



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

Project **FP7-318652 / BioASQ**

Deliverable **D4.1**

Distribution **Public**



<http://www.bioasq.org>

## **Evaluation Framework Specifications**

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

Status: Final (Version 1.1)

June 2013

**Project**

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

**Deliverable**

Deliverable type	Report
Distribution level	Public
Deliverable Number	D4.1
Deliverable title	Evaluation Framework Specifications
Contractual date of delivery	M6 (March 2013)
Actual date of delivery	June 2013
Relevant Task(s)	WP4/Task 1.3
Partner Responsible	UPMC
Other contributors	UJF, AUEB, NCSR
Number of pages	32
Author(s)	Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres and Patrick Gallinari
Internal Reviewers	Axel-C. Ngonga Ngomo
Status & version	Final
Keywords	evaluation framework, specifications

---

## Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Purpose . . . . .	1
1.2	Who will read this . . . . .	1
1.3	Project Description . . . . .	2
<b>2</b>	<b>Description of the Evaluation Framework</b>	<b>3</b>
2.1	General Overview . . . . .	3
2.2	Registration and Access to the Platform . . . . .	4
2.3	Web Interface . . . . .	4
2.4	Administration . . . . .	6
2.5	Technical Support . . . . .	6
<b>3</b>	<b>Task 1A: Large-scale on-line biomedical semantic indexing</b>	<b>8</b>
3.1	Description of the Task . . . . .	8
3.2	Datasets and MeSH hierarchy . . . . .	8
3.3	Evaluation Procedure . . . . .	9
3.3.1	Challenge Evaluation Process . . . . .	9
3.3.2	Performance and evaluation measures . . . . .	9
3.4	Specification Guidelines . . . . .	11
3.4.1	Web services . . . . .	11
3.4.2	Submitting results . . . . .	13
3.4.3	Database . . . . .	15
<b>4</b>	<b>Task 1B: Biomedical Semantic Question Answering</b>	<b>17</b>
4.1	Description of the task . . . . .	17
4.2	Distribution of Data . . . . .	17
4.3	Evaluation process and measures for Task 1b Phase A . . . . .	18
4.3.1	Mean precision, mean recall, mean F-measure . . . . .	19
4.3.2	Mean average precision and geometric mean average precision . . . . .	20
4.4	Evaluation process and measures for Task 1b Phase B . . . . .	22
4.4.1	Evaluating ‘exact’ answers . . . . .	22
4.4.2	Evaluating ‘ideal’ answers . . . . .	23

4.5	Specification Guidelines . . . . .	26
4.5.1	Database schema . . . . .	26
4.5.2	Web services for Task 1B Phase A . . . . .	27
4.5.3	Web services for Task 1B Phase B . . . . .	28

---

## List of Figures

---

2.1	The BioASQ Participants' Area home page in <a href="http://bioasq.lip6.fr">http://bioasq.lip6.fr</a> . . . . .	5
2.2	The BIOASQ Admin Interface where administrators can view and edit the information stored in the BIOASQ Participants Area database. . . . .	7
3.1	Task 1A schedule, showing the beginning of the challenge and the dates the test sets will become available. . . . .	10
3.2	A simple hierarchy. . . . .	10
3.3	The BIOASQ Participants Area for task 1A with available test sets, the active upload form and the download links and information for the user bioasq. . . . .	14
3.4	Database schema and foreign key connections between the tables used in Task 1A. . . .	15
4.1	An article-offset pair example. Article 1 has $n$ characters and a golden snippet starting at offset 3 and ending at offset 10. . . . .	20
4.2	The database schema and the foreign key connections between the tables that will be used in task 1B. . . . .	26

---

## List of Tables

---

4.1	Evaluation measures for Phase A of Task 1b. . . . .	21
4.2	Evaluation measures for the ‘exact’ answers in Phase B of Task 1b. . . . .	23
4.3	Criteria for the manual evaluation of the ‘ideal’ answers in Phase B of Task 1b. . . . .	25
4.4	Evaluation measures for the ‘ideal’ answers in Phase B of Task 1b. . . . .	26

---

## Introduction

---

### 1.1 Purpose

This document describes the evaluation framework and the infrastructure that will be used during the BioASQ challenges. More specifically, this document specifies the characteristics of the following components:

- Data preparation for both training and testing
- Tasks description
- Evaluation system
  - Evaluation and performance measures
  - Testing procedure
  - Web service specifications
- Web interface infrastructure for the challenge
  - Registration process
  - Web services and functionalities for the participants
  - Support and guidelines

### 1.2 Who will read this

- European Union
- Consortium

## 1.3 Project Description

BioASQ is a project that will initiate a challenge for information retrieval systems in the biomedical sector. Two tasks along with subtasks are going to be organised. The targets are to test automatic annotation of biomedical documents, as well as information retrieval methods for question answering.

### Short Description of the Tasks

A short description of the tasks that will be organised during the first year of the challenge follows:

- *Task 1A: Large-scale on-line biomedical semantic indexing (begins on the 15th of April).* Large-scale semantic indexing will be evaluated on the whole of MEDLINE. In particular, participants will be asked to classify incoming documents before the human curators do. BioASQ will distribute new unclassified MEDLINE documents every week and participants will have a limited response time to estimate the MeSH terms.
- *Task 1B: Introductory biomedical semantic QA (begins on the 3rd of June).* Benchmarks containing development and evaluation questions, as well as golden standard (reference) answers, will be developed. The gold answers will be produced by a team of biomedical experts from research teams around Europe. Established methodologies from QA, summarisation, and classification will be followed to produce the benchmarks and evaluate the participating systems. The task will run in two phases:
  - *Phase A:* BIOASQ will transmit questions from the benchmark while participants will have to respond with concepts, snippets and triples in limited time.
  - *Phase B:* BIOASQ will transmit questions and concepts, snippets and triples. Participants will respond with facts, summaries, etc. The evaluation will be based on gold answers while a small percentage will be evaluated manually from the biomedical experts.

A second version of the challenge will run in the second year of the project. Based on the comments and the feedback of the first year minor changes to the process may occur. There will be two tasks again with the same objectives:

- **Task 2A:** Large-scale on-line biomedical semantic indexing
- **Task 2B:** Biomedical semantic QA

In addition, one of the BIOASQ's main goal is to make the challenge sustainable after the end of the project. For this purpose, a special BIOASQ social network is being developed that will support the construction of new benchmarks and evaluation campaigns while the platform that supports the tasks will be designed to work independently.



---

## Description of the Evaluation Framework

---

### 2.1 General Overview

The BIOASQ Participants' Area (hereafter platform) general purpose and functionality can be summarized as follows:

- Provide an interface for the administrators of the challenge
- Provide functionality to register and gain access to the challenge:
  - Authentication process
  - Profile editing
  - Password reset
- Provide data and details for each task
  - Using a web interface
  - Using an API
- Provide the evaluation system for each task:
  - Provide functionality for downloading test sets
  - Provide training and dry-run data for the test sets
  - Provide functionality for uploading test results
  - Computation of the evaluation measures
  - Browsing challenge results
- Provide tools for pre-processing data
- Provide technical support during the competition

Each of the above mentioned points will be discussed in the following paragraphs while their function will be analysed for each task of the challenge.

## 2.2 Registration and Access to the Platform

The platform is publicly available in the URL <http://bioasq.lip6.fr>. Figure 2.1 shows the homepage of the platform. In order to gain full access to the platform's functionality users have to register first. Each team that will participate in the challenge will be considered as a single user of the platform. During registration, users have to fill a form and provide:

- Username (unique for each user)
- Password
- Valid e-mail address (unique for each user)
- Institution/enterprise (optional)
- Select the tasks they are planning to participate in
- Choose if they want to receive information and announcements about the BIOASQ challenge in the e-mail address they provided

After completing the form an e-mail with an activation link is generated automatically and is sent to the e-mail address that the user has provided. Clicking on the activation link, will activate the user's account and he will be able to log in. The activation link remains active for two days. If it expires before the account activation the process has to be repeated. Once logged in, a user will be able to use all the available resources that will be provided from BIOASQ Team for the tasks of the challenge (data sets, web services, etc.).

