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

Project **FP7-318652 / BioASQ**

Deliverable **D1.6**

Distribution **Public**

http://www.bioasq.org

Final Report

Georgios Paliouras and Anastasia Krithara

Status: Final (Version 1.0)

March 2015

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	D1.6
Deliverable title	Final Report
Contractual date of delivery	M24 (September 2014)
Actual date of delivery	March 2015
Relevant Task(s)	WP1/Task 1.2
Partner Responsible	NCSR “D”
Other contributors	TI, UJF, ULEI, UPMC, AUEB-RC
Number of pages	32
Author(s)	Georgios Paliouras and Anastasia Krithara
Internal Reviewers	BioASQ consortium
Status & version	Final
Keywords	final report, reports

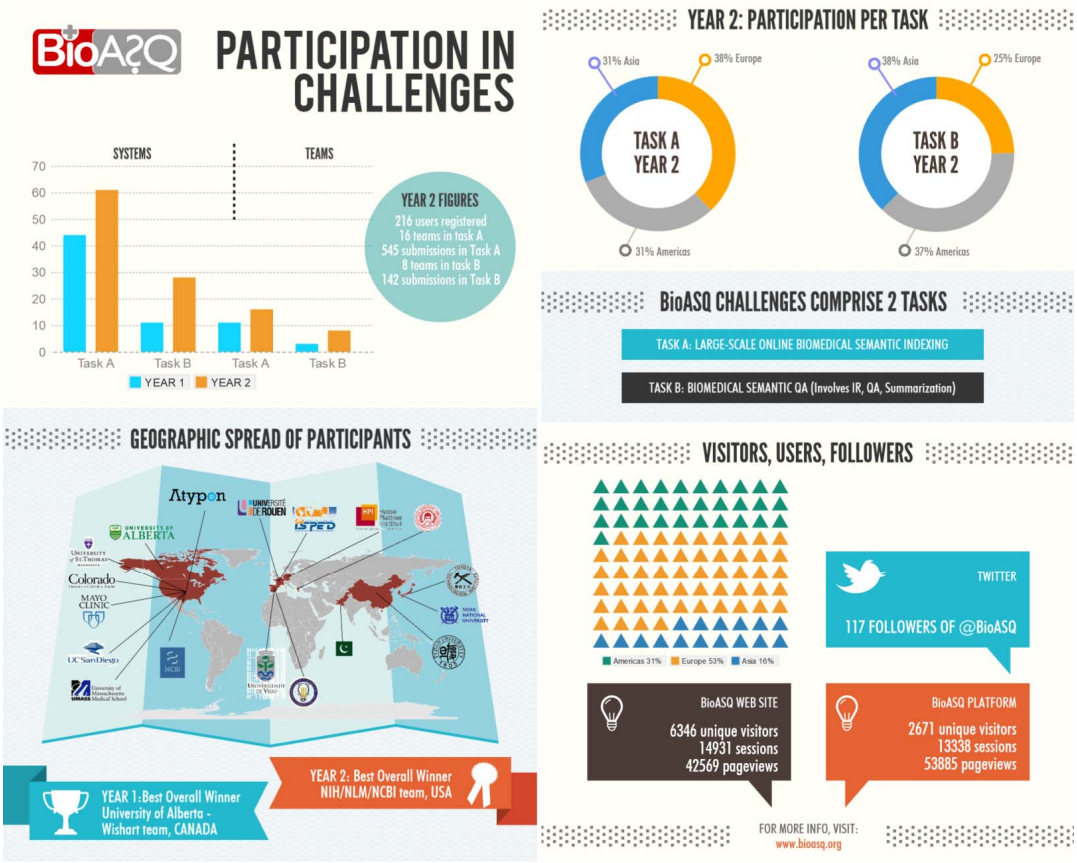
Executive Summary

Motivation: Every day, more than 3000 new articles are published in biomedical journals. That averages to more than 2 articles every minute! MEDLINE currently comprises more than 20 million articles, while the number and size of non-textual biomedical data sources also increases rapidly. This wealth of new knowledge plays a central role in the progress achieved in biomedicine and its impact on public health. However, ensuring that this knowledge is used for the sake of the patients in a timely manner is a demanding task. The BIOASQ project aimed to push research towards highly precise biomedical information access systems. The project achieved this goal by establishing a series of challenges (competitions), in which systems from teams around the world compete. BIOASQ provides data, software and the evaluation infrastructure for the challenge. By these means, the project ensures that the biomedical experts of the future can rely on software tools to identify, process and present the fragments of the huge space of biomedical resources that address their personal questions.

BioASQ format: BIOASQ comprises two tasks. In Task A systems are required to automatically assign MESH terms to biomedical articles, thus assisting the indexing of biomedical literature. Systems participating in the task are given newly published MEDLINE articles, before the NLM curators have assigned MESH terms to them. The systems assign MESH terms to the documents, which are then compared against the terms assigned by the NLM curators. Task B focuses on obtaining precise and comprehensible answers to biomedical questions. The systems that participate in Task B are given English questions written by biomedical experts that reflect real-life information needs. For each question, the systems are required to return relevant articles, snippets of the articles, concepts from designated ontologies, RDF triples from Linked Life Data, an 'exact' answer (e.g., a disease or symptom), and a paragraph-sized summary answer. Hence, this task incorporates traditional information retrieval, with question answering from text and structured data, as well as multi-document text summarization.

Objectives and Main achievements:

Advancing the state-of-the-art in large-scale semantic indexing and question answering. By participating in BIOASQ, systems are pushed to their limits in terms of scalability, efficiency, accuracy, coverage, and conciseness of responses. Typically, the participating systems combine and improve state-of-the-art methods in several of the research areas. As an example of the impact of BIOASQ in the semantic indexing area, in both the first and the second challenge, the best system consistently outperformed the Medical Text Indexer (MTI), which has been developed by NLM especially for this task and is used to recommend MESH terms to NLM curators. Furthermore, as announced recently by NLM, MTI itself was improved in the second year, by incorporating ideas from the winning system of the first BIOASQ challenge.



Successful organisation of the challenge. The benchmarks that are provided by BIOASQ include very large document collections, as well as databases, knowledge bases, ontologies, and other structured data. Given this organisational complexity, one of the main objectives of BIOASQ was to ensure the timely and successful organisation of the competition, attracting a large number of key players as participants. In the most recent BIOASQ challenge, 216 users and 142 systems registered in order to participate in the challenge, while 25 teams (with 95 systems) finally submitted their results.

Establishment of BIOASQ as a reference point in biomedical question answering. BIOASQ is the first international series of challenges for biomedical question answering and has managed in the first two years of existence to attract the attention of key players in the field, either as challenge participants or as members of its advisory board. Furthermore, a social network of biomedical experts has been formed, starting with the experts who contributed data to the BIOASQ benchmarks. This network will continue to exist after the end of the project, providing a platform for maintaining and extending the BIOASQ benchmarks, based on contributions and evaluation by peers.

Building foundations for further competitions. Beyond the social network of experts, BIOASQ's heritage includes reusable infrastructure for creating benchmark data and running challenges. The main components of this infrastructure include tools for annotating data, tools for assessing the results of participating systems, benchmark repositories, evaluation services, etc. The existence of this infrastructure facilitates the sustainability of BIOASQ challenges beyond the end of the project at low cost. Additionally, by providing all these components under open source licenses, new benchmarks and competitions, possibly in different domains, can be easily organised.

Contents

List of Figures

List of Tables

1

Project Context and Objectives

Every day, more than 3000 new articles are published in biomedical journals. That averages to more than 2 articles every minute! Figure 1.1 illustrates the exponential growth of the biomedical literature, in the last 60 years. MEDLINE currently comprises more than 20 million articles, while the number and size of non-textual biomedical data sources is also increasing rapidly. Linked Life Data[1] alone provides more than 10 billion RDF statements of biomedical information.

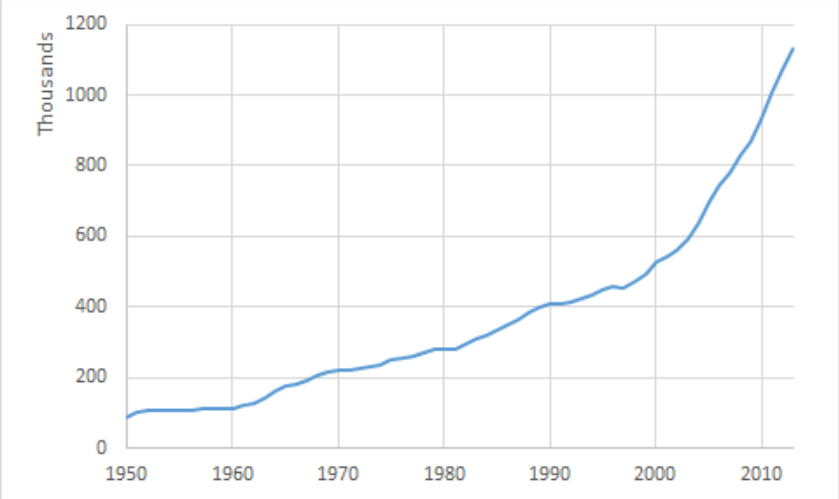


Figure 1.1: Number of articles indexed by MEDLINE (PUBMED) per year.
Source: http://dan.corlan.net/medline-trend.html

This wealth of new knowledge plays a central role in the progress achieved in biomedicine and its impact on public health. However, managing this large amount of data is a challenge in itself. Ensuring that this knowledge is used for the sake of the patients in a timely manner is an even more demanding task for both computer scientists and biomedical experts. The BIOASQ project, which started on

[1] http://linkedlifedata.com/

October 1st 2012 and ran for 2 years, aimed to push research in information technology towards highly precise biomedical information access systems. The project achieved this goal by establishing a series of challenges (competitions), in which systems from teams around the world compete. BIOASQ provides data, software and the evaluation infrastructure for the challenge. By these means, the project ensures that the biomedical experts of the future can rely on software tools to identify, process and present the fragments of the huge space of biomedical resources that address their personal questions. Questions like the ones shown in Tab.1.2.

Q: Are there any DNMT3 proteins present in plants?
A: Yes. The plant DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) is a homolog of the mammalian de novo methyltransferase DNMT3. DRM2 contains a novel arrangement of the motifs required for DNA methyltransferase catalytic activity.
Q: What is the methyl donor of DNA (cytosine-5)-methyltransferases?
A: S-adenosyl-L-methionine (AdoMet, SAM) is the methyl donor of DNA (cytosine-5)-methyltransferases. DNA (cytosine-5)-methyltransferases catalyze the transfer of a methyl group from S-adenosyl-L-methionine to the C-5 position of cytosine residues in DNA.
Q: Which species may be used for the biotechnological production of itaconic acid?
A: In 1955, the production of itaconic acid was firstly described for Ustilago maydis. Some Aspergillus species, like A. itaconicus and A. terreus, show the ability to synthesize this organic acid and A. terreus can secrete significant amounts to the media. Itaconic acid is mainly supplied by biotechnological processes with the fungus Aspergillus terreus. Cloning of the cadA gene into the citric acid producing fungus A. niger showed that it is possible to produce itaconic acid also in a different host organism.
Q: How do histone methyltransferases cause histone modification?
A: Histone methyltransferases (HMTs) are responsible for the site-specific addition of covalent modifications on the histone tails, which serve as markers for the recruitment of chromatin organization complexes. There are two major types of HMTs: histone-lysine N-Methyltransferases and histone-arginine N-methyltransferases. The former methylate specific lysine (K) residues such as 4, 9, 27, 36, and 79 on histone H3 and residue 20 on histone H4. The latter methylate arginine (R) residues such as 2, 8, 17, and 26 on histone H3 and residue 3 on histone H4. Depending on what residue is modified and the degree of methylation (mono-, di- and tri-methylation), lysine methylation of histones is linked to either transcriptionally active or silent chromatin.

Table 1.2: Example questions from the BioASQ question answering task.

