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

Project **FP7-318652 / BioASQ**

Deliverable **D4.5**

Distribution **Public**

http://www.bioasq.org

Evaluation Framework Specification - 2nd version

Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Thierry Artieres, Eric Gaussier and Patrick Gallinari

Status: Final (Version 1.0)

December 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	D4.5
Deliverable title	Evaluation Framework Specification - 2nd version
Contractual date of delivery	M15 (December 2013)
Actual date of delivery	December 2013
Relevant Task(s)	WP4/Task 4.1
Partner Responsible	UPMC
Other contributors	UJF, AUEB, NCSR “D”
Number of pages	13
Author(s)	Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Thierry Artieres, Eric Gaussier and Patrick Gallinari
Internal Reviewers	Axel Ngonga Ngomo
Status & version	Final
Keywords	BioASQ, platform, challenge operation

Executive Summary

The document discusses the specifications of the BIOASQ Participants Area, publicly available under http://bioasq.lip6.fr. The BIOASQ Participants Area allows participants to register and participate in the evaluation procedures of the challenge. The deliverable is structured as following:

- First, we provide a short description of the existing technology in the BioASQ Participants Area as well as of the functionality that currently the platform provides to potential participants of the challenge. We highlight the main achievements of the 1st year's implementation and we also provide information on the technology and the architecture that was used. The description finishes with the presentation of the evaluation measures that are calculated in each of the phases of the challenge.
- The document continues by introducing the changes we will implement in the platform for the second year. Those belong to two different categories: (i) minor changes to the templates and the logic of the platform, and, (ii) implementation of subsystems for the second year of the challenge.

The main addition during the second year will be the implementation of the oracles. The oracles will guarantee the sustainability and the long-term exploitation of the BIOASQ facilities by the scientific community. Researchers will be able to eveluate their systems after the end of the BIOASQ challenge using the datasets that were created during the challenge. They will be able to submit results for the past tests of the challenge and get as feedback the calculated evaluation measures for the results they submitted. There will be no limits in oracle usage frequency. After submitting results they will be notified with an email for their performance and they will be able to see their ranking among the systems that participated in the official part of the challenge and those that submitted results in the oracle.

Contents

1 Introduction **1**
1.1 The second year of the BioASQ Challenge . . . 1
1.2 The online platform . . . 1

2 The BioASQ Participants Area **3**
2.1 Functionality for the participants . . . 3
2.2 Functionality for the BIOASQ team . . . 4
2.3 The architecture . . . 4
2.4 The evaluation framework . . . 6
2.4.1 The datasets . . . 6
2.4.2 The evaluation measures . . . 7
2.5 Planned additions . . . 9

3 The oracles **10**
3.1 The specifications . . . 10
3.2 A quick tour . . . 11

List of Figures

2.1 The platform homepage under bioasq.lip6.fr 4
2.2 The MTV pattern followed for the development of the platform. 5
2.3 The architecture of the platform. 6

3.1 A possible implementation of the BIOASQ form of the oracle. 12
3.2 A possible implementation of the BIOASQ results page. 12

List of Tables

2.1 Flat and hierarchical measures that will be used for assessing the participating systems in Task 2a. The measures that will be used for selecting the winners are in bold font. . . 7
2.2 Evaluation measures for Phase A of Task 1b. 8
2.3 Evaluation measures for the ‘exact’ answers in Phase B of Task 1b. 8
2.4 Criteria for the manual evaluation of the ‘ideal’ answers in Phase B of Task 1b. 9
2.5 Evaluation measures for the ‘ideal’ answers in Phase B of Task 1b. 9

1

Introduction

1.1 The second year of the BioASQ Challenge

The second year of the BIOASQ challenge will have two tasks, identical to the first year's tasks:

- Task 2a: Large Scale Online Biomedical Semantic Indexing: The task will examine the ability of systems to perform large scale multi-label classification. The systems will be evaluated on the whole of PubMed[1]. They will be asked to classify incoming sets of documents within a limited time window by attaching MeSH[2] terms on each one of them. Their performance will be assessed using the MeSH terms that the PubMed annotators select for each article.
- Task 2b: Biomedical Semantic Question Answering: The task will examine the ability of systems to perform semantic annotation, information retrieval, question answering and summarization. Given a question, systems will have to respond with relevant annotations, an exact and a paragraph-sized answer. Systems will be able to participate partially in the task. Systems that produce only exact answers for certain questions will not be excluded by the evaluation, they will be evaluated only on the answers they submitted.

