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

Project **FP7-318652 / BioASQ**

Deliverable **D4.7**

Distribution **Public**

http://www.bioasq.org

Evaluation infrastructure software for the challenges 2nd version

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

Status: Final (Version 1.0)

March 2014

Project

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

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D4.7
Deliverable title	Evaluation infrastructure software for the challenges 2nd version
Contractual date of delivery	M18 (March 2014)
Actual date of delivery	March 2014
Relevant Task(s)	WP4/Task 4.2
Partner Responsible	UPMC
Other contributors	UJF, AUEB-RC
Number of pages	9
Author(s)	Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Thierry Artieres, Eric Gausier, Patrick Gallinari
Internal Reviewers	Prodromos Malakasiotis and John Pavlopoulos
Status & version	Final
Keywords	BioASQ, platform, database, web services, templates, latex, reports

Executive Summary

Goal of the document is to describe the overall architecture of the online BioASQ Participants Area. The first version of the document covered the functionality developed during the first year. This is an extension that provides the details for the implementation and the integration of the Oracle. The Oracle was developed from scratch during the second year of the BioASQ project and is now integrated in the online BioASQ Participants Area. It is accessible in http://bioasq.lip6.fr/oracle/. The infrastructure of BioASQ Participants Area developed during the first year provides the means to the participants for entering the BioASQ challenge. The implementation of the oracle extends the scope of this BioASQ Participants Area by acting towards making it sustainable, functional and useful after the end of the challenge.
In general, participants using the BIOASQ Participants Area can:

- Download data and tools for the tasks of the BIOASQ challenge.
- Submit their results for each task.
- Find guidelines and technical support for the challenge.
- View the evaluation measures for each task of the challenge.

Now, with the integration of the Oracle, participants can test their systems in an off-challenge mode. They can submit files with results for the released datasets of the BioASQ challenge. Having submitted their results, they can receive feedback for their performance immediately. The returned information contains the scores of the submitted results (based on the evaluation measures of the challenge) as well as the ranking of the submitted results compared to that of the systems participated in the challenge and others that have used the oracle off-challenge. There are no limits in usage frequency of the Oracle and users can select whether they wish their results to be saved and/or become publicly available.

The oracle provides the users the ability to use the existing infrastructure for testing their systems in a long-term basis. In addition, it was developed to be robust and extendable in terms of data and evaluation measures, highlighting the sustainability of the BioASQ infrastructure after the end of the project.

Contents

List of Figures

List of Tables

1

Introduction

Organizing the series of the BIOASQ challenge requires frequent interaction on behalf of the participants with the BIOASQ team. The BIOASQ challenge consists of two tasks:

- *Task A: Large-scale on-line biomedical semantic indexing.* Large-scale semantic indexing is evaluated on the whole of MEDLINE.[1] In particular, participants are asked to classify incoming documents before the human curators do. BioASQ distributes new unclassified MEDLINE documents every week and participants have a limited response time to estimate the MeSH[2] terms of the distributed articles.
- *Task B: Introductory biomedical semantic QA.* Benchmarks containing development and evaluation questions, as well as golden standard (reference) answers, are developed. The gold answers are produced by a team of biomedical experts from research teams around Europe. Established methodologies from QA, summarisation, and classification are followed to produce the benchmarks and evaluate the participating systems The task runs in two phases:
  - *Phase A:* BIOASQ releases questions from the benchmark; participants have to respond with concepts, documents, snippets found in the documents and triples for each question in limited time.
  - *Phase B:* BIOASQ releases questions and concepts, documents, snippets found in the documents and triplets; participants have to respond with facts, summaries, etc. The evaluation is based on gold answers while a small percentage is evaluated manually from the biomedical experts.