### User Authentication

The user authentication will use hash-based message authentication code (HMAC). During the authentication process the user will input his password, the hash function will be applied to it and the output will be compared with the hash output stored in the platform's database. In addition, in order to prevent programs from registering to the system, users will be asked to fill a captcha when filling the registration form in the platform. A reset-password option is also available.

### Profile Editing


Logged in users can edit the personal information they have submitted apart from their username. While editing their profile they can also register systems for participating in the tasks of the challenge. Each user can register up to five systems. Systems are added by filling a "System name" and a "System description" form. The "System name" is the unique identifier of the system and will be used while uploading test results while the "System description" is a unique system description publicly available and displayed in the "Results" section.

## 2.3 Web Interface

The platform's web interface will be created using the Django Framework ([Django](http://www.djangoproject.com/)). Django is a high-level Python <sup>1</sup> web framework that supports rapid development and clean, pragmatic design. Because Django was developed to support a fast-paced newsroom environment, it was designed to make common

---

<sup>1</sup><http://www.python.org>



A challenge in large-scale  
biomedical semantic indexing  
and question answering

[Home](#) | [Log in](#) | [Register](#)

[Guidelines](#) | [Submitting](#) | [Results](#) | [FAQ](#) | [Contact Us](#)

## BioASQ Participants Area

### Overview

The BioASQ Participants Area will enable users during the challenge to:

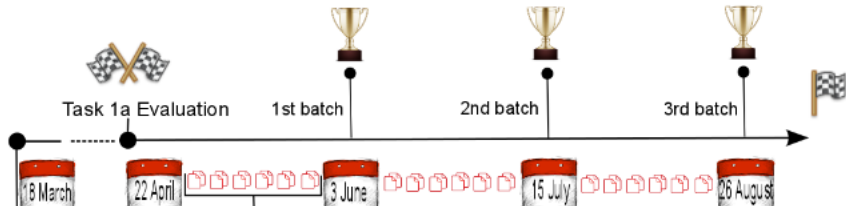
- Access guidelines for each challenge task
- Register to the challenge
- Download data for each Task
- Upload your answers
- Review your evaluation results
- Participate to the forum for discussion about the challenge

You can register [here](#).

Further information about the BioASQ project can be found at the [BioASQ's official website](#).

### Task 1A Schedule

Below you can see the BioASQ Task 1A schedule:



The diagram illustrates the Task 1A schedule as a horizontal timeline. It begins with a calendar icon for '18 March' and a dashed line leading to 'Task 1a Evaluation'. This is followed by another calendar icon for '22 April', then a series of red document icons representing data batches. A vertical line marks '3 June', which is labeled '1st batch' with a trophy icon above it. Further along, another vertical line marks '15 July', labeled '2nd batch' with a trophy icon. A final vertical line marks '26 August', labeled '3rd batch' with a trophy icon. The timeline ends with a checkered flag icon.

Figure 2.1: The BioASQ Participants' Area home page in <http://bioasq.lip6.fr>

web-development tasks fast and easy. It offers many features and there is a vibrant on-line community that supports it. It is based on the DRY principle (Don't Repeat Yourself) and tries to achieve loose coupling between the application's components as this makes them reusable and gives the developer the ability to make changes and improvements in the code without affecting the entire project. The DRY principle as well as the loose coupling between the BIOASQ platform components are essential, as the platform will be developed incrementally; Task 1A functionality will be developed first and will be supported from a MySQL <sup>2</sup> database. The functionality of Task 1B will be developed separately, it will be also supported from MySQL and will be added later. Finally, many of the main Python advantages (short code, easy to read, easy to learn) are important for the project as more than one developers from different scientific fields will co-operate to produce the output.

## 2.4 Administration

Administrators will be able to use the web interface as well as an administrator interface in order to perform the extra operations that are required. More specifically, they will be able to:

- Add/remove content in the web interface.
- Access the participants' personal details.
- Post, delete or edit answers in the forum. Regular users can only create topics and post comments on them.
- Create new test sets for the tasks, by simple filling a form.
- Trigger the evaluation functions for the tasks by simply calling a url.
- Receive e-mails when errors or unexpected events occur.
- Access a log table and monitor the users' activity.

Figure 2.2 shows the home page of the administrator interface as well as the database tables that contain the required data for Task 1A.

## 2.5 Technical Support

Users will be able to submit questions and discuss about the BIOASQ challenge using the following functionality that is part of the platform:

1. Detailed Guidelines for each task
2. FAQ
3. E-mail Help Desk
4. Discussion Fora (The access will be restricted to registered users only)

---

<sup>2</sup><http://www.mysql.com>

**BioASQ Administration Site**

### Site administration

Auth	
<b>Groups</b>	+ Add    ✎ Change
<b>Users</b>	+ Add    ✎ Change
Forum	
<b>Categories</b>	+ Add    ✎ Change
<b>Forums</b>	+ Add    ✎ Change
<b>Posts</b>	+ Add    ✎ Change
<b>Topics</b>	+ Add    ✎ Change
Registration	
<b>Registration profiles</b>	+ Add    ✎ Change
Sites	
<b>Sites</b>	+ Add    ✎ Change
Test	
<b>Articles</b>	+ Add    ✎ Change
<b>Bioasq_baselines</b>	+ Add    ✎ Change
<b>Details</b>	+ Add    ✎ Change
<b>Evaluation Measures</b>	+ Add    ✎ Change
<b>Systems per User</b>	+ Add    ✎ Change
<b>Test Result files</b>	+ Add    ✎ Change
<b>Test Results</b>	+ Add    ✎ Change
<b>Upload Information/Log</b>	+ Add    ✎ Change
Uploads	
<b>Documents</b>	+ Add    ✎ Change

#### Recent Actions

**My Actions**

- ✖ test\_result\_file object  
Test Result file
- ✖ 3  
Detail
- ✖ 3  
Detail
- ✖ 3  
Detail
- ✎ 2  
Detail
- ✎ eval\_meas object  
Flat Evaluation Measure
- ✎ eval\_meas object  
Flat Evaluation Measure
- ✎ eval\_meas object  
Flat Evaluation Measure
- ✎ eval\_meas object  
Flat Evaluation Measure
- ✎ eval\_meas object  
Flat Evaluation Measure

Figure 2.2: The BIOASQ Admin Interface where administrators can view and edit the information stored in the BIOASQ Participants Area database.

---

## Task 1A: Large-scale on-line biomedical semantic indexing

---

### 3.1 Description of the Task

This task will simulate the process that is followed by the curators of the MEDLINE (NLM, b) article database. Articles are uploaded in the database daily. The curators process them manually and classify them using the MeSH (Medical Subject Headings) (NLM, a) vocabulary as labels. MeSH are organised in hierarchies. A given heading may appear more than once in the hierarchies. When annotated, each article receives 10-15 subject headings and subheadings. During the task, sets of article abstracts will be provided that will not have been classified manually. The target of the participating systems will be to classify the abstracts using the MeSH hierarchies and return the MeSH headings they estimated. When the manual annotations from the MEDLINE curators will become available, they will be used to evaluate the classification performance of the participating systems. In this sense the evaluation measures will be calculated incrementally. The performance measures that are going to be used are:

- The multi-label version of micro F-measure
- Lowest Common Ancestor-F-measure (Kosmopoulos et al., 2013)

### 3.2 Datasets and MeSH hierarchy

Registered users will be able to download the necessary data for the task and submit their results. In particular, we will provide:

- Training data set; articles with their title, their abstract, the journal they were published, the year they were published and the MeSH labels they have received<sup>1</sup>. These data will be available both in raw and vectorized format<sup>2</sup>. Using this data set is optional; participants can train their classifiers using unrestricted resources.
- Dry-run data set. Before the beginning of the official challenge we will provide a dry-run data set which will not be considered in the ranking. In this way, participants will have the opportunity to become familiar with the process and the steps of the challenge.

---

<sup>1</sup>It is possible that some MeSH terms are not covered.

<sup>2</sup>The vectorized format will be distributed as a Lucene index.

