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

Project **FP7-318652 / BioASQ**

Deliverable **D4.4**

Distribution **Public**

http://www.bioasq.org

Report on challenge operation and technical support 1

Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Thierry Artieres, Eric Gaussier and Patrick Gallinari

Status: Final (Version 1.0)

October 2013

Project

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

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D4.4
Deliverable title	Report on challenge operation and technical support 1
Contractual date of delivery	M13 (October 2013)
Actual date of delivery	October 2013
Relevant Task(s)	WP4/Task 4.2
Partner Responsible	UPMC
Other contributors	UJF
Number of pages	14
Author(s)	Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Thierry Artieres, Eric Gaussier and Patrick Gallinari
Internal Reviewers	Axel Ngonga Ngomo
Status & version	Final
Keywords	BioASQ, platform, challenge operation

Executive Summary

The deliverable describes the first year of the BIOASQ challenge as it was monitored in BIOASQ Participants Area[1]. The BIOASQ Participants Area is an online platform that was developed to provide the necessary functionality for data exchange, evaluation and participant support during the BIOASQ challenge.
During the first year of the BioASQ challenge, a series of test datasets pertaining to both Task 1A[2] and Task 1B[3] were released by the means of the platform. Participants downloaded the datasets and responded with the required answers and results that were produced by their systems. The system answers were evaluated against correct, human answers and based on this evaluation the winners of the first year of BIOASQ challenge were announced.
This deliverable contains detailed information and statistics with respect on the following:

- The datasets for both tasks of the challenge for the first year of the challenge,
- The participation in both tasks of the challenge,
- The support that was provided from the BIOASQ team in the participants and
- The problems that occurred in the operation as well as the actions that were proposed to overcome them.

[1] http://bioasq.lip6.fr
[2] The Task 1A datasets are available online in http://bioasq.lip6.fr/Tasks/1a/
[3] The Task 1B datasets are available online in http://bioasq.lip6.fr/Tasks/1b/phaseB/

Contents

List of Figures

List of Tables

1

Introduction

BIOASQ initiated a series of challenges on biomedical semantic indexing and question answering. The motivation behind the challenge is to push for a solution to the information access problem biomedical experts face and concerns their difficulty to synthesize and filter quickly, accurate and specialized information that comes from large and fast-growing sources.
The project organised two tasks during its first year. During each task the BIOASQ team released test sets following a predefined and announced schedule. Participants were allowed to download the test sets and submit their results using the online BIOASQ Participants Area[1] (hereafter platform) within a limited time window. A short description of the tasks of the challenge follows. For a more detailed description, please visit http://www.bioasq.org.

Task 1A

Task 1A, entitled "Large scale online biomedical semantic indexing", deals with large scale classification of biomedical documents onto ontology concepts. It simulates the process that is followed in PubMed[2] by human curators. PubMed is a public, online database hosted in the US where new articles are uploaded in a daily basis and are annotated with concepts from the MeSH[3] hierarchy. The gap between the submission of an article in PubMed and its annotation is used by the BIOASQ team in order to release test sets that consist of non-annotated articles. Participants submit their system's estimations for the annotations of those articles. The evaluation of the participant systems is performed when the human annotations from PubMed become available. The articles and the deadlines for the test sets are selected in a way that prevents cheating and ensures a short annotation period.
There were nineteen test sets released during the first year of the challenge. Firstly, a "dry-run" test was released so that participants could familiarize themselves with the process of downloading the test set and submitting results in a limited time window. The rest of the test sets were split into three groups of six test sets each. The first official test set was released a week after the "dry-run" test set, on April 22nd, 2013. Figure 1.1 shows the schedule of Task 1A. The date of the first test set of each test group is

[1] The BIOASQ Participants Area is deployed under http://bioasq.lip6.fr
[2] http://www.ncbi.nlm.nih.gov/pubmed
[3] http://www.ncbi.nlm.nih.gov/mesh

marked in the figure. The task's last, official test dataset release was on the 26th of August, 2013.