More information on the tasks of the second year of the challenge is available in the official site of the BIOASQ challenge[3].

1.2 The online platform

The series of the BIOASQ challenges require the exchange of data between the challenge participants and the organizers. The developed mechanisms that cover those needs are integrated in an online platform, publicly available at http://bioasq.lip6.fr, the BIOASQ Participants Area. The BIOASQ Participants Area provides mechanisms for the participants to find information and support regarding the challenge as well as to participate in the tasks. On the other hand, the BIOASQ team can

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.ncbi.nlm.nih.gov/mesh
[3] http://bioasq.org

administrate the challenge, release the benchmark datasets and provide the necessary mechanisms that will allow the evaluation of the participating systems via the platform.

2

The BioASQ Participants Area

The BIOASQ Participants Area (hereafter platform) provides the necessary functionality for registered users (called participants) to enter the BIOASQ challenge. Figure 2.1 shows its homepage. It was originally developed during the first year of the BIOASQ challenge. Since then it has been used to support both tasks of the BIOASQ challenge. Throughout the first year minor changes have occurred to the first version to help the participants to understand the processes of the challenge. This section summarizes the main functionalities of the platform. The details can be found in Balikas et al. (2013b); Malakasiotis et al. (2013). For the development of the platform we used Django[1], which is a high-level Python Web framework.

2.1 Functionality for the participants

The platform provides a set of mechanisms to the registered users of the challenge. After subscribing to the platform, they gain access to the following:

- the BIOASQ benchmark datasets; they consist of training and test datasets which are available for downloading after their release,
- detailed guidelines describing the BIOASQ tasks,
- tools that have been developed to help participants process the datasets,
- mechanisms for submitting results; they include:
  - HTML forms available as long as there are active tests, and
  - Web services for submitting results in an automated way,
- tables for browsing the evaluation results,
- the "BioASQ Discussions Area", which is a forum about the BIOASQ challenge, and

[1] More information on Django on https://www.djangoproject.com/.

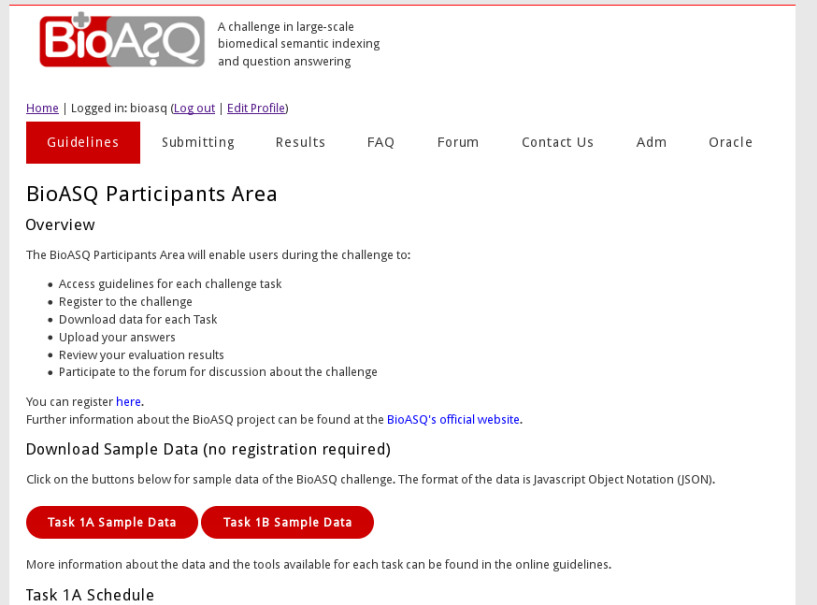

Figure 2.1: The platform homepage under `bioasq.lip6.fr`

- an e-mail help desk for contacting the organizing team.

The presented functionality has been integrated to the platform to the platform during the first year. For more information please consult Balikas et al. (2013a).

2.2 Functionality for the BIOASQ team

The platform has been developed to offer functionality not only to participants but also to the BIOASQ team, who are the administrators of the challenge. The administrators, using the platform can achieve the following:

- create datasets for Task 1a,
- release the datasets for both tasks,
- trigger the evaluation procedures to update the evaluation measures,
- monitor the challenge participation, and
- contact the participants using an e-mail list.