- Tools for pre-processing the documents (extraction of vocabulary, stemming etc.).
- Download links for the MeSH hierarchies and the indexing that will be used for the classification. Apart from the indexing, we will also provide the hierarchies in parent-child format. For example, the line:  
D001570 D000876  
is to be read as node D001570 is a parent of D000876.
- Test sets that will be available both in raw and vectorized format. Users will be able to obtain them either using the web interface or an API. The first test will be publically available in the 22th of April, 2013.
- Upload form, which will be active for users to submit their results before the expiration of the test set.

### 3.3 Evaluation Procedure

#### 3.3.1 Challenge Evaluation Process

New articles become available every day in MEDLINE. The manual annotation that is performed by curators takes time. As a result, there is a gap between a document's release and its classification, which is exactly what we will take advantage of during Task 1A. Three batches of tests will be released. Each batch will consist of six test sets. A new test set will be available every week. It will consist of article abstracts that were published in MEDLINE the last six days and will not have been annotated manually at that time. The test set will be available for download either from the web interface or using an API. Similarly, users will upload their results either using the web interface or the API. They will have to respond within a fixed time window starting from the release of the test.

The platform will check periodically for the MeSH annotations from the MEDLINE curators of the articles. Each time that new annotations are available the evaluation function will be triggered and the evaluation measures in the platform will be updated. In addition to viewing the performance of their systems, users will be able to see the percentage of the annotated articles. The final scores, that will decide the winner of each batch, will be calculated using the scores of the systems' four best attempts. Systems, that will participate in less than four test sets in a batch, will be evaluated in those test sets, but they will not be considered for the prizes. Figure 3.1 shows the schedule for Task 1A BIOASQ challenge.

#### 3.3.2 Performance and evaluation measures

As the task concerns the classification in a hierarchical setting, flat evaluation measures (that ignore the presence of relations among the classes) are not sufficient for a proper evaluation of classification systems. For example, let us assume the simple hierarchy depicted in Figure 3.2, where the predicted class is denoted  $P_1$  and three true classes are respectively  $T_1$ ,  $T_2$  and  $T_3$ . In the case of flat measures (e.g. accuracy) the classification system will fail even though several true classes on the path from the root have been recognized (note that  $T_2$ ,  $T_3$  with  $P_1$  have common ancestors - bicolored nodes in the figure). On the other hand, a hierarchical measure takes into account these relations and assigns a value accordingly. For example, using the hierarchical version of precision (Kiritchenko et al., 2006), calculated as  $\frac{|An(C_p) \cap An(C_t)|}{|An(C_p)|}$ , where  $C_p$  is the set of predicted categories,  $An(C_p)$  is the set of ancestors of  $C_p$ ,  $C_t$  is the set of true categories and  $An(C_t)$  is the set of ancestors of  $C_t$ , we get  $\frac{2}{7}$ .

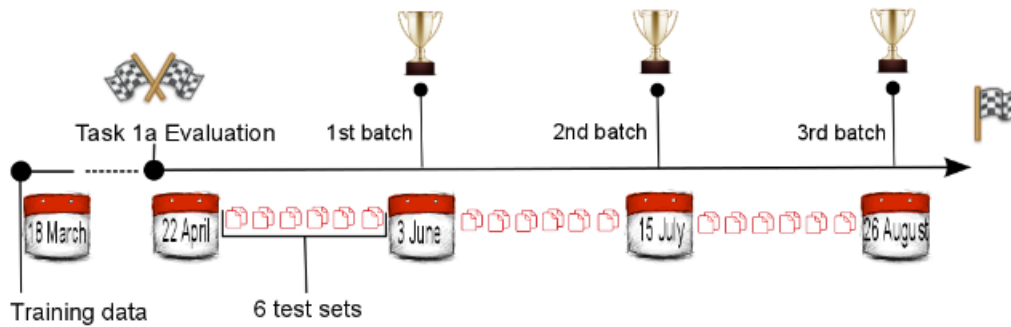


Figure 3.1: Task 1A schedule, showing the beginning of the challenge and the dates the test sets will become available.

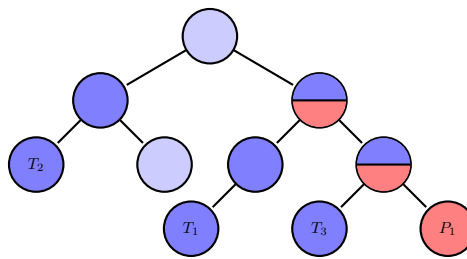


Figure 3.2: A simple hierarchy.

For the assessment of the systems participating in Task 1a, one flat and one hierarchical measure will be used. Specifically, the flat micro-F1 measure will be used which is a label-based measure [Tsoumakas et al. \(2010\)](#):

$$MiF1 = \frac{2 * MiP * MiR}{MiP + MiR},$$

where  $MiP$  and  $MiR$  are the micro-precision and micro-recall measures calculated as follows:

$$MiP = \frac{\sum_{i=1}^{|C|} tp_{c_i}}{\sum_{i=1}^{|C|} (tp_{c_i} + fp_{c_i})}$$

$$MiR = \frac{\sum_{i=1}^{|C|} tp_{c_i}}{\sum_{i=1}^{|C|} (tp_{c_i} + fn_{c_i})}$$

where  $tp_{c_i}$ ,  $fp_{c_i}$  and  $fn_{c_i}$  are respectively the true positives, false positives and false negatives for class  $c_i$ .

From the family of hierarchical measures the Lowest Common Ancestor - F measure ( $LCaF$ ) ([Kosmopoulos et al., 2013](#)) will be used:

$$LCaF = \frac{2 * LCaP * LCaR}{LCaP + LCaR} \quad (3.1)$$

where the corresponding precision and recall measures ( $LCaP$  and  $LCaR$  respectively) are calculated as follows:

$$LCaP = \frac{|\hat{Y}_{aug} \cap Y_{aug}|}{|\hat{Y}_{aug}|} \quad (3.2)$$



$$LCaR = \frac{|\hat{Y}_{aug} \cap Y_{aug}|}{|Y_{aug}|} \quad (3.3)$$

where  $Y_{aug}$  and  $\hat{Y}_{aug}$  are augmented sets of the true and the predicted classes respectively, based on the hierarchical relations. Specifically, in the case of *LCaF* these sets are constructed as follows:

1. First for each class  $y$  in the set of true classes  $Y$  the lowest common ancestor with respect to the set of predicted classes  $\hat{Y}$  is calculated:

$$LCA(y, \hat{Y}) = \arg \min_m \gamma(m, y),$$

where  $\gamma(u, v)$  denotes the distance between the nodes  $u$  and  $v$  in the graph.

2. Symmetrically, for each class in  $\hat{Y}$  the  $LCA(\hat{y}, Y)$  is computed.
3. Then two graphs,  $G_t$  and  $G_p$ , are defined containing the shortest paths from each  $y \in Y$  to  $LCA(y, \hat{Y})$  for  $G_t$  and  $\hat{y} \in \hat{Y}$  to  $LCA(\hat{y}, Y)$  for  $G_p$ .
4. Finally, Equation 3.1 is applied to the sets of the nodes defined by the two graphs.

*LCaF* assigns the minimum cost and thus curing handles over-penalization of errors that occurs in multi-label problems with DAG hierarchies.

## 3.4 Specification Guidelines

### 3.4.1 Web services

There are two sets of web services that are going to be used. They will be RESTful implemented and JSON<sup>3</sup> based. The first set of web services will be used during the creation of test sets and the evaluation of the results. The second set will enable users to download and upload the necessary data for the tasks.

#### Internal web services

- **getCitationsAfterDateWithoutMesh**: Each article in MEDLINE is uniquely identified by an integer, namely PMID. The platform will make a POST request to the web service's URI with a date parameter and a list of PMIDs to be excluded from the test set. The service will return a JSON string that will consist of the non-annotated articles from MEDLINE from the date parameter on, after excluding the PMIDs of the list. The articles that will be contained in the JSON must be "In-Process"<sup>4</sup>. The JSON string will contain the PMIDs of the articles, their titles and their abstracts. The selected articles will come from pre-selected journals from a wide variety of biomedical fields. In the same time the preselected journals will have a short average annotation time<sup>5</sup>. An example of the format of a JSON string that could be returned follows:

<sup>3</sup><http://json.org>

<sup>4</sup>PubMed "In-Process" records provide basic citation information and abstracts while these records are reviewed for accuracy of the bibliographic data and assigned subject headings if the subject of the article is within the scope of MEDLINE.