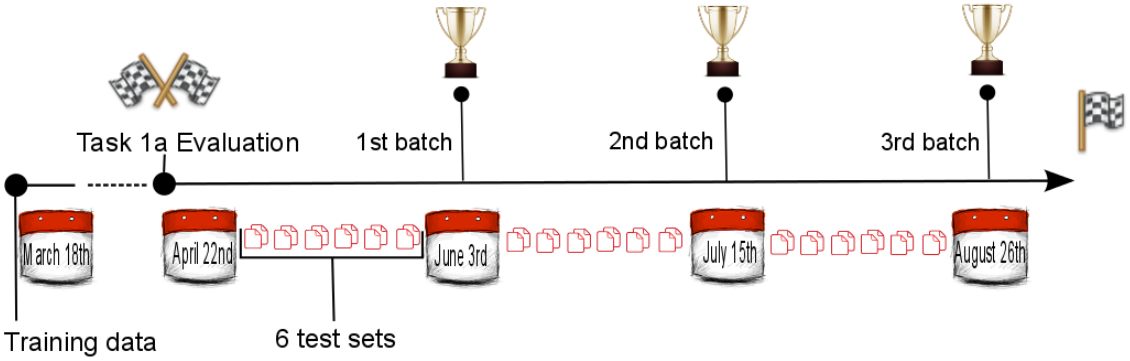

Figure 1.1: Task 1A schedule. Each goup of tests consists of six test sets. The date of the first test set of each group is marked in the figure.

Task 1B

Task 1B, entitled "Biomedical Semantic Question Answering", examines system's ability to annotate questions with concepts from relevant ontologies and return "exact" and paragraph sized, "ideal" answers. A network of ten experts around Europe was established to create a benchmark dataset of around 300 questions using an "Annotation Tool" developed from the BIOASQ team for this reason. The task was organised in two phases:

- Phase A: The BIOASQ team released questions from the benchmark datasets. The participating systems had to respond with relevant concepts from designated terminologies and ontologies, relevant articles in English from designated article repositories, relevant snippets from the relevant articles, and relevant RDF triples from designated ontologies.
- Phase B: The BIOASQ team released questions and gold (correct) relevant concepts, articles, snippets, and RDF triples from the benchmark datasets. The participating systems had to respond with exact answers (e.g., named entities in the case of factoid questions) and ideal answers (paragraph-sized summaries), both written in English. For the synthesis of the answers, using the provided gold annotations was sufficient. However, users were also allowed to use the annotations their systems estimated in Phase A.

The process of releasing test sets was repeated three times. The datasets for Phase B, were released after the expiration of Phase A. Figure 1.2 shows the schedule that was followed for Task 1B. In the figure, the dates the data for each phase were made available are marked. For example, the first test dataset concerning phase A of Task 1B was released on 26th of June, 2013. More information on the process followed during Task 1B can be found in the guidelines in Androutsopoulos et al. (2013). More information about the "Annotation Tool" is available in Ngonga Ngomo et al. (2013).

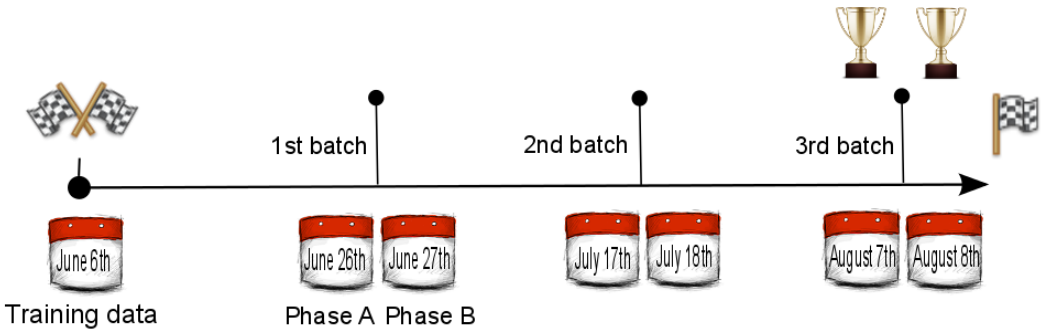

Figure 1.2: Task 1B schedule. Three test sets were released. The dates of both phases of each test set are marked in the figure.

