by the biomedical semantic search engines is the maintenance of the underlying ontologies, to reflect changes and advances in the biomedical domain. Though the problem, also known in the literature as *ontology evolution* Leenheer and Mens (2008), has been studied for a long time, in the biomedical domain it is far from being solved. The intrinsic difference of the biomedical domain compared to other disciplines is the exponential pace at which new facts and findings are communicated via newly published articles. Thus, the cost of maintaining manually the underlying ontologies is extremely large, given the large number of new articles indexed weekly (especially by *PubMed*).

Despite the aforementioned problems, the use of biomedical knowledge bases is constantly expanding, assisting biomedical workers in harnessing and processing the knowledge in the domain. The bottlenecks in this process are: (a) the standardization of the ontologies' representations, so that concept labels across different ontologies refer to the same entities, (b) the methodology of using ontologies for specific applications, and, (c) the large scale of the data in the biomedical databases that the systems have to process, in order to retrieve information or answer questions. In the following we briefly discuss these aspects, namely how ontologies are expected to aid in the organization of the biomedical knowledge, examples of how they are actually used nowadays in real applications, and which formats have prevailed over the years for the organization of the biomedical knowledge.

The extremely large scale of data and information in the biomedical domain has been motivating the organization of biomedical knowledge with the use of ontologies and has been the focus point of many initiatives and activities in the biomedical domain (Bodenreider and Stevens (2006)). Domain ontologies represent a conceptualization of how things are organized in reality in the underlying domain (Guarino (1998)). An advantage of such a formal representation is that the labels used to describe concepts, and their properties, provide an actual language for the community of this domain, with which they can talk about domain knowledge, and exchange data. Moreover, sharing the same conceptualization of things allows researchers to communicate new facts and knowledge referring to the same concepts that may be found with different labels across several different data sources. In a way, producing formal definitions of concepts and their properties in the biomedical domain, allows the biomedical knowledge workers to handle knowledge computationally in a manner comparable to that in which we handle numeric data.

More formally, an ontology is a set of logical axioms which model the reality of the domain. With the advent of *description logics* (*DL*) (Baader et al. (2004)) and *OWL DL*, which is the description logic flavor of *OWL*, the task of designing and implementing formally ontologies has become easier, as the ontology engineers may express the ontology concepts and their relations without losing computational completeness, and in parallel retain decidability of reasoning systems. In practice *DL* has become the leading formalism for representing ontologies, a task which nowadays is also supported by many popular ontology editors such as *Protégé*[1] and *OBO-Edit*[2]. In addition, many large biomedical ontologies have adopted this formalism, such as *GALEN* (Rector and Rogers (2006)), which was also the first biomedical ontology to be developed in *DL*, the *NCI Thesaurus*, and *SNOMED CT* , and for several years now, research on how other biomedical ontologies may be translated to *DL* has been conducted (Hahn and Schulz (2003); Soualmia et al. (2004); Jupp et al. (2012)).

With *OWL DL* and similar formalisms, e.g., lightweight Description Logic $\mathcal{EL}++$, becoming standard ontology languages, the way to actual applications of biomedical ontologies has been paved. Coherent formalization of biomedical ontologies facilitates harmonisation, integration and re-usability (Smith and Brochhausen (2010)). Such formalization can also unravel multiple ontological perspectives of biomedical entities and enrich the results of the query process (Jupp et al. (2012)). For example, Hoehndorf *et* al. (2012) pinpoint the contribution of biomedical ontologies to drug repurposing (also known as drug repositioning). They propose the integration of pharmacogenomic databases with formalized

[1] http://protege.stanford.edu/

[2] http://oboedit.org/

ontological information in order to shed light on hidden drug-pathway-disease associations (Hoehndorf et al. (2012)). Biomedical ontologies have also been applied successfully to automated protein classification and annotation tasks (Wolstencroft et al. (2006)). In other examples, biomedical ontologies are used to model of patient information and clinical data (Bouamrane et al. (2011)), gene product annotations (Ashburner et al. (2000)), analysis of high-throughput data (Whetzel et al. (2006)) and search (Tsuruoka et al. (2008)). Therefore, the role of ontologies in the current biomedical landscape has become extremely important, and under this scope, we discuss in the next section the rationale through which biomedical ontologies can also play an important role in biomedical question answering, and, more precisely, how they will be used for the *BioASQ* challenges.

1.3 Requirements of the *BioASQ* challenge task

Although research on biomedical *Question Answering* (*QA*) has boomed in recent years (Athenikos and Han (2010); Cairns et al. (2011); Cao et al. (2011)), current systems focus on particular resources; for example, *MedQA* (Lee et al. (2006)) uses *MeSH* and the *Gene Ontology* only. By contrast, biomedical knowledge workers need to synthesise and filter information from multiple, extremely large and fast-growing sources. Existing search engines (e.g., *PubMed*[3], *GoPubMed*[4], *EBIMed*[5]) only partially address this need. They also focus on a limited range of resources (e.g., only *PubMed* articles and concepts from the *GO* or *MeSH*), whereas multiple sources (e.g., specialised drug databases and ontologies) often need to be combined (Athenikos and Han (2010)). Semantic indexing, i.e., annotating resources with concepts from established semantic taxonomies or, more generally, ontologies, provides a means to combine multiple sources and facilitates matching questions to answers. Current semantic indexing, however, is largely performed manually, and needs to be automated to cope with the vast amount of new information that becomes available daily. At the same time, both current semantic indexing and *QA* methods require a significant push to reach a level of quality and efficiency acceptable by biomedical experts. *BioASQ* intends to push towards that direction: motivating the development of efficient and effective semantic indexing and *QA* methods for the biomedical domain, and establishing an evaluation framework and benchmarks for biomedical *QA*.

