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

Project **FP7-318652 / BioASQ**

Deliverable **D2.3**

Distribution **Public**

http://www.bioasq.org

Exploitation and dissemination strategy

Authors: Michael Alvers, George Tsatsaronis, Matthias Zschunke and Axel-Cyrille Ngonga Ngomo

Status: Final (Version 1.0)

May 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	Ms Martina Eydner

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D2.3
Deliverable title	Exploitation and dissemination strategy
Contractual date of delivery	M6 (March 2013)
Actual date of delivery	May 2013
Relevant Task(s)	WP2/Task 2.3
Partner Responsible	TI
Other contributors	TI, ULEI
Number of pages	6
Author(s)	Authors: Michael Alvers, George Tsatsaronis, Matthias Zschunke and Axel-Cyrille Ngonga Ngomo
Internal Reviewers	UJF
Status & version	Final
Keywords	Exploitation Plan, Dissemination Strategy

Executive Summary

This deliverable focuses on specifying how the results of BIOASQ are to be exploited both commercially and scientifically. Especially, 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 strategy for the future management of the social network is set in place.

Exploitation beyond the end of the project is recognized as the key enabler for the success of BIOASQ, and therefore the planned exploitation activities aim to create value within all the participating organizations and thus to improve their competitive advantages. Scaling the results up to commercial offerings ensures profitability through economies of scale. In this direction, 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.

Chapter 1 focuses on scientific exploitation and discusses the planned scientific paper submissions, the dissemination through teaching material, participation in scientific events, and additionally planned software releases.

Chapter 2 focuses on the commercial exploitation of BIOASQ starting with the direct outcomes of the project, such as the BIOASQ dataset and the services that are being developed. Specific policies and pricing 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.

Contents

List of Figures

List of Tables

1

Scientific Exploitation and Dissemination

The BIOASQ exploitation and dissemination targets of the academic partners are mainly **excellence building**, **knowledge transfer**, **education** and later **research** in BIOASQ-related areas. In the following, the suggested activities are discussed.

1.1 Scientific Publications

In the course of the BIOASQ project, the academic partners will aim to publish innovations that will result from the project. These publications will mainly be of scientific nature and will target renowned scientific publication outlets (i.e., international peer-reviewed conferences and journals). The BIOASQ publications will contribute to demonstrating the advancement of the expertise and excellence of the research group involved in the project. The BIOASQ consortium has already released two scientific papers at top conferences (Tsatsaronis et al., 2012; Ngonga Ngomo et al., 2013) and plans to release at least three more. By these means, we aim to facilitate a sustainable development of the research group. In addition to scientific publications, we also aim to release informal publications, including blog posts, tweets, *Facebook* and *LinkedIn* entries, through the project's web dissemination channels. Also, several public deliverables will be released, including a roadmap report, which will propose ways to advance biomedical semantic indexing and QA beyond the end of the project. Furthermore, we will release teaching material, which will be used during summer schools such as the upcoming versions of *IASLOD*[1], as well as presentations and videos on platforms such as *SlideShare*[2], *SlideWiki*[3] and *Videolectures*[4].

1.2 Events

In addition to organizing the BIOASQ workshops, the academic partners plan to participate in a great variety of events with presentations and invited talks. The BIOASQ workshops will aim to attract both

[1] http://semanticweb.kaist.ac.kr/2012lodsummer/
[2] http://www.slideshare.net
[3] http://slidewiki.org/
[4] http://videolectures.net/

participating researchers and non-participating yet interested researchers and companies. The workshops will be made public via the BIOASQ leaflet and website and communicated throughout a variety of channels such as mailing lists, blogs, tweets, etc. During the workshops, the project results (benchmark creation, tools, social network, etc.) will be advertised. As part of the BIOASQ dissemination activities, the first workshop has already been scheduled to take place in September 2013, collocated with the *Conference and Labs of the Evaluation Forum* (*CLEF 2013*), to be held in Valencia, Spain.

1.3 Collaborations

During the two years of the project, the relations among the partners of the project and the members of the advisory board will be strengthened and can lead to further collaborations in future projects. Also, the organization of the challenges and the workshops as well as the participation to the events mentioned in the previous section, will give 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 have already started comprise the *Memorandum of Understanding* which has been mutually signed between BIOASQ and the *VISCERAL* project, the discussions that have started with the *IBM Watson* team and with the PortDial 2[5] project, as well as the interaction with *NLM* for the creation of an additional baseline for Task 1a, which is already in place.