2

Challenge operation

The chapter contains a description of the challenge from the scope of data and participation as it is recorded in the online platform. The platform integrates the necessary functionality the participants used during the challenge. In addition, it is where the test sets of both tasks of the challenge are released. As a result, monitoring its activity enables us to infer about the challenge operation. The chapter is organised as follows:

- the first section summarizes the functionality that is integrated in the platform,
- the second and the third sections describe the data and the participation in Task 1A,
- the fourth and the fifth sections describe the data and the participation in Task 1B and
- the sixth section provides information on the support of the users.

An overview of the traffic in the platform is presented in Figure 2.1 that comes from Google Analytics. There were around fifty unique visitors in the platform on Mondays that the test datasets of Task 1A were released.

Figure 2.1: Google Analytics: Traffic in the platform from the beginning of the challenge.

2.1 Functionality of the platform

The platform was designed to be user-friendly. A key concept was to make it simple enough so that participants wouldn't need much time to exchange data, receive support or check the performance of their

systems by browsing the evaluation measures. Another key concept was to provide ways to automate the process of downloading and submitting results. For this reason, web services were developed so that users could programmatically download the test sets and submit results saving time and effort.
A short description of the provided functionality follows:

- Registration: Participants can register to the platform by filling a simple form. Registered users gain access to the datasets, to the forum, to the BIOASQ announcements and to the result submission forms and web services.
- BIOASQ datasets: The datasets include training sets for each task of the challenge and test sets that are released periodically. They are served as JSON[1] (JavaScript Object Notation) strings which are platform-independent and human-readable. The datasets can be accessed via links in the platform and via web services. The datasets remain available in the platform after their expiration but submitting results is disabled.
- Results submission: Participants can submit their results for the active testsets by using a simple form in the platform or by using web services. Participants can submit results more than once for a particular system before the dataset deadline.
- Support: There are detailed, online guidelines that address the main points of the challenge. There is a BIOASQ Discussions Area where registered participants can discuss about the challenge. There is also a contact form that participants can use to contact the BIOASQ team directly.
- Evaluation results: Participants can browse tables that contain the evaluation measures for their systems and compare their system's performance with other users.

More information on the platform functionality is provided in Balikas et al. (2013a).

2.2 Datasets for Task 1A

In the context of Task 1A a large training dataset and eighteen test datasets, organised in three test groups, were released. The datasets are available both in text format and in a pre-processed format obtained with Apache Lucene framework[2]. More information on the pre-processed benchmark datasets and the format used for data exchange is available in Partalas et al. (2013). Table 2.1 provides information about the training dataset. The training dataset is a collection of more than ten million articles from MEDLINE published after 1949. The articles come from 8,915 different scientific journals. Each article has 12.55 MeSH labels in average.

Articles	10,876,004
Total labels	26,563
Labels per article	12.55
Size in GB	22

Table 2.1: Properties of the training data for Task1a.

The official test datasets of the BIOASQ challenge were released on Mondays starting from the 22nd of April, 2013. The articles of the test sets come from 1,805 pre-selected journals. The list of the journals in

[1] www.json.org
[2] http://lucene.apache.org/

available in http://bioasq.lip6.fr/journals/. The journals were selected by the BIOASQ team based on statistics about the annotation period of their articles. It was essential that the annotation period of an article was short, so that participants can receive feedback on their system's performance shortly after the submission of their results to improve it. Table 2.1 shows the size of each test, the number of the annotated articles during the writing of this deliverable and finally the average number of MeSH concepts that the annotators in MEDLINE gave in each of the annotated articles. The second test dataset for example consists of 845 articles. 701 out of 845 articles have been annotated with an average number of 11.56 MeSH concepts.

Testset	Articles	Annotated Articles	Labels per article
1	1,942	1,543	10.00
2	845	701	11.56
3	793	706	10.87
4	2,408	586	10.27
5	6,742	4,194	11.70
6	4,556	2,503	11.67
7	5,012	1,658	12.39
8	5,590	1,658	11.48
9	7,349	2,100	12.93
10	4,674	1,552	12.37
11	8,254	2,556	12.18
12	8,626	2,284	13.20
13	7,650	2,002	12.58
14	10,233	2,880	13.07
15	8,861	2,274	12.44
16	1,986	1,118	10.81
17	1,750	1,024	10.70
18	1,357	530	11.14
total	88,628	31,869	12.01

Table 2.2: Statistics on the test datasets of Task1a.