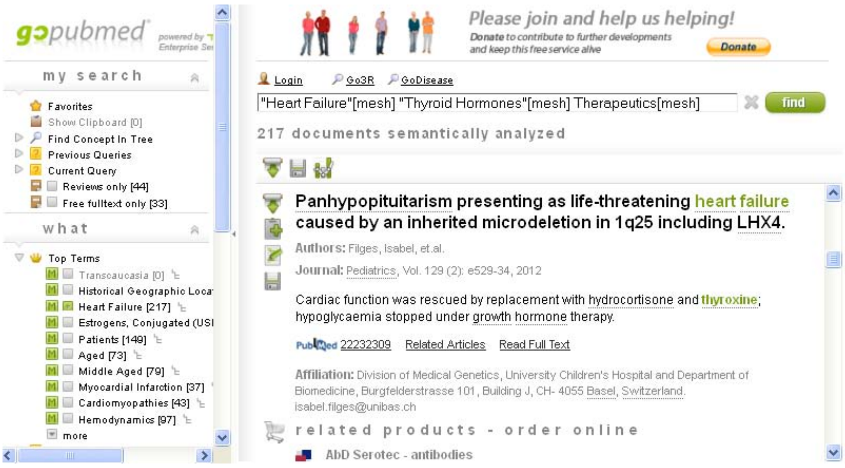
To illustrate the challenges that a modern biomedical *QA* system faces, we present below a case study, which is part of a larger scenario from the *PONTE* project[6]. The larger scenario, called *THIRST*, pertains to the design of a Clinical Trial Protocol (*CTP*) regarding the safety and feasibility of synthetic thyroid (*TH*) replacement therapy with a triiodothyronine analogue (*Liotir*) in patients with *ST-Elevation Myocardial Infarction* (*STEMI*), both in the acute (in-hospital period) and chronic phase (after hospital discharge) of coronary artery disease and its association with cardiac function and outcome. In addition, *THIRST* examines the effects of *TH* replacement therapy on the clinical outcome in terms of major (cardiac and non cardiac death, reinfarction) and minor (recurrence of angina, coronary revascularization, and hospital re-admission) events. During the *THIRST* scenario, the *Principal Investigator* (*PI*) of the *CTP* design formulates a “hypothesis”, in effect a target to be proven, based on which a new clinical trial can potentially start. The target is *“Evaluate the safety and the effects of TH treatment in patients with acute myocardial infarction”*. The target requires concrete answers to several questions; we show below two of the questions (Q1, Q2). The questions are produced by the *PI* and his/her colleagues, and in effect capture their information needs using natural language.

[3] http://www.ncbi.nlm.nih.gov/pubmed/

[4] http://www.gopubmed.com/web/gopubmed/

[5] http://www.ebi.ac.uk/Rebholz-srv/ebimed/

[6] http://www.ponte-project.eu/



"Heart Failure"[mesh] "Thyroid Hormones"[mesh] Therapeutics[mesh]

217 documents semantically analyzed

Panhypopituitarism presenting as life-threatening heart failure caused by an inherited microdeletion in 1q25 including LHX4.

Authors: Filges, Isabel, et.al.

Journal: Pediatrics, Vol. 129 (2): e529-34, 2012

Cardiac function was rescued by replacement with hydrocortisone and thyroxine; hypoglycaemia stopped under growth hormone therapy.

PubMed 22232309 Related Articles Read Full Text

Affiliation: Division of Medical Genetics, University Children's Hospital and Department of Biomedicine, Burgfelderstrasse 101, Building J, CH- 4055 Basel, Switzerland. isabel.filges@unibas.ch

Figure 1.2: Using GoPubMed to find the answer to question Q1.

Q1 What is the role of thyroid hormones administration in the treatment of heart failure?

Q2 What is the relation of thyroid hormones levels in myocardial infarction?

Unfortunately, the questions cannot be submitted directly to current bibliographic databases (e.g., *PubMed*). To retrieve the scientific articles that provide the answers to the questions, the *PI* and his/her team have to translate the questions to collections of terms, in effect concepts from the taxonomy used by the curators of the bibliographic database (e.g., *MeSH* headings in the case of *PubMed*). The terms (concepts) that correspond to Q1 and Q2 are shown below as T1 and T2, respectively. Note that this translation process is not trivial, as the original questions may, for example, contain synonymous to the concepts of the taxonomy terms. Furthermore, additional terms (concepts) may have to be added, e.g., topically related terms, hypernyms etc., to increase the recall of document retrieval (find additional relevant documents) and its precision (i.e., avoid documents corresponding to different concepts, e.g., due to sense ambiguity).

T1 heart failure thyroid hormone treatment

T2 myocardial infarction thyroid hormone

T1 and T2 are submitted to a document retrieval engine as queries, and the engine returns a (possibly long) list of documents. The *PI* and his/her team have to study these documents to find snippets that contain information that answer their questions. As an example of the advantages, but also of the limitations that current state-of-the-art biomedical semantic search engines offer, Figure 1.2 shows a screenshot from *GoPubMed*, where the terms of T1 are used as *MeSH* filters. The user still gets 217 documents, which he/she has to read to manually extract the answer to Q1. The major advantage of such systems is, of course, their ability to filter the 21 millions of *Medline* documents using the specified concepts, reducing the search space of the human reader to just a few hundreds, yet the engine cannot directly produce answers to questions like Q1 and Q2.

BioASQ aims to push towards solutions to the problems illustrated in the above scenario. It will set up a challenge on biomedical semantic indexing and *QA*, which requires the challenge participants to semantically index content from large-scale biomedical sources (e.g., *Medline*) and to assemble data from multiple heterogeneous sources (e.g., scientific articles, ontologies, databases) in order to compose informative answers to biomedical natural language questions. In particular, the *BioASQ* challenge evaluates the ability of systems to perform:

- large-scale classification of biomedical documents onto ontology concepts, to automate semantic indexing,
- classification of biomedical questions on the same concepts,
- integration of relevant document snippets, database records, and information (possibly inferred) from knowledge bases, and
- delivery of the retrieved information in a concise and user-understandable form.

Benchmarks containing development and evaluation questions and gold standard (reference) answers will be developed during the project. The gold standard answers will be produced by a team of biomedical experts from research teams around Europe. Established methodologies from *QA*, summarisation, and classification will be followed to produce the benchmarks and evaluate the participating systems.