For more information about the BIOASQ challenge and the details of the tasks consult http://bioasq.org and Balikas et al. (2013b). More information about the preparation of the questions for Task 2B can be found in Malakasiotis et al. (2013).
The evaluation framework and the required functionality for the participants is provided through an on-line participants area that was developed to support the BIOASQ challenge. The details of the imple-

[1] More information about MEDLINE can be found in http://www.nlm.nih.gov/bsd/pmresources.html
[2] More information about the MeSH hierarchies can be found in http://www.ncbi.nlm.nih.gov/mesh

mentation are presented in Balikas et al. (2013a). In general, the BIOASQ Participants Area (hereafter platform):

- supports message based communication when interacting and exchanging data.
- supports automated communication using web services.
- is modular, as functionality and add-ons will be developed.
- offers the organisers the ability to supervise the challenge in a robust way.
- is easy to be used from participants

The above mentioned part of functionality was developed and integrated in the online platform during the first year of the challenge. By the end of the tasks of the first year a few recommendations were gathered and the platform along with its content was refined. The goal of those modifications was to increase the platform's support towards the participants of the challenge. As a result, it enables participants to find information about the challenge and browse the guidelines of the tasks efficiently. Having achieved those, the next step was to develop a subsystem for the platform that would provide the means to the users to develop and train their systems easier.

To this direction an Oracle was developed and is currently integrated in the online platform. It is accessible at `http://bioasq.lip6.fr`. The main purposes it serves follow:

1. It provides to the participants the means and the infrastructure for developing and tuning their systems faster.
2. It tries to make the BioASQ infrastructure and data available so that they will keep supporting and promoting this research area even after the end of the challenge.

The following list enumerates the steps that a user has to do in order to use the Oracle. At the same time, it illustrates the functionality it offers:

1. Download the available data of the challenge.
2. Select the task and the test set you intend to submit results for.
3. Submit a file with results.
4. Browse over the scores for the correspond evaluation measures for your submission and compare your system's performance with that of other available systems.
5. If you want, you can save the scores of your system and/or make them available to the other users of the oracle.
6. Repeat the above without any restrictions in usage frequency.

In the rest of this document the ideas behind the implementation of the oracle are presented. Also, some figures are presented as a quick tour to the developed facilities.

2

Overall architecture

This chapter describes the implementation of the Oracle and its integration in the platform. The Oracle was developed in the context of the BioASQ Participants Area, which uses the Model-View-Controller (MVC) software design pattern. Details of this design paradigm and information for the development of the platform are presented in Balikas et al. (2013a). Shortly, the platform is organized as a set of subsystems, each of which is responsible for offering a particular piece of functionality either to the users or to the organisers of the challenge. Each of the subsystems follows the Model-Template-View design paradigm, which is slightly different from the MVC pattern. Its goal is to separate the logic, the data and the appearance of the platform. The advantages of its use are important: the software becomes extensible and flexible. Many teams can work in parallel and development of extra functionality (adding more subsystems) comes without the cost of the unavailability of the platform. This occurs because the application's subsystems work independently in a way that any partial unavailability or failure cannot create a problem to the other subsystems. Recall that the previously existing subsystems were:

- "Task1a", "Task1b-PhaseA" and "Task1b-Phase B" offer the 'core' functionality required for each one of the tasks of the first year and can be modified according to the requirements of other years (or even projects).
- "Registration" offers the functionality for registering in the platform.
- "Forum" creates a BIOASQ Discussion Area.
- "Django Admin" creates the interface where administrators can supervise the BIOASQ challenge.

The updated list of subsystems is:

- "Task2a", "Task2b-PhaseA" and "Task2b-Phase B" offer the 'core' functionality required for each one of the tasks of the second year.
- "Registration" offers the functionality for registering in the platform.
- "Forum" creates a BIOASQ Discussion Area.
- "Django Admin" creates the interface where administrators can supervise the BIOASQ challenge.

- "Oracle" which is behind the functionality that the Oracle offers.

Note that concerning the renaming of the sytems "Task1a", "Task1b-PhaseA" and "Task1b-Phase B" to "Task2a", "Task2b-PhaseA" and "Task2b-Phase B" was decided for illustration reasons. The new names reflect the updates in the URLs and the head titles, which are required to distinguish the second edition of the challenge from the first. The core functionality they offer remained the same.

