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

Project	**FP7-318652 / BioASQ**
Deliverable	**D2.11**
Distribution	**Public**

http://www.bioasq.org

Exploitation and Dissemination Plan

Authors: Michael Alvers, George Tsatsaronis, Matthias Zschunke, Axel-Cyrille Ngonga Ngomo, Aris Kosmopoulos and Christianna Armeniakou

Status: Final (Version 1.0)

October 2014

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	Ms Martina Eydner

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D2.11
Deliverable title	Exploitation and Dissemination Plan
Contractual date of delivery	M24 (September 2014)
Actual date of delivery	October 2014
Relevant Task(s)	WP2/Task 2.3
Partner Responsible	TI
Other contributors	ULEI, NCSR “D”
Number of pages	20
Author(s)	Authors: Michael Alvers, George Tsatsaronis, Matthias Zschunke, Axel-Cyrille Ngonga Ngomo, Aris Kosmopoulos and Christianna Armeniakou
Internal Reviewers	UJF, NCSR ’D’
Status & version	Final
Keywords	Exploitation Plan, Dissemination Plan

Executive Summary

This deliverable specifies the exploitation and dissemination plan of the BioASQ project results. The report focuses in the business as well as the scientific and academic exploitation. The planned commercial use of the project results by the business partner of the project (namely TI) and the planned scientific use of the results by all partners are explicated. In addition, the plan for the future management of the social network, the annotation tool and the evaluation platform is set in place.

As the planned exploitation activities aim to create value within all the participating organizations and thus to improve their competitive advantages in the market, the present deliverable also presents the cases where research results will be exploited for the internal development and support of new products and services. These products and services will lead to a competitive advantage of the participating organizations and will substantially contribute to the benefit of the targeted users.

In particular, the deliverable starts with a summary of the achieved project results and arrays the major achievements and success stories of the results so far in the industry, as well as in the academia. It also sets the basis for the exploitation plan, as it arrays the novelties of the achieved results. Then a discussion about the competition is presented, at the level of relative initiatives, such as related challenges, and the level of products and services which are already in the market and perform similar functionalities to the BioASQ components. The scientific exploitation of the results is analyzed as well as the different ways the consortium is planning to disseminate and exploit the results further in the academic world, such as further usage of the tools for new benchmarks, usage in other related projects, and public release of the BioASQ software components.

The focus is then shifted on the business exploitation of the BioASQ results. It envisages the business aspects of the produced results and discusses the further development needed that can transform those into competitive products in the market. In this accord, specific policies are suggested for consumption of the developed services both from potential academic clients, but also from the industry. In addition the potential of exploiting the research results and creating novel technologies and products is analyzed in detail.

Contents

List of Figures

List of Tables

1 Introduction

The BIOASQ exploitation and dissemination targets of the academic partners are mainly **excellence building**, **knowledge transfer**, **education** and later **research** in BIOASQ-related areas. These targets are built on the BIOASQ results. Direct results of the project fall into two main categories: software and services, including platforms and tools, and know-how and technologies. In addition to the direct results, there are indirect results of the project, which constitute success stories, and can generate further ideas and exploitable outcomes. In the remaining of this chapter, we discuss in brief both direct and indirect results of the project, and give a brief introduction of their potential impact.

1.1 Direct results of the BIOASQ project

The two-year BIOASQ project has managed to put together in a successful manner all the resources that were required to bring into life a unique challenge on biomedical question answering. The challenge engulfed heterogeneous areas of expertise, related to the life sciences as well as to computer science. From the computer science perspective, the organised challenge addressed researchers and groups that were related to one or more of the following areas: classification, document retrieval, machine learning, information extraction, information retrieval, natural language generation, named entity recognition, named entity disambiguation, passage retrieval, QA from structured and unstructured information, relation extraction, reasoning, textual entailment, text summarization, and, last but not least, semantic indexing. From the life sciences perspective the BIOASQ challenge utilized top experts, which were responsible for the timely preparation of the benchmark questions and answers, in the following areas: cardiovascular endocrinology, psychiatry, psychophysiology, pharmacology, drug repositioning, cardiac remodelling, cardiovascular pharmacology, computational genomics, pharmacogenomics, comparative genomics, molecular evolution, proteomics, mass-spectometry, and protein evolution.

Under this mosaic of disciplines, the BIOASQ consortium achieved much more than the mere organisation of a biomedical QA challenge: (i) it produced efficient software, tools and services that are re-usable, exploitable and are opening novel business opportunities in the after-project era, and, (ii) it created a technical "precedent" on how systems can address efficiently biomedical QA, and gave to the community the benchmarks for future evaluation, and the tools and pipelines to continue producing more benchmark sets at a low cost for future challenges.

1.1.1 Software, services, platforms and tools

The main effort of the BIOASQ project was put into the creation of efficient software, tools and services to support the organization and execution of the challenge. The annotation and assessment tools were produced for the creation of the benchmark questions and answers by the team of the biomedical experts. They offer the ability to search efficiently in a large spectrum of biomedical resources, including all of the indexed *PubMed* articles, open access articles, as well as a plethora of knowledge bases that provide information for drugs, diseases and targets, i.e., *MeSH*, *GO*, *Uniprot*, *Jochem*, *Disease Ontology*, and the 6 billion of triples produced by the *Linked Life Data* project. The tools offer in addition the ability to its users to synthesize questions and answers, analyze and mark relevant retrieved results, and highlight snippets of the articles that help formulate and answers to the questions. The annotation and assessment tools are supported by *JSON*-based services which retrieve the results on given queries from the underlying resources, and the social network, which enables discussions among the experts, and validation of the the answers. The BIOASQ Participants Area (i.e the Platform) is another tool that was developed during the project. It provides mechanisms for the exchange of data between the challenge participants and the organizers. In addition, the participants to find information and support regarding the challenge. The BIOASQ team can administrate the challenge, release the benchmark datasets and provide the necessary mechanisms that will allow the evaluation of the participating systems via the platform.