The tasks included in the BIOASQ challenges help advance the state of the art in two fields. BIOASQ Task A aims at improving the automatic classification of biomedical documents. Here, systems are required to automatically assign MESH terms to biomedical articles, thus assisting the indexing of biomedical literature. Systems participating in the task are given newly published MEDLINE articles, before the NLM curators have assigned MESH terms to them. The systems assign MESH terms to the documents, which are then compared against the terms assigned by the NLM curators. BIOASQ Task B focuses on obtaining precise and comprehensible answers to biomedical questions. The systems that participate in Task B are given English questions written by biomedical experts that reflect real-life

information needs. For each question, the systems are required to return relevant articles, snippets of the articles, concepts from designated ontologies, RDF triples from Linked Life Data, an 'exact' answer (e.g., a disease or symptom), and a paragraph-sized summary answer. Hence, this task incorporates traditional information retrieval, with question answering from text and structured data, as well as multi-document text summarization. Figure 1.2 places BIOASQ in research context, by providing an overview of the areas that are of relevance to the BIOASQ challenges. Researchers working in any of these areas, with an additional interest in helping biomedical research progress, are encouraged to participate in the challenges.

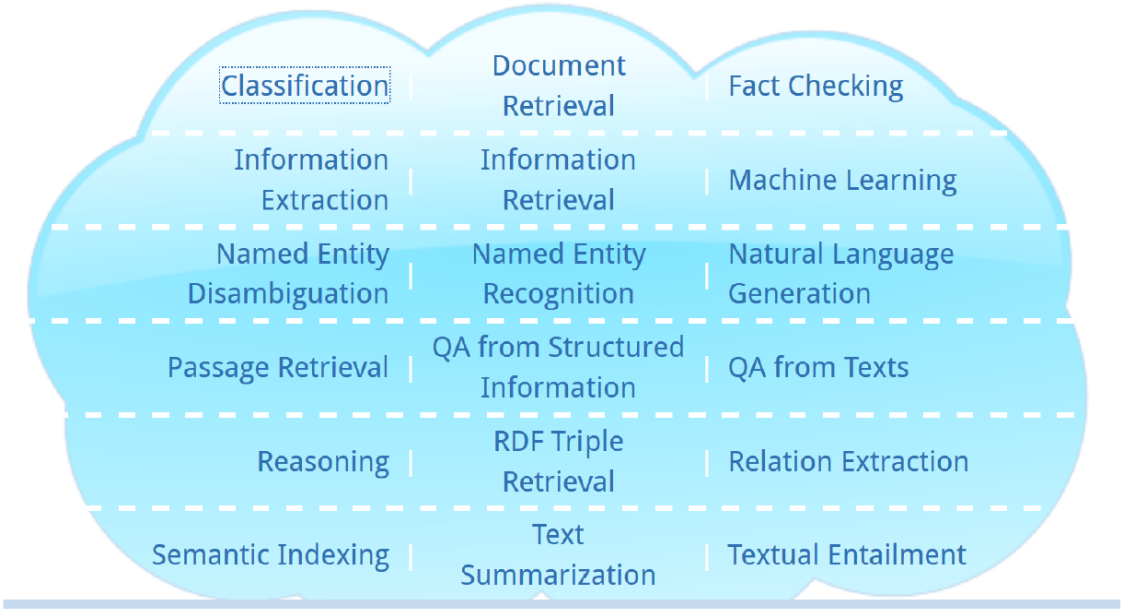


Figure 1.2: Research areas of relevance to BIOASQ

By establishing the new series of challenges, BIOASQ aimed at achieving the following objectives:

Objective 1: Advancing the state-of-the-art in large-scale semantic indexing and question answering. By participating in BIOASQ, systems are pushed to their limits in terms of scalability, efficiency, accuracy, coverage, and conciseness of responses. Typically, the participating systems combine and improve state-of-the-art methods in several of the areas of Fig. 1.2. This helps improve the state of the art in these areas, and semantic indexing and question answering in particular. As an example of the impact of BIOASQ in the semantic indexing area, Fig. 1.3 shows the performance of the best system, the baseline system MTI (Medical Text Indexer) of NLM and the average of all of BIOASQ participants. In both the first and the second challenge, the best system consistently outperformed MTI, which has been developed especially for this task and is used to recommend MESH terms to NLM curators. Furthermore, as announced recently by NLM, MTI itself was improved in the second year, by incorporating ideas from the winning system of the first BIOASQ challenge.[2]

Objective 2: Successful organisation of the challenge. The benchmarks that are provided by BIOASQ include very large document collections, as well as databases, knowledge bases, ontologies, and other structured data. Given this organisational complexity, one of the main objectives of BIOASQ

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

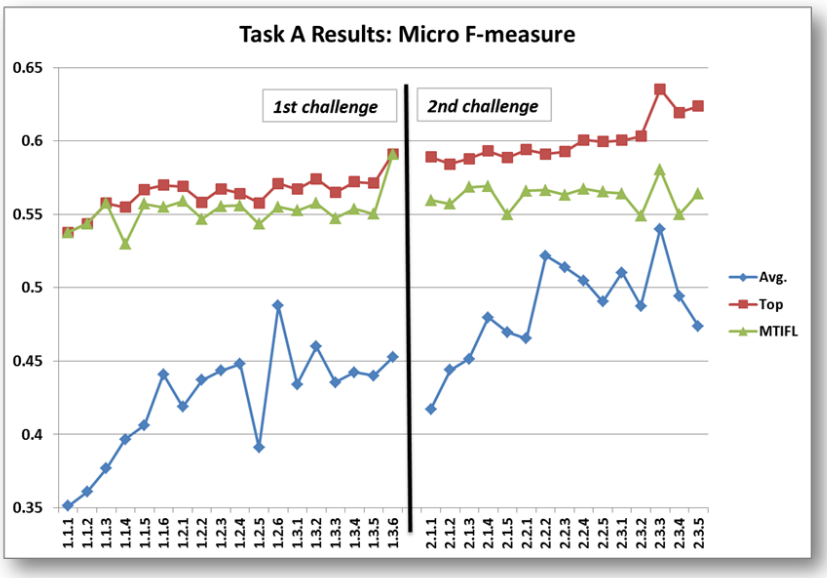


Figure 1.3: Increase in semantic indexing performance in the first two BIOASQ challenges.

was to ensure the timely and successful organisation of the competition, attracting a large number of key players as participants. In the most recent BIOASQ challenge, 216 users and 142 systems registered in order to participate in the challenge, while 25 teams (with 95 systems) finally submitted their results. Figure 1.4 illustrates the geographic spread of BIOASQ participation.

Figure 1.4: Geograpic spread of BIOASQ participants.

Objective 3: Establishment of BIOASQ as a reference point in biomedical question answering. Although question answering has a long history in artificial intelligence and computational linguistics, biomedical question answering in particular is a relatively new field with distinct characteristics (**?**), most notably the heterogeneity and exponential growth of the underlying information sources, but also the very extensive use of domain-specific terminology. BIOASQ is the first inter-

national series of challenges for biomedical question answering and has managed in the first two years of existence to attract the attention of key players in the field, either as challenge participants or as members of its advisory board.[3] Furthermore, a social network of biomedical experts has been formed, starting with the experts who contributed data to the BIOASQ benchmarks. This network will continue to exist after the end of the project, providing a platform for maintaining and extending the BIOASQ benchmarks, based on contributions and evaluation by peers.

Objective 4: Building foundations for further competitions. Beyond the social network of experts, BIOASQ's heritage includes reusable infrastructure for creating benchmark data and running challenges. Figure 1.5 summarises the main components of this infrastructure, which include tools for annotating data, tools for assessing the results of participating systems, benchmark repositories, evaluation services, etc. The existence of this infrastructure facilitates the sustainability of BIOASQ challenges beyond the end of the two-year project at low cost. Additionally, by providing all these components under open source licenses, new benchmarks and competitions, possibly in different domains, can be easily organised.

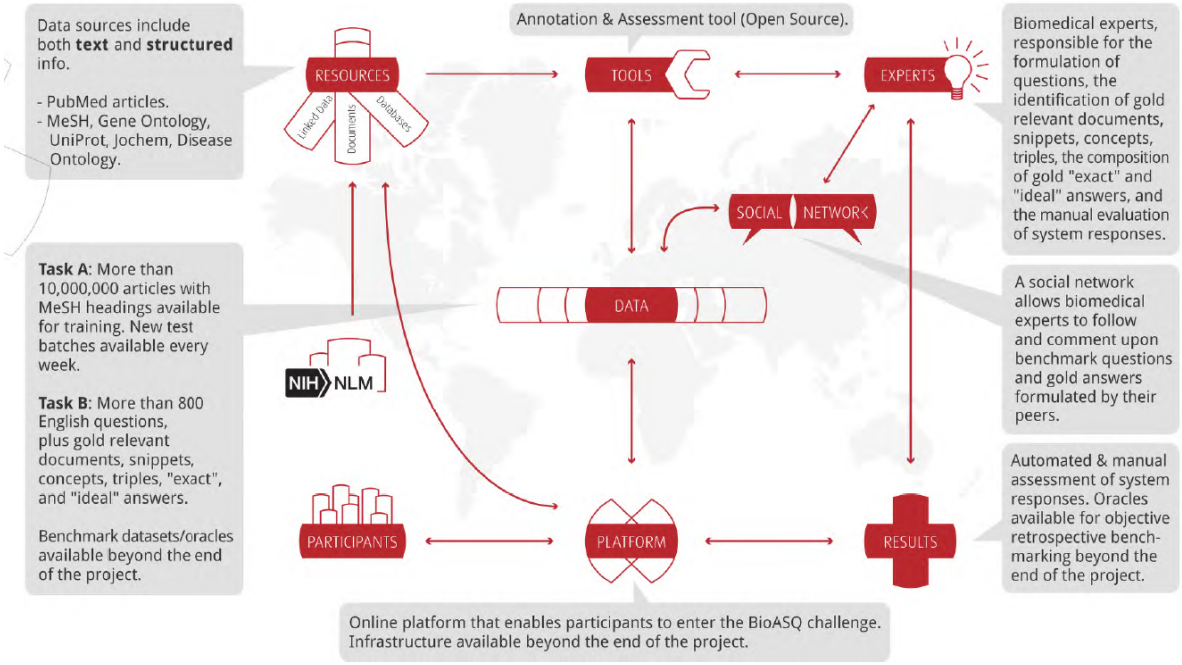


Figure 1.5: The BIOASQ infrastructure.

In the following sections, this report provides details of the main results of the project, addressing the above-mentioned objectives.

[3] http://www.bioasq.org/project/advisory-board

2

Main Results of the Project

2.1 BioASQ Challenges

BIOASQ assesses the performance of information systems in supporting two tasks that are central in the biomedical question answering process: (a) the indexing of large volumes of unlabeled data, primarily scientific articles, with biomedical concepts, (b) the processing of biomedical questions and the generation of answers and supporting material. These two tasks of the challenge are presented briefly in the following two subsections.

2.1.1 Task A: Large-scale biomedical semantic indexing

BIOASQ Task A requires systems to automatically assign MESH terms to biomedical articles added to the MEDLINE database, thus assisting the indexing of biomedical literature. MEDLINE and the associated search engine PUBMED are very valuable sources of information for biomedical researchers, as they provide unified access to a very large proportion of the biomedical literature worldwide. Currently MEDLINE articles are indexed manually by a large team of expert curators who collaborate with NLM. As the volume of the biomedical literature is increasing ever faster (see Fig. 1.1) this manual approach becomes impractical and may soon even be impossible. Automating, at least partially, this process would be of great service to the biomedical community. The NLM has developed a special tool for this purpose, which is called Medical Text Indexer (MTI) and is used to recommend MESH terms to the NLM curators. However, there is still much room for improving the performance of MTI-like indexing systems.

Systems participating in Task A are given newly published MEDLINE articles, before the NLM curators have assigned MESH terms to them. The systems assign MESH terms to the documents, which are then compared against the terms assigned by the NLM curators. Thus, as the manual annotations become gradually available, the scores of the systems are updated. In this manner, the evaluation of the systems participating in Task A is fully automated on the side of BIOASQ and thus can run on a weekly basis throughout the year. NLM also provides kindly the results of MTI, both as a baseline against which the performance of other systems can be compared, but also for other systems that may want to use this information as input.

The performance of the systems taking part in Task A is assessed with a range of different measures. Some of them are variants of standard information retrieval measures for multi-label classification problems (e.g., precision, recall, f-measure accuracy). Additionally, measures that use the MESH hierarchy to provide a more refined estimate of the systems' performance are used. Some of these methods have been developed in BIOASQ and are presented in (**?**).