1.4 Teaching

An important aspect of disseminating expertise and knowledge gained within the project is through the curricula of students studying at the academic institutions participating in the project. Although these activities are not directly part of the project, they are important for disseminating best-practices and knowledge, as well as for preparing students to work with innovative relevant technologies. Theses, both at undergraduate and postgraduate level related to the project, an also be assigned to students. Also, the topics of the project will influence some of the lectures, seminars and practical work which are held at the academic partners' institutions.

1.5 Software Releases

In addition to the software releases that constitute formal project deliverables, the BIOASQ consortium will be releasing a number of software components and/or intermediate releases as open-source to the wider public. These include the BIOASQ annotation tool and underlying technologies such as *SPARQL2NL*[6] as well as corresponding extensions of the *TBSL Question Answering Engine*[7].

Moreover, the BIOASQ social network will be released in the middle of the project and ensure the sustainability of the project even after its completion. The main vision here is that the BIOASQ social network will facilitate the creation of new benchmarks even after the end of the BIOASQeffort. To this end, the consortium will rely on a *Distributed Social Network Technologies* (Tramp et al., 2012) architecture instead of using centric architectures such as those of *Elgg*[8]. The main advantage of this approach is that the network will not require any central management. Thus, it will be ensured to survive after the end of the project. The network will also be used and tested during the course of the project.

[5] http://sites.google.com/site/portdial2/
[6] http://github.org/AKSW/SPARQL2NL, http://sparql2nl.aksw.org/demo
[7] http://autosparql-tbsl.dl-learner.org/
[8] http://elgg.org

In particular, the data compiled by the consortium will be released on the social network after the end of each challenge. The users of the network will be allowed to comment and discuss the data, therewith generating insights that will allow improving the datasets. The software releases will be announced at mailing lists and blogs. In addition, the source code and installation archives will be available at major open-source project repositories such as *GitHub*[9].

1.6 Datasets

The corpus created in the BIOASQ project of questions, answers and evidences (concepts, triples, snippets and related *PubMed* documents) will contain at least 600 questions with the accompanying *gold standard* answers. The BIOASQ data will focus specifically in the life sciences which constitutes the advantage the BIOASQ consortium has; in producing for the first time such a high-quality question answering dataset for the specific domain.

In addition, during the project an evaluation platform will be implemented. After the end of the challenge, the platform will be available to the research community in order to serve as a framework for experimental evaluation for large scale information retrieval systems. Specifically, for the tasks of large scale annotation (tasks 1a and 2a) the platform will be available for users (after registration) in order to upload results and test their systems. The platform provides an easy way for the evaluation of such systems and thus it will help researchers to assess their systems under a common benchmark. Additionally, the platform will provide a web service to the users for the creation of new evaluation tests for both tasks *a* and *b*. The datasets that will be used during the challenges will be available for experimentation, while it will also be possible to create new datasets through the social network. 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. Finally, the evaluation tools will be used during the challenges for tasks 1a, 2a, 1b, and 2b will be compiled into a package and made available under an open-source licence.

1.7 Licensing

With regards to the licensing scheme for the release of the BIOASQ components (software) and datasets, overall, the BIOASQ consortium has been clearly committed to the open source approach for the projects results. The partners have come into an agreement to follow a common licensing approach for the components and datasets developed within the projects 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. More specifically, 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, it will ensure 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 coordinator for obtaining a different license, which will include the assurances which distributors typically find in commercial distribution agreements. The exact details of the

[9] http://github.com

selected licensing scheme will be decided at a later stage in the lifetime of the project.

2

Commercial Exploitation and Advertisement

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.

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

[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).