The developed services within the BIOASQ project constitute a great technical capital of the consortium. Besides the services that are responsible to search for concepts, documents and triples, there are also services that enable participants to submit results, retrieve data sets, and evaluate automatically participants' answers for the challenge tasks, employing a wide plethora of evaluation measures. These latter services are also part of the evaluation platform that was implemented within the project, which is responsible to deliver solutions for quantifying the quality of system responses within the challenge.

1.1.2 Indexed Resources and Benchmark datasets

The *BioASQ* 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*.

Besides the indexed resources, the produced benchmark datasets are also of great importance to the community. In the case of Task a, i.e., Tasks 1a and 2a, the BIOASQ consortium, in collaboration with *NLM*, has implemented a novel way to evaluate semantic indexing of biomedical articles using the *MeSH* hierarchy, based on the process followed by *NLM*. On this basis, the benchmark sets creation for Task a is fully automated. In the case of the Task b, i.e., Tasks 1b and 2b, the corpus created comprises questions, answers and evidences (concepts, triples, snippets and related *PubMed* documents) for 800 questions

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

with the accompanying gold standard answers. It is for the first time, to the best of our knowledge, that such a high-quality question answering dataset has been produced for the specific domain, at such a scale.

1.1.3 Technologies and know-how: Progress beyond the State-Of-the-Art

The needs of the BIOASQ challenge organisation and execution, as well as the competition between the participants within the challenge, in the light of the prizes that were promised to the winners, created an ideal setup for progressing the SoA in two directions: (i) novel solutions for semantic indexing and biomedical QA that were developed by the participants, and, (ii) novel evaluation measures and respective implementations for both BIOASQ tasks.

With regards to (i), the satisfactory number of participants in both iterations of the challenge enabled high-level competition. In challenge 1 there were 46 registered systems from 11 participating teams in Task 1a, and 7 systems from 4 teams in Task 1b. In challenge 2 a total of 18 teams participated in Task 2a using 61 registered systems, and 8 teams participated in Task 2b using 15 systems. Besides the increased participation between the first and the second iteration of the challenge, the quality of the techniques improved also significantly. By the end of the second iteration, the participating systems may offer semantic indexing solutions with a Micro-F1 score of slightly more than 60%, while the *NLM* indexer performs slightly lower, at the levels of $55 - 59\%$. In parallel, systems may answer questions with an accuracy that is higher than 90% for the Yes/No questions, and list and factoid questions with an accuracy that has high variance, but which can reach at the levels of 50%. In parallel, for the evaluation of the semantic indexing, novel techniques were developed Kosmopoulos et al. (2014), which take into account the structure of the underlying hierarchies. In the case of the systems evaluation, novel measures were also developed to address the problem of evaluating the summary questions.

1.2 Indirect results of the BIOASQ project

Besides the measurable results of the project, there are also some other aspects that are important and exploitable, and occurred indirectly via the challenge. First, the BIOASQ challenge created a new community with increased awareness and interest on QA and semantic indexing techniques, which, judging from the measurable results of the project, has big dynamics. Second, several of the project technologies developed from the participants attracted increased interest from big stakeholders, such as *NLM*.

1.2.1 Community

In contrast to other related initiatives, such as the *BioCreative* and the *BioNLP*, BIOASQ did not start from a community, but rather created one through its project life-cycle. This is probably the most intriguing characteristic of the project; as timely as ever, it brought into the surface the major problems that modern systems face towards automated QA in the domain, created the infrastructure to address them, and became the source of a community that aspired the concept. Nowadays, the BIOASQ project has hundreds of followers in popular social network media, and an increasing number of researchers and groups that are eager for the next series of the challenge. These dynamics constitute a solid basis not only for the sustainability of the effort, but also for the potential impact and the breakthrough business opportunities that can emerge.

1.2.2 Success Stories

Many of the developed techniques did surprisingly well in the two tasks. The winners of Task 1a, performed very well, and their techniques did not go unnoticed. *NLM* has by now adopted in their automated annotation pipeline ideas from the winning approach Mork et al. (2014). In addition, the whole effort raised the awareness of one of the biggest stakeholders in the field, which is the team behind *IBM* Watson. In the first BIOASQ workshop Dr. Jennifer Chu-Caroll of the IBM Watson team, in her invited talk, highlighted the intriguing characteristics of the QA in the domain, and the importance of the developed techniques in such a large-scale setup.

1.3 Potential impact of BIOASQ technologies

Based on the success stories of the project, there are many interesting aspects of the developed technologies that carry great potential impact for the future business opportunities. The top rated systems which were able to improve substantially over the *MTI* baseline follow different approaches. The first ranked systems followed a hybrid approach mixing an information retrieval phase and a learning-to-rank procedure. The second best rated systems presented followed a pure machine learning approach employing flat classification schemes using SVMs and combining several systems with ensemble methods. Also the hierarchical approaches that presented in the competition achieved good results having low complexity due to the use of the hierarchical structure.

While the former approaches are able to provide better results the latter enjoy faster training and inference times (very crucial for on-line search engines like *GoPubMed*). So, potentially both technologies could be used in order to boost the prediction capabilities of a search engine where the first can be employed in an off-line scenario for improving the annotations of the articles in the database. The technologies of the learning-to-rank systems can be integrated in the front-end of the search engine in order to provide accurate and fast results to the users. In addition, the approaches developed and submitted in the framework of Task b, may be used as a basis to develop QA expansions of *GoPubMed*. Based on this observation, *GoPubMed* could be among the first search engines to launch a fully fledged QA for the biomedical domain in the search engine market. More details on the potential impact and business opportunities of the proposed approaches in the challenges will be presented in Chapter 4.