2.3 Participation in Task 1A

Table 2.3 shows the participation as it can be inferred from the submission of results for each test set. Participants of the BIOASQ challenge are allowed to participate in the challenge with a maximum of five systems. This decision was made in order to help research teams participate with more than a system, since they usually test multiple implementations of an algorithm or several algorithms at the same time. For example, for the second test of the challenge nine users submitted results. However, there are nineteen different result files, since most users submit results with more than one system. In addition, the second test was downloaded 28 times since its release. A team usually downloads the data more than once. For example downloading first the raw format and then their pre-processed description in Apache Lucene format would count as two downloads. The same result would be obtained if a team consists of two people who use the same identification in the platform and each one downloads the datasets.

Testset	Users	Systems	# of downloads
1	7	19	32
2	9	26	28
3	8	24	25
4	8	24	22
5	7	20	22
6	8	23	19
7	8	23	20
8	7	22	20
9	8	23	32
10	7	23	27
11	7	24	31
12	6	18	26
13	7	25	26
14	7	25	32
15	8	27	25
16	8	29	26
17	8	28	19
18	9	33	16

Table 2.3: Statistics on the test datasets of Task1a.

Table 2.4 shows the username, the affiliation and the number of the registered systems for each team that participated in Task 1A. In total, twelve different teams submitted results, at least once, for the challenge. For example, one of the teams that participated is "Kota" and comes from the Toyota Technological Institute. "Kota" has registered five systems for Task 1A. Regarding the origin of the teams that participated in Task 1A, the majority of the teams come from U.S.A.

Team username	# of registered systems	Affiliation
chyc	3	Fudan University, China
jgmork	2	U.S. National Library of Medicine, USA
tsoumakas	4	Aristotle University of Thessaloniki, Greece
abdedsai	5	Universit de Rouen, France
wishart	5	University of Alberta, Canada
ansonkahng	4	UCSD, USA
imran	1	not specified
chris_funk	4	University of Colorado, USA
fribadas	5	University of Vigo, Spain
maoy	5	NCBI, USA
Kota	5	Toyota Technological Institute, Japan
mcteam	5	Mayo Clinic, USA

Table 2.4: Affiliations of the teams that submitted data for Task1a.

2.4 Datasets for Task 1B

The datasets for Task 1B of the BIOASQ challenge were created by a team of ten biomedical experts around Europe. More information on the team of the biomedical experts can be found in Polychronopoulos et al. (2103). The biomedical experts formulated questions depending on their field of specialization and analysed them by producing annotations, exact and ideal answers. The questions were created online using an "Annotation Tool" and an "Assessment Tool" the BIOASQ team developed for that reason. The production of the final version of the datasets that are used for the evaluation of the systems involves two steps:

- First, the experts formulated the questions they considered interesting and used the "Annotation Tool" to annotate them using documents, snippets, concepts and triples from designated ontologies. That was the first version of the golden dataset.
- After the first phase of the challenge where systems were requested to return documents, snippets, concepts and triples, the experts used the "Assessment Tool" to update the first version of the golden dataset. This final version is updated with annotations that the experts may had lost during the first round of annotation. During the assessment of the system responses of the first round, the experts also evaluated the ideal answers of the systems as described in Deliverable Balikas et al. (2013b) in order to be used for the official evaluation of the ideal answers.

More information on the process followed for the creation of the questions and on the "Annotation Tool" is available in Ngonga Ngomo et al. (2013). More information on the biomedical resources is available in Tsatsaronis et al. (2013). Table 2.5 provides information on the dataset the experts created for the task.

Dataset	Size	Avg. # of documents	Avg. # of snippets	Avg. # of concepts	Avg. # of triples
training	29	10.31	14.00	4.82	3.67
1	100	14.89	19.89	8.30	21.87
2	100	14.66	20.24	7.58	5.56
3	82	14.47	17.06	6.24	4.50
total	311	14.28	18.70	7.11	9.00

Table 2.5: Statistics on the test datasets of Task1B.