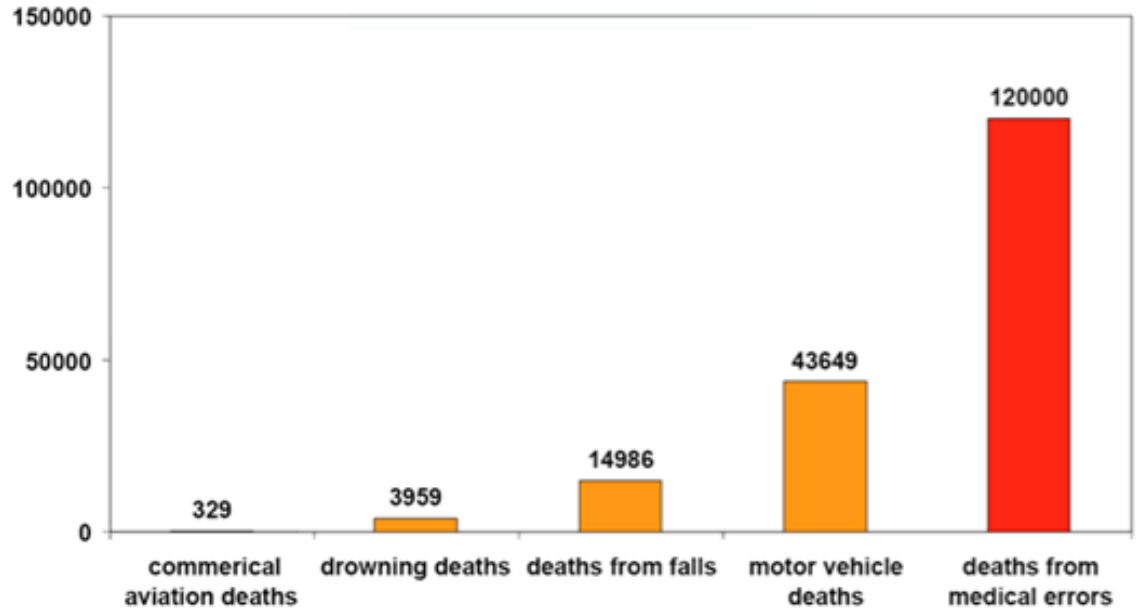

Figure 2.1: Comparison between death rates based on the type of cause of death. Data from *U.S.A.*, 1999.

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 will develop their own exploitation plan throughout the project, the final version will serve as input to their product and solution management organization and enable them to develop a first business case for their development decisions. In the following we analyze one such direction, namely the enhancement of the *GoPubMed* semantic search engine of TI with question answering features.

2.1.1 Current State of the Art Systems and Technologies

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, 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).

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* (**?**), *Semantic Web Search Engine* (*SWSE*), *WikiDB*, *Sindice* (**?**), *Watson* (**?**), *Falcons* (**?**), and *CORESE* (**?**). 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* (**?**) and large efforts such as *Freebase* (**?**), which provides an environment for authoring formal statements.

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* (**?**), *Hakia*[2] and *Answer Bus* (**?**). 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

[2] http://www.hakia.com/
[3] http://clusty.com
[4] http://carrotsearch.com/

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* (**?**), *GoWeb* (**?**)[5], *EBIMed* (**?**) and *XplorMed* (**?**). According to previously published studies (**??**), 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 (**?**). *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. (**?**) 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 (**?**) 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.

2.1.2 BIOASQ Technologies into *GoPubMed*

Given the success of *GoWeb* in benchmark evaluations (**?**) for answering questions in the biomedical domain, as well as that of *GoPonte* (**?**), 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 (**?**). We plan to build on the experiences of the participating systems of Tasks 1a and 2a to optimize the annotation process, and on the experiences of the systems that participate in Tasks 1b and 2b in order to learn how to extract the most useful answers for a given question, aided by the performed annotations.

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

2.2 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., 2012). 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 2.1, where each query is considered a service request (call) to any of the developed services[6]. 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 2.1: Suggested pricing of the BIOASQ developed services. Differentiation between universities and industry is applied.

2.3 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 site of the Bioinformatics group of our partner Prof. Michael Schroeder. In addition, BIOASQ will be advertised as the platform for biomedical question-answering in general on all trade fair participations like *CeBIT 2013* and *DMG*.

[6] 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

Bibliography

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

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.

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.

S. Tramp, P. Frischmuth, T. Ermilov, S. Shekarpour, and S. Auer. An Architecture of a Distributed Semantic Social Network. *Semantic Web Journal*, Special Issue on The Personal and Social Semantic Web, 2012. URL http://www.semantic-web-journal.net/sites/default/files/swj201_4.pdf.

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*, 2012.