<sup>5</sup>It is possible that some articles will not be annotated until the end of the challenge.

```
{
  "result": {
    "articlesPerPage": 5,
    "documents": [
      {
        "documentTitle": "A novel method for detecting antigen-specific
human regulatory T cells",
        "documentAbstract": "Antigenic epitopes recognized by FoxP3(+)
regulatory T cells (Treg) are poorly defined, largely due
to a lack of assays for determining Treg specificity...",
        "PMID": "22265970"
      },
      .
      .
    ]
  }
}
```

- **CheckForMeSH**: The platform will make a POST request to the web service's URI with a JSON that will consist of an array of PMIDs. Server-side, the service will check if the articles with the provided PMIDs have been annotated in MEDLINE. A JSON string with the PMIDs along with their MeSH annotations will be returned, as in the following example:

```
{
  "result": {
    "articlesPerPage": 1,
    "documents": [
      {
        "PMID": "22265970",
        "meshHeadings": ["D081256", ..., "D045211"]
      },
      .
      .
      {
        "PMID": "22265978",
        "meshHeadings": ["D081487", ..., "D045278"]
      }
    ]
  }
}
```

If the returned results outnumber the 5% of the size of the test set, the evaluation function will be triggered to update the evaluation measures of the participants' systems.

### Public web services

- **DownloadTestSet**: Users will be able to make a GET request to the web service's URI with their authentication credentials. The web service, after checking that the user is registered, will return a JSON string with the PMIDs, the titles and the abstracts of the selected test set. The vectorized and the raw descriptions of the test set will be available in different URIs. The test set will also be available for downloading after the expiration date of the test set. However, after the expiration of the test set users will not be able to submit their results. The format of the JSON string that will be returned is the following:

```
[
  {
    "documentTitle": "Title",
    "documentAbstract": "Abstract",
    "PMID": "22511223"
  },
  {
    "documentTitle": "Title",
    "documentAbstract": "Abstract",
    "PMID": "22511224"
  },
  .
]
```

```

    .
    { "documentTitle": "Title", "documentAbstract": "Absrtact",
      "PMID": "22511225" }
  ]}

```

- **UploadTestResults:** Users will be able to make a POST request to the web service's URI with a JSON string containing their answers and their authentication credentials. Users will be able to upload their answers for the active test set as many times as they want before the test expires. Each time a user updates his answers the old answers will be erased. In addition, a log will be kept with the time stamps of the uploads. The following JSON string illustrates the format of the data in the JSON string in the POST request:

```

    .
    { "username": "your_username", "password": "your_password",
      "system": "name_of_your_system",
      "documents": [ { "labels": [index1, ..., indexN1], "PMID": "22511223" },
                     { "labels": [index1, ..., indexN2], "PMID": "22511224" },
    .
    .
    { "labels": [index1, ..., indexN3], "PMID": "22511225" }
  ]}

```

Users will be encouraged to use the web services during the challenge in order to automate the evaluation procedure. Scripts that perform the POST and GET requests to the web services and can be integrated to the user's code are available in the website <http://bioasq.lip6.fr>. In case that a user prefers uploading results

### 3.4.2 Submitting results

Submitting results will be possible either using the web services or the web interface. The upload will be considered when it begins before the expiration of the test set. The platform before saving the results checks that:

- the system in the JSON string belongs to the user.
- the PMIDs in the provided JSON belong to the active test set.
- there are MeSH indices for every article of the test set.
- the MeSH indices exist.
- the test set is active in the beginning of the upload.

When all of the above are valid, the results are saved in a file and the path of the file in the database. A message informing the user that the results were saved is returned. If anything of the above fails, a message with information about the error(s) is returned.

In order to provide the users the ability to check the data the platform has saved, download links for their results per system and per test set are provided. Figure 3.3 shows the BIOASQ Participants Area for Task 1A with the dry-run and two official test sets available, the active upload form, and the download links and information for the user bioasq uploads.

### Available test sets

Currently, 3 **tests sets** are available. The first test (Test 1) is the dry-run while the others are the official test sets of the BioASQ Task 1A challenge.

- Dry run test set, available [in raw format](#) and [in vectorized format](#). Active for uploads from April 18, 2013, 2 p.m. until April 20, 2013, 6 p.m.
- Test set of Test batch 1, week 1, available [in raw format](#) and [in vectorized format](#). Active for uploads from April 22, 2013, 5 p.m. until April 23, 2013, 2 p.m.
- Test set of Test batch 1, week 2, available [in raw format](#) and [in vectorized format](#). Active for uploads from April 29, 2013, 5 p.m. until April 30, 2013, 2:08 p.m.

### Submit your results

You can select a file from your computer, that contains the test results and upload it by clicking the Upload button below. The results will be considered for the Test 3 and the system you will select. Submitting results will be open until April 30, 2013, 2:08 p.m..

**Warning:** Selecting a system for which you have already uploaded results will replace the previous results.

**Attention:** The process of uploading results may take several minutes.

Select a file:

System name

### Download the results you have submitted

In the following table you can find information and links for the results you have submitted.

Test set	System Name.	Download link	Date/Time
2	bioasq_baseline	<a href="#">2-bioasq_baseline.json</a>	April 22, 2013, 5:55 p.m.
1	bioasq_baseline	<a href="#">1-bioasq_baseline.json</a>	April 18, 2013, 7:12 p.m.
3	bioasq_baseline	<a href="#">3-bioasq_baseline.json</a>	April 29, 2013, 4:48 p.m.

Copyright © 2013, the BioASQ project



Figure 3.3: The BIOASQ Participants Area for task 1A with available test sets, the active upload form and the download links and information for the user bioasq.