From the table we can see that the experts produced 311 questions in total. Those questions were semantically annotated with documents, snippets from those documents, concepts and triples. The average number of concepts and triples is based on the questions that have this type of annotations since experts could not locate them for every question. In contrast, documents and snippets were used to annotate every question.

2.5 Participation in Task 1B

Tables 2.6 and 2.7 show the participation in phase A and phase B of Task 1B respectively. Again, users could participate in the challenge with more than one systems to test and compare their methods without having to create different accounts in the platform. The participation in this task was lower than in Task 1A, since this task was more difficult and demanding. However, from the number of downloads of the first datasets we can infer that the task was interesting since many teams downloaded and inspected the

data. The low participation and download activity on the third dataset probably occurred due to the fact that it was released during August (the schedule of the Task is available in Figure 1.2), when many of the participating teams were on vacation.

Testset	Users	Systems	# of Downloads
1	2	3	30
2	2	4	13
3	1	2	5

Table 2.6: Participation in phase A of Task 1B.

Testset	Users	Systems	# of Downloads
1	2	4	23
2	2	5	10
3	1	2	3

Table 2.7: Participation in phase B of Task 1B.

Table 2.8 shows that affiliations of the users that participated in Task 1B. From those teams, "Wishart" participated in both phases of the Task 1B, "mcteam" in the first phase and "Kota" in the second phase. These particular teams participated also in the first task of the challenge.

Team username	# of registered systems	Affiliation
wishart	5	University of Alberta, Canada
Kota	5	Toyota Technological Institute, Japan
mcteam	5	Mayo Clinic, USA

Table 2.8: Affiliations of the teams that submitted data for Task1a.

2.6 Providing support to the users

2.6.1 The Guidelines

In order to help participants understand the requirements and the process of the challenge the BIOASQ team composed detailed guidelines. The guidelines for participating in the challenge are available online under http://bioasq.lip6.fr and no registration is required. The guidelines cover different topics concerning the participation in the challenge:

- Registration. The process is automated, after filling a form the participant receives a confirmation e-mail with an activation link. Clicking on the link registers the participant giving him access in the full functionality of the platform.
- The tasks of the challenge. There are separate guidelines with respect to the two tasks of the challenge. The provided information covers:

- the schedule,
- the data sources with statistics when there are available,
- the evaluation process,
- the provided tools,
- the benchmark datasets that are released during the challenge and their format,
- the format of the system answers and
- code snippets written in Python, that implement the data exchange between a system and the BIOASQ platform using the provided web services.

2.6.2 Participation in the BIOASQ Discussions

BIOASQ Discussions is a component integrated in the platform under http://bioasq.lip6.fr/forum/. It consists of forums where registered users can discuss the problems they face in the challenge, contact the organisers or ask for other participants' opinions. BIOASQ Discussions was organised in four forums, each one with a different topic:

- BIOASQ General,
- BIOASQ-Task 1A,
- BIOASQ-Task 1B/Phase A and
- BIOASQ-Task 1B/Phase B.

Figure 2.2 shows the main page of the BIOASQ Discussions and the available forums. In total, 22 posts were made in the forums, mainly asking for clarifications on the schedule of the challenge and on the format of the data. In every case, the BIOASQ team replied fast and after taking into account the feedback from the participants updated the guidelines or produces the necessary documents to clarify each aspect of the discussed topics.

2.6.3 The contact form

The contact form that was available in the platform was used twice, from users that had a problem while registering in the platform. The administrators of the challenge helped the participants to register successfully in the platform and access the available resources.

Figure 2.2: The BIOASQ Discusison Area. There are four forums where participants can discuss about the challenge.