Having described this setting, the Oracle was developed as a separate subsystem of the platform. In the following subsections each layer of the subsystem will be presented. [1]

The view

The rationale behind the Oracle sub-application is to let the users of the platform use the developed facilities in an off-challenge mode. This means that users would be able to submit results and evaluate their systems even if there is no active test set. They can download the tests, process them and then evaluate their solutions. The functions in the view layer of the subsystem allow users to:

- Select a task and a test set,
- Submit results, and
- Receive the scores based on the corresponding evaluation measures.

Internally, when a user submits results the platform performs a series of validation procedures in order to assure that the user has submitted results for the correct test set and in the required format. If a problem occurs, a notification message appears in the platform informing the user about the situation and prompting the user to try again. If the format and the submitted data seem to have the required structure the evaluation functions are triggered. Those functions are the same used for the evaluation of the participating systems. Note that for the test sets of the first year the results are calculated using the version of 2013 of the MeSH terms, whereas for the second year test sets the MeSH 2014 will be used. The returned to the user results consist of:

- the evaluation scores of the system the user submitted,
- the evaluation scores of the systems that participated in the test set during the official challenge, and
- the evaluation scores of systems that participated in the Oracle and selected to be publicly available.

Those results are returned to the user and are displayed as HTML tables. Then the user can select if his results will be stored. Only one result per system and test set can be stored in the database. In case the user selects to keep his results, any older results for the selected system and test set are overwritten. For the results he saves, he can also select whether they will be publicly available. In this case the performance will appear in the Oracle results table to every user. Otherwise, if the user selects to keep them private, they will not appear in the results table to every user of the platform, but only to him (the platform detects who is logged in and presents the appropriate content to him).

In addition, the platform provides the ability to users to keep track of their performance. To achieve that is sends an e-mail to the participant after his submission. The e-mail contains the scores based on the corresponding evaluation measures of his submission.

[1] The functionality of the Oracle was developed using the Python programming language and the Django framework.

Name	Type
User	Foreign Key (Users)
Test id	Foreign Key (Tests)
Accuracy	Float
Example based Precisionebp	Float
Example Based Recall	Float
Example Based F-measure	Float
Macro Precision	Float
Macro Recall	Float
Macro F-measure	Float
Micro Precision	Float
Micro Recall	Float
Micro F-measure	Float
Hierarchical Precision	Float
Hierarchical Recall	Float
Hierarchical F-measure	Float
LCA Precision	Float
LCA Rrecall	Float
LCA F-measure	Float
timestamp	Date/Time
is_visible	Boolean

Table 2.1: The table that is used for saving the scores for the Task A of the challenge

Name	Type
User	Forein Key (Users)
Test id	Foreign Key (Tests)
System	Foreing Key (System)
timestamp	Date/Time
Comment	Text

Table 2.2: The table used for logging the activity in the Oracle

The model

The model layer describes the information that will be saved for the subsystem. The only information required for the "Oracle" is the scores for the corresponding evaluation measures of the participating systems and a log to keep track of the activity and the failures. Tables 2.1 and 2.2 show the SQL tables that are used to store the necessary information for the Oracle. The column "Type" indicates the type of the data that is saves in those tables. When a foreign key relation is indicated, it means that this cell is a foreign key that points to the primary key of the table in the parenthesis. For example, Test id: Foreign Key (Tests) indicates that the field points to the table "Tests" where information about the released tests of the challenge is maintained. Those tables have been presented in the Balikas et al. (2013a).

The template

The template is what the user sees in his browser. The templates of the Oracle sub-application contain HTML files written in a way that the fields in the tables will be populated automatically with the performance of the users system. HTML tables are also generated for showing the results. Figures 2.1 and 2.2 show some snapshots of the Oracle. In particular, Figure 2.1 depicts the form for submitting results with the corresponding drop-down menus. This is the home page of the Oracle. Figure 2.2 shows how do participants receive the evaluation measures of their submissions. The performance of their current submission is highlighted in yellow. The performance of other systems who used the Oracle and have publicly available results is high-lighted in purple. Also, the performance of the systems that participated in the official part of the challenge is given without any highlight.