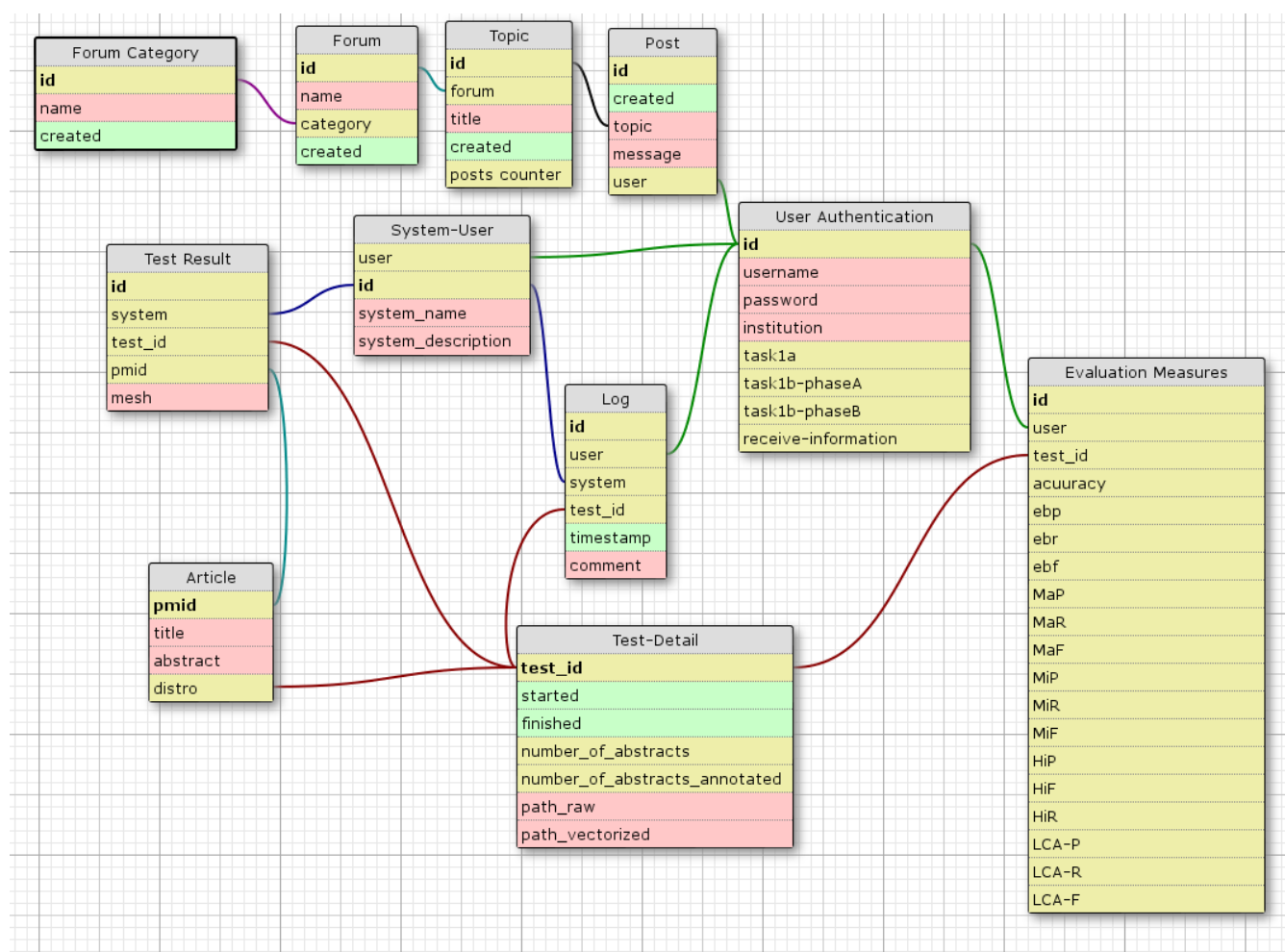


Figure 3.4: Database schema and foreign key connections between the tables used in Task 1A.

### 3.4.3 Database

The database that will support Task 1A is MySQL. The effective maximum table size for MySQL databases is determined by the operating system constraints on file sizes. In Linux this is 4 TB, which is more than enough for the database tables that will be created during the challenge. Furthermore, the user experience in the platform will be good as it requires only a few queries for the page content to be displayed. On the other hand, the evaluation procedure, which requires some extra queries and processing power will be triggered once a the week to update the evaluation results. The database will store the following information in its tables:

- The User Authentication table will store the necessary information for each user.
- The System-User table will contain the systems that each user has registered.
- The Test-Details table will contain information for each test set; start/end datetime stamps and the number of articles in each test set.
- The Article table will contain the information about each article in the test sets.

- The Test Results table will contain the answers that the systems have uploaded for the test sets.
- The Flat Measures and Hierarchical Measures tables will be updated with the flat and the hierarchical evaluation measures that will be computed for the participating systems for each test set.
- The Log Table will keep information each time a user uploads results.
- The Categories, Forums, Post and Topics tables will store the necessary information for the forum functionality.

Figure 3.4 shows the database schema as well as the foreign key connections between the database tables.

---

## Task 1B: Biomedical Semantic Question Answering

---

### 4.1 Description of the task

BIOASQ Task 1B will take place in 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. The designated terminologies, ontologies and article repositories are described by [Tsatsaronis et al. \(2013\)](#). The system responses of Phase A will be automatically compared against golden responses constructed by the BIOASQ team of biomedical experts; consult [Malakasiotis et al. \(2013\)](#) for information on how the golden responses will be constructed.

**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). The ‘exact’ and ‘ideal’ answers of the systems will be automatically compared against golden ‘exact’ and ‘ideal’ answers constructed by the BIOASQ team of biomedical experts; again, consult [Malakasiotis et al. \(2013\)](#) for information on how the golden ‘exact’ and ‘ideal’ answers will be constructed. A sample of the ‘ideal’ answers of the systems will also be manually evaluated by the biomedical experts.

Phase B will take place immediately after Phase A. In both phases, the participants will have very limited time to submit their responses, to make it difficult for participants to produce their responses manually.

### 4.2 Distribution of Data

During Task 1B, registered users will be able to download through the web interface the following set of *indexed* ontologies:

- The MeSH ontology

- The UniProt database (the Swiss-Prot component)
- The Jochem ontology, for the purpose of covering drugs
- The Disease Ontology, for the purpose of covering diseases
- The Go Ontology

In addition, a training set of 30 annotated questions produced by the biomedical experts will be available for downloading. The questions that will consist the test sets will be accessible both from the web interface as direct downloads and from web services. The web services will be RESTful implemented and JSON based. The data will be served as JSON strings to the systems. In particular, in task 1B the BIOASQ team will provide:

1. *Phase A*: The body and the type of the question
2. *Phase B*: The body and the type of the question, as well as the golden responses of Phase A

Systems will have to respond within a limited time frame. Uploading the answers will be done either using the web interface or using the API. In the first case, users will have to upload a file that will contain the JSON string, while in the second case they will have to POST the JSON string to a specified URI.

### 4.3 Evaluation process and measures for Task 1b Phase A

In Phase A, the participants will be provided with English questions  $q_1, q_2, q_3, \dots, q_n$ . For each question  $q_i$ , each participating system will be required to return:

**A list of relevant concepts**  $c_{i,1}, c_{i,2}, c_{i,3}, \dots$  from the designated terminologies and ontologies. The list should be ordered by decreasing confidence, i.e.,  $c_{i,1}$  should be the concept that the system considers most relevant to the question  $q_i$ ,  $c_{i,2}$  should be the concept that the system considers to be the second most relevant etc. A single concept list will be returned per question and participant, and the list may contain concepts from multiple designated terminologies and ontologies. The returned concept list will actually contain unique concept identifiers (obtained from the terminologies and ontologies), rather than terms (words or phrases).

**A list of relevant articles** (documents)  $d_{i,1}, d_{i,2}, d_{i,3}, \dots$  from the designated article repositories. Again, the list should be ordered by decreasing confidence, i.e.,  $d_{i,1}$  should be the article that the system considers most relevant to the question,  $d_{i,2}$  should be the article that the system considers to be the second most relevant etc. A single article list will be returned per question and participant, and the list may contain articles from multiple designated repositories. The returned article list will actually contain unique article identifiers (obtained from the repositories).

**A list of relevant text snippets**  $s_{i,1}, s_{i,2}, s_{i,3}, \dots$  from the returned articles. Again, the list should be ordered by decreasing confidence. A single snippet list will be returned per question and participant, and the list may contain any number (or no) snippets from any of the returned articles  $d_{i,1}, d_{i,2}, d_{i,3}, \dots, d_{i,k}$ . Each snippet will be represented by the unique identifier of the article it comes from and the offsets (character positions in the article) of the snippet's beginning and end (offsets of the first and last characters).

**A list of relevant RDF triples**  $t_{i,1}, t_{i,2}, t_{i,3}, \dots$  from the designated ontologies. Again, the list should be ordered by decreasing confidence. A single triple list will be returned per question and participant, and the list may contain any triples from multiple designated ontologies.



For practical purposes (e.g., to avoid malicious attacks to the BIOASQ servers), the maximum allowed number of returned concepts, articles, snippets, and triples per question and system will be limited (e.g., to 100 concepts, 100 articles, 100 snippets, and 100 triples per question and system).

For each question  $q_i$ , the BIOASQ team of biomedical experts will have constructed the gold (correct) sets of concepts, articles, snippets, and triples, as discussed by [Malakasiotis et al. \(2013\)](#). Once the responses of the participating systems have been submitted, the biomedical experts will also inspect the top  $k$  concepts, articles, snippets, and triples of each system, i.e., the  $k$  concepts, articles, snippets, and triples that each system is most confident about, in order to add to the corresponding golden sets any correct (relevant) items that the biomedical experts had missed, but the systems managed to retrieve. The value of  $k$  will be determined once the number of participating systems is known. Facilities similar to the annotation tool discussed by [Ngonga Ngomo et al. \(2013\)](#) will be made available to the biomedical experts to help them inspect the system responses.

For each system, the lists of returned concepts, articles, snippets, and triples of all the questions will then be evaluated using the *mean average precision* (*MAP*) measure, defined below, which is widely used in information retrieval to evaluate ranked lists of retrieved items (see [Manning et al. \(2008\)](#)). We will also use the *geometric mean average precision* (*GMAP*), which places more emphasis on improvements in low performing queries (see [Robertson \(2006\)](#) and [Sanderson \(2010\)](#)). For the sake of completeness, we will also compute the *mean precision*, *mean recall*, and *mean F-measure* of each system, also defined below, but the official scores for Phase A will be based on *GMAP*.