3

Support actions during the first year of the challenge

During the first year of the challenge there were a lot of decisions to be made between the partners of the BIOASQ consortium and much to be done in terms of software development. The efforts were focused on producing user friendly software and supporting the involved parties e.g. biomedical experts, that would use the software. In this chapter we present the main problems along with the support actions on behalf of the BIOASQ team. The occurred problems belong to the following axis:

- The development and documentation of the software. The development of the software followed strict deadlines as the beginning of the challenge was scheduled for the sixth month of the project. After developing the software, detailed documentation had to be available so that users had enough time for familiarizing themselves and experimenting.
- The creation and exchange of the benchmark datasets. The BIOASQ project initiates a series of challenges which are based on data exchange. The BIOASQ team releases data and systems respond with the appropriate answers. The main parts of this process that require effort involve producing the benchmark datasets, specifying their format and providing the resources to support participants.

In the rest of the deliverable we discuss the main problems that were faced.

3.1 Software testing

To avoid problems the developed software was tested by the members of the consortium extensively and "dry-run" tests were organised before the official beginning of the challenge. The "dry-run" tests were beneficial for both sides of the challenge:

- The participants became familiar with the process of the challenge, the data format, the deadlines and the ways of submitting data.
- On the other hand, the BIOASQ team had the chance to test the software in real conditions and improve its functionality and user-friendliness.

3.2 Guidelines

After producing the software and the tools, it was essential to write a clear and high-quality documentation explaining their use. As already mentioned, the BIOASQ team created detailed online guidelines. The guidelines were inspected by the members of the consortium and were updated every time a problem was reported. To this direction, the feedback from the BIOASQ Discussions was used, where participants reported their problems and difficulties. The guidelines will be elaborated and modified before the second year of the challenge.

3.3 Data format

The main problem that participants faced in the beginning of the challenge was to understand the JSON format that the BIOASQ team used for data exchange. Most of the teams were familiar with exchanging data using the XML format and had never used the JSON data encoders and decoders. As a result, the BIOASQ team updated and improved the guidelines by describing in depth the JSON format and also provided users with example JSON files. Those actions had significant effects on the participant's ability to use JSON strings and after the first test set no problem was recorded regarding JSON manipulation.

3.4 Creation of datasets

The BIOASQ team faced problems when creating the test datasets for phase A. Using the web services for creating the datasets, that are described in Balikas et al. (2013b), requires a date argument and a list of articles to be excluded from the dataset. The service works as follows: It queries the PubMed database for articles that

- have been uploaded in the database after the argument date,
- don't belong in the exclude list,
- belong to a certain list of pre-selected journals,
- satisfy a number of internal criteria in PubMed, such as they don't have comments or they are not republished.

Due to the fact that the criteria are many, and the complexity of the queries was becoming bigger week after week since more articles had been released as tests, managing to satisfy all proved tricky and required many tests and validations, before achieving the requested quality. In addition, the creation of the datasets depends on the availability of the web services of PubMed for interacting with the PubMed databases. There is always a high demand for use of those services and there are cases that are unavailable due to technical reasons.

Bibliography

I. Androutsopoulos, P. Malakasiotis, G. Tsatsaronis, M. Zschunke, and G. Balikas. Resources and Services for Task 1B. 2013.

G. Balikas, I. Partalas, N. Baskiotis, E. Gaussier, T. Artieres, and P. Gallinari. Evaluation Infrastructure. Technical Report D4.3, BioASQ Deliverable, 2013a.

G. Balikas, I. Partalas, A. Kosmopoulos, S. Petridis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, N. Baskiotis, E. Gaussier, T. Artieres, and P. Gallinari. Evaluation Framework Specifications. Technical Report D4.1, BioASQ Deliverable, 2013b.

A.-C. Ngonga Ngomo, N. Heino, R. Speck, T. Ermilov, and G. Tsatsaronis. Annotation Tool. Technical Report D3.3, BioASQ Deliverable, 2013.

I. Partalas, G. Balikas, N. Baskiotis, T. Artieres, E. Gaussier, and P. Gallinari. Pre-processed Benchmark Set 1. Technical Report D4.2, BioASQ Deliverable, 2013.

D. Polychronopoulos, Y. Almirantis, A. Krithara, and G. Paliouras. Expert Team. (D3.1), 2103.

G. Tsatsaronis, M. Zschunke, M. R. Alvers, and C. Plonka. Report on existing and selected datasets. Technical Report D3.2, BioASQ Deliverable, 2013.