Producing sufficient and concise answers from this wealth of information that exists in the biomedical domain is a challenging task for traditional search engines, which largely rely on term (keyword) indexing. Obtaining the required information is made even more difficult by non-standard terminology and the ambiguity of the technical terms involved. Therefore, indexing at the semantic (concept) level, rather than at the level of keywords only, is particularly important. Biomedical concept taxonomies or, more generally, ontologies are abundant and they provide concept inventories that can be used in semantic indices. Hierarchical classification algorithms (Silla and Freitas (2011)) can classify documents and questions onto the concepts of these inventories, facilitating the matching of questions, documents, and also structured data (e.g., *RDF* triples) that already have explicit semantics based on the same concepts.

Figure 1.3 provides an overview of the biomedical semantic indexing and *QA* architecture adopted by *BioASQ*. To the best of our knowledge, this architecture subsumes all the existing relevant approaches, but no single existing biomedical search system currently instantiates all the components of the architecture. Hence, the architecture can be seen as a broader framework for the future systems that *BioASQ* hopes to push towards. Starting with a variety of data sources (lower right corner of the figure), semantic indexing and integration brings the data into a form that can be used to respond effectively to domain-specific questions. A semantic *QA* system associates ontology concepts with each question and uses the semantic index of the data to retrieve the relevant pieces of information. The retrieved information is then turned into a concise user-understandable form, which may be, for example, a ranked list of candidate answers (e.g., in factoid questions, like *"What are the physiological manifestations of disorder Y?"*) or a collection of text snippets, ideally forming a coherent summary (e.g., in *"What is known about the metabolism of drug Z?"*).

In the duration of the project two versions of the *BioASQ* challenge will run and will be organized into tasks, depicting the steps of the methodology. The first version if the challenge comprises two tasks:

Task 1a: Large-scale biomedical semantic indexing

This task is based on the standard process followed by *PubMed* curators, who manually index biomedical articles. The participants will be asked to classify new abstracts, written in English, as they become available online, before *PubMed* curators annotate (in effect, classify) them manually; at any point in time there is usually a backlog of approximately $10,000$ non-annotated abstracts. The classes (concepts) will come from *MeSH*; they will be the subject headings currently used to manually index the abstracts. As new manual annotations become available, they will be used to evaluate the classification performance of participating systems (which will classify articles before they are manually annotated) using standard IR measures (e.g., precision, recall, accuracy), as well as hierarchical variants of them (Brucker

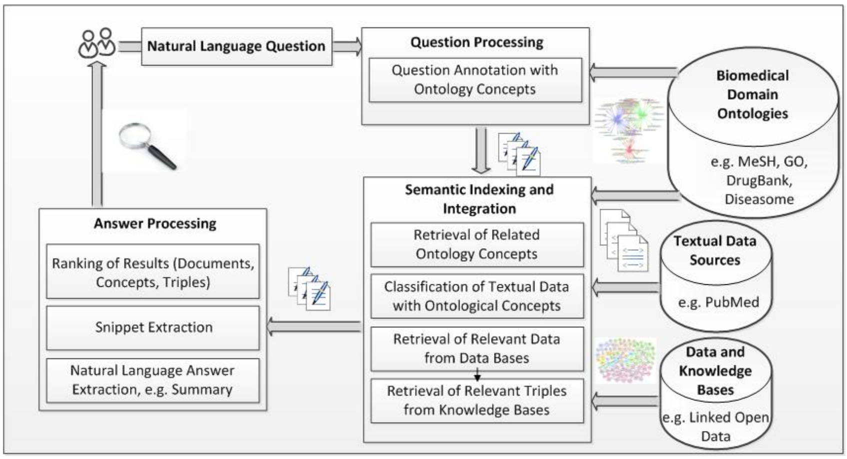

Natural Language Question
Question Processing
Question Annotation with Ontology Concepts
Biomedical Domain Ontologies
e.g. MeSH, GO, DrugBank, Diseasome
Semantic Indexing and Integration
Retrieval of Related Ontology Concepts
Classification of Textual Data with Ontological Concepts
Retrieval of Relevant Data from Data Bases
Retrieval of Relevant Triples from Knowledge Bases
Textual Data Sources
e.g. PubMed
Data and Knowledge Bases
e.g. Linked Open Data
Answer Processing
Ranking of Results (Documents, Concepts, Triples)
Snippet Extraction
Natural Language Answer Extraction, e.g. Summary

Figure 1.3: Overview of semantic indexing and question answering methodology adopted in *BioASQ*.

et al. (2011)). The participants will be able to train their classifiers, using the whole history of manually annotated abstracts.

Task 1b: Introductory biomedical semantic *QA*

This task aims to be an introductory step towards biomedical semantic *QA* for state-of-the-art generic IR and *QA* systems. It will be based on benchmarks created specifically for *BioASQ* with the help of biomedical experts. The task will take place in two phases:

Annotate questions, retrieve relevant snippets and triples. In the first phase of Task 1b, the participants will be provided with questions written in English and will be asked to (i) semantically annotate the questions with concepts from a particular set of ontologies, and (ii) retrieve data (text snippets from *PubMed* articles written in English, knowledge base triples, etc.) that are relevant to the questions (possibly as revealed by the semantic annotations of Task 1a and the semantic annotations of the questions) from designated sources. The system responses will be compared against gold responses provided by the human experts, using standard IR measures.

Find and report answers. In the second phase of Task 1b, the questions and gold responses of the first phase will be provided as input and the participants will be asked to report answers found in the input snippets and triples. In effect, this phase assumes a perfect first-phase system, which is available to obtain relevant snippets and triples. The competing systems will be required to output ranked lists of candidate answers (e.g., names or numbers) in the case of factoid questions, or sets of text snippets and/or triples in the case of questions that ask for summaries. The answers of the systems will be compared against gold answers constructed by biomedical experts, using evaluation measures from *QA* and summarisation, such as *mean reciprocal rank* (Voorhees (2001)), *ROUGE* (Lin (2004)), *Basic Elements* (Tratz and Hovy (2008)), and other automatic summary evaluation measures (Giannakopoulos et al. (2009)). Systems that opt to provide partial responses (e.g., report only triples and no snippets) will be evaluated partially.