### 4.3.1 Mean precision, mean recall, mean F-measure

Given a set of golden items (e.g., articles), and a set of items returned by a system (for a particular question in our case), precision ( $P$ ) and recall ( $R$ ) are defined as follows:

$$P = \frac{TP}{TP + FP} \quad (4.1)$$

$$R = \frac{TP}{TP + FN} \quad (4.2)$$

where  $TP$  (true positives) is the number of returned items that are also present in the golden set,  $FP$  (false positives) is the number of returned items that are not present in the golden set, and  $FN$  (false negatives) is the number of items of the golden set that were not returned by the system. The  $F_\beta$  measure is the weighted harmonic mean of  $P$  and  $R$ , defined as follows:

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R} \quad (4.3)$$

For  $\beta = 1$ , the same weight is assigned to both precision and recall, and the resulting measure, often called simply *F-measure*, is defined as follows:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (4.4)$$

Given a set of queries (in our case, questions)  $q_1, \dots, q_n$ , the *mean precision*, *mean recall*, and *mean F-measure* of each system is obtained by averaging its precision, recall, and *F-measure* for all the queries.

In BIOASQ, we will compute the mean precision, mean recall, and mean *F-measure* of the concepts, articles, snippets, and triples returned by each system. In the case of snippets, a complication is that a returned snippet may overlap with one or more golden snippets, without being identical to any of them. To take this into account, in the case of snippets we modify the definitions of precision and recall. Figure

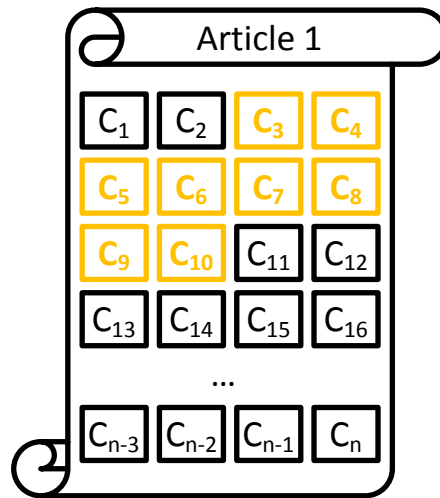


Figure 4.1: An article-offset pair example. Article 1 has  $n$  characters and a golden snippet starting at offset 3 and ending at offset 10.

4.1 illustrates what we mean by article-offset pairs. A snippet is determined by the article it comes from and by the offsets (positions) in the article of the first and last characters of the snippet. We can also think of the snippet as a set of (article, offset) pairs, one pair for each character of the snippet. In the example of Figure 4.1, Article 1 has  $n$  characters and a golden snippet starting at offset 3 and ending at offset 10. Let us call  $S$  the set of all the article-offset pairs of all the characters in the snippets returned by a system for a particular question,  $G$  the set of all the article-offset pairs of all the characters in the golden snippets of the question, and let  $|s|$  denote the cardinality of a set  $s$ . The definitions of precision ( $P_{snip}$ ) and recall ( $R_{snip}$ ) for snippets are:

$$P_{snip} = \frac{|S \cap G|}{|S|} \quad (4.5)$$

$$R_{snip} = \frac{|S \cap G|}{|G|} \quad (4.6)$$

In effect,  $P_{snip}$  divides the size (in characters) of the total overlap between the returned and golden snippets by the total size of the returned snippets, whereas  $R_{snip}$  divides the size of the total overlap by the total size of the golden snippets. The definitions of  $F_\beta$ , mean precision, mean recall, and mean  $F$ -measure for snippets are the same as the corresponding definitions for concepts, articles, and triples, but they use  $P_{snip}$  and  $R_{snip}$  instead of  $P$  and  $R$ .

### 4.3.2 Mean average precision and geometric mean average precision

Precision, recall, and  $F$ -measure do not consider the order of the items returned by a system for each query. Recall that in BIOASQ we require the lists of concepts, articles, snippets, and triples that a system returns for each question to be ordered (ranked) by decreasing confidence. To take the ordering of a particular returned list (for a particular question) into account, it is common in information retrieval

Retrieved items	Unordered retrieval measures	Ordered retrieval measures
concepts	mean precision, recall, $F$ -measure	<b><i>MAP, GMAP</i></b>
articles	mean precision, recall, $F$ -measure	<b><i>MAP, GMAP</i></b>
snippets	mean precision, recall, $F$ -measure	<b><i>MAP, GMAP</i></b>
triples	mean precision, recall, $F$ -measure	<b><i>MAP, GMAP</i></b>

Table 4.1: Evaluation measures for Phase A of Task 1b.

to compute the (non-interpolated) *average precision* ( $AP$ ) of the list, defined as follows:

$$AP = \frac{\sum_{r=1}^{|L|} P(r) \cdot rel(r)}{|L_R|} \quad (4.7)$$

where  $|L|$  is the number of items in the list,  $|L_R|$  is the number of relevant items,  $P(r)$  is the precision when the returned list is treated as containing only its first  $r$  items, and  $rel(r)$  equals 1 if the  $r$ -th item of the list is in the golden set (i.e., if the  $r$ -th item is relevant) and 0 otherwise.<sup>1</sup> In BIOASQ, especially when computing the average precision of a list of *snippets*,  $P(r)$  will be taken to be the snippet precision  $P_{snip}$  (as in Section 4.3.1) when the returned list of snippets is treated as containing only its first  $r$  snippets; and  $rel(r)$  will be taken to be 1 if the  $r$ -th returned snippet has a non-zero overlap (shares at least one article-offset pair) with at least one golden snippet of the particular question.

By averaging  $AP$  over a set of queries (in our case, questions)  $q_1, \dots, q_n$ , we obtain the *mean average precision* ( $MAP$ ), defined as follows:

$$MAP = \frac{1}{n} \cdot \sum_{i=1}^n AP_i \quad (4.8)$$

where  $AP_i$  is the average precision of the list returned for query (question)  $q_i$ . In our case, each system will receive four  $MAP$  scores, for the lists of concepts, articles, snippets, and triples, respectively, that it returned for all the questions.

The *geometric mean average precision* ( $GMAP$ ), defined below, is very similar to  $MAP$ , but it uses the geometric instead of the arithmetic mean, which places more emphasis on improvements in low performing queries, as already noted.

$$GMAP = \sqrt[n]{\prod_{i=1}^n (AP_i + \epsilon)} \quad (4.9)$$

An alternative way to more easily compute  $GMAP$  is by using the following equation:

$$GMAP = \exp\left(\frac{1}{n} \cdot \sum_{i=1}^n \ln(AP_i + \epsilon)\right) \quad (4.10)$$

In both versions of  $GMAP$ ,  $\epsilon$  is a small number added to handle cases where  $AP_i = 0$ . As with  $MAP$ , in BioASQ each system will receive four  $GMAP$  scores, for the lists of concepts, articles, snippets, and triples, respectively, that it returned for all the questions. The official scores for Task 1b Phase A will be based on  $GMAP$ , as already noted. Table 4.1 summarizes the evaluation measures of Phase A; the official measures are shown in bold.

<sup>1</sup> $AP$  approximates the area under a recall–precision curve; consult Robertson (2006).

## 4.4 Evaluation process and measures for Task 1b Phase B

In Phase B, the participants will be provided with the same questions  $q_1, \dots, q_n$  as in Phase A, but this time they will also be given the golden (correct) lists of concepts, articles, snippets, and triples of each question. For each question, each participating system will have to return an ‘ideal’ answer, i.e., a paragraph-sized summary of relevant information. In the case of yes/no, factoid, and list questions, the systems will also have to return ‘exact’ answers; for summary questions, no ‘exact’ answers will be returned. Consult [Malakasiotis et al. \(2013\)](#) for a discussion of the types of questions that will be used in BIOASQ, and the nature of ‘exact’ and ‘ideal’ answers in BIOASQ, and the nature of exact and ideal answers. The participants will be told the type of each question.