2

Competition

2.1 Research and development aiming at the search market

The impact of having an intelligent *"avatar"* able to answer medical questions is definitely immense. A comparison between death rates is shown in Figure 2.1[1]. The figure shows the tremendous needs for better information providing systems: death through aviation (329), drowning (3, 959), falling (14, 986), traffic accidents (43, 649) and, finally, medical errors (120, 000).

It is hard to estimate how many deaths through medical errors could have been avoided by better information at the right time at the right place, but it is fair to assume that the share is rather big. If the numbers from *U.S.A.* are scaled up to the world, we end up to an 18-fold increase, not taking into account the fact that underdeveloped countries have a much less sophisticated infrastructure than the *U.S.A*. So an estimated number of around 3 million deaths per year should be a very good motivation to invest in research in the area of medical information systems especially in automated question-answering systems. These facts alone constitute a very strong basis to invest and build upon research and development of novel semantic-enabled technologies and question answering systems in the biomedical domain.

Such an effort requires continuous analysis of transfer opportunities in order to ensure the best possible outcome. Furthermore, detailed investigation into the possible economic benefits and impact of the expected research results needs to be conducted, along with continuous evaluation of research results against the user requirements/needs throughout the project with the help of the partners and adjustment of the project, when necessary.

Despite the fact that all consortium partners have been developing their own exploitation plan throughout the project, this deliverable will enable the development of a first business case. In the following we analyze one such direction, namely the enhancement of the *GoPubMed* semantic search engine of TI with question answering features.

In such a direction, semantic search plays a crucial role. Semantic search may be essentially supported by two categories of approaches: (1) searching structured documents and reasoning over them and, (2) searching unstructured documents and, possibly, attempting to extract knowledge and reason over it. The knowledge extraction step of the latter uses combinations of natural language processing,

[1] U.S.A. 1999 - more recent data are not available. Sources: (1) Philadelphia Enquirer (9/12/99), (2) The Institute of Medicine 1999 report, (3) "To err is human", Richardson et al. (Richardson, 2006).

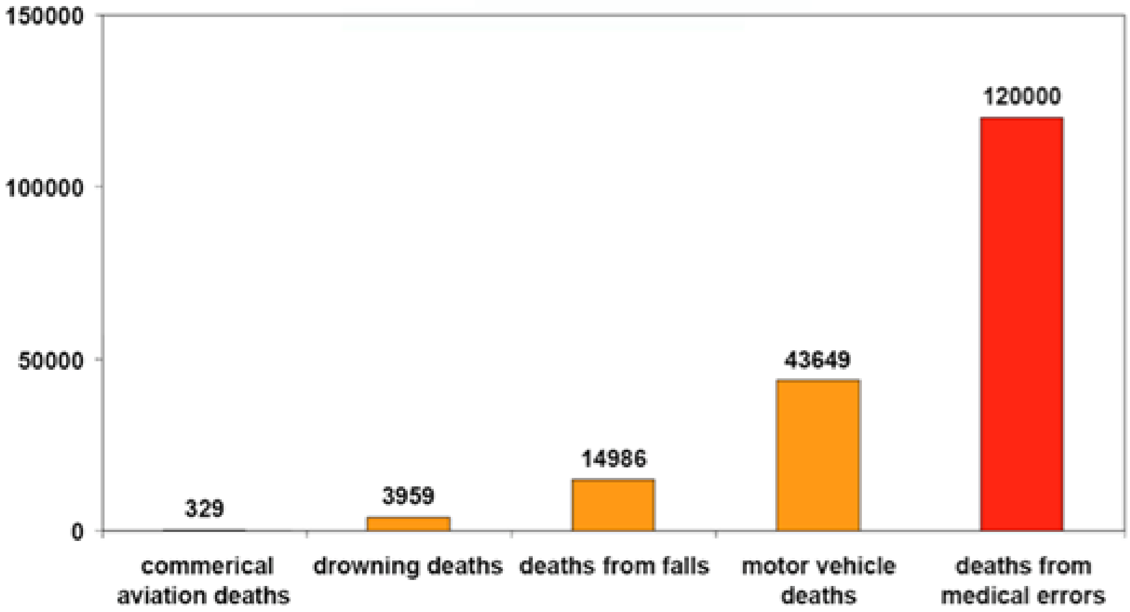

Figure 2.1: Comparison between death rates based on the type of cause of death. Data from *U.S.A.*, 1999.

information retrieval, text-mining, and ontologies. In the following, we give a short overview of representative engines of both categories. We distinguish between engines or tools that perform searching and ranking of structured knowledge, which are approaches of category (1), and other approaches that deal with unstructured text, which constitute category (2).

2.2 Searching and ranking structured knowledge

Searching and Ranking Formal *RDF/OWL* Statements To facilitate machine-readability and knowledge processing, a set of standards, query languages, and the semantic stack were proposed by the *W3C*. The stack comprises at the base unique identifiers and *XML* as common markup language. On top of *XML*, it defines the *Resource Description Framework* (*RDF*) to capture subject-predicate-object triples. Furthermore, there is the modelling language *RDFS* and the query language *SPARQL*. The basic class definitions and triples of *RDF* are extended at the next level by the Web ontology language *OWL*, which provides description logic as modelling language and by a rule layer.

Besides the expressiveness of *OWL*, mark up for vocabularies and meta-data emerged such as *Simple Knowledge Organisation Systems* (*SKOS*), *Dublin Core1*, *Friend of a Friend* (*FOAF*) and the *Semantically-Interlinked Online Communities Project* (*SIOC*). Additionally, there are formats to embed semantic annotations within Web documents, such as embedded *RDF* (*eRDF*), *Microformats2* or *RDFa*. All of the above standards serve the need to formally represent knowledge and facilitate reasoning over it. They require explicit statements of knowledge. As a consequence, the amount of such structured data is still small in comparison to the unstructured data, but still, there are many research works that attempt to search and rank knowledge following some of the aforementioned standards.