The datasets of Task 2a will be created following the process introduced for datasets of Task 1a of the challenge. For more information on this process please consult Balikas et al. (2013b). Apart from the desired behavior of participants, there are cases where certain actions can lead to errors. In those cases, a standard 500 page is displayed and administrators receive an email with information about the problem that occurred.

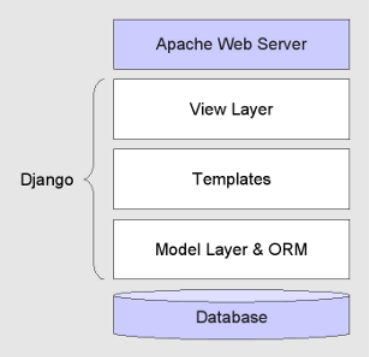

Figure 2.2: The MTV pattern followed for the development of the platform.

2.3 The architecture

One of the most popular paradigms in software engineering is the Model-View-Controller (MVC) pattern. During the development of the platform a slightly different pattern was followed called Model-Template-View (MTV) pattern and is proposed in the documentation of Django. Figure 2.2 presents the basic architecture of the MTV pattern. MTV separates the process of the development of an application into three layers:

1. The model layer provides an abstraction layer (the models) for structuring and manipulating the data of an application.
2. The template layer provides the rendering of the information that is presented to the user by the means of a designer-friendly syntax.
3. The view layer, which is responsible for encapsulating the logic behind the application and for processing a user's request and returning the appropriate response.

Each one of the layers is independent from the others. For example, the template layer, which describes the way that the data are presented to the user of the software, is separated from the model layer, which describes the way that the data are stored in the database. This loose coupling of the layers offers the developer the flexibility to adapt the application to the current needs and easily modify it according to future needs. For example, in case that the developing team is not satisfied with the user interface of an application or they are scheduling major changes in the way that the content is presented, changes would occur only in the template layer, leaving the others unchanged.
Figure 2.3 shows the architecture of the platform during the first year of the challenge. There are five subsystems, each following the MTV paradigm:

1. Task 1a
2. Task 1b-PhaseA
3. Task 1b-PhaseB

} core functionality of the challenge

4. Registration
5. Forum

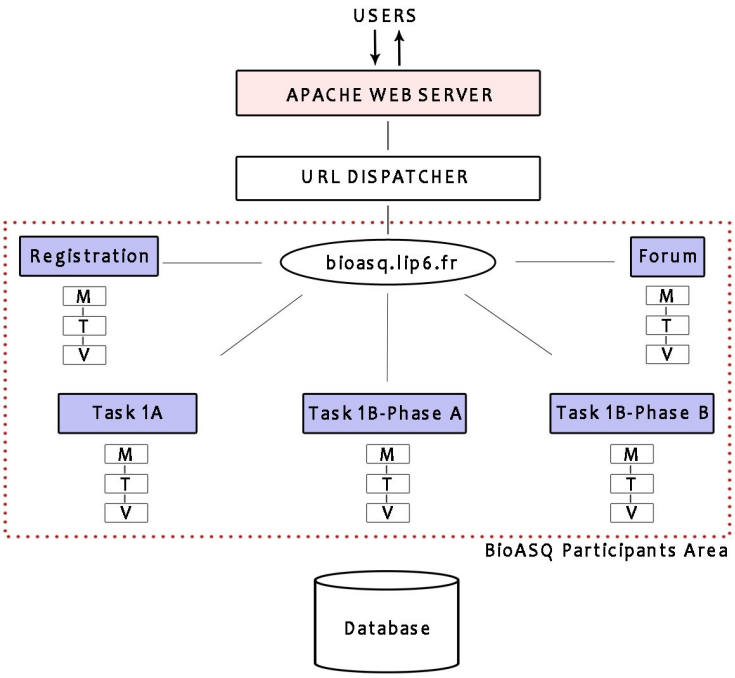

Figure 2.3: The architecture of the platform.

The three first offer the core functionality required for each one of the phases of the challenge while the other systems offer the functionality required for user registration and support. Each of the subsystems follow the MTV pattern in order to make it easier to make changes and extensions to the platform. More information about the architecture and the technical details of the implementation are available in deliverable D 4.3.

2.4 The evaluation framework

2.4.1 The datasets

The section discusses the benchmark datasets that will be provided during the challenge. Information is given with respect to the format and the information they will contain. Details about the data and the process of each task are also available in the platform in the "Guidelines" tab.