2.1.2 Task B: Biomedical question answering

BIOASQ Task B takes place in two phases, called Phase A and Phase B. In Phase A, the participants are given English questions formulated by biomedical experts. For each question, the participating systems have to retrieve relevant documents (from PUBMED), relevant snippets (passages) of the relevant documents, relevant concepts (from five designated ontologies), and relevant RDF triples (from the Linked Life Data platform). Subsequently, in Phase B the participants are given the relevant documents, snippets, concepts, and triples that the experts themselves have identified (using tools developed in BIOASQ), and they are required to return 'exact' answers (e.g., names of particular diseases or genes) and 'ideal' answers (a paragraph-sized summary of the most important information of Phase A per question). The responses of the systems are evaluated both automatically (e.g., using Mean Average Precision, ROUGE against gold responses provided by the experts) and manually by the experts.

Task B was designed to be ambitious and realistic. Most notably, the questions reflected real information needs of biomedical experts (see examples in Tab. 1.2. The task aimed also to promote a broader view of Question Answering (QA), which integrates Information Retrieval (including Passage Retrieval), QA for document collections, QA for structured data, and multi-document summarization, among other technologies. Unlike most search engines, which accept keyword queries and return lists of documents, BIOASQ Task B requires the participating systems to accept syntactically well-formed and often quite complex English questions, and to return concise answers ('exact' and 'ideal' answers) again in English, along with the sources (documents, snippets, concepts, triples) that the answers are based on. In interviews conducted at the end of BIOASQ, the experts who formulated the questions, authored the gold responses, and evaluated the system responses confirmed that QA systems of this kind could, with further improvements, be of significant value in biomedicine.

At the same time, BIOASQ Task B encourages teams who do not work on end-to-end QA systems to participate. For instance, a system that is able to answer only factual questions will be assessed only on this subtask. Similarly, systems that do not provide exact answers, but focus only on the retrieval of material that is of relevance to a question can participate and be assessed only on phase A of the task. Equivalently, systems which deal only with the formation of answers, rather than the retrieval of related material, can use the material provided by the BIOASQ biomedical experts and participate only in Phase B. This flexibility of the task facilitates the participation of diverse research teams, with a common interest to contribute in the biomedical QA process. Thus, the challenge provides also a forum for exchange of ideas and potential collaboration between teams with complementary expertise.

2.2 Datasets and Knowledge Resources

One of the main contributions of BIOASQ to the research community is the generation of benchmark datasets on which research teams can test their methods and systems. Partly these datasets are created by the team of BIOASQ experts and partly are based on publicly available data, such as MEDLINE abstracts. In order to enrich the information found in the scientific literature and associate it to biomedical knowledge, BIOASQ also uses a number of carefully selected knowledge resources, including biomedical ontologies and databases. The use of data and knowledge resources, as well as the process of

generating new datasets within BIOASQ is briefly described in the following subsections.

2.2.1 BioASQ Team of Biomedical Experts

The biomedical expert team was established during the first two months of the BIOASQ project. Several experts had been considered from a variety of institutions across Europe. The final selection of ten experts was based on the need to cover the broad biomedical scientific field, representing as much as possible, medicine, biosciences and bioinformatics. All the members of the biomedical team hold senior positions in universities, hospitals or research institutes in Europe. Their primary fields of research interests are the following: cardiovascular endocrinology, psychiatry, psychophysiology, pharmacology, drug repositioning, cardiac remodeling, cardiovascular pharmacology, computational genomics, pharmacogenomics, comparative genomics, molecular evolution, proteomics, mass-spectometry, and protein evolution.

The principal job of the biomedical expert team is the composition of the Question Answering (QA) benchmark dataset which is used during the BIOASQ challenge Task B. This is achieved with the use of tools that have been developed in BIOASQ for this purpose and have been made publicly available. Using these tools, the experts compose their questions and identify relevant material in the designated resources. This material includes:

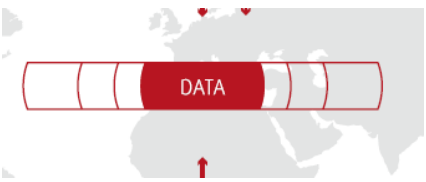
- documents from MEDLINE and relevant snippets out of these documents,
- concepts from designated ontologies, and,
- statements (RDF triples), from selected life science knowledge bases.

In the first two challenges, more than 800 questions were composed and made available, together with the material collected by the experts and the corresponding gold answers. The datasets generated so far can be used as training data for the next BIOASQ challenge.

The members of the BIOASQ biomedical expert team were also asked to manually assess the responses of the participating systems. In addition to scoring the systems' answers, during this process they had the opportunity to enrich and modify the gold material that they provided, thus improving the quality of the benchmark dataset.

A number of physical meetings have been organised with the experts, which not only helped in training the experts, but also pointed to improvements of the data generation process, which were implemented. Having stabilised the process, one of the planned steps for the future is to open the data generation effort to volunteers outside the core BIOASQ expert team. The social network that was developed and deployed within BIOASQ allows the more experienced experts to help newcomers in providing high quality data for the benchmarks.

2.2.2 Datasets used or generated by BioASQ



Task A

As mentioned in section 2.1.1, Task A of the challenge concerns the classification of MEDLINE articles into MESH categories, whereby each article is typically assigned to many categories (multi-label classification). Two large datasets have been generated for the two BIOASQ challenges that have been completed. The training data provided by BIOASQ contains for each article its title, its abstract as it appears in MEDLINE and the MESH labels assigned to it by NLM curators. In the testing phase of the challenge the released test sets contain the title and the abstract of the corresponding article without any information about the labels.

Once the systems that participate in the challenge have provided their responses, their performance is assessed as the NLM curators provide manually labels to the same articles. In order for this assessment process to be completed within a reasonable period of time, with the help of NLM, BIOASQ has selected a subset of the journals covered by MEDLINE that are manually curated typically within a period of two months, following the publication of the articles. The list of selected journals is available on the BIOASQ challenge Platform.

Both training and test data for Task A are provided in raw format (plain text) and in a pre-processed form (in a vectorized format) using the Lucene library. The former helps assess the effect of text feature extraction in the classification process, while the latter allows the comparison of systems on the basis of the same features. The training data is further provided in two versions, containing either all the English journals covered by MEDLINE or only the ones selected by BIOASQ. Additionally, on the challenge Platform (`http://participants-area.bioasq.org/`) the participants can find resources that can help them in improving the performance of the classifiers, such as deep vectors extracted from the articles and the results of NLM's Medical Text Indexer on the test data.

The following tables provide basic statistics about the Task A datasets in the first two BIOASQ challenges. Table 2.1 presents the statistics of the training data for Task A.

	Training set 2013	Training set 2014	Reduced set 2014
# of articles	10,876,004	12,628,968	4,458,300
Avrg. labels/article	12.55	12.72	13.20
MeSH labels	26,563	26,831	26,631
Size zip/unzip (raw)	5.1Gb/18Gb	6.2G/20.31Gb	1.9Gb/6.4Gb
Size zip/unzip (Lucene)	4.8Gb/6.2Gb	4.4G/6.2Gb	1.3Gb/1.9Gb

Table 2.1: Statistics of the training data provided to the participants for Task A during the two editions of the BIOASQ competition. The reduced set consists of articles from the journal selected for BIOASQ.

During the test phase several unlabeled datasets were provided to the participants in different batches. This resulted in 33 test-sets with a total of 161,058 documents. In addition to the official test sets, a new off-challenge test set is continuously being issued on a weekly basis. These test sets are useful for those participants preparing for the next challenge or those who want to just test a new method.

Task B

As mentioned in section 2.1.2, Task B is divided into two phases. For each question, the experts provided related documents, snippets, concepts and triples, in order to assess the systems that participated in phase A. Furthermore, the experts provided exact and ideal answers for the assessment of phase B.

The following tables provide basic statistics about the data generated in the first two BIOASQ challenges. For the first challenge, 311 questions were provided, together with the associated material. These

data were given as training material for the second challenge, while an extra 500 questions and related material were generated for the second challenge. The following tables provide basic statistics about the two challenges. In particular, Tab. 2.2 provides information about the number of questions and the average number of documents, snippets, concepts and triples generated per question. Table 2.3, on the other hand, provides statistics per type of question.

Challenge	Size	# of documents	# of snippets	# of concepts	# of triples
1	311	14.28	18.70	7.11	9.00
2	500	11.83	14.92	5.93	116.30[1]

Table 2.2: Statistics on the training and test datasets of Task B: numbers of documents, snippets, concepts and triples refer to averages.

Challenge	Yes/No	Factoid	List	Summary	Total
1	85	59	93	74	311
2	152	135	119	94	500
total	237	194	212	168	811

Table 2.3: Statistics per question type for the two challenges.

All the data generated for the first two BIOASQ challenges are provided on the challenge Platform (http://participants-area.bioasq.org/) .

2.2.3 Knowledge resources used by BioASQ

The selection of resources for the BIOASQ challenge follows the triangle *drug-target-disease* which defines the prime information axes for medical investigations. The basic principle is shown in Fig. 2.1[2].

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. Based on this "*knowledge-triangle*" the selected resources for the BIOASQ challenges are presented next.

Drugs: Jochem (**?**), 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. Given the variety and the population of the different resources in it, JOCHEM is currently one of the largest biomedical resources for drugs and chemicals.

Targets: 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, molecular function, and biological process. The *Universal*

[1]The average was based on the 20% of questions that were annotated with triples.
[2]The figure is courtesy of Strategic Medicine Inc.

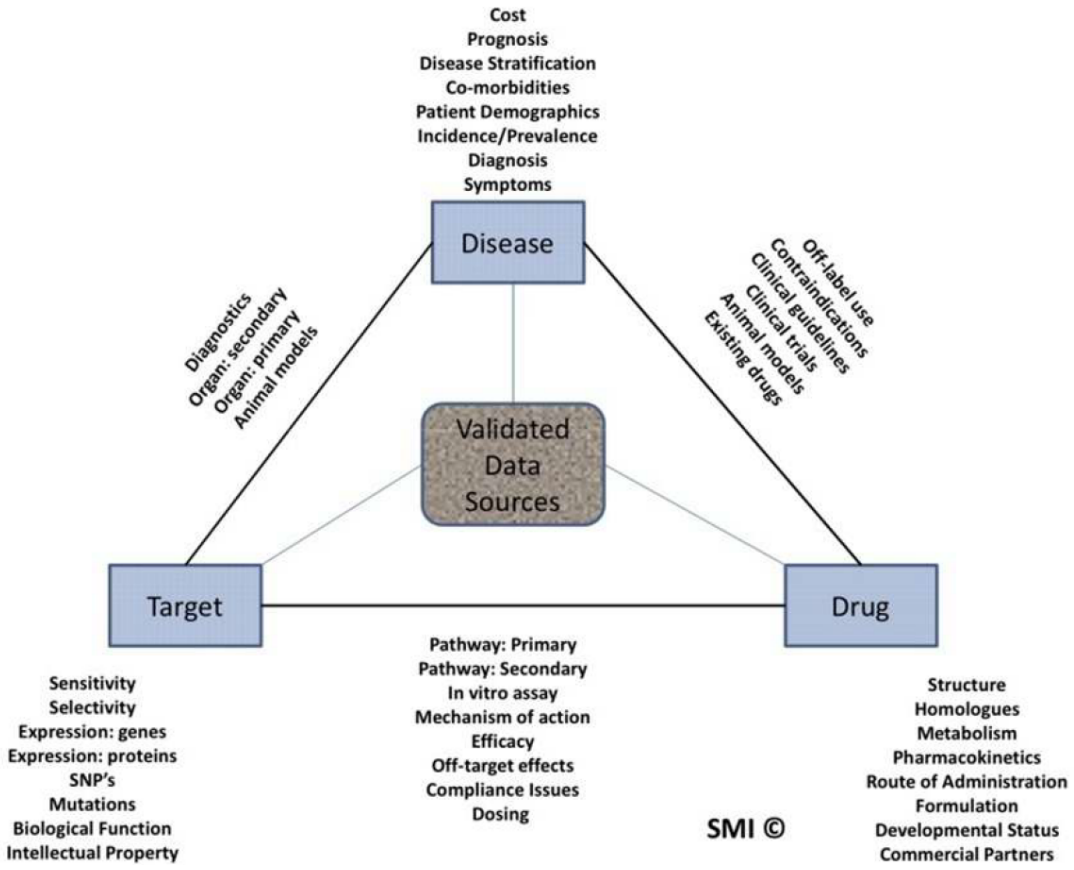


Figure 2.1: The principle of drug-target-disease information axes followed in the selection of resources for BIOASQ.