Known state-of-the-art search engines in this category of approaches are: *Swoogle* (Ding et al., 2004), *Semantic Web Search Engine* (*SWSE*), *WikiDB*, *Sindice* (Tummarello et al., 2007), *Watson* (dAquin et al., 2007), *Falcons* (Cheng et al., 2008), and *CORESE* (Dieng-Kuntz and Corby, 2005). They include existing *RDF* repositories and crawl the internet for formal statements, e.g., *OWL* files. A search retrieves a list of results with *URIs*. For *SWSE* and *Falcon* the result is enriched with a description and a filtering mechanism for result types. *CORESE* uses conceptual graphs for matching a query to its databases. *WikiDB* is slightly different from the others in that it extracts formal knowledge implicit in meta tags of *Wikipedia* pages and converts it into *RDF* offering querying with *SPARQL*. As mentioned, the above systems are limited by the availability of structured documents, a problem addressed by approaches such as the *Semantic Media Wiki* (Völkel et al., 2006) and large efforts such as *Freebase* (Bollacker et al., 2008), which provides an environment for authoring formal statements.

2.3 Searching and ranking unstructured text

Keyword-based Search with Synonym Expansion Traditional search engines like *Google*, *Yahoo!* and *Bing* have the largest coverage but they miss the explicit usage of ontological background knowledge. They only present a long list of results. This works very well for simple retrieval of documents, but is limited for complex tasks, e.g., answering questions, or attaining a view of a knowledge field that was previously unknown. The greatest advantages of those engines is simplicity, wide coverage and wide adoption. They do not offer text annotations with ontological background knowledge, but they expand terms with their synonyms, which increases their recall levels.

Natural Language Processing of Queries and Text In this category of engines the aim is to process query and text using natural language processing methods (*NLP*). Known such engines are *START* (Katz et al., 2006), *Hakia*[2] and *Answer Bus* (Zheng, 2002). They all use techniques such as stemming, concept identification, and deep/shallow parsing to understand documents. The main disadvantage of the used methods is the computationally-intensive language specific *NLP* techniques. In addition, natural language may be very complex, and the state-of-the-art frontier of the research in natural language processing still struggles to address difficulties in automated natural language understanding.

Keyword-based Search by Performing Clustering In an effort to organize the results thematically and semantically, but in an unsupervised manner, the engines *Yippy*[3] (previously known as *Clusty*, and developed by *Vivisimo*), and *Carrot*[4] cluster search results and label them with phrases, which are offered as related queries. *Yippy* and *Carrot* are not semantic search engines in a strict sense, since these phrases are not part of an ontology or vocabulary. However, they do have the benefit of being generally applicable, since the labels, i.e., the thematic categories, are extracted on-the-fly based on the results, and there is no need for a pre-existing conceptualization of the results' domain. The major disadvantage is the fact that clustering is a hard problem and, frequently, demands the setting of a number of non-obvious parameters (thresholds, number of clusters, etc.). Estimating these parameters automatically and efficiently is still an open problem in the area of data mining.

Ontology-based Search by Performing Advanced Text Mining Within this category of approaches, we find engines that use background knowledge in the form of a domain ontology. Examples of such engines are *GoPubMed* (Doms and Schroeder, 2005), *GoWeb* (Dietze and Schroeder, 2009)[5], *EBIMed*

[2] http://www.hakia.com/

[3] http://clusty.com

[4] http://carrotsearch.com/

[5] *GoPubMed* and *GoWeb* are both developed by Transinsight.

(Rebholz-Schuhmann et al., 2007) and *XplorMed* (Perez-Iratxeta et al., 2003). According to previously published studies (Doms and Schroeder, 2005; Dietze and Schroeder, 2009), engines of this category are more successful in retrieving information within their domains, compared to any other type of search engine. *GoPubMed* and *EBIMed* use the *GeneOntology* and the *Medical Subject Headings* (*MeSH*). *XplorMed* filters by eight *MeSH* categories and extracts topic keyword co-occurrences. *GoWeb* issues queries to *Yahoo!* and indexes the snippets semantically with ontology terms. These are then offered to filter results by concepts.

***Wikipedia*-based Annotation Systems** An alternative to the underlying ontology used by ontology-based search engines, is the use of *Wikipedia* categories, and the ability to annotate Web documents with these concepts. In this direction, there is a lot of research that suggests a methodology to annotate unstructured text with *Wikipedia* information. A representative example of such an approach can be found in (Mihalcea and Csomai, 2007). *Wikify!* enriches any input text with links to the *Wikipedia* encyclopedic knowledge. The approach is based on keyword extraction from the input text, and the mapping of the keywords (linking) to the respective *Wikipedia* articles. In a very similar direction, Paci et al. (Paci et al., 2010) link the extracted keywords of input text to *Wikipedia* articles, in order to utilize these links for mapping the texts to ontology concepts using *Wikipedia*-based measures of semantic relatedness. In a slightly different direction, in (Pipitone and Pirrone, 2010) the authors annotate *Wikipedia* articles, to turn them into semantic *Wiki* articles. Though approaches like the aforementioned are efficient, it is difficult to utilize them for an on-line engine, as tasks like *keyword extraction* and *part of speech tagging* are impossible to apply in real time.

3

Scientific Exploitation and Dissemination

The BIOASQ exploitation and dissemination targets of the academic partners were mainly **excellence building**, **knowledge transfer**, **education** as well as **research** in BIOASQ-related areas. In the following, we present how we achieved these goals within the project and aim to build upon them in the future.