Task 2a

For task 2a the BIOASQ team will provide the following:

- training datasets,
- test datasets, and
- the MeSH terms, according to MeSH version 2014.

The training and test datasets will be provided in three formats:

- in raw text format, as JSON (Java Script Object Notation)[2],
- in a vectorized description, as an Apache Lucene index [3], and
- using another vectorized description, LibSVM[4].

The MeSH 2014 terms will be provided in two formats:

- a list with the MeSH heading and their indexing, and,
- a mapping in a parent-child format that is frequently used for hierarchies.

Task 2b

Concerning task 2b we will provide the following:

- training dataset, which will be the complete dataset of the first year's task 1b,
- test datasets, that will be created from the BIOASQ biomedical experts,
- mechanisms for interacting with predefined ontologies, described in Malakasiotis et al. (2013).

The datasets of Task 2b will be provided in raw text format as JSON strings.

2.4.2 The evaluation measures

There will be no changes on the evaluation measures that will be calculated during the second year of the BIOASQ challenge. Apart from the official evaluation measures that will decide the winners of the challenge, several other will be calculated for reasons of completeness. The following paragraphs present the evaluation measures for each of the tasks of the challenge. For more details, please consult Balikas et al. (2013b), where the measures are presented in detail.

Task 2a

In Task 2a, participant systems will be provided with a set of unclassified artciles. They will have to return the MeSH terms they estimated for each one of the articles of the test set. The winners will be selected based on their performance on Micro-F measure (flat measure) and Lowest Common Ancestor-F measure (hierarchical measure). For completeness, several other flat and hierarchical measures will be reported. Table 2.1 presents the measures that will be used during the second BIOASQ challenge.

Flat	**Hierarchical**
Micro precision, recall, **F-measure**, Example-based precision, recall, F-measure, Macro precision, recall, F-measure, Accuracy	Hierarchical precision, recall, F-measure, **Lowest Common Ancestor** precision, recall, **F-measure**

Table 2.1: Flat and hierarchical measures that will be used for assessing the participating systems in Task 2a. The measures that will be used for selecting the winners are in bold font.

[2] www.json.org
[3] http://lucene.apache.org/core/
[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Retrieved items	Unordered retrieval measures	Ordered retrieval measures
concepts	mean precision, recall, F-measure	*MAP*, ***GMAP***
articles	mean precision, recall, F-measure	*MAP*, ***GMAP***
snippets	mean precision, recall, F-measure	*MAP*, ***GMAP***
triples	mean precision, recall, F-measure	*MAP*, ***GMAP***

Table 2.2: Evaluation measures for Phase A of Task 1b.

Question type	Participant response	Evaluation measures
yes/no	'yes' or 'no'	**accuracy**
factoid	up to 5 entity names	strict and lenient accuracy, ***MRR***
list	a list of entity names	**mean** precision, recall, ***F*-measure**

Table 2.3: Evaluation measures for the 'exact' answers in Phase B of Task 1b.

Task 2b

Task 2b consists of two phases:

- Phase A (annotate questions, retrieve relevant articles, snippets, triples): In this phase, participants will be provided with biomedical questions written in English and will be asked to: (i) semantically annotate the questions with concepts from a set of designated terminologies and ontologies; and (ii) retrieve relevant articles, text snippets and RDF triples from designated article repositories and ontologies.
- Phase B (find and report 'exact' and 'ideal' answers): In this phase, the questions and golden responses of Phase A (correct concepts, articles, snippets, triples) will be provided as input. The participants will be asked to report 'exact answers' (e.g., named entities in the case of factoid questions) and 'ideal answers' (paragraph-sized summaries).

Phase A

The participating systems will be evaluated separately for each type of annotation they will retrieve. The official evaluation measure of the BIOASQ challenge for each type of annotation the systems will retrieve will be the geometric mean average precision (GMAP). For reasons of completeness more evaluation measures will be calculated. Table 2.2 summarizes the evaluation measures of Phase A; the official measures are shown in bold.

Phase B

Phase B requires two types of answers for each questions: 'exact' answers and 'ideal' answers. We first discuss how 'exact' answers will be evaluated in Phase B, by considering in turn yes/no, factoid, and list questions. For each kind of question different measures will be computed. Table 2.3 summarizes the kinds of responses and the evaluation measures that will be used in Phase B. Again, the official measures of the challenge are displayed in bold.