Task 2a: Large-scale biomedical semantic indexing

This task will be the same as Task 1a, improved according to the feedback that will have been collected.

Task 2b: Biomedical semantic *QA*

This task aims to combine the two phases of Task 1b. The participants will be provided with a fresh set of questions and will be asked to (i) annotate them with concepts and retrieve relevant data (snippets and triples) from designated sources, as in the first phase of Task 1b; and (ii) find and report answers, as in the second phase of Task 1b, but now without assuming that a perfect first-phase system is available to

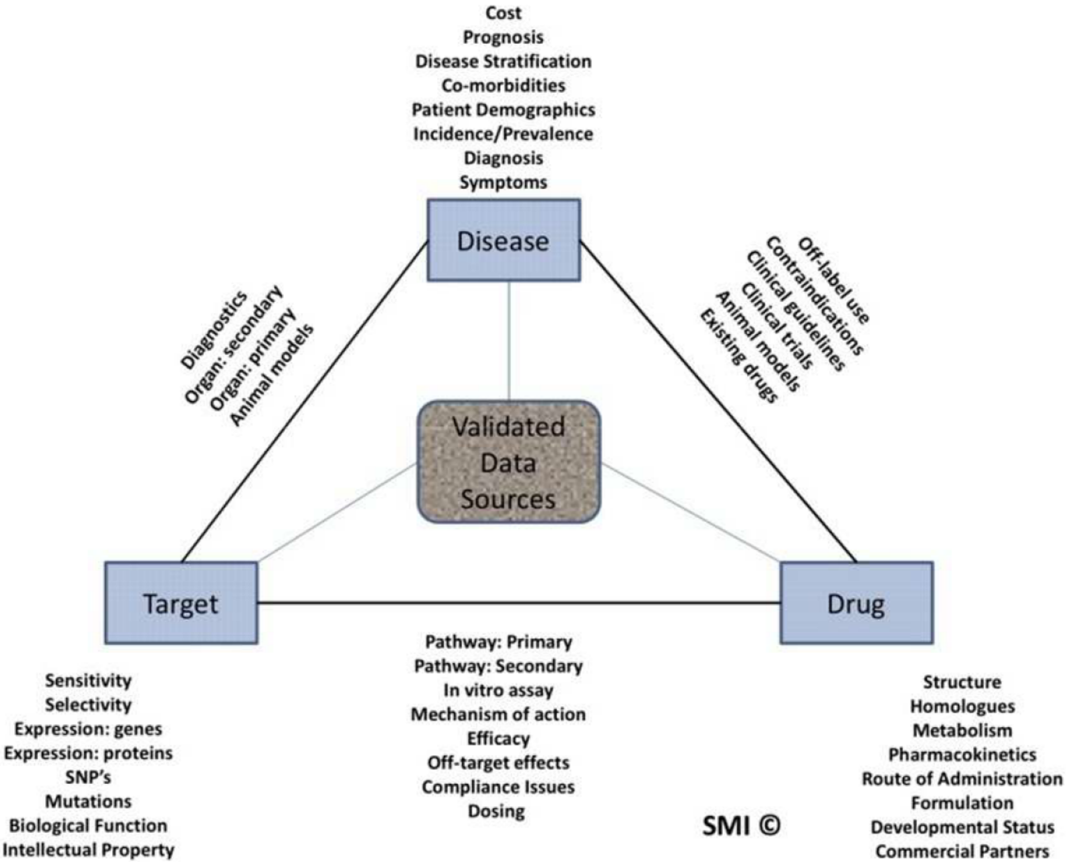


Figure 1.4: The principle of drug-target-disease information axes followed in *BioASQ*.

obtain relevant snippets and triples, i.e., no gold responses for (i) will be provided. Again, systems that opt to provide partial responses will be evaluated partially.

For all of the *BioASQ* tasks the reader may find a discussion of related competitions, projects and initiatives in Tsatsaronis et al. (2012b). Regarding tasks 1a and 2a, the resources that will be used will be the documents indexed by *Medline*, and the *MeSH* concepts from which the document classes are drawn. For the tasks 1b and 2b the *BioASQ* challenges aim to cover a wide range of biomedical concepts, through the use of ontologies and linked data that describe several facets of the domain. The selection of resources for these tasks follows the triangle *drug-target-disease* which defines the prime information axes for any medical investigation. The main principle is shown in Figure 1.4[7].

This "*knowledge-triangle*" supports the conceptual linking of biomedical knowledge databases and the processing of the related resources. Based on this processing, systems can address questions that combine any path connecting the vertices of the triangle, provided that they can also annotate with accuracy the natural language questions with ontology concepts. Examples of the questions that may be addressed if resources are in principle connected as shown in Figure 1.4 follow, seen from all three perspectives (drugs, targets, diseases).

[7] The figure is courtesy of Strategic Medicine Inc.

Disease Focus

Q1 Are there known distinguishing characteristics of patients who progress from <disorder X> to <disorder Y>?

Q2 What are the current clinical guidelines about treatment of <disorder X>?

Q3 What are the parameters used to make the diagnosis of <disorder X>?

Target Focus

Q4 What is known about the selectivity of <target X> as a target for current drug treatments?

Q5 Is there data about normal levels of gene and/or protein expression levels for <target X>?

Q6 What are known side-effects of acting on <target X>?

Drug Focus

Q7 What are current clinical uses for <drug X> and what are the contraindications for its use?

Q8 What are the alternatives for administering <drug X>?

Q9 What is the primary and secondary target of <drug X>?

Based on this "*knowledge-triangle*" (drug-target-disease), in the next section we describe popular and widely used available resources that cover all parts of the triangle, as well as resources that are built for general purpose usage, e.g., clinical term dictionaries.