3.1 Publications

In the course of the BIOASQ project, the academic partners have developed innovative solutions to address the requirements that resulted from the project. These solutions resulted in publications that were presented at top conferences (see e.g. (Tsatsaronis et al., 2012b; Ngonga Ngomo et al., 2013; Babbar et al., 2013b,a; Ngonga Ngomo et al., 2014)) and journals (see e.g., (Babbar et al., 2014)).[1] Moreover, we published proceedings of all BIOASQ workshops and made them freely available at http://ceur-ws.org. The BIOASQ publications clearly demonstrate the advancements that were achieved through the project. These advancements had a world-wide impact, as all PudMed users benefit from better annotations of their documents.[2] In addition, a comparison of the results of the BIOASQ challenges suggest that we are able to push the accuracy of existing bio-medical question answering systems by ca. 5%.

Concerning future publications, different articles are already planned and some of them are already under preparation. In particular, a common journal article providing an overview of both BIOASQ and Visceral is under preparation. The article will focus on the impact of system competitions to the biomedical community. It will summarize the two competitions, the resources created (i.e. tools, data, evaluation measures etc.), as well as future directions of such competitions.

A journal article providing an overview of Task b is also planned. This article will focus exclusively on Task b, providing more details about it, and it will cover both years (Tasks 1b and 2b). The issues to be covered include: a detailed discussion of the subtasks, benchmark data, and evaluation measures of Tasks 1b and 2b; inter-annotator agreement studies; evaluation results and technology overview of

[1] A complete list can be found at http://bioasq.org/project/relevant_bibliography.

[2] See http://www.nlm.nih.gov/news/indexer_challenge.html.

the participating systems and baselines; evaluation results of ensembles that combine the outputs of participating systems; correlations of automatic evaluation measures with manual evaluation scores for 'ideal' answers; highlights and recommendations of the interviews with the biomedical experts that authored questions and evaluated system responses.

In an attempt to address complaints of the biomedical experts that the 'statements' (pseudo-English renderings of RDF triples) were difficult to make sense of, AUEB-RC experimented with generating texts with its NaturalOWL system [http://www.jair.org/papers/paper4017.html] from Disease Ontology, one of the ontologies used in BioASQ. NaturalOWL generates texts (descriptions of entities and classes) from OWL ontologies. AUEB-RC semi-automatically created an OWL version of Disease Ontology and conducted experiments with biomedical experts indicating that NaturalOWL can produce higher quality texts from the new version of the ontology, compared to the pseudo-English descriptions of diseases that are included in the original (OBO) form of the ontology. The results of these experiments are to be included in future publications of AUEB-RC, along with experiments with other biomedical ontologies. The OWL version of the Disease Ontology will be made publicly available on the BioASQ site, along with automatically generated high-quality English sentences for each OWL statement of the ontology, and automatically generated high-quality texts for each disease of the ontology. NaturalOWL is already publicly available [http://nlp.cs.aueb.gr/software.html].

AUEB-RC is already co-operating with the Institute for Language and Speech Processing, Research Centre "Athena" (http://www.ilsp.gr/) to develop a state of the art system for Task b, building upon and extending the query-focused summarizer that was used as the baseline for 'ideal' answers in Tasks 1b and 2b, and exploiting the BioASQ benchmarks. Work on this system is already acting as a strong link between the two research groups, and is expected to lead to a system that will participate in BioASQ3, research publications, and future research grants.

These results will build the basis for future collaborations and acquisitions by the research groups involved in BioASQ. In addition to scientific publications, we also aim to continue releasing informal publications, including blog posts, tweets, *Facebook* and *LinkedIn* entries, through the project's web dissemination channels. We have released teaching material, which has been used during summer schools such as *IASLOD*[3] and the Reasoning Web Summer School,[4] as well as presentations and videos on platforms such as *SlideShare*,[5] and *Youtube*[6]. We will aim to continue using these channels to disseminate the results of future BioASQ challenges.

3.2 Events

BioASQ (co-) organized a number of events over the last two years, of which the most important were the BioASQ workshops. In particular, the 2013 BioASQ workshop was held at the Technical University of Valencia (Universitat Politcnica de Valncia), Spain, as a post-conference workshop after CLEF 2013. The 2014 BioASQ Workshop was part of the CLEF 2014 conference and was held at the University of Sheffield, UK. These workshops were able to attract both participating researchers and non-participating yet interested researchers and companies. The growth in the number of participants by approx. 30% suggest a growing interest in the subject matter of the project and challenges. Thus, we aim to continue BioASQ events even after the end of the project and have already applied for a BioASQ 2015 event with the setting of the CLEF conference. The workshops were made public via the BioASQ leaflet and website and communicated throughout a variety of channels such as mailing lists,

[3] http://semanticweb.kaist.ac.kr/2012lodsummer/

[4] http://rw2014.di.uoa.gr/

[5] http://www.slideshare.net

[6] https://www.youtube.com/channel/UCLG0adw5SLQCcQIff5DQ8Ig

blogs, tweets, etc. During the workshops, the project results (benchmark creation, tools, social network, etc.) were advertised. We will continue to use the same channels to push BIOASQ further in the future.