For each question (yes/no, factoid, list, summary), each participating system of Phase B will also have to return a single paragraph-sized text summarizing the most relevant information of the retrieved concepts, articles, snippets, and triples of Phase A. The returned 'ideal' answer is intended to approximate

Criterion	Explanation	Score
information recall	All the necessary information is reported.	1–5
information precision	No irrelevant information is reported.	1–5
information repetition	The answer does not repeat the same information multiple times.	1–5
readability	The answer is easily readable and fluent.	1–5

Table 2.4: Criteria for the manual evaluation of the 'ideal' answers in Phase B of Task 1b.

Question type	Participant response	Evaluation measures
any	paragraph-sized text	*ROUGE-2*, *ROUGE-SU4*, **manual scores**

Table 2.5: Evaluation measures for the 'ideal' answers in Phase B of Task 1b.

a short text that a biomedical expert would write to answer the question (e.g., including prominent supportive information), whereas the 'exact' answers are only 'yes'/'no' responses, entity names, or lists of entity names; and there are no 'exact' answers in the case of summary questions. The 'ideal' answers will be evaluated both manually i.e the BIOASQ biomedical experts will inspect a number of answers the systems retrieved, and automatically i.e by computing automated evaluation measures. Table 2.4 summarizes the criteria that will be used in the manual evaluation of the 'ideal' answers. Table 2.5 summarizes the evaluation measures of Phase B. The official measures are shown in bold. For more information on the evaluation measures please consult Balikas et al. (2013b).

-

2.5 Planned additions

Having presented the functionality and the architecture of the platform, we continue by presenting the functionality that will be added in the platform for the second year of the project. There are two groups of extensions planned:

- Extensions that are necessary in order to facilitate the challenge for the second year (e.g. extend the navigation menus), and
- Extensions that reflect the workflow of WP4 as described in the DOW document of the project and will add some extra functionality to the platform.

The first group of extensions will simply make some modifications to the templates and the logic behind the platform to prepare it for the second year. Extending the navigation menus and providing extra tabs for browsing the results of the second year are typical examples of the modifications needed.
The second group of extensions includes the oracle for the participants. The oracle is a mechanism that can be used from the participants of the challenge to improve their systems, by checking their performance against past tests of the challenge. From the architecture point of view the oracle will be an extra subsystem in the Figure 2.3. There will be no limits in usage frequency and the evaluations from the use of the oracle will not be taken into account for the prizes of the challenge. More information about the oracle is available in Chapter 3.

3

The oracles

One of the goals of the BIOASQ project is to create a sustainable challenge after the end of the project and also provide the means for researchers to test their methods. The platform of the challenge has been designed to be automated and user-friendly serving this purpose. However, the functionality currently offered to the users provides them means for short-term research only; researchers can submit results only while new tests are released and are active. The idea of the oracle is to enable researchers to test their systems after the end of the challenge in an off-challenge mode. Using the oracle, participants of the challenge can submit results for the past tests of the challenge and get as feedback the calculated evaluation measures for the results they submitted. There will be no limits in oracle usage frequency. After submitting results users will be notified with an email for them to be able to keep an archive of their performance.

3.1 The specifications

The oracle will be developed to provide users with functionality to test their systems and receive feedback about their performance outside the official BIOASQ challenge. They will be able to submit results for past BIOASQ test sets and receive the evaluation measures for those tests. The process that will be followed internally can be described as following:

1. The main content of the oracle homepage will contain a form. The form menus will be instantiated with the released BIOASQ test sets, so that the user can select a task and a test set.
2. After selecting them, he will choose a file from his computer containing a JSON string with the results.
3. The system will calculate the automated evaluation measures for each task and will display them to the user. The user will also be notified with an e-mail, so that he can keep an archive of his performance.
4. The system will be able to keep in the database up to one result per test and per system. After submitting the results and browsing his results a user will be able to decide if he will replace his

old results (if they exist). Apart from saving the results he will be able to select if he wants them public, so that other users of the platform can see the performance of his system when they submit results.

5. The results will be displayed in tables populated with evaluation measures of:
   - the systems that participated during the official challenge,
   - the systems that will have participated in the oracle and their users that will have selected to keep their performance public, and,
   - the systems that have submitted results for this testset and belong to the same user, even if they are not public.