2

Biomedical Resources for Question Answering

An overwhelming amount of biomedical text information is available reporting on the vast quantities of biomedical discoveries and studies. Most of this literature is stored to digital libraries or biomedical literature databases such as *Medline*). *Medline* is the largest biomedical bibliographic text database, and currently has nearly 23 million articles indexed. Addressing the resulting information overload, attempts are made to transform text information into machine-understandable knowledge for knowledge management. For this purpose, text mining techniques have been used, together with ontologies that organize biomedical knowledge. In the following we discuss the most popular resources in the biomedical domain frequently used for question answering (Athenikos and Han (2010)).

2.1 Drugs

2.1.1 Jochem

Jochem (Hettne et al. (2009)), the Joint Chemical Dictionary,, is a dictionary for the identification of small molecules and drugs in text, combining information from *UMLS*, *MeSH*, *ChEBI*, *DrugBank*, *KEGG*, *HMDB*, and *ChemIDplus*. The resources were chosen on the basis of free availability. They are downloadable terminology databases containing small molecules from human studies. Given the variety and the population of the different resources merged in *Jochem*, it is currently one of the largest biomedical resources for drugs and chemicals.

2.1.2 Drug Ontology

The *Drug Ontology* is developed by the nosology project at Stanford Center for Biomedical Informatics Research. This project seeks new therapeutic uses and adverse effects of drugs by identifying diseases that have gene expression profiles similar to those of the known indications and adverse effects of drugs. The objectives of the *Drug Ontology* are: (a) to define a core set of concepts and relationships that allows integration of information from multiple sources, (b) to provide classification services along multiple axes, and, (c) to provide links to external sources so that data not in the ontology can be queried from these sources. The *Drug Ontology* contains the *Pharmacogenomics knowledge base*, *DrugBank*, and the *US National Drug Data File - Reference Terminology*.

2.1.3 ATC Ontology

The *Anatomical Therapeutic Chemical Classification System* (*ATC*) is used for the classification of drugs. It is controlled by the *WHO Collaborating Centre for Drug Statistics Methodology* (*WHOCC*), and was first published in 1976. This pharmaceutical coding system divides drugs into different groups, according to the organ or system on which they act and/or their therapeutic and chemical characteristics. Currently, the *International Classification of Diseases* (*ICD* - explained in Section 2.3.2) contains no reference to any external classification of medical substances. However, guidelines for its 10th revision (*ICD-10*) section "*Poisoning by drugs, medicaments and biological substances*" (T36-T50) could contain suggestions on how to combine, or replace *ICD* codes with specific substance codes, preferably using an internationally widespread and recommended standard system such as the *ATC*.

2.2 Targets

2.2.1 Gene Ontology

The *Gene Ontology* (*GO*) is currently the most successful case of ontology use in bioinformatics and provides a controlled vocabulary to describe functional aspects of gene products. The ontology covers three domains: *cellular component*, the parts of a cell or its extracellular environment; *molecular function*, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and *biological process*, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

2.2.2 UniProt

The *Universal Protein Resource* (*UniProt*) provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. Its protein knowledge base consists of two sections: *Swiss-Prot*, which is manually annotated and reviewed, and contains approximately 500 thousand sequences, and *TrEMBL*, which is automatically annotated and is not reviewed, and contains approximately 23 million sequences. The primary mission of the *Universal Protein Resource* (*UniProt*) is to support biological research by maintaining a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledge base, with extensive cross-references and querying interfaces freely accessible to the scientific community. In particular, the *Swiss-Prot* component of *UniProt*, it is a high-quality, manually annotated, non-redundant protein sequence database which combines information extracted from scientific literature and biocurator-evaluated computational analysis. The aim of *Swiss-Prot* is to provide all known relevant information about a particular protein. Annotation is regularly reviewed to keep up with current scientific findings. The manual annotation of an entry involves detailed analysis of the protein sequence and of the scientific literature.

2.2.3 SuperTarget and Matador

SuperTarget (Günther et al. (2008)) integrates drug-related information about medical indication areas, adverse drug effects, drug metabolization, pathways and *Gene Ontology* terms of the target proteins. The database contains more than $2,500$ target proteins, which are annotated with about $7,300$ relations to $1,500$ drugs; the vast majority of entries have pointers to the respective literature source. A subset of these drugs has been annotated with additional binding information and indirect interactions and is available as a separate resource called *Matador*.

2.3 Diseases

2.3.1 Disease Ontology

The *Disease Ontology* (*DO*) contains data associating genes with human diseases, using established disease codes and terminologies. Approximately $8,000$ inherited, developmental and acquired human diseases are included in the resource. The *DO* semantically integrates disease and medical vocabularies through extensive cross-mapping and integration of *MeSH*, *ICD*, *NCI*'s thesaurus, *SNOMED CT* and *OMIM* disease-specific terms and identifiers. The *DO* is utilized for disease annotation by major biomedical databases (e.g., *Array Express*, *NIF*, *IEDB*), as a standard representation of human disease in biomedical ontologies (e.g., *IDO*, *Cell line* ontology, *NIFSTD* ontology, *Experimental Factor* Ontology, *Influenza* Ontology), and as an ontological cross-mappings resource between *DO*, *MeSH* and *OMIM* (e.g., *GeneWiki*). *DO* has been incorporated into open source tools (e.g., *Gene Answers*, *FunDO*) to connect gene and disease biomedical data through the lens of human disease.

2.3.2 ICD-10

The *International Classification of Diseases* (*ICD*) is the standard diagnostic tool for epidemiology, health management and clinical purposes. It supports the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problems, and to classify diseases recorded on many types of health and vital records including death certificates and health records. *ICD-10* is the 10th revision of *ICD*. It codes diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. The code set allows more than $14,400$ different codes and permits the tracking of many new diagnoses. It is mainly used in the medical domain and provides a format for reporting causes of death on the death certificates. The reported conditions are then translated into medical codes through the use of the classification structure and the selection and modification rules contained in the applicable revision of the *ICD*, published by the *World Health Organization* (*WHO*).