3.3 Collaborations

During the two years of the project, the relations among the partners of the project and the members of the advisory board were strengthened and led to novel proposals and common endeavors. In addition, the organization of the challenges and the workshops as well as the participation to the events mentioned in the previous section gave the consortium the opportunity to meet researchers and companies of the domain, both participating and non-participating to the challenges, and to lay the foundations for possible future collaborations. Examples of such collaborations which were started during the project the *Memorandum of Understanding* which has been mutually signed between BIOASQ and the *VISCERAL* project, the QALD-4 benchmark supported by the project PortDial 2[7] project, the interaction with *NLM* for the creation of baselines for Task 1a and Task2a and the collaboration with the NLP Lab at UNED[8] with the QA Track of CLEF 2014. These collaborations will be continued in the future within QALD-5, the QA Track at CLEF 2015 and project proposals, especially within the upcoming Horizon 2020 rounds. One particular collaboration that turned out to be very beneficial for both parties was that of BIOASQ with the US National Library of Medicine (NLM). NLM has created the Medical Text Indexer (MTI) to help MEDLINE curators in associating Mesh Terms with MEDLINE abstracts. MTI has been used in BIOASQ as a baseline system, that the challenge participants tried to outperform. We observed a significant improvement of the baseline system in the challenge, while NLM benefited from the ideas used in the participating systems to improve MTI (see the corresponding announcement of NLM: http://www.nlm.nih.gov/news/indexer_challenge.html).

3.4 Software Releases

Throughout the project, we released the results of our scientific endeavors as open software. In addition to the software releases that constituted formal project deliverables, the BIOASQ consortium has releases underlying technologies such as *SPARQL2NL*[9] as well as corresponding extensions of the *TBSL Question Answering Engine*[10]. The annotation and assessment tools as well as the BIOASQ social network were released at the project's GitHub page[11]. We release two versions of the first tool as well as continuous updates of the social networks after real tests with real users. Thus, the frameworks created by the consortium are now mature, user-friendly frameworks that can be used for upcoming competitions without any major changes. The data compiled by the consortium is now released on the social network (see http://sn.bioasq.org) and can be updated, reviewed and corrected by domain experts. The BIOASQ Platform is another software that was developed during BIOASQ project. It provides mechanisms for the participants to find information and support regarding the challenge as well as to participate in the tasks.In addition, the BIOASQ organizers can administrate the challenge, release the benchmark datasets and provide the necessary mechanisms that will allow the evaluation of the participating systems via the platform. The BIOASQ Platform includes also the BIOASQ oracle. The latter provides its users the opportunity to test their systems using past BIOASQ datasets. Both tasks of the BIOASQ challenge are available in the oracle. In this way, participants can continue improving their

[7] http://sites.google.com/site/portdial2/
[8] http://nlp.uned.es/
[9] http://github.org/AKSW/SPARQL2NL, http://sparql2nl.aksw.org/demo
[10] http://autosparql-tbsl.dl-learner.org/
[11] https://github.com/BioASQ

systems by using the datasets and the infrastructure generated during the BIOASQ challenge to evaluate their progress in an off-challenge mode. The BIOASQ oracle will remain active after the end of the project.

Finally, the evaluation tools that have been available during the challenges are now available under the Apache open-source licence on the GitHub site of the project[12]. More specifically, the evaluation software is composed by a package that implements the hierarchical classification measures used in Tasks 1a and 2a (Kosmopoulos et al., 2014) and a package for the measures used in both tasks of the challenge. This package also provides functionalities for vectorizing text collections into LibSVM format.

3.5 Datasets

The corpus created in the BIOASQ project of questions, answers and evidences (concepts, triples, snippets and related *PubMed* documents) contains 800 questions with the accompanying *gold standard* answers. The BIOASQ data focuses specifically in the life sciences, which constitutes the advantage the BIOASQ consortium has as it produces for the first time such a high-quality question answering dataset for the specific domain. The evaluation platform implemented during the project will remain alive and is already available to the research community in order to serve as a framework for experimental evaluation for large-scale information retrieval systems[13]. The platform provides an easy way for the evaluation of such systems and thus it will continue to help researchers to assess their systems under the BIOASQ benchmarks. Additionally, the platform will continue to provide a web service to the users for the creation of new evaluation tests for both tasks *a* and *b*. We also maintain the social network alive, through which it can now extend the datasets as well as create new datasets. These datasets will be useful resources to the research community as they will test the capabilities of state-of-the-art systems to handle large masses of data. Moreover, they will be used as basis for future projects aiming to develop question answering systems.

3.6 Licensing

With regards to the licensing scheme for the release of the BIOASQ components (software) and datasets, the BIOASQ consortium is clearly committed to the open-source approach for the project's results. The partners have agreed to follow a common licensing approach for the components and datasets developed within the project's lifetime, given their common strategies as well as their willingness to boost the exploitation potential of the BIOASQ project either as a whole or in individual components. The developed software tools are available at `http://github.com/bioasq`. Anyone who is developing and distributing open-source applications under the *GPL* is free to use the BIOASQ components under a *GPL* and/or a *GPL*-compatible license, with the only exception being the services developed by TI, for which special license must be obtained, as explained in the next chapter. Through its copyleft feature, our licensing model ensures that four important principles will be met: the freedom to use the software and datasets for any purpose, the freedom to change the software to suit specific needs, the freedom to share the software and datasets, and the freedom to share the changes implemented. In cases that any interested party does not want to either combine or distribute any of the BIOASQ software or datasets with their own software under the *GPLv3*, and hence among others do not want to openly release the source code of their proprietary solution, they should contact the BIOASQ organization that developed

[12] https://github.com/BioASQ

[13] See `http://github.com/bioasq` for the code. The platform is available at `http://bioasq.lip6.fr`

the tool or the BIOASQ coordinator for obtaining a different license, which will include the assurances which distributors typically find in commercial distribution agreements.

4

Commercial Exploitation, Dissemination, and Roadmap for Future Business Development

BIOASQ aims at the advancement of medical and biological question-answering systems. The unique opportunity is to foster the developments towards the ultimate goal of having a system which gives the right answer (according to the current state of research in the field) to a given question. It can be foreseen that in the next 5-10 years many groups worldwide will dive into this research topic which has a strong practical flavor, e.g., it may help to arrive at better treatment plans especially for rare diseases. Towards the exploitation of the BIOASQ results in this direction, we foresee the transfer of the project results into development, product, and service organizations of the partners. In the following we analyze the directions of the commercial exploitation.