Protein Resource (UNIPROT) provides a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. Its protein knowledge base consists of two sections: SWISSPROT, 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. In BIOASQ the SWISSPROT component of UNIPROT is used.

Diseases: Disease Ontology (DO) contains data associating genes with human diseases, using established disease codes and terminologies. Approximately 8 thousand 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.

General Purpose: The Medical Subject Headings Hierarchy (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 ((**?**)). The primary MESH terms (called) descriptors

are organized into 16 trees, and are approximately 28 thousand. MESH is the main resource used by PUBMED to index the biomedical scientific bibliography in MEDLINE.

Document Sources The primary corpora for question answering (QA) in the biomedical domain are accessible through PUBMED and PUBMED Central. PUBMED, a service provided by the National Library of Medicine (NLM), contains over 23 million citations from MEDLINE, a bibliographic database of biomedical literature, and other biomedical and life science journals dating back to the 1950s. PUBMED Central (PMC) is a digital archive of full-text biomedical and life-science articles. The full text of all PMC articles is freely available. The archive contains approximately 3 million items.

Linked Data BIOASQ Task B requires 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 LINKED LIFE DATA platform. LINKED LIFE DATA is a data warehouse that syndicates large volumes of heterogeneous biomedical knowledge in a common data model. 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.

2.2.4 Data indexing and retrieval services

The BIOASQ resources have been indexed by Transinsight 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[3]. The concept names (labels), their synonyms and their relations have been indexed in separate LUCENE indices. For the document resources, also LUCENE indices are used, applying the standard LUCENE analyser for the English language.

In order to facilitate quick retrieval, high throughput techniques have been developed for indexing large amounts of text and large-scale knowledge bases, as well as searching efficiently documents, concepts, and triples given keyword queries.

The API through which the resources may be accessed, is based onJSON. For each resource, a respective service is implemented in a uniqueURL. Each URL request opens a session and may request the results, given a query, e.g., a concept. The reply 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.

The following is a list of the services that have been developed:

- a service for accessing the MESH ontology, returning the list of concepts related to an entity,
- similar services for accessing GO, JOCHEM, *Disease Ontology* and the UNIPROT database,
- a service for accessing the PUBMED indexed documents (titles and abstracts), returning document entries, together with MESH annotations, where available,
- a service for accessing the PUBMED Central full text articles, returning also the full text of the article,

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

- a service for accessing the LINKED LIFE DATA platform triples, returning triples that match a set of keywords and a matching score.

2.3 BioASQ Platform

BIOASQ provides an infrastructure to allow challenge participants to acquire training and test data, submit their results and be informed about the performance of their systems, in comparison to other systems. This infrastructure is made available through the BIOASQ Participants Area (or Platform), an online platform (http://participants-area.bioasq.org/) that provides graphical interfaces and web services for the realisation of the challenges. The BIOASQ Platform provides various information (e.g. participation guidelines), support material (e.g. deep vectors) and services (e.g. discussion forum) for the participants. Additionally, it provides a range of services for in-challenge and off-challenge evaluation of systems in the two tasks (e.g. the BIOASQ oracles). Finally, the Platform allows the organisers to administer the challenge. Most importantly, perhaps, the software for the BIOASQ Platform is provided under an open-source license, in order to help in the organisation of other challenges, either in the area of biomedical information access or in different areas.

2.3.1 Functionalities provided by the Platform

The Platform offers a set of functionalities to the participants of the challenge (hereafter users) and also to the BIOASQ organisers. Figure 2.2 depicts the homepage of the Platform as well as the navigational menu that users use to browse the offered resources.

For its users, the Platform provides the following features :

1. Participation in the challenge :
   - downloading the BIOASQ benchmark datasets, consisting of training and test data,
   - mechanisms for submitting results, including HTML forms for manual, and Web services for automated submission,
   - the BioASQ Oracles, where the users can evaluate their systems in an off-challenge mode,
   - tables for browsing the evaluation results,
2. Support in participation:
   - detailed guidelines describing the BIOASQ tasks,
   - tools that have been developed to help participants process the datasets,
   - the "BioASQ Discussions Area", which is a forum about the BIOASQ challenge, and
   - an e-mail help desk for contacting the organising team.

On the other hand, the administrators of the challenge, using the Platform, can:

- create datasets for the semantic indexing task (Task A) automatically, by interacting with the services described in 2.2.4,
- release the test datasets for both tasks,

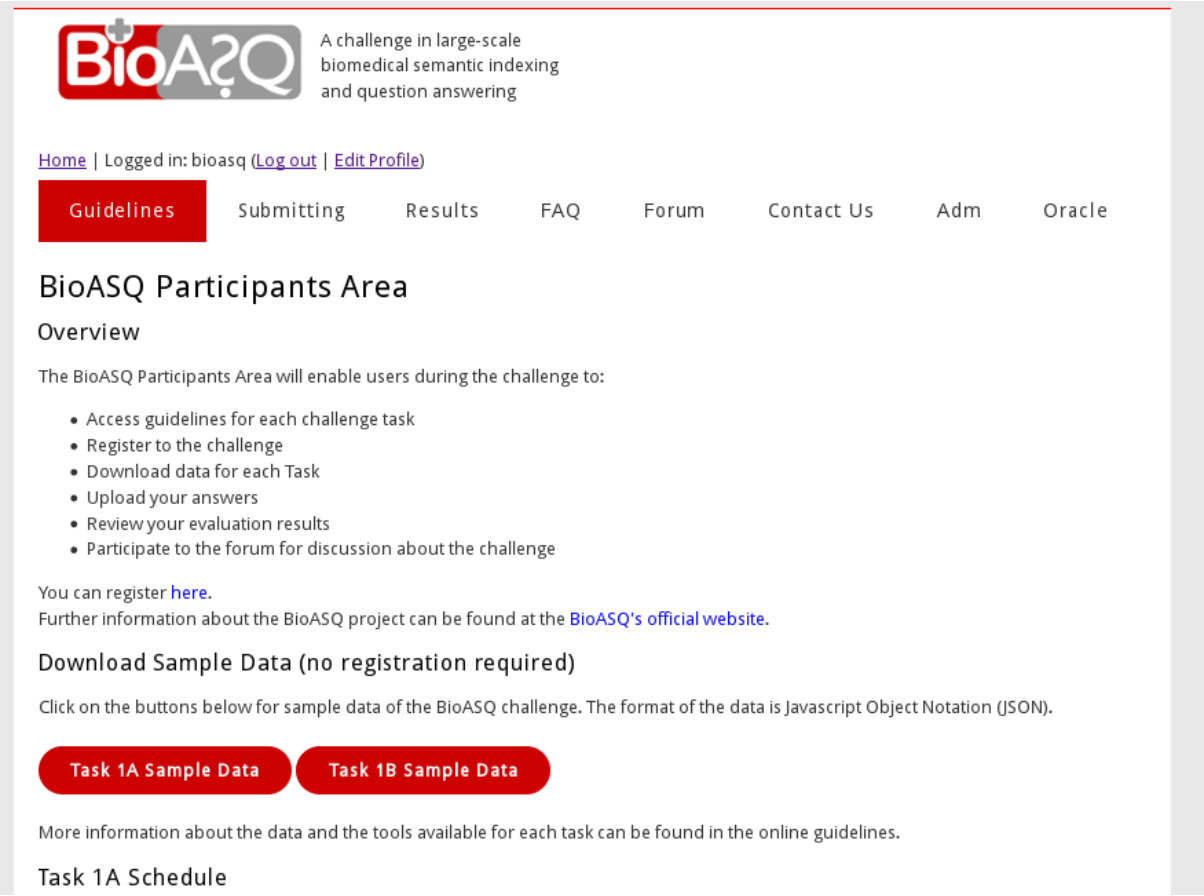


Figure 2.2: The Platform homepage at http://participants-area.bioasq.org/

- trigger the evaluation procedures to update the evaluation results,
- monitor the challenge participation, and
- contact the participants using an e-mail list.

2.3.2 Evaluation measures and procedure

One of the main purposes of BIOASQ was to provide a reliable basis for assessing the performance of biomedical semantic indexing and question answering systems. This is achieved by a carefully designed evaluation procedure and selected evaluation measures for the different tasks and subtasks. The evaluation procedure was designed with the following criteria in mind:

1. Fairness and objectivity. To the extent possible, BIOASQ provides equal opportunities to all participants and independent assessment of the results by external experts.
2. Flexibility and accessibility. Particular effort has been made to allow participants who are working on a subtask of the whole question answering process to participate and be evaluated for that

subtask. Furthermore, by splitting the assessment into batches, BIOASQ allows interested participants to enter the competition at different times.

3. Sustainability. By simplifying and where possible automating the data generation and assessment process, the cost of running the challenge has been reduced considerably.

The following subsections provide a brief account of the evaluation setting for each of the two tasks.

Task A

The evaluation for the semantic indexing task is performed online on a weekly basis. From the time that the test set is announced, participants need to provide their results within 21 hours. When the annotations become available from the MEDLINE curators, in the following weeks, the performance of each system is calculated using standard information retrieval measures as well as hierarchical ones.

The official evaluation period for the first two challenges was divided into three batches containing 5-6 test sets each. In total, 18 test sets were announced for the first and 15 for the second BIOASQ challenge. Based on their performance in individual tests, different winners were selected for each batch. Beyond the official evaluation period, additional test sets are announced weekly. In total so far more than 40 such off-challenge test sets were announced.

The winners of each batch are decided on the basis of their average ranking in the four test sets, in which they have done best in each batch. Therefore, a participant may miss or skip participation in one of the tests, without lowering his/her chances of winning.

The selection of the evaluation measures was based on a thorough study of measures for multi-label and hierarchical classifiers. Among the flat measures, Micro F-measure (MiF) was selected from those proposed in (**?**), while several others are reported for completeness (**?**). For the hierarchical measures, many of those existing in the literature are used, but also a new one, called Lowest Common Ancestor F-measure (LCA-F) was developed, which has several advantages, as explained in (**?**).

Figure 2.3 presents and example of the ranking of the systems on the BIOASQ Platform for a test set. The systems are ranked according to the MiF measure, while several other measures are presented for completeness.

Task B

Task B was split into three and five independent batches for the first and the second BIOASQ challenge respectively. Around 100 questions were provided for each batch. Given the questions, the systems that participate in phase A of the task need to provide relevant documents, snippets, concepts and triples (also called statements) within 24 hours. Then, phase B starts, where the participating systems are given the documents, snippets, concepts and triples considered correct by the experts and are asked to provide exact and ideal answers within another 24 hours. Different winners are selected for each batch and each subtask, e.g. document retrieval.

The evaluation of each subtask in phase A is automated and the winners are determined using mean average precision (MAP). A number of other measures are included for completeness, as explained in (**?**). In phase B, automated evaluation according to various measures is also provided, but the selection of the winners is based on the manual assessment of each answer by the expert who generated the question. The expert provides four scores for each answer, according to its readability, recall, precision and lack of repetition. The winner for each batch is selected on the basis of the average ranking of the systems according to the four criteria and for all the questions of the batch.