### 4.4.1 Evaluating ‘exact’ answers

We first discuss how ‘exact’ answers will be evaluated in Phase B, by considering in turn yes/no, factoid, and list questions.

#### Evaluating the ‘exact’ answers of yes/no questions

For each yes/no question, the ‘exact’ answer of each participating system will have to be either ‘yes’ or ‘no’. The response will be compared against the golden ‘exact’ answer (again ‘yes’ or ‘no’) that the BIOASQ team of biomedical experts will have associated with the question. For each system, we will compute the *accuracy* ( $Acc$ ) of its responses to yes/no questions. Assuming that there are  $n$  yes/no questions, accuracy is defined as follows, where  $c$  is the number of correctly answered yes/no questions.

$$Acc = \frac{c}{n} \quad (4.11)$$

#### Evaluating the ‘exact’ answers of factoid questions

For each factoid question, each participating system will have to return a list of up to 5 entity names, ordered by decreasing confidence. The BIOASQ team of biomedical experts will have associated with each factoid question a single golden entity name, as well as possible synonyms of that name. Once the responses of the participating systems have been submitted, the biomedical experts will also inspect the entity names returned by the participating systems for the factoid questions, in order to add synonyms they may have missed when preparing the golden answers.

We will measure the *strict accuracy* ( $SAcc$ ) and *lenient accuracy* ( $LAcc$ ) of each system for factoid questions. Strict accuracy counts a question as correctly answered if the golden entity name (or a synonym of that name) is the first element of the list returned by the system. By contrast, lenient accuracy counts a question as correctly answered if the golden entity name (or synonym) is included, not necessarily as the first element, in the list returned by the system. In the definitions below,  $n$  is the number of factoid questions,  $c_1$  is the number of factoid questions that have been answered correctly when only the first element of each returned list is considered, and  $c_5$  is the number of factoid questions that have been answered correctly in the lenient sense, when all the elements of the returned list are considered.

$$SAcc = \frac{c_1}{n} \quad (4.12)$$

$$LAcc = \frac{c_5}{n} \quad (4.13)$$

Strict and lenient accuracy will be measured for completeness. The official measure for the ‘exact’ answers of factoid questions will be the *mean reciprocal rank* ( $MRR$ ), which is often used to evaluate

Question type	Participant response	Evaluation measures
yes/no	‘yes’ or ‘no’	<b>accuracy</b>
factoid	up to 5 entity names	strict and lenient accuracy, <b><i>MRR</i></b>
list	a list of entity names	<b>mean</b> precision, recall, <b><i>F-measure</i></b>

Table 4.2: Evaluation measures for the ‘exact’ answers in Phase B of Task 1b.

factoid questions in question answering challenges; consult, for example, Voorhees (2001). In the definition below, for each factoid question  $q_i$  we search the returned list looking for the topmost position that contains the golden entity name (or one of its synonyms). If the topmost position is the  $j$ -th one, then  $r(i) = j$ ; otherwise  $r(i) \rightarrow +\infty$ , i.e.,  $\frac{1}{r(i)} = 0$ .

$$MRR = \frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{r(i)} \quad (4.14)$$

In effect, *MRR* rewards systems that manage to include the golden responses (or their synonyms) higher in the returned lists.

### Evaluating the ‘exact’ answers of list questions

For each list question, each participating system will have to return a list of entity names, jointly taken to constitute a single answer (e.g., the most common symptoms of a disease); for practical purposes, the maximum allowed size of each returned list may be limited (e.g., up to 100 names, each one up to 100 characters). The BIOASQ team of biomedical experts will have associated with each list question a golden list of entity names, also providing possible synonyms for each entity name of the golden list; consult Malakasiotis et al. (2013) for details. Again, once the responses of the participating systems have been submitted, the biomedical experts will also inspect the lists returned by the participating systems for the list questions, in order to add synonyms they may have missed.

For each list question, the list returned by the system will be compared against the golden list by computing its *precision* ( $P$ ), *recall* ( $R$ ), and *F-measure* ( $F_1$ ), as in Section 4.3.1. Here  $TP$  is the number of entities that are mentioned both in the returned and the golden list;  $FP$  is the number of entities that are mentioned in the returned, but not in the golden list; and  $FN$  is the number of entities that are mentioned in the golden, but not in the returned list. If the same entity is mentioned using different synonyms in the returned and golden lists, it will be counted as having been mentioned in both lists. If an entity is mentioned multiple times, possibly using different synonyms, in the returned list, it will be counted only once.

By averaging precision, recall, and *F-measure* over the list questions, we will obtain the *mean average precision*, *mean average recall*, and *mean average F-measure* score of each system for list questions. The official measure for list questions will be mean *F-measure*. Table 4.2 summarizes the kinds of responses and the evaluation measures that will be used in Phase B.

### 4.4.2 Evaluating ‘ideal’ answers

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; recall that the correct concepts, articles, snippets, and triples of each question will be provided to the participants of Phase B. The returned ‘ideal’ answer is intended to approximate 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 BIOASQ team of biomedical experts will have associated each question with a golden ‘ideal’ answer. The maximum allowed length (in characters) of each ‘ideal’ answer to be produced by a system will be set to be equal to the maximum length of all the golden ‘ideal’ answers (of all the questions). The same maximum allowed length will be used for all the questions to avoid revealing the expected length of each particular ‘ideal’ answer.

The ‘ideal’ answers of the systems will be evaluated both manually (by the BIOASQ team of biomedical experts) and automatically (by comparing them to the golden ‘ideal’ answers). The official scores will be based on the manual evaluation; the automatic evaluation will be performed mostly to explore how well automatic evaluation measures (e.g., from multi-document text summarization) correlate with the scores of the biomedical experts.

### Manual evaluation of ‘ideal’ answers

The questions the participating systems will have to answer will be grouped according to their difficulty into three groups, i.e., very difficult, difficult, easy. The difficulty of the questions will be determined by studying the performance of all the participating systems in Phase A. We expect that we will be able to classify the questions into the three groups of difficulty by examining how well the systems of Phase A managed (on average) to retrieve concepts, documents, snippets, and triples.

The BIOASQ team of biomedical experts will inspect the ‘ideal’ answers of the participating systems for a sample of  $m$  questions. The  $m$  questions will be from all the three groups of difficulty, with an approximately uniform distribution across the three groups. We also note that each question will have been formulated by a single biomedical expert (or by a pair of biomedical experts, for some questions) and that the biomedical expert who formulated each question should be the one to assess the ‘ideal’ answers of the systems for that question. Hence, care will also be taken to ensure that the sample of  $m$  questions contains approximately the same number of questions from each biomedical expert, in order to balance the workload of the biomedical experts during the manual evaluation. Moreover, the participating systems will be required to order their ‘ideal’ answers (for all the questions, one ‘ideal’ answer per question) by decreasing confidence, indicating how confident they are that each ‘ideal’ answer is close to an ‘ideal’ answer that a biomedical expert would produce. This ordering will not affect the manual evaluation, but it will help us to better analyse the results.

Each one of the top  $m$  ‘ideal’ answers of each system will be inspected by a biomedical expert, who will be asked to evaluate the answer in terms of *information recall* (the ‘ideal’ answer reports all the necessary information), *information precision* (no irrelevant information is reported), *information repetition* (the ‘ideal’ answer does not repeat the same information multiple times, e.g., when sentences of the ‘ideal’ answer that have been extracted from different articles convey the same information), and *readability* (the ‘ideal’ answer is easily readable and fluent). An 1–5 scale will be used in all four criteria (1 for ‘very poor’, 5 for ‘excellent’). Table 4.3 summarizes the criteria that will be used in the manual evaluation of the ‘ideal’ answers in Phase B. A sample of the ‘ideal’ answers will be evaluated by more than one biomedical experts to measure the inter-annotator agreement.