4.1 Next generation search services

The first step for commercialising the *BioASQ* results, would be to identify, and conduct an initial assessment on the results and components that have the better commercialisation potential. From this respect, the search services attract the main focus and will gather the support of the efforts to coach the related processes that can lead to commercialisation of the respective results.

In the framework of *BioASQ* a valuable set of services has been developed which support the challenge. High throughput techniques have been developed for indexing large amounts of text and large-scale knowledge bases, and searching efficiently documents, concepts, and triples given keyword queries. In addition, novel services which are able to annotate efficiently unstructured text with ontological concepts have been developed.

This set is the basis of the next generation search engines and services, based on the concept of semantic-enabled technologies. Key applications of these services can be efficient search in the biomedical domain, and also search in unstructured text that has not been traditionally included in such engines, such as patents and web pages. At the present state, the services perform their tasks individually, but when combined, they can potentially be integrated into a single product, i.e., an advanced semantic search engine for the biomedical domain, which can also analyze patents and web text. A key advantage is the versatile nature of the implemented services, which allow an easy expansion of the underlying resources, e.g., adding new domain ontologies and more document sources. Challenges in this direction, such as the mapping of the underlying resources, will constitute a great obstacle, but not one that cannot be overcome. Another key advantage is the novelty of the solutions provided by the challenge participants to solve tasks a and b. A prominent example of a success story that stems as a result of the novelty of the participating systems, is the fact that the method created by the first year's challenge task a winner, was adopted by the NLM in order to improve the automated suggestion of MeSH headings for PubMed articles to the professional indexers Mork et al. (2014).

Taking the concept of the search in the biomedical domain at the next level, the *BioASQ* consortium also envisages the creation of novel methods that are based in these services, such as relation extraction, with the aim to support automated hypothesis generation and validation. For this latter aspect, the existing technologies will have to be enhanced with entity disambiguation techniques, and relation extraction and/or learning techniques. The *BioASQ* team, comprising people with high expertise in these areas, could take this additional step and create novel methods that can assist biomedical hypothesis creation and validation.

4.2 Putting everything together towards an integrated QA engine for the biomedical domain

An engine that can answer with high accuracy questions in the biomedical domain, would constitute a tool of high value and importance, not only from the research perspective, but also in the industry. The value of such an engine is known to many a stakeholders; QA engines such as IBM Watson, have started considering such directions[1]. The *BioASQ* consortium has a strategic advantage in this market. The experience of the participating systems, and the associated technologies, once in a state that the IPR issues are clarified, could constitute a mature basis for implementing such an integrated engine, with the aim to automatically answer biomedical questions.

In fact, the infrastructure is already in place, e.g., services and indexes. What is missing is a strategic view of the overall product, that could give it a focus, tight enough to ensure high accuracy, but in tandem wide enough to attract attention and demand. Task b is principled by the ”drug-target-disease” triangle to allow a variety of hypotheses being formulated as natural language questions. A QA engine in the biomedical domain could not be efficient enough at this stage and with these techniques to address all domain questions, but the results of task b can show us the way to identify the domain subjects and respective question patterns in which the existing technologies can succeed with high accuracy.

4.3 A concrete exploitation plan: BIOASQ technologies into *GoPubMed*

Given the success of *GoWeb* in benchmark evaluations (Dietze and Schroeder, 2009) for answering questions in the biomedical domain, as well as that of *GoPonte* (Tsatsaronis et al., 2012a), we plan to utilize the same principles and expand *GoPubMed* by modules that annotate in real time Web documents with *Wikipedia* categories and *UMLS* concepts. A great advantage of these methods is that they may utilize their background knowledge to annotate unstructured text with existing domain knowledge. Though a fast and efficient annotation technique is needed to perform the task in real time, the problem can also be seen as a text classification one, where the categories are the ontology concepts, and the instances are the fetched Web documents. In this direction, there are many solutions for the annotation process, that may also address the ambiguity of the terms in the unstructured text, despite the technologies developed

[1] http://bioasq.org/sites/default/files/workshop1/bioasq_chu-caroll.pdf

within the two iterations of the BioASQ challenges. We plan to build on the experiences of the participating systems of Task a to optimize the annotation process, and on the experiences of the systems that participate in Task b in order to learn how to extract the most useful answers for a given question, aided by the performed annotations.

4.4 Exploitation of Developed Infrastructure

Though it is hard to make an exact estimate, there are indications that worldwide there are hundreds of research groups which work in the area of natural language question answering (Hirschman and Gaizauskas, 2001). The services within the BioASQ project, namely the services that return related concepts, related documents and snippets, and related triples given a query, are of great value for research groups or industry who wish to develop question answering systems (Tsatsaronis et al., 2012b). In this case, judging from respective experience of similar service usage for the development of semantic search technologies for clients of *Transinsight GmbH*, server access is estimated between $5,000$ and $10,000$ Euros per industry user per annum. These estimations are based on the pricing per query shown in the last row of Table 4.1, where each query is considered a service request (call) to any of the developed services[2]. Having $5-10$ users would generate between $12,500$ and $25,000$ Euros per annum from the exploitation of the services.

	All numbers in Euros	**Universities**	**Industry**
Services	**Server Access**	0.10 per query (first 1,000 free)	0.50 per query

Table 4.1: Suggested pricing of the BioASQ developed services. Differentiation between universities and industry is applied.