2.3.3 Diseasome

Diseasome describes characteristics of disorders and disease genes linked by known disordergene associations. It publishes a network of $4,300$ which explores all known phenotype and disease gene associations, indicating the common genetic origin of many diseases. The list of disorders, disease genes, and associations between them was obtained from the *Online Mendelian Inheritance in Man* (*OMIM*), a compilation of human disease genes and phenotypes.

2.4 General Purpose

2.4.1 The Medical Subject Headings Hierarchy

Medical Subject Headings (*MeSH*) is a hierarchy of terms maintained by the *United States National Library of Medicine* (*NLM*) and its purpose is to provide headings (terms) which can be used to index scientific publications in the life sciences, e.g., journal articles, books, and articles in conference proceedings. The indexed publications may be then searched through popular search engines, such as *PubMed* or *GoPubMed*, using the *MeSH* headings to filter semantically the results. This retrieval methodology seems to be in some cases beneficial, especially when precision of the retrieved results is important (Doms and Schroeder (2005)).

MeSH includes three types of data: (i) *descriptors*, also known as *subject headings*, (ii) *qualifiers*, and, (iii) *supplementary concept records*. *Descriptors* are the main terms that are used to index scientific publications. The *descriptors* are organized into 16 trees, and as of 2013 they are $26,853$[1]. They include a short description or definition of the term, and they frequently have synonyms, known as *entry terms*. *Qualifiers*, also known as *subheadings*, may be used additionally to narrow down the topic of each of the *descriptors*. In total there are approximately 80 *qualifiers* in *MeSH*. *Supplementary concept records*, approximately $214,000$ in the most recent *MeSH* release, describe mainly chemical substances and are linked to respective *descriptors* in order to enlarge the thesaurus with information for specific substances. *MeSH* is the main resource used by *PubMed* to index the biomedical scientific bibliography in *Medline*.

2.4.2 SNOMED CT

The *Systematized Nomenclature of Medicine* (*SNOMED CT*) is developed by the *College of American Pathologists*, and was formed by the convergence of *SNOMED RT* and *Clinical Terms* Version 3 (formerly known as the *Read Codes*). *SNOMED CT* is the most comprehensive biomedical terminology recently developed in native description logic formalism. It contains approximately $300,000$ entries. Each *SNOMED CT* concept is described by a variable number of elements, among which are roles (or semantic relations) that connect the concept to other *SNOMED CT* concepts. *SNOMED CT* consists of eighteen independent hierarchies reflecting, in part, the organization of previous versions of *SNOMED* into axes such as *Diseases*, *Drugs*, *Living Organisms*, *Procedures* and *Topography*.

2.4.3 UMLS

The *Unified Medical Language System* (*UMLS*) started by *NLM* as a long-term R&D project in 1986 and provides a mechanism for integrating all the major biomedical vocabularies such as the *MeSH*, and the *Systematized Nomenclature of Medicine Clinical Terms* (*SNOMED CT*). Ultimately, *UMLS* aims to pave the way for the development of intelligent biomedical systems that can read biomedical and healthcare data, comprehend the meaning of them and further make inferences from them. *UMLS* consists of three knowledge sources: *Metathesaurus*, *Semantic Network*, and the *SPECIALIST* lexicon. The *Metathesaurus* is a very large vocabulary database (nearly 5GB in text) whose data are collected from various biomedical thesauri. Currently, *Metathesaurus* contains more than 1 million biomedical concepts (meanings), 5 million unique concept names from nearly 150 different source vocabularies, and more than 17 million relationships between concepts. Practically this means that each meaning (concept) may be mapped to more than one label (concept name), which in this case implies that they are synonyms. The *Semantic Network* of *UMLS* consists of semantic types and semantic relations. Semantic types are simply categories for concepts so that every *Metathesaurus* concept is assigned to at least one semantic type as a category. Semantic relations are relationships (e.g., diagnoses) between semantic types instead of concepts. Thus, in the *Semantic Network*, semantic types are nodes and semantic relations are links between nodes. The Semantic Network currently contains approximately 130 semantic types and 50 distinct semantic relations. Both the semantic types and the semantic relations are hierarchically arranged from most general to most specific as are *MeSH* Descriptors. The *SPECIALIST* lexicon has

[1] The most recent *MeSH* version is released as *MeSH* 2013.

been designed for the *SPECIALIST Natural Language Processing* (*NLP*) System. In order to supply the *NLP* system with lexical information, the lexicon contains general English words and many biomedical terms. Those words and biomedical terms come from the *American Heritage Word Frequency Book*, English dictionaries such as *Longmans Dictionary of Contemporary English*, the *UMLS Test Collection of MEDLINE abstracts*, and *Dorlands Illustrated Medical Dictionary*.

2.5 Document Sources

The primary corpora for text-based *QA* in the biomedical domain are accessible through *PubMed* and *PubMed Central*. *PubMed*, a service provided by the *National Library of Medicine* (*NLM*), under the *U.S. National Institutes of Health* (*NIH*), contains over 23 million citations from *Medline*, a bibliographic database (DB) of biomedical literature, and other biomedical and life science journals dating back to the 1950s. It is accessible through the *National Center for Biotechnology Information* (*NCBI*). *PubMed Central* (*PMC*) is a digital archive of full-text biomedical and life science articles. The full text of all *PubMed Central* articles is freely available. As of July 2011, the archive contains approximately 2.2 million items, including articles, editorials and letters.

2.6 Linked Data

The *BioASQ* tasks 1b and 2b require the usage of biomedical data expressed as triples, e.g., *subject-predicate-object* structured facts, extracted from biomedical resources or bibliography. In this direction, the *Linked Life Data* project provides the *LinkedLifeData* platform. *LinkedLifeData* is a data warehouse that syndicates large volumes of heterogeneous biomedical knowledge in a common data model. The platform uses an extension of the *RDF* model that is able to track the provenance of each individual fact in the repository and thus update the information. It contains currently more than 8 billion statements, with almost 2 billion entities involved. The statements are extracted from 26 biomedical resources, such as *PubMed*, *UMLS*, *DrugBank*, *Diseasome*, and *Gene Ontology*. The statements are publicly available, and the project provides also a wide list of instance mappings.