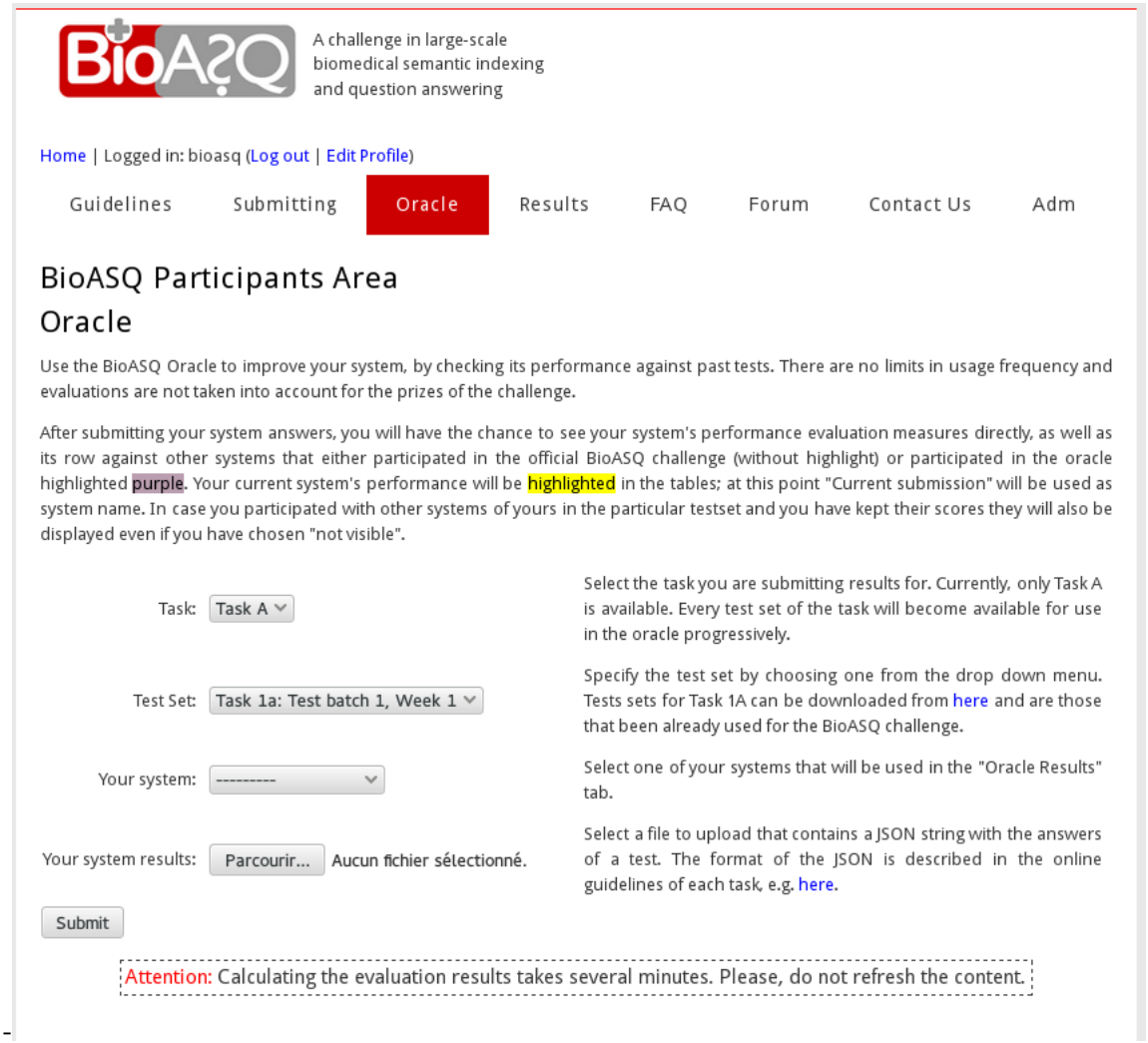

Figure 2.1: The home page of the Oracle. Your form for submitting results with the dropdown menus for selecting task and tests are also presented.