+ Test batch 1, week 1

Annotated articles:3319/4440

Flat Measures

System Name	MiF	EBP	EBR	EBF	MaP	MaR	MaF	MiP	MiR	Acc.
Asclepius	0.5890	0.5882	0.6060	0.5814	0.5961	0.4434	0.4323	0.5904	0.5876	0.4241
Mixing with MTI	0.5726	0.5089	0.6872	0.5665	0.4978	0.5728	0.5038	0.4989	0.6720	0.4070
L2R-n2	0.5718	0.6361	0.5576	0.5733	0.5891	0.4482	0.4422	0.6322	0.5219	0.4159
Default MTI	0.5704	0.5895	0.5797	0.5659	0.5490	0.5166	0.4813	0.5828	0.5585	0.4082
L2R-n1	0.5671	0.6490	0.5422	0.5702	0.6019	0.4328	0.4324	0.6470	0.5047	0.4129
L2R-n3	0.5659	0.6642	0.5312	0.5692	0.6144	0.4233	0.4259	0.6588	0.4960	0.4122
MTI First Line Index	0.5605	0.6213	0.5368	0.5557	0.5803	0.4745	0.4558	0.6132	0.5162	0.3994
Antinomyra SYS3	0.5346	0.5427	0.5409	0.5257	0.6215	0.2530	0.2652	0.5421	0.5272	0.3709
Antinomyra SYS4	0.5345	0.5401	0.5447	0.5262	0.5838	0.2791	0.2885	0.5399	0.5292	0.3712
EO_Sys1	0.5270	0.5647	0.5104	0.5099	0.5234	0.2790	0.2655	0.5473	0.5083	0.3567
Limited sample	0.5214	0.5629	0.5008	0.5030	0.5111	0.2701	0.2562	0.5450	0.4997	0.3507

Figure 2.3: The ranked list of participants from the BIOASQ Platform for the first test set of Task A of the second challenge.

Figure 2.4 presents the ranking of the systems on the BIOASQ Platform for a single batch of phase A. The systems are ranked based on the MAP measure and several other measures are presented for completeness.

2.3.3 Evaluating a system after the end of a challenge

Beyond the support of the challenges themselves, the BIOASQ Platform offers its users a way to evaluate their systems in an off-challenge mode. One way to achieve this is the weekly announcement of new test sets for task A, which is done continuously. A different service provided by the Platform is through the BIOASQ Oracles (http://participants-area.bioasq.org/oracle/), one for each task. The BIOASQ Oracles constitute a sustainable way of evaluating off-challenge system submissions for the released BIOASQ datasets. Using the Oracles the participants can submit results for already released test datasets and receive as feedback:

1. the scores of their system according to the corresponding evaluation measures, and
2. the ranking of their system compared to the systems that participated in the official part of the challenge and other submissions that were made public through the Oracles.

Figure 2.5 depicts the Oracle submission form, where the user provides information about the requested submission.

An Oracle is provided for each of the two challenge tasks, the evaluation is automated and the results are provided in real time. Obviously, the manual scores for task B are not provided. The scores

Test batch 1

Documents

System Name	Mean precision	Mean Recall	Mean F-Measure	MAP	GMAP
SNUMedinfo1	0.0658	0.6114	0.1131	0.2794	0.0773
SNUMedinfo3	0.0661	0.6155	0.1136	0.2762	0.0744
SNUMedinfo2	0.0655	0.6030	0.1125	0.2727	0.0686
SNUMedinfo4	0.0663	0.6154	0.1139	0.2675	0.0694
SNUMedinfo5	0.0667	0.6161	0.1146	0.2596	0.0662
Top 100 Baseline	0.2477	0.3994	0.2269	0.1766	0.0076
Top 50 Baseline	0.2495	0.3671	0.2275	0.1735	0.0062
main system	0.0447	0.2453	0.0713	0.1108	0.0018
Biomedical Text Ming	0.2595	0.1919	0.1678	0.1040	0.0017
Wishart-S2	0.1223	0.1248	0.0837	0.0601	0.0002
Wishart-S1	0.1447	0.1082	0.0858	0.0543	0.0002
UMass-irSDM	0.0275	0.0558	0.0339	0.0271	0.0002
All-Figdoc-UMLS	0.0275	0.0558	0.0339	0.0087	0.0002
All-Figdoc	0.0260	0.0528	0.0319	0.0082	0.0001

Figure 2.4: The ranked list of participants in the BIOASQ Platform for the first batch of Task B (phase A) of the second challenge.

are rendered in tables among the other publicly available results. Following a successful submission, the users are also informed about the scores of their submission by e-mail. The users can choose to store the scores of their submissions in the BIOASQ database. They can also choose to make their scores public for further use by others.

2.4 BioASQ Tools

One of the key requirements towards supporting the easy development and extension of the BIOASQ benchmarks was the provision of a suite of tools that would allow non-computer scientists to easily create, maintain, update, evaluate and alter benchmark datasets for biomedical question answering. A considerable amount of thought and engineering, aligned with continuous feedback from the BIOASQ biomedical expert team, went into creating a simple yet complete suite of tools for this purpose. The results of this development are the BIOASQ annotation and assessment tools (**?**), which are presentedbelow.

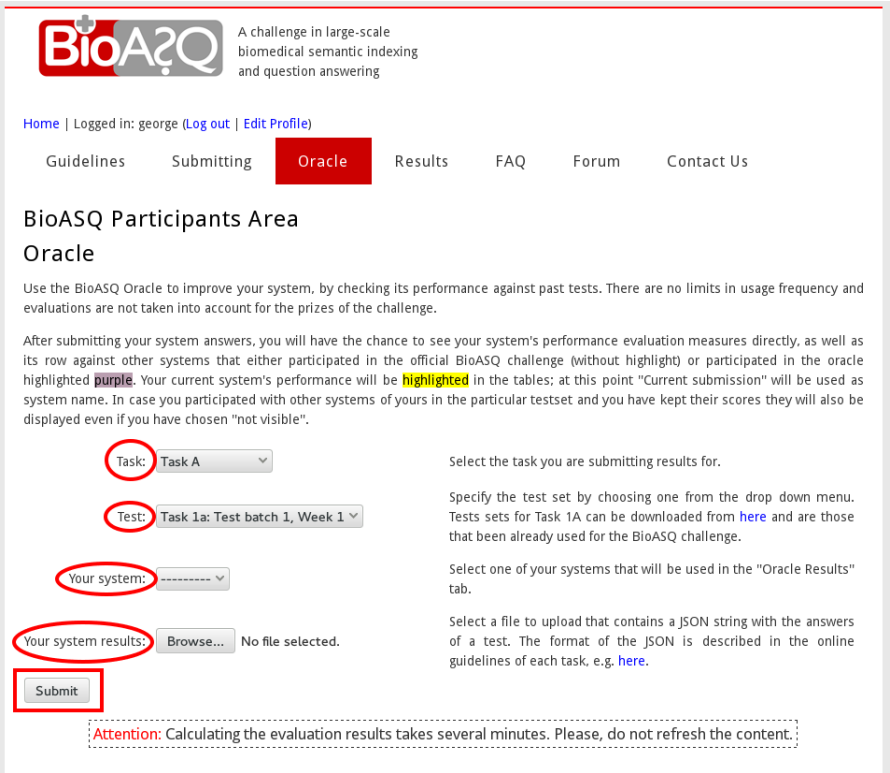


Figure 2.5: The oracle submission form, available at http://participants-area.bioasq.org/oracle/

2.4.1 BioASQ Annotation Tool

The main aim of the annotation tool, short BAT, was to support the creation of benchmarks for the BIOASQ challenges. To this end, the annotation tool had to provide the necessary functionality to create questions, gather and select information that allowed generating answers, as well as annotating relevant fragments of large documents that could be used to derive answers to the created questions. Two further premises governed the development of the tool as well as its interface. First, BAT was designed to be easy to use even for non-experts. Moreover, we assumed that a team of users is working on each benchmark. These premises led us to implement a simple five-step-paradigm: authenticate, search, select, annotate and store. The *authentication* ensures that each question created by a certain expert can be assigned to this given expert. After the authentication, the expert can choose to either work on a question he/she created in a prior session or to create a new one. The subsequent *search* window (see Figure 2.6) allows *selecting* information necessary to answer the question at hand.

To support potentially relevant sources of information, BAT implements interfaces to different search services of which each is connected to a different data source (see section 2.2.4). One of the main features of BAT is that it supports data sources of different types, i.e., unstructured, semi-structured or structured. Given that we could not expect domain experts to be familiar with Semantic Web standards such as RDF, BAT also implements an innovative natural language generation approach which allows converting RDF into natural language. The SPARQL2NL framework implements this technology and was presented in (**?**). The final steps of the benchmark (Fig. 2.7) creation consists of (a) creating an answer for the selected question manually, (b) annotating this answer with the selected information and

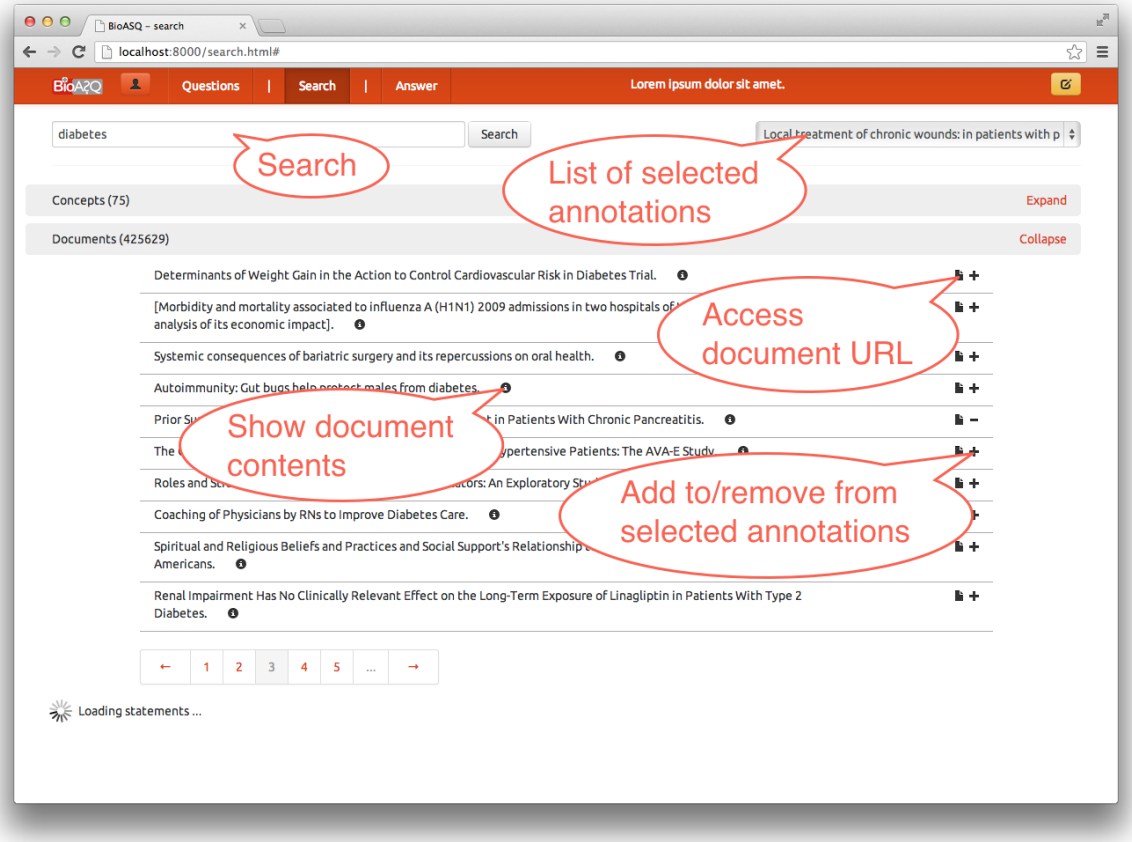


Figure 2.6: Search window. Users can search for information that can help answering the question they posed, as well as select the relevant results return by several search engines.

(c) storing the question in the persistent storage.

The iterative development of the annotation tool led to a framework that was widely accepted by the BIOASQ biomedical expert team. The user-driven evaluation conducted during the project suggests that the tool is easy to use. Furthermore, a study of the behaviour of domain experts when creating benchmarks led to key insights that will govern the creation of future benchmarks and the evaluation of tools in future campaigns. In particular, a study of the set of queries generated by experts to answer the same questions made clear that indeed "many roads lead to Rome". This suggests that semantic methods might be more useful than syntactic ones when evaluating question answering systems on unstructured data sources. Moreover, the limited use of RDF data during the annotation process suggests that the large number of duplications in structured data is still a hindrance towards their use in question answering by humans. Here, the improvement of existing natural-language generation techniques seems to be a valid avenue towards making this type of data more usable for humans. While the tool was designed especially for BIOASQ, generic design principles were adopted, which led to an extensible and portable framework that can now be used in most projects pertaining to annotating unstructured data with data from different types and sources.