### Automatic evaluation of ‘ideal’ answers

The ‘ideal’ answers returned by the systems will also be automatically evaluated using *ROUGE*; consult Lin (2004). Roughly speaking, *ROUGE* counts the overlap between an automatically constructed summary and a set of reference (golden) summaries constructed by humans. There are several different versions of *ROUGE*. *ROUGE-N* is, defined below, uses word  $n$ -grams when computing the overlap



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 4.3: Criteria for the manual evaluation of the ‘ideal’ answers in Phase B of Task 1b.

between an automatically constructed summary  $S$  and a set  $Refs$  of reference summaries:

$$ROUGE-N(S|Refs) = \frac{\sum_{R \in Refs} \sum_{g_n \in R} C(g_n, S, R)}{\sum_{R \in Refs} \sum_{g_n \in R} C(g_n, R)} \quad (4.15)$$

In the definition above,  $g_n$  is a word  $n$ -gram,  $C(g_n, S, R)$  is the number of times that  $g_n$  co-occurs in  $S$  and a reference summary  $R$ , and  $C(g_n, R)$  is the number of times  $g_n$  occurs in reference  $R$ .

*ROUGE-S* uses skip bigrams, instead of  $n$ -grams, when computing the overlap. A skip bigram is any pair of words, maintaining the order of the two words and ignoring any intermediate words. *ROUGE-SU* is similar to *ROUGE-S*, but it also counts unigrams (individual words) that occur both in  $S$  and  $Refs$ . The most widely used versions of *ROUGE* are *ROUGE-2* and *ROUGE-SU4*, which have been found to correlate well with human judgements, when multiple reference summaries are available per question; consult [Lin \(2004\)](#). *ROUGE-2* is *ROUGE-N* with  $n = 2$ ; and *ROUGE-SU4* is a version of *ROUGE-SU* with the maximum distance between the words of any skip bigram limited to 4.

In BIOASQ, we will use *ROUGE-2* and *ROUGE-SU4*, with  $S$  being an ‘ideal’ answer constructed by a system and  $Refs$  being any of the following:

- The golden ‘ideal’ answer of the particular question  $S$  was constructed for. Recall that there will be only one golden ‘ideal’ answer per question, and this may not allow *ROUGE-2* and *ROUGE-SU4* to correlate well with the scores of the manual evaluation.
- The correct snippets of Phase A for the particular question that  $S$  was constructed for.
- Pseudo-natural language renderings of the correct RDF triples for the particular question that  $S$  was constructed for.
- All the ‘ideal’ answers returned by the systems for the particular question that  $S$  was constructed for.
- All the ‘ideal’ answers returned by the systems for the particular question that  $S$  was constructed for, excluding returned ‘ideal’ answers that (i) were not manually evaluated by biomedical experts and (ii) were manually evaluated, but did not receive a score of at least 3 in all of the criteria of Table 4.3.
- Combinations of the above.

Table 4.4 summarizes the evaluation measures of Phase B; the official measures are shown in bold. We may also consider measures based on  $n$ -gram graphs ([Giannakopoulos et al. \(2008\)](#)), measures that do not require reference summaries [Louis and Nenkova \(2013\)](#) and combinations of measures ([Giannakopoulos and Karkaletsis \(2013\)](#)) to examine if we can use them in the second BIOASQ challenge.

Question type	Participant response	Evaluation measures
any	paragraph-sized text	<i>ROUGE-2</i> , <i>ROUGE-SU4</i> , <b>manual scores</b>

Table 4.4: Evaluation measures for the ‘ideal’ answers in Phase B of Task 1b.

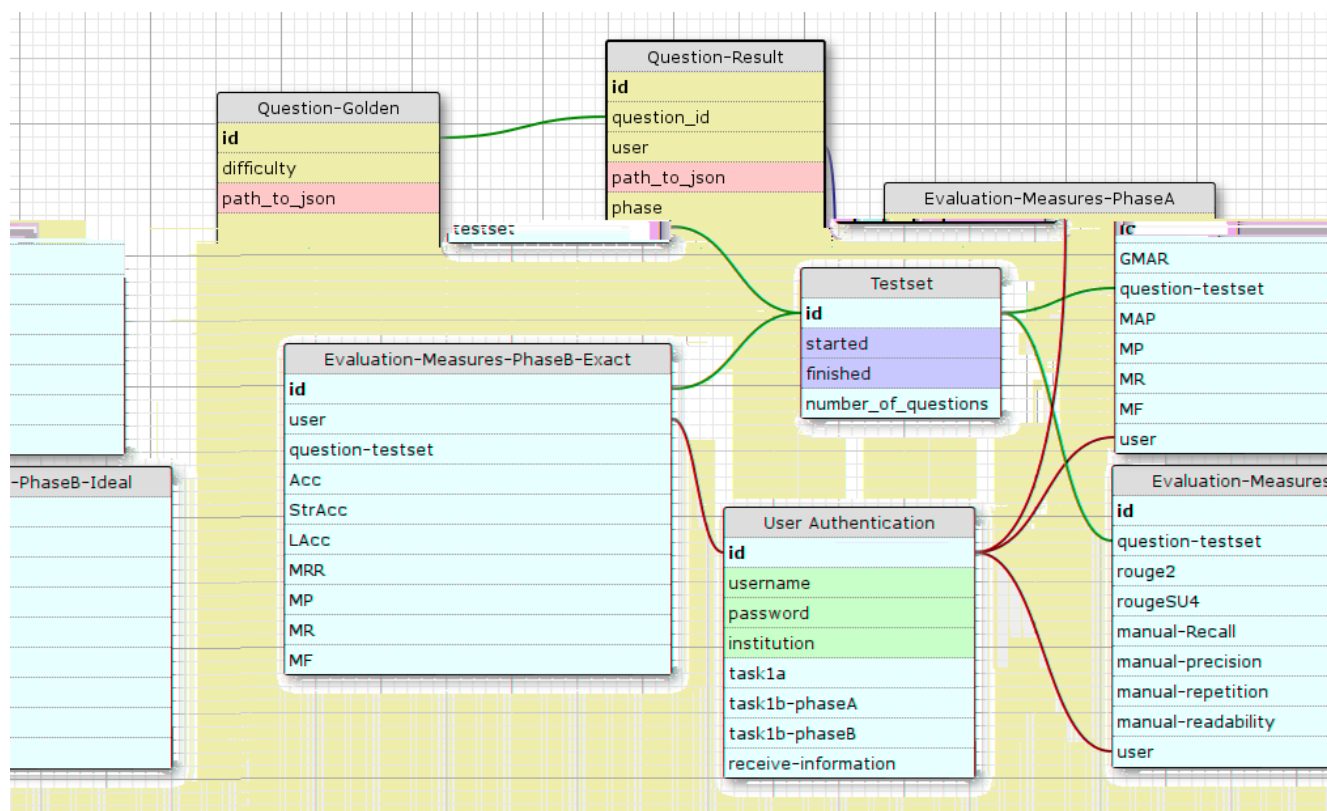


Figure 4.2: The database schema and the foreign key connections between the tables that will be used in task 1B.

## 4.5 Specification Guidelines

### 4.5.1 Database schema

Data for both phases of Task 1B will be stored in MySQL. The questions with the golden data produced by the biomedical experts will be saved in the database as files containing the JSON strings with the data. When necessary, the files will be processed and the test sets will be created. Three tables are necessary for saving the data:

- the Golden-Question table, where the questions will be saved along with the golden annotations produced by the biomedical experts.
- the Question-Results table, where the participants submitted results will be saved.
- the Evaluation-Measures tables, one for each phase of the task, where the performance and evaluation measures will be saved.



Figure 4.2 shows the database schema and the foreign key connections between the tables that are going to be used during Task 1B.

## 4.5.2 Web services for Task 1B Phase A

### Test set creation and download

In phase A of Task 1B, participants will be provided with questions from the database. In order to produce the test sets, internally, a function will select from the database the ids the types and the bodies of the questions produced from the biomedical experts. These data will create JSON strings that will be served to the participating systems. The format of the string will be the following:

```
{ "questions": [
  { "questionId": "0001",
    "questionType": "decisive"
    "questionBody": "Do CpG Islands colocalize with
transcription start sites",
    .
    .
  ] }
```

where:

- `questionId` will be the unique identifier of each question
- `questionBody` will be the question as it was formulated by the biomedical experts
- `questionType` will be the type of the question with values:
  - "desicive", for yes/no questions.
  - "factoid", for factoid questions.
  - "list