-

Results

Annotated documents: 726 out of 845.

Please, take a look at the results below and fill the following form:

Keep my results visible: ☑ If enabled, your uploaded results will be visible in the oracle to any registered user. Otherwise, it will be visible only to you.

Save my score: ☑ If enabled, it will replace the previous score for the selected system and testset in the BioASQ database.

Submit

Flat Measures

System	MiF	Acc.	EBP	EBR	EBF	MaP	MaR	MaF	MiP	MiR
MTI First Line Index	0.5437	0.3730	0.6026	0.5040	0.5247	0.5482	0.4468	0.4335	0.6033	0.4948
MeSH Indexing Pre	0.5339	0.3691	0.5146	0.5937	0.5238	0.4884	0.4576	0.4323	0.5024	0.5696
MeSH Indexing	0.5303	0.3662	0.5168	0.5842	0.5202	0.4807	0.4491	0.4276	0.5035	0.5600
MeSH Indexing New	0.5231	0.3606	0.4861	0.6072	0.5135	0.4525	0.4732	0.4412	0.4729	0.5853
MeSH Indexing Ref	0.5209	0.3551	0.4627	0.6202	0.5083	0.4413	0.4822	0.4439	0.4627	0.5959
MeSH Indexing Add	0.5203	0.3545	0.4622	0.6194	0.5077	0.4383	0.4818	0.4435	0.4622	0.5952
Wishart-S3	0.4706	0.3077	0.5381	0.4393	0.4568	0.5295	0.3330	0.3212	0.5198	0.4300
Wishart-S2	0.4705	0.3067	0.5083	0.4616	0.4564	0.5021	0.3555	0.3372	0.4908	0.4518
Wishart-S4	0.4665	0.3038	0.5648	0.4173	0.4516	0.5582	0.3105	0.3054	0.5453	0.4077
Current Submission	0.4652	0.3019	0.5642	0.4149	0.4496	0.5553	0.3103	0.3045	0.5441	0.4062
testing2	0.4652	0.3019	0.5642	0.4149	0.4496	0.5553	0.3103	0.3045	0.5441	0.4062
MCTeamMM	0.4598	0.3059	0.4977	0.4419	0.4492	0.5103	0.2824	0.2831	0.4977	0.4273
Wishart-S1	0.4588	0.3052	0.4966	0.4556	0.4545	0.5107	0.3380	0.3270	0.4966	0.4264
Galen	0.4480	0.2858	0.3869	0.5123	0.4273	0.4557	0.3680	0.3282	0.3983	0.5119
MCTeamSR	0.4406	0.2946	0.5868	0.3717	0.4371	0.6142	0.2120	0.2217	0.5868	0.3527

Figure 2.2: The Model-View-Template pattern that was followed during the development of the platform

3

Future work

The document described the architecture and the design decisions behind the implementation of the Oracle. The described functionality concerns Task A of the challenge. It is already integrated in the online BioASQ Participants Area. The Oracle is implemented so that extesions in terms of data and/or evaluation measures are easy. The following list presents some examples of extensions that are already scheduled:

- Provide the means to participants to use the Oracle for Task B. In particular, participants will be able to submit results for both phases of Task B and receive as feedback the scores of the corresponding evaluation measures of the task.
- Make available to the Oracle the benchmark dataset of the second year of the challenge (after the end of the challenge).

In terms of evaluation measures, the Oracle uses the already implemented evaluation scripts for the challenge. Extending the set of the evaluation measures whose scores are calculated, would only require to update the existing evaluation scripts and allow the extra measures to be saved in the SQL tables presented earlier in the document.

Bibliography

G. Balikas, I. Partalas, N. Baskiotis, T. Artieres, E. Gausier, and P. Gallinari. Evaluation Infrastructure. Technical report, 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.