The code as well as a manual of the annotation tool can be found at https://github.com/BioASQ/AnnotationTool. A video showing how to use the tool is part of the project showcase available at http://bioasq.org/project/showcase.

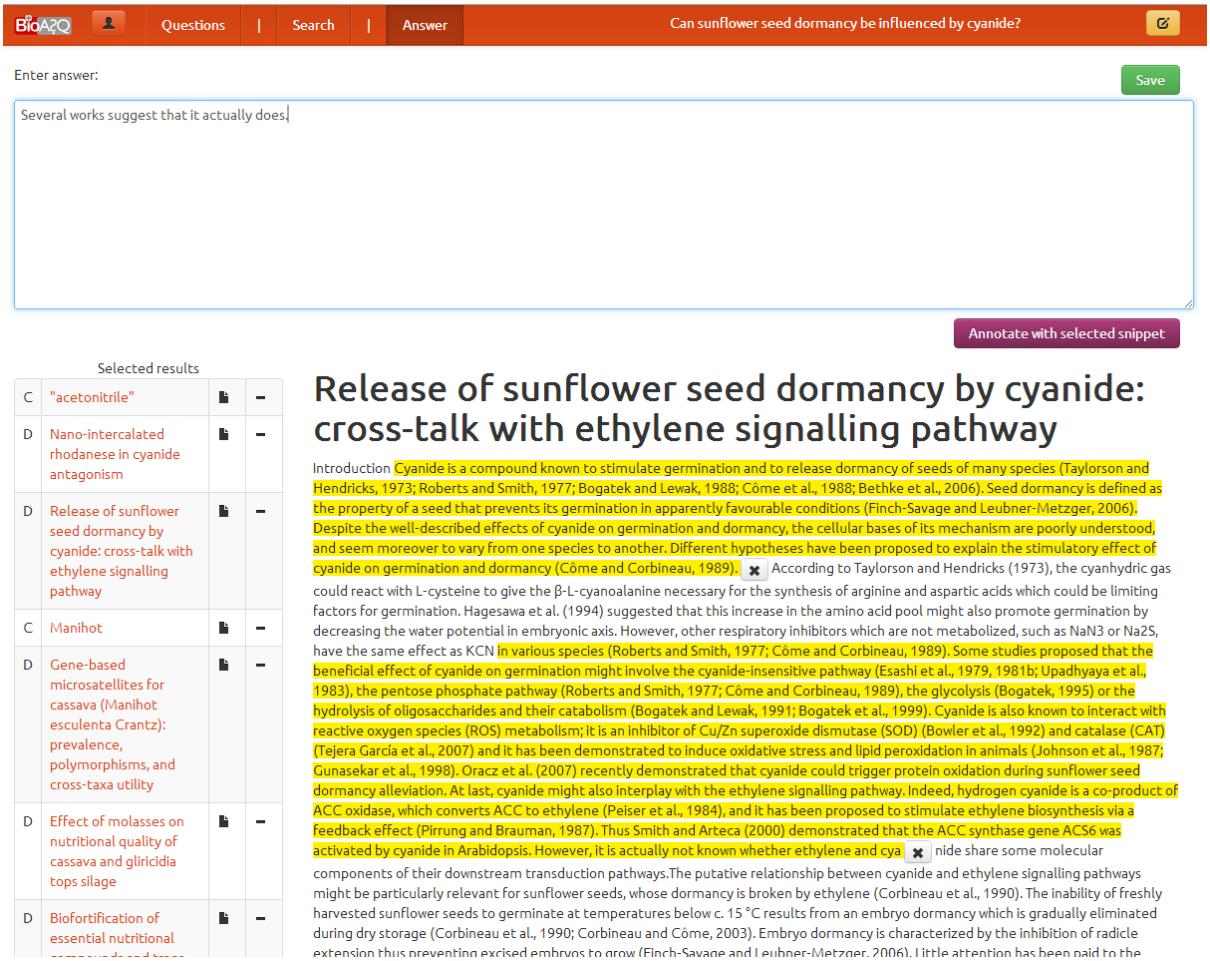


Figure 2.7: Annotation window. The selected snippets are marked in yellow. The list of results on the left gives an overview of the selected concepts, documents and statements used as information sources to create the answer. The question is seen on the top right.

2.4.2 BioASQ Assessment Tool

The assessment tool was designed to be a companion to the annotation tool as was implemented by reusing most of its functionality. The aim of the assessment tool was to enable the domain experts to evaluate the answer of other parties to the questions they generated. The tool can thus be used for two purposes:

1. It can be used for the manual evaluation of answers provided by participating systems.
2. It can also be used to perform an inter-annotator agreement study. In this case, domain experts are provided with answers generated by other (anonymous) domain experts and are asked to evaluate them.

The design of the interface is such that the users can always see the gold answers/annotations to questions that they are asked to review (see Fiugre 2.8). Moreover, the interface is dynamic and can adapt to different question types. Finally, all information sources that were used by an entity (tool, other expert,

etc.) to answer the question can also be reviewed. By these means, domain experts can perform an informed assessment.

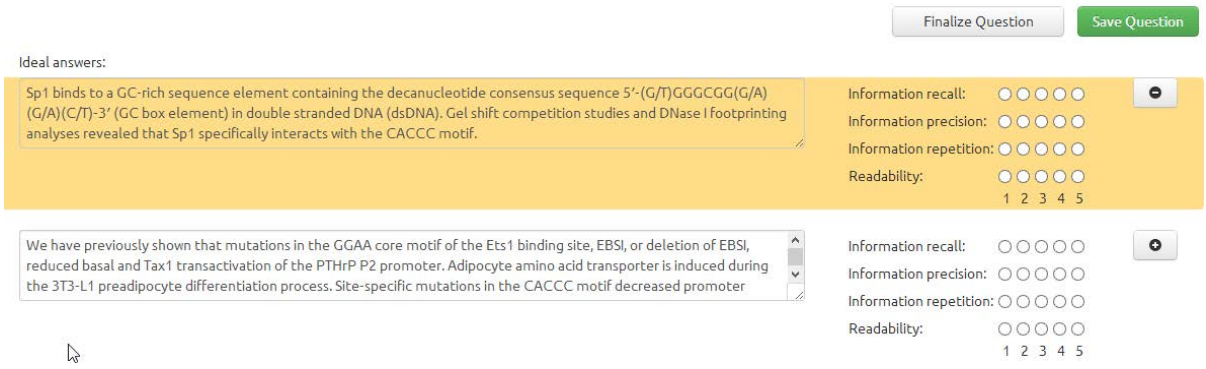


Figure 2.8: Annotation tab for system answers. The answer from the gold standard is at the top.

The assessment tool played a key role during both the benchmark creation and the quality assurance of the results generated by the experts during the BIOASQ challenges. Especially, we were able to confirm some of the behavioral patterns observed, based on the statistical and log data generated by means of the annotation tool. In particular, experts tended to mark the answers generated by other experts mostly as correct. However, the data used by different experts to get to the same answers was very commonly different. Moreover, the assessment tool allowed the experts to improve their own gold answers and associated material, based on that provided by the systems. This led to an improvement of the benchmark datasets that were provided publicly.

2.4.3 BioASQ Social Network

Once created, benchmarks are often regarded as monolithic, unchangeable resources. Yet, even manually curated complex benchmarks can contain errors which were not detected by the domain experts while creating the benchmark. We wanted to go beyond this view on benchmarks and provide both experts and the community the means to amend, extend and improve the BIOASQ benchmarks. To this end, we deployed a dedicated social network around the BIOASQ benchmark dubbed BISON. BISON is an asynchronous social network that allows not only humans but also benchmark entries to be agents. By these means, BISON allows members of the social network to subscribe to (i.e., follow), comment on and refer to questions (Figs. 2.10 and 2.9). Based on these commentaries, the experts in charge of the data can iteratively improve their questions to create updated versions of the benchmark. Note that the network also implements standard asynchronous network functionality as known from Twitter such as private messaging and following other experts.

BISON was used during the creation of the benchmark datasets for the second BIOASQ challenge and helped improve the quality of the data. Opening the framework to the community could be the key towards the development of a novel type of benchmark, i.e., community-driven, incremental benchmarks. Furthermore, it could lower considerably the costs of benchmark creation, through crowdsourcing.

2.5 BioASQ Workshops and Publications

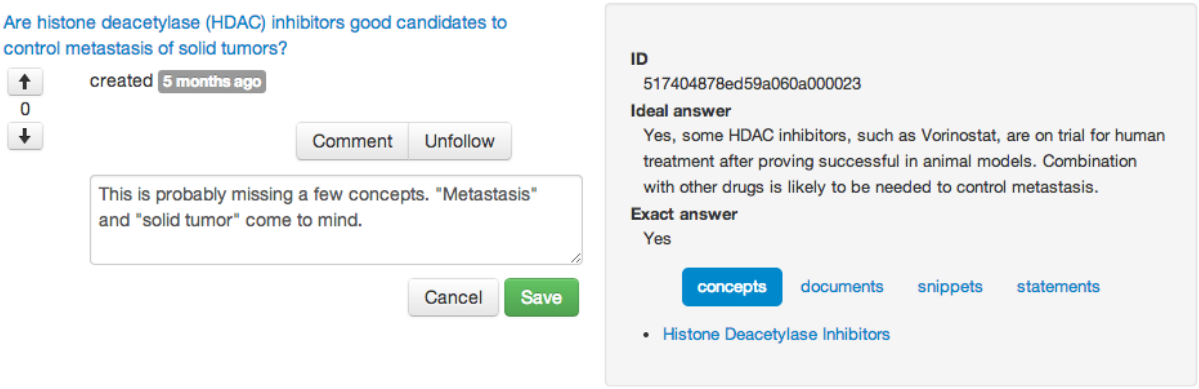


Figure 2.9: Question window in the social network. This window allows reading questions as well as their annotations, voting for the quality of the questions, as well as commenting on the questions.

2.5.1 Workshops

The BIOASQ workshops were organized as events that allowed the participants to present their results, as well as exchange expertise pertaining to biomedical question answering. The first workshop took place in Valencia (Spain) as a satellite event of the Cross-Language Evaluation Forum (CLEF 2013). The schedule of the workshop included 4 full paper presentations, 3 short paper presentations, 2 invited talks and the BIOASQ award session. The most interesting findings pertaining to the workshop suggest that approaches such as learning-to-rank still perform best for semantic indexing, while the performance of current question answering frameworks on BIOASQ Task B still left significant room for improvement (**?**). The best performing systems were awarded prizes sponsored by BIOASQ and Transinsight. More information about the winners can be found at http://www.bioasq.org/participate/first-challenge-winners.

The second workshop was integrated in the Cross-Language Evaluation Forum (CLEF 2014), which took place in Sheffield (UK). The workshop revealed the increased difficulty of Task A, as the accuracy of the baseline MTI system, provided by NLM, was improved by 4.5%, thank to the results of the first BIOASQ challenge.[4] This improvement did not stop the participating systems from outperforming the state of the art (**?**). Participation also increased, particularly in the harder Task B of the challenge. With the increase in the number of participants, several new approaches were presented in the 9 talks that made up the program. Again, prizes sponsored by BIOASQ and Transinsight were awarded to the winners. More information about the winners can be found at http://www.bioasq.org/participate/second-challenge-winners. Figure 2.11 provides highlights from the award session at the workshop.