3

Selected Resources for BioASQ Tasks

3.1 Resources

Based on the *BioASQ* project needs, discussed in Chapter 1, in the following we describe the resources that will be used for the *BioASQ* tasks. The selection was made following simple criteria: (a) the selected resources should be publicly available, (b) the selected resources should cover as widely as possible the three main vertices of the *drug-target-disease* triangle, to allow formulation of interesting questions, (c) the format of the resources that are ontologies should be widely acceptable and usable for representing ontologies and concept properties, such as *OBO*, *OWL*, or *SKOS*, (d) the selected resources should preferably have low degree of overlap, and when overlap exists, mapping between the overlapped resources should be available, or could be derived. High degree of overlap would lead systems to perform multiple annotations with the same concepts, which is not desired. Furthermore, high overlap can overwhelm the biomedical experts who will engineer the benchmark questions.

For the purposes of *BioASQ* tasks 1a and 2a, we have indexed the *PubMed* data, e.g., approximately 23 million entries with titles and abstracts of the papers, as well as the *MeSH* ontology, from which the class labels (*MeSH* headings) are drawn. Regarding tasks 1b and 2b, in addition to the aforementioned, we have indexed:

- the *Jochem* ontology, for the purpose of covering drugs
- the *UniProt* database (the *Swiss-Prot* component), for the purpose of covering targets
- the *Gene Ontology*, also for the purpose of covering targets
- the *Disease Ontology*, for the purpose of covering diseases
- the *LinkedLifeData* triples, for the purpose of covering the needs of tasks 1b and 2b regarding extracted facts (approximately 8 billion statements, which also include all of the statements extracted from *UMLS*)
- approximately $800,000$ full text articles from *PubMed Central*, for the purpose of expanding the *PubMed* document source with searchable full text articles. The articles are offered through a particular agreement with *TI*.

3.2 Accessibility of the Resources via Services

All of the aforementioned resources have been indexed by *TI* and are provided through respective Web services. The ontological resources have been converted to proper *OBO* files, i.e., files formatted following the *OBO Foundry* Flat File Format Specification for ontologies[1]. The concept names (labels), their synonyms and their relations have been indexed in separate *Lucene* indexes. For the document resources, also *Lucene* indexes are used, applying the standard *Lucene* analyzer for the English language.

The *API* through which the resources may be accessed, is based on *JSON*. For each resource, a respective service is implemented in a unique *URL*. Each *URL* request opens a session and may request the results, given a query, e.g., a concept, using *HTTP-POST* and a parameter *json*. The reply (the value of the *json* parameter) is a *JSON* object that contains the results for the given query. In the case of the ontological resources, the result list contains concepts from the respective ontology, and in the case of the document sources, the result list contains citations from *Medline* (title, and abstract), or full text articles from *PubMed Central*.

A list of the services that have been developed follows, with a short description of the input and output parameters used for accessing the resources and getting results.

- In the URL: `http://www.gopubmed.org/web/bioasq/mesh/json`, a service for accessing the *MeSH* ontology, with input parameter "*findEntity*", and output parameter "*findings*", which contains the list of related concepts (a list of "*concept*" entries with "*label*" entries), given the query submitted with the input parameter. Additional information is provided inside each "*label*" entry in the *JSON* object, such as "*termId*" and "*uri*" of the concept. In addition, inside each "*concept*" entry, the offsets in which the query keywords matched each returned concept are provided.

- In the URL: `http://www.gopubmed.org/web/bioasq/go/json`, a service for accessing the *GO* ontology, with the same input and output parameters as aforementioned.

- In the URL: `http://www.gopubmed.org/web/bioasq/uniprot/json`, a service for accessing the *UniProt* database, with the same input and output parameters as aforementioned.

- In the URL: `http://www.gopubmed.org/web/bioasq/jochem/json`, a service for accessing the *Jochem* ontology, with the same input and output parameters as aforementioned.

- In the URL: `http://www.gopubmed.org/web/bioasq/doid/json`, a service for accessing the *Disease Ontology*, with the same input and output parameters as aforementioned.

- In the URL: `http://www.gopubmed.org/web/bioasq/pubmed`, a service for accessing the *PubMed* indexed documents (titles and abstracts), with the same input parameters as aforementioned, and the output parameter containing "*documents*" entries in the returned *JSON* object. Each entry has a "*pmid*" element, which is the *PubMed* id of the indexed citation, a "*documentAbstract*" entry, and a "*title*" entry. In addition, the *MeSH* annotations are provided when available.

- In the URL: `http://www.gopubmed.org/web/bioasq/pmc/json`, a service for accessing the *PMC* full text articles, with the same input parameters and output parameters as aforementioned, with the only difference being that the articles returned contain in addition the full text.

[1] `http://www.obofoundry.org/`

- In the URL: `http://www.gopubmed.org/web/bioasq/linkedlifedata/triples`, a service for accessing the *LinkedLifeData* platform triples. The input parameter is "*findTriples*", and accepts any keywords as query. The output parameter contains a list of "*triples*" entries. Each entry has in turn a "*subj*", "*pred*", "*obj*" and "*score*" field, representing the subject, the predicate and the object of the triple, and the matching score given the input query.

4

Summary

As part of task *3.2* of the *BioASQ* project, this deliverable provides a summary of the research landscape in the biomedical domain, pertaining to the usage of ontologies, databases and thesauri. Additionally, it associates the use of biomedical resources to the needs of the *BioASQ* challenges. We provided a description of the the most widely used biomedical resources, based on which parts of the *knowledge-triangle* (*drug-target-disease*) they cover, and also resources mentioned more generic. Finally, we summarized the criteria according to which we selected a subset of the discussed resources, and we listed the biomedical resources that are indexed for the purposes of the *BioASQ* challenges, along with the respective APIs developed for their access. The indexed data described in this deliverable will constitute the information basis both for the medical experts to design the benchmark questions, but also for the participants to design and implement their question answering systems.