Implementing the oracles for each task of the BIOASQ challenge and providing participants the ability to compare their performance with other system submissions will provide the functionality for improving their systems and achieving higher scores during the challenge. For the evaluation of the systems we will use:

- *Task 2a*: The released articles that are in the test sets receive MeSH terms manually from the annotatots in National Library of Medicine (NLM). This process is time-consuming as the annotation process is incremental. The BIOASQ database containing the annotations is updated periodically with the new MeSH terms. The evaluation measures will be calculated every time based on the available MeSH terms for the released testsets.
- *Task 2b*: The evaluation measures will be calculated using the gold datasets that the BIOASQ biomedical experts have created during the first year of the challenge.

3.2 A quick tour

The section is a quick tour over the proposed implementation for a better understanding of the idea. Figure 3.1 shows how the form could be designed. Selecting a system of the registered of each user is necessary because user can submit results with more than one systems. Imagine a user who uses two different implementations; he would like to be able to know which of his systems performs better in the biomedical field. Figure 3.2 shows how the results table can be implemented. The participant can browse over his current submission (highlighted yellow), his submissions for other systems (highlighted purple) and submissions from the official challenge as well as from other participants of the oracle with public results.

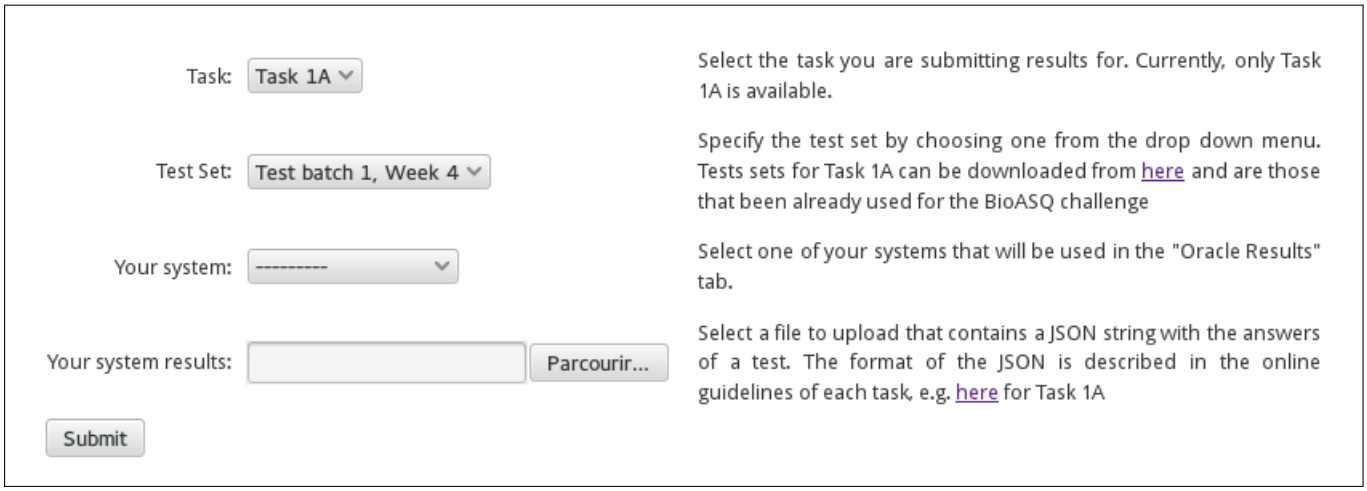

Figure 3.1: A possible implementation of the BIOASQ form of the oracle.

MCTeamSR8	0.2780	0.5580	0.4032	0.4464	0.3677	0.2354
UCD-CMgg	0.2630	0.3941	0.6156	0.4617	0.2171	0.3623
UCD-CMr	0.2623	0.3287	0.6886	0.4282	0.2059	0.4028
Current Submission	0.2246	0.3524	0.4176	0.3627	0.2210	0.2597
testing2	0.2246	0.3524	0.4176	0.3627	0.2210	0.2597
UCD-CMd	0.1606	0.6857	0.1534	0.2260	0.3793	0.1091

Figure 3.2: A possible implementation of the BIOASQ results page.

Bibliography

G. Balikas, I. Partalas, N. Baskiotis, T. Artieres, E. Gaussier, and P. Gallinari. Evaluation Infrastucture. Technical Report D 4.3, 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.

P. Malakasiotis, I. Androutsopoulos, Y. Almirantis, D. Polychronopoulos, and I. Pavlopoulos. Tutorials and Guidelines. Technical Report D3.4, BioASQ Deliverable, 2013.