2.5.2 Journal special issue

In order to capture the state of the art after the second BIOASQ challenge, a supplement of the Journal of Biomedical Semantics.[5] was organised. The text of the call for papers can be found at http://bioasq.org/project/bioasq-special-issue. The primary aim of this special issue is to enable the teams that participate in BIOASQ to present

[4] http://www.nlm.nih.gov/news/indexer_challenge.html
[5] http://www.jbiomedsem.com/

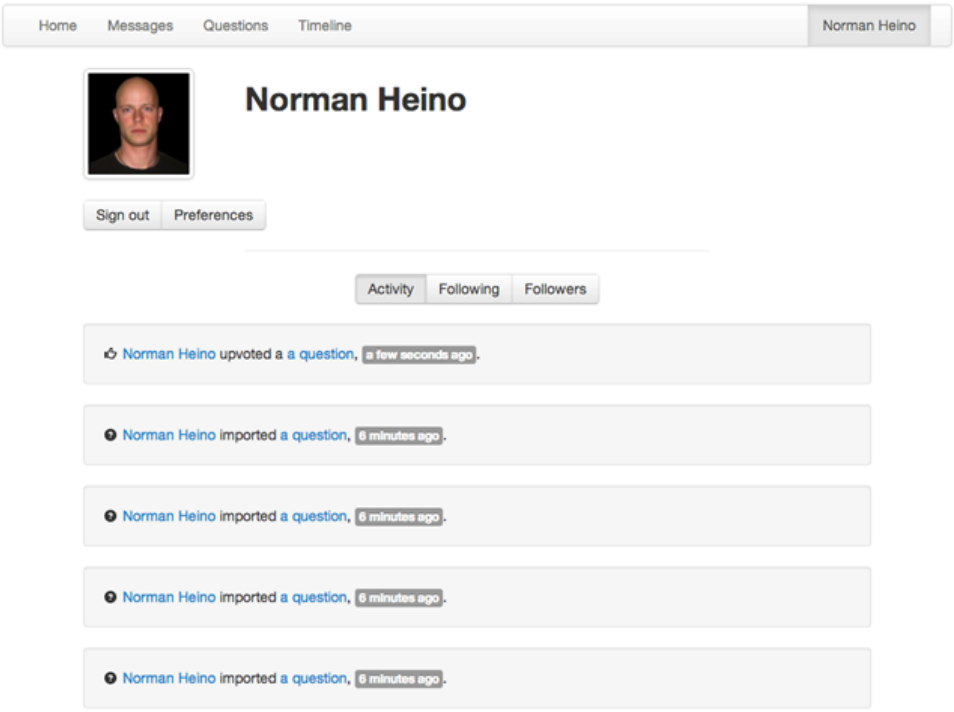


Figure 2.10: Fragment of a user window.

the core scientific improvements that they have achieved (**?**). Yet, the call was designed in a more open manner and requested contributions from any group/individual who are contributing to the state of the art in areas of research related to BIOASQ. Overall, we expect contributions from areas as diverse as large-scale hierarchical text classification, textual entailment and concept-driven data transformation. The common ground of these contributions is their interest in addressing the problem of biomedical question answering. The diversity of the relevant areas indicate the complexity of the task.

2.5.3 Publications and public talks

Despite it being a fairly new initiative, BIOASQ has become well known in the relevant community. This was achieved by a combination of various dissemination activities of the BIOASQ consortium. Among these, many targeted the research community, where the challenge participants come from. During the course of the project, members of the BIOASQ consortium presented BIOASQ in more than 30 events. Overall, the talks can be divided into three main categories:

- A large number of talks were related to publications. Although BIOASQ was not a research project, it engendered a considerable body of publications that mostly presented the insights won by the consortium while building the infrastructure required to run the BIOASQ challenges. For example, the verbalization technology of BIOASQ was presented at the prestigious World Wide Web conference (**?**). Furthermore, the BIOASQ tool suite was the subject of a presentation and a demo at the well-known Language Resources Evaluation Conference (LREC) (**?**). Careful studies of the measures used to evaluate the tools were presented in papers such as (**?**). A list of publications related to BIOASQ can be found at http://www.bioasq.org/project/public_documents.

Figure 2.11: Prizes awarded to winners of the second challenge.

- The consortium also gave *dissemination talks* at relevant international events and conferences. In particular, members of the consortium presented the project at the Extended Semantic Web Conference, the World Wide Web Conference, the European Data Forum, the European Conference on Information Retrieval, the AAAI Fall symposia and many others.
- Finally, due to the impact of BIOASQ in the community, a number of *invited talks* were also given by members of the BIOASQ consortium. The venues of such invited talks included the the Language Resources Evaluation Conference (LREC), the National Library of Medicine and the ISO15926 and Semantic Technologies meeting. An important type of invited talk concerned talks at Summer Schools, such as the Reasoning Web Summer School 2014, where we presented some of the insights obtained from running the challenges to European scholars and PhD students.

In order to reach the widest possible audience, the various announcements and calls for participation to the challenges and workshops were distributed to a number of scientific mailing lists. This has led to the impressive number of registrations in the BIOASQ Platform, which itself has now led to the formation of a much more focused audience for the BIOASQ announcements.

In addition to the research community, BIOASQ has reached out to the wider public through a series of press releases, most of which were translated into 4 languages: English, French, German and Greek. Following the end of the project, a special article is being prepared about the results of the project, to be published in one of the researcheu magazines. Finally, the presence of BIOASQ in the social media, particularly on Twitter (@BioASQ account) has also helped in communicating the results of the BIOASQ challenges.

3

Impact

Due to the organisation of the challenges and the associated workshops, the impact of BIOASQ in the research community is measurable in a number of different ways. The following sections provide a summary of these findings. Additionally, the plans for exploiting the results produced so far, as well as the potential continuation and expansion of BIOASQ are discussed.

3.1 Mobilization of the related research community

As explained in section 2.5.3, the emphasis of the dissemination efforts of the BIOASQ consortium has been on promoting awareness about the new series of challenges in the research community and attracting participation. To this end a number of announcements and presentations were made in different fora. The result was a stable increase in the number of people following the developments of BIOASQ and participating. Figure 3.1 provides various facts about the participation in the challenges and the visibility of BIOASQ on the Web and on Twitter.

The increase in the number of participants between the first and the second challenge indicates a very positive momentum which will help establish BioASQ as a reference point for biomedical semantic indexing and question answering. It is worth noting that in the complex Task B of BIOASQ the number of teams that participated has more than doubled in the second year. Furthermore, there is an even distribution of teams in three continents, despite the fact that BIOASQ was a European initiative and both BIOASQ workshops took place in Europe.

The visibility of the BIOASQ Web site and the BIOASQ platform, according to Google Analytics, is also quite impressive, as they had several thousands of unique visitors and tens of thousands of pageviews in the two years of the project. The geographic spread here has a clear European bias, due to the 6 European partners of the project and their affiliates. The number of followers of the @BioASQ Twitter account is also increasing, providing a direct communication channel for announcements related to the BIOASQ challenges. The presence of BIOASQ on Twitter has recently been boosted also by the announcement made about BIOASQ on the @nlm_news account, which is followed by more than 28 thousand Twitter users.[1]

[1] https://twitter.com/nlm_news/status/512260089517207553

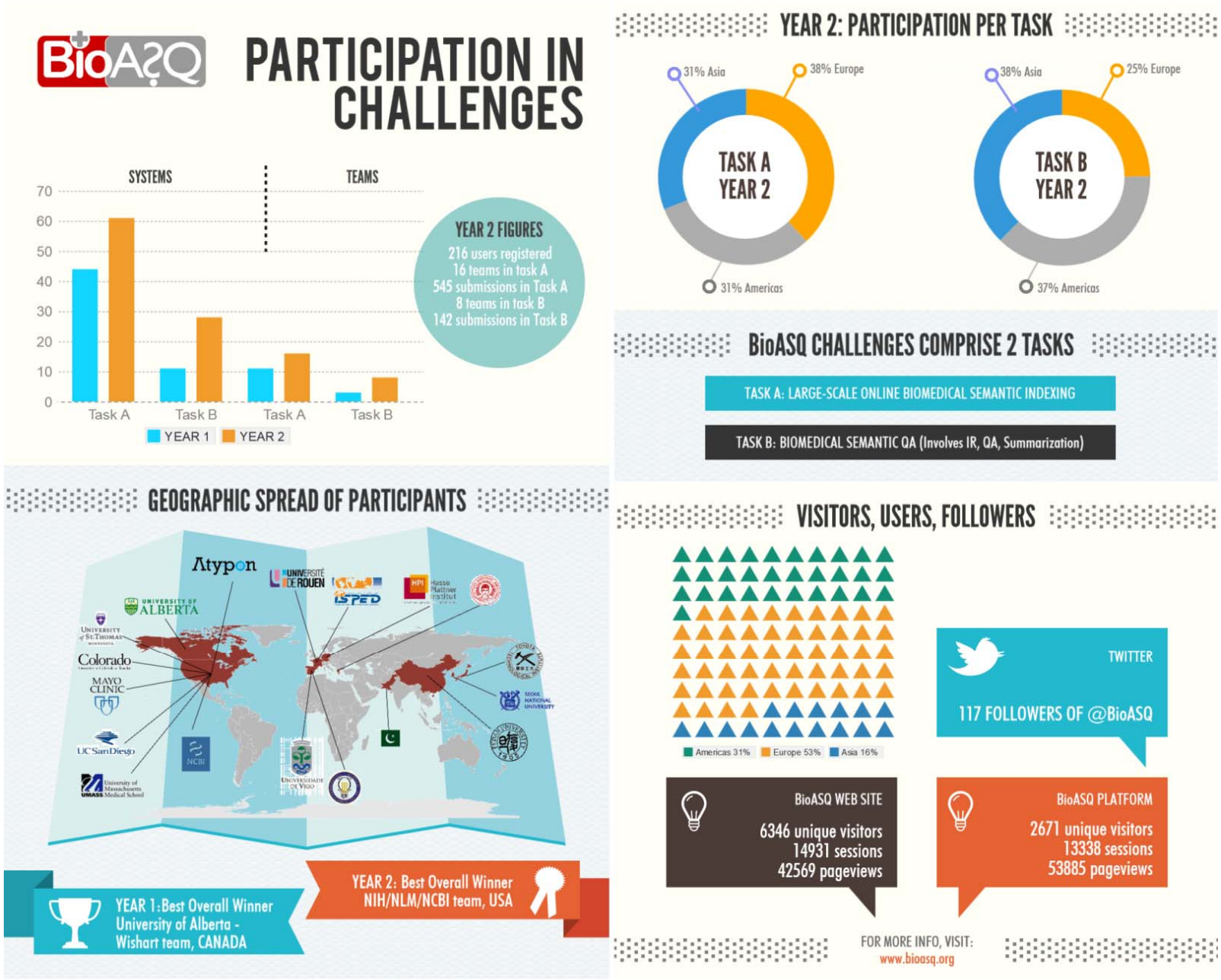


Figure 3.1: Facts about the participation in the BIOASQ challenges and the visibility of BIOASQ.

3.2 Improving the state-of-the-art performance

The participation of key players in the two BIOASQ challenges has helped establishing the state-of-the-art performance in biomedical semantic indexing and question answering in an objective manner. The continuation of BIOASQ, which is facilitated by the infrastructure and tools developed during the project, will certainly help push the state-of-the-art performance to new grounds. Comparing the results of the first two challenges, the improvement is already noticeable (Fig. 3.2). The most impressive result is clearly in Task A, i.e. large-scale biomedical semantic indexing, where we have observed a steadily increasing improvement upon the NLM Medical Text Indexer (MTI). This improvement has also helped already to improve MTI itself,[2] hopefully leading to more efficient and of better quality indexing of the biomedical literature. Improvements have also been noticed in the much harder biomedical question answering task, BIOASQ Task. It is particularly encouraging that the BIOASQ biomedical expert team have assessed the ideal answers provided by the participating systems as being very good. The average manual scores were above 4 out of 5 (above 80% in Fig. 3.2). However, there is still much room for improvement and the future challenges of BIOASQ, as well as the benchmark datasets that it provides, will hopefully push in that direction.

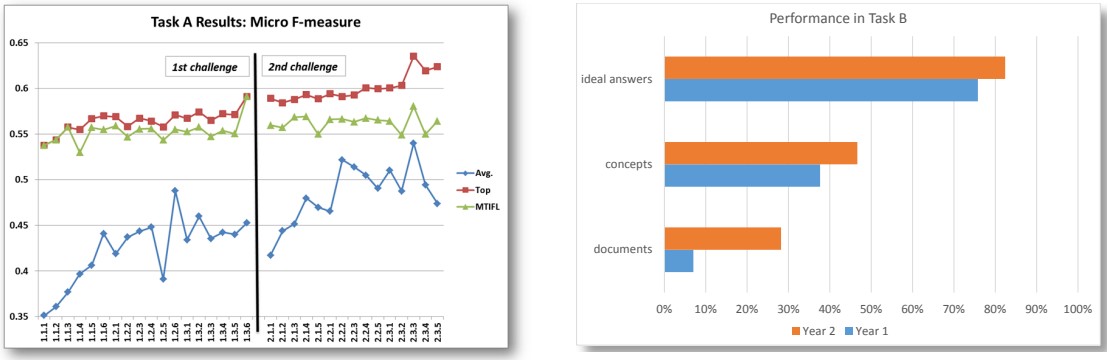


Figure 3.2: Improvement of performance in the two tasks of BIOASQ.