4.5 Advertisement

The dissemination of the idea will be done through several channels. The most promising is a banner on *GoPubMed*'s landing page. About $20,000$ visitors (page impressions) per day will have a good reach out to the relevant *"crowd"* in the biomedical domain. Other channels are the BioASQ web site, the web site of *Transinsight GmbH*, and the web sites of challenge participants. In addition, BioASQ will be advertised as the platform for biomedical question-answering in general on all trade fair participations like *CeBIT* and *DMG*, in addition to the presence in science fora, such as *CLEF* and other major conference.

4.6 Future Directions and Steps

As future directions, the *BioASQ* consortium will assess the commercial potential of these business opportunities, and evaluate each based on the innovation of the underlying technology, the market, the team that can lead the efforts, and the expected financial outcome.

[2] At this stage we do not distinguish the cost between different services, though in the future such differentiation might exist, e.g., different pricing for the triples service and different for the documents service

More precisely, the evaluation of the technology will take place at the level of examining the progress beyond the state of the art, and the value of the scientific publications and references associated with these technologies. In addition, products and/or services will be carefully designed, with a view on the customer's/client's benefit, and more specifically focusing on the added value for the consumers of the services, and the unique selling proposition (lower price, better performance, better service, advanced capabilities). Further dependencies will also be examined, e.g., legal restraints, interdependency with other software and intellectual property rights. Finally, for each of the aforementioned potentials, a proof of concept will be sought, as well as the technical feasibility and a proof of market.

Regarding the market evaluation, the target groups must be identified, as well as the characteristics of the intended market, e.g., size, market entry barriers (technological and legal), value and growth of the market, competitors. Besides these, the marketing strategy must be clarified, with regards to the overall sales and distribution concept, the distribution channels, and associated costs appropriate for planned turnover.

Finally, regarding the financial assessment, profit and loss statements have to be produced, with specific focus on the forecast of liquidity, and the investment needs have to be determined, e.g., total amount of investment, split, own contributions.

Bibliography

R. Babbar, I. Partalas, É. Gaussier, and M. Amini. On flat versus hierarchical classification in large-scale taxonomies. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013a.

R. Babbar, I. Partalas, É. Gaussier, and M. Amini. Maximum-margin framework for training data synchronization in large-scale hierarchical classification. In *Neural Information Processing - 20th International Conference, ICONIP*, 2013b.

R. Babbar, C. Metzig, I. Partalas, E. Gaussier, and M.-R. Amini. On Power Law Distributions in Large-Scale Taxonomies. *SIGKDD Explorations*, 16(1), 2014.

K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*, pages 1247–1250, 2008.

G. Cheng, W. Ge, and Y. Qu. Falcons: searching and browsing entities on the semantic web. In *WWW*, pages 1101–1102, 2008.

M. dAquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing Knowledge on the Semantic Web with Watson. In *Proceedings of the 5th International Workshop on Evaluation of Ontologies and Ontology-based tools, co-located with ISWC*, 2007.

R. Dieng-Kuntz and O. Corby. Conceptual graphs for semantic web applications. In *ICCS*, pages 19–50, 2005.

H. Dietze and M. Schroeder. Goweb: a semantic search engine for the life science web. *BMC Bioinformatics*, 10(S-10):7, 2009.

L. Ding, T. W. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *CIKM*, pages 652–659, 2004.

A. Doms and M. Schroeder. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Research*, 33:783–786, 2005.

L. Hirschman and R. J. Gaizauskas. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300, 2001.

B. Katz, G. C. Borchardt, and S. Felshin. Natural language annotations for question answering. In *FLAIRS Conference*, pages 303–306, 2006.

A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, pages 1–46, 2014. ISSN 1384-5810. doi: 10.1007/s10618-014-0382-x. URL http://dx.doi.org/10.1007/s10618-014-0382-x.

R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM*, pages 233–242, 2007.

J. G. Mork, D. Demner-Fushman, S. Schmidt, and A. R. Aronson. Recent enhancements to the NLM medical text indexer. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 1328–1336, 2014. URL http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-MorkEt2014.pdf.

A.-C. Ngonga Ngomo, L. Bühmann, C. Unger, J. Lehmann, and D. Gerber. Sorry, I don't speak SPARQL – Translating SPARQL Queries into Natural Language. In *Proceedings of WWW*, 2013.

A.-C. Ngonga Ngomo, N. Heino, R. Speck, and P. Malakasiotis. A tool suite for creating question answering benchmarks. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

G. Paci, G. Pedrazzi, and R. Turra. Wikipedia-based approach for linking ontology concepts to their realisations in text. In *LREC*, 2010.

C. Perez-Iratxeta, A. J. Pérez, P. Bork, and M. A. Andrade. Update on xplormed: a web server for exploring scientific literature. *Nucleic Acids Research*, 31(13):3866–3868, 2003.

A. Pipitone and R. Pirrone. A framework for automatic annotation of Wikipedia articles. In *Proceedings of the 6th Workshop on Semantic Web Applications and Perspectives*, 2010.

D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. EBIMed - text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):237–244, 2007.

W. C. Richardson. To err is human: building a safer health system. In *To err is human: building a safer health system*. National Acad. Press, 2006.

G. Tsatsaronis, K. Mourtzoukos, V. Andronikou, T. Tagaris, I. Varlamis, M. Schroeder, T. Varvarigou, D. Koutsouris, and N. Matskanis. PONTE: A context-aware approach for automated clinical trial protocol design. In *Proceedings of the 6th International VLDB Workshop on Personalized Access, Profile Management, and Context Awareness in Databases*, 2012a.

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

G. Tummarello, R. Delbru, E. Oren, and R. Cyganiak. Sindice.com: A Semantic Web Search Engine. Technical report, 2007. URL http://sindice.com/.

M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *WWW*, pages 585–594, 2006.

Z. Zheng. Answerbus question answering system. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 399–404, 2002.