Bibliography

The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD). http://www.who.int/classifications/atcddd/en/. Accessed: December 2012.

Disease ontology. http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page. Accessed: December 2012.

Diseasome. http://diseasome.eu/. Accessed: December 2012.

The Gene Ontology. http://www.geneontology.org/. Accessed: December 2012.

The WHO International Classification of Diseases. http://www.who.int/classifications/icd/en/. Accessed: December 2012.

The joint chemical dictionary (JOCHEM). http://www.biosemantics.org/index.php?page=Jochem. Accessed: December 2012.

Linked Life Data. http://linkedlifedata.com/. Accessed: December 2012.

Medical Subject headings. http://www.nlm.nih.gov/mesh/. Accessed: December 2012.

PubMed NCBI. http://www.ncbi.nlm.nih.gov/pubmed/. Accessed: December 2012.

NCI Thesaurus. http://ncit.nci.nih.gov/. Accessed: December 2012.

OWL Web Ontology Language Guide. http://www.w3.org/TR/owl-guide/. Accessed: December 2012.

SNOMED CT. http://www.ihtsdo.org/snomed-ct/. Accessed: December 2012.

Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/. Accessed: December 2012.

UniProt. http://www.uniprot.org/. Accessed: December 2012.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

S. J. Athenikos and H. Han. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, pages 1–24, 2010.

F. Baader, I. Horrocks, and U. Sattler. Description logics. In *Handbook on Ontologies*, pages 3–28. 2004.

O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, 2006.

M.-M. Bouamrane, A. L. Rector, and M. Hurrell. Using owl ontologies for adaptive patient information modelling and preoperative clinical decision support. *Knowl. Inf. Syst.*, 29(2):405–418, 2011.

F. Brucker, F. Benites, and E. Sapozhnikova. An empirical comparison of flat and hierarchical performance measures for multi-label classification with hierarchy extraction. In *Proceedings of the 15th International Conference on Knowledge-based and Intelligent Information and Engineering systems*, pages 579–589, 2011.

B. Cairns, R. Nielsen, J. Masanz, J. Martin, M. Palmer, W. Ward, and G. Savova. The MiPACQ Clinical Question Answering System. In *Proceedings of the AMIA Annnual Symposium*, Washington, DC, 2011.

Y. Cao, F. Liu, P. Simpson, L. D. Antieau, A. Bennett, J. J. Cimino, J. W. Ely, and H. Yu. AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44(2):277–288, 2011.

A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, pages 783–786, 2005.

G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. Summarization System Evaluation Revisited: N-gram Graphs. *ACM Transactions on Speech and Language Processing*, pages 5:1–5:40, 2009.

N. Guarino. Formal ontology in information systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*, pages 3–15, 1998.

S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(Database-Issue):919–922, 2008.

U. Hahn and S. Schulz. Towards a broad-coverage biomedical ontology based on description logics. In *Pacific Symposium on Biocomputing*, pages 577–588, 2003.

K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. van Mulligen, J. Kleinjans, and J. A. Kors. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22):2983–2991, 2009.

R. Hoehndorf, M. Dumontier, and G. V. Gkoutos. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*, 28(16):2169–2175, 2012.

S. Jupp, R. Stevens, and R. Hoehndorf. Logical Gene Ontology Annotations (GOAL): exploring gene ontology annotations with OWL. *J Biomed Semantics*, 3 Suppl 1:S3, 2012.

M. Lee, J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. Ely, and H. Yu. Beyond information retrieval-medical question answering. In *Proceedings of the AMIA Annual Symposium*, pages 469–73, 2006.

P. D. Leenheer and T. Mens. Ontology evolution. In *Ontology Management*, pages 131–176. 2008.

C. W. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop "Text Summarization Branches Out"*, 2004.

A. L. Rector and J. Rogers. Ontological and practical issues in using a description logic to represent medical concept systems: Experience from galen. In *Reasoning Web*, pages 197–231, 2006.

C. N. Silla and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining Knowledge Discovery*, pages 31–72, 2011.

B. Smith and M. Brochhausen. Putting biomedical ontologies to work. *Methods Inf Med.*, 49(2):135–140, 2010.

L. F. Soualmia, C. Golbreich, and S. J. Darmoni. Representing the MeSH in OWL: Towards a Semi-Automatic Migration. In *Proceedings of the KR Workshop on Formal Biomedical Knowledge Representation*, pages 81–87, 2004.

S. Tratz and E. Hovy. Summarization evaluation using transformed basic elements. In *Proceedings of the 1st Text Analysis Conference*, 2008.

G. Tsatsaronis, N. Macari, S. Torge, H. Dietze, and M. Schroeder. A maximum-entropy approach for accurate document annotation in the biomedical domain. *BMC Journal of Biomedical Semantics*, 3 Suppl 1:S2, 2012a.

G. Tsatsaronis, M. Schroeder, G. Paliouras, Y. Almirantis, I. Androutsopoulos, E. Gaussier, P. Gallinari, T. Artieres, M. Alvers, M. Zschunke, and A. Ngonga. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *Proceedings of the AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*, pages 92–98, 2012b.

Y. Tsuruoka, J. ichi Tsujii, and S. Ananiadou. FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, 24(21):2559–2560, 2008.

E. M. Voorhees. The TREC question answering track. *Natural Language Engineering*, pages 361–378, 2001.

P. L. Whetzel, H. E. Parkinson, H. C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S.-A. Sansone, C. J. Taylor, J. White, and C. J. S. Jr. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22 (7):866–873, 2006.

K. Wolstencroft, P. W. Lord, L. Tabernero, A. Brass, and R. Stevens. Protein classification using ontology classification. In *ISMB (Supplement of Bioinformatics)*, pages 530–538, 2006.