3.3 Setting benchmarks

BIOASQ aims to become a reference point for research in biomedical semantic indexing and question answering. For this purpose, it provides an opportunity for objective and comparative assessment of new methods in this area. The datasets created by BIOASQ were designed to be used as benchmarks from the start. In Task A, the semantic indexing task, BIOASQ builds upon the most widely used bibliographic resource in the area of biomedicine, namely MEDLINE, with the valuable assistance and support of the NLM. The benchmark datasets for Task A provide a representative subset of MEDLINE, chosen in a way that it facilitates realistic and objective assessment of the systems that participate in the task. It is for these reasons that BIOASQ has attracted the key teams working in the area of biomedical semantic indexing and has motivated new researchers to start working on this task.

The question answering (QA) benchmark datasets of Task B are also unique and particularly useful. To the best of our knowledge, they are they only ones that contain questions reflecting real information

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

needs of biomedical experts. They are also unique in that they combine both unstructured data (articles, snippets) and structured data (RDF triples, concepts from ontologies). Hence, they can be used to combine research on QA for structured and unstructured data, which are currently rather separate. Furthermore, unlike most previous QA benchmarks that include only 'exact' answers, the benchmark datasets of Task B also include 'ideal' answers (in effect, summaries), which are particularly useful for research on multi-document summarization. The gold documents and snippets of the benchmarks are also useful for Information Retrieval and Passage Retrieval experiments, whereas the gold RDF triples and concepts may be useful in triple retrieval and concept-to-text Natural Language Generation. Researchers working on paraphrasing and textual entailment may also wish to measure the degree to which their methods can improve the performance of biomedical QA systems, by helping match questions to differently phrased document snippets.

Most importantly, by making the BIOASQ challenge sustainable after the end of the project, the continuation of the effort and the enrichment of the benchmarks created so far is ensured. For Task A this enrichment is automated and continuous, as new test sets are announced every week. Furthermore, in the third BIOASQ challenge, we foresee at least as big a dataset for Task B as in the second challenge.

3.4 Educating biomedical experts

An equally important goal of BIOASQ was to bring the biomedical community in a more direct contact with the computer science teams working to help them with their information access needs. This coupling is of mutual benefit to the two communities. First, the computer scientists understand the real problems on which they need to focus their energy. At the same time, biomedical experts learn about the technologies that can help them find more efficiently the answers to their questions. This valuable coupling of the two communities was built into BIOASQ since its inception. During the project a sizable team of biomedical experts was directly involved in the shaping of the BIOASQ tools, the formation of the benchmark datasets and the assessment of the participating systems. These experts were selected among others for their ability to act as multipliers in their local or thematic communities. The long-term goal is to educate a larger population of biomedical experts, who may want to get involved in BIOASQ and help improve the tools that they will be using in the future.

The biomedical experts who have been involved in BIOASQ so far participated in a number of training sessions, which aimed at achieving the following:

- Familiarization with the annotation and assessment tools used during the formulation and assessment of biomedical questions respectively. This step also involved familiarization of the experts with the specific types of questions used in the challenge, i.e. factoid, yes/no, list and summary questions. Concerning the assessment task, experts were asked to evaluate the material returned by the participating systems in terms of several measures (information recall, precision, repetition and readability of the composed answer) and use the information provided by the systems to potentially enrich the benchmark data. At the same time, the experts commented on and contributed towards the shaping of the BIOASQ tools themselves.

- Familiarization with the resources used in BioASQ, both MEDLINE and various structured sources. The aim was to help the experts understand the data provided by theses sources, in response to different questions set by the experts.

- Resolution of issues that come up during the question composition and assessment tasks. This was a continuous process that extended beyond the training sessions themselves. Continuous support was provided to the experts through a dedicated mailing list, while the experts could also interact

with each other and provide feedback on the data being created, through the BIOASQ social network.

This core team of experts, who are now very familiar with the BIOASQ purposes and infrastructure, interact with a large network of other biomedical scientists, many of which are undergraduate and postgraduate students. Educating these younger scientists and involving them in the process of generating benchmarks and assessing the performance of information systems is the next important step for BIOASQ. Young scientists have a lot to offer and a lot to gain from being involved in this interaction:

- They have fresh ideas and information requirements that may not be evident yet.
- The volume of data that is generated could increase.
- They are more familiar with new technologies and will thus direct technology to new paths.
- They will learn about technologies that will become available in the near future.
- They will get involved in multi-disciplinary research.

This process is greatly facilitated by the BIOASQ tools. In particular, the BIOASQ social network incorporates user roles (senior and junior members) and provides rewarding mechanisms to those contributing the most to the data generation process. Using these tools, the BIOASQ team is already planning pilot studies with members of the biomedical expert team who are in academia.

3.5 Potential for commercialization

BIOASQ has adopted an open-source and open-data approach to the material created during the project. This was essential fro achieving the impact that was sought. Nevertheless, many of the results and by-products of BIOASQ could form the basis for innovative commercial initiatives. In particular the following directions for potential commercial exploitation of BIOASQ results and know-how have been identified.

Next generation search services 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, as well as 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 as baseline systems.

This technology can form the basis of the next generation semantic search engines and services. Key applications of these services can be efficient search in the biomedical domain, and also search in sources that are not 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. A key advantage of such a product would be 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. Another key advantage would be the direct access to novel ideas, as proposed by the challenge participants. A prominent example of such a success story is the improvement of the Medical Text Indexer (MTI) by NLM.

Taking semantic biomedical search to the next level, BIOASQ could also form the basis new services, such as the extraction of relations from structured and unstructured medical sources. Such services

could support automated hypothesis generation and validation. In order to achieve this, existing technologies will have to be enhanced with entity disambiguation techniques, relation extraction and learning techniques. The BIOASQ team comprises people with high expertise in these areas and is thus at a position to progress in the direction of automated hypothesis generation and validation in biomedicine.

A biomedical question answering engine An engine that can answer with high accuracy questions in the biomedical domain, would constitute a tool of high value and importance. The value of such an engine is acknowledged by key players, such as IBM Watson, who are investing in this direction[3]. The BIOASQ consortium has a strategic advantage in this market. The experience of the participating systems, and the associated technologies, could constitute a mature basis for implementing such an integrated engine, with the aim to automatically answer biomedical questions.

Much of the required infrastructure is already in place, e.g., services and indexes. Based on this infrastructure a strategic view of the overall product is needed, in order to give it a focus, tight enough to ensure high accuracy, but in tandem wide enough to attract attention and demand. The "drug-target-disease" principle, on which the BIOASQ question answering approach is based, allows a variety of hypotheses being formulated as natural language questions. Although, a generic biomedical question answering engine is unlikely to be accurate enough at this stage, the results of BIOASQ task B can direct us to sub-domains and question patterns in which the existing technologies can be sufficiently accurate.

3.6 BioASQ 3 and beyond

The infrastructure and tools developed in the duration of the project facilitates the continuation of the BIOASQ series of challenges at low cost. Therefore, the short-term goals include the organisation of the third BIOASQ challenge (BIOASQ3) in 2015. Task A already runs weekly in an off-challenge mode, fully automated, i.e. batches of newly published articles continue to be released, system responses are collected and evaluated against the MESH headings provided by NLM curators. For Task B, new questions will have to be formulated by biomedical experts, but the required effort and cost will be much lower compared to previous years, since all the tools to be used by the experts and the evaluation infrastructure are now fully developed. Experiments conducted during BIOASQ also indicate that automated evaluation scores of 'ideal' answers (based on BLEU and ROUGE) correlate well with manual evaluation scores provided by the experts. Hence, the time-consuming manual evaluation of 'ideal' answers may be unnecessary in future challenges, probably beyond BIOASQ3.

A medium-term goal is to improve and extend the BIOASQ challenge to better reflect user needs. Towards this direction, we interviewed biomedical experts to better understand how they currently search (e.g., what information they search for, where they search, how they search, what problems they encounter). The interviewees were the same people who had authored the questions and gold responses of the BIOASQ benchmarks for Task B. They had also participated in the manual evaluation of the responses of the participating systems. Hence, during the interviews we were also able to discuss the extent to which the experts thought that the BIOASQ challenge matched their needs, which types of required answers are most useful in practice and, more generally, how the challenge could better match their needs. The conclusions of the interviews, along with recommendations for future challenges are reported in the BIOASQ Roadmap and provide many ideas and recommendations for improving BIOASQ in the future.

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

Also in medium term, tighter collaboration with related efforts will be sought. Some collaboration has already started in the context of the CLEF QA lab, where three different challenges in the area of QA meet. Discussions in this context have led to ideas about new tasks within BioASQ, such as automated answering of biomedical student examinations and question answering from biomedical linked data. These tasks could be introduced gradually in future BIOASQ challenges, beyond BIOASQ3. Furthermore, discussions with the BIOCREATIVE and the BIONLP efforts have started and will hopefully lead to closer collaboration in the near future.

A longer-term goal would be to port BIOASQ to other scientific domains. Widely used document repositories (with a role similar to PUBMED) also exist, for example, in Economics and Social Sciences, where concept taxonomies (with a role similar to MESH headings) also exist (e.g., JEL codes).[4] In these domains, BIOASQ Task A would require newly published documents to be mapped to the correct concepts (e.g., JEL codes), much as in BIOASQ. Task B would require benchmark questions and gold responses (relevant documents, snippets, concepts, triples, 'exact' and 'ideal' answers) to be constructed with the help of domain experts. Another particularly interesting domain is European Law. The EURLEX repository provides free access to EU directives, regulations, decisions, international agreements, reports etc. and its documents are indexed with 6700 hierarchically organized EUROVOC descriptors.[5] The documents are in 24 languages, and the EUROVOC descriptors are also available in most European languages, which opens up also the possibility of organizing a multi-lingual challenge similar to BIOASQ, but for European law. Another possible domain would be European cultural heritage, where meta-data of collections throughout Europe are aggregated by the European Library and EUROPEANA.[6]

[4] Consult, for example, http://repec.org/, http://www.ssrn.com/, and https://www.aeaweb.org/econlit/jelCodes.php.

[5] See http://eur-lex.europa.eu and eurovoc.europa.eu/.

[6] See www.theeuropeanlibrary.org/ and http://www.europeana.eu/.

4

Further Information about BioASQ

BioASQ online

http://twitter.com/bioasq

http://plus.google.com/104709672946762321818

http://www.linkedin.com/groups/BioASQ-4801043?home=&gid=4801043

http://www.youtube.com/channel/UCLG0adw5SLQCcQIff5DQ8Ig

GitHub https://github.com/BioASQ

Contact Details of beneficiaries

Demokritos
National Centre for Scientific Research

TRANSINSIGHT

UNIVERSITE JOSEPH FOURIER
SCIENCES.TECHNOLOGIE.SANTÉ

UNIVERSITÄT LEIPZIG

UPMC Universite Pierre et Marie Curie Paris 6
SORBONNE UNIVERSITÉS

AUEB-RC
Athens University of Economics and Business Research Centre

Coordinator
George Paliouras
paliourg@iit.demokritos.gr
http://users.iit.demokritos.gr/~paliourg/
Tel. +30 210 650 3158

Michael Alvers
malvers@transinsight.com
http://transinsight.com/
Tel. +49 351 796 5780

Eric Gaussier
eric.gaussier@imag.fr
http://ama.liglab.fr/~gaussier/
Tel. +33 6 821 979 88

Axel-Cyrille Ngonga Ngomo
ngonga@informatik.uni-leipzig.de
http://aksw.org/AxelNgonga.html
Tel. +49 341 973 2341

Patrick Gallinari
patrick.gallinari@lip6.fr
http://lip6.fr/Patrick.Gallinari
Tel. +33 1 447 273 70

Ion Androutsopoulos
ion@aueb.gr
http://www.aueb.gr/users/ion/
Tel. +30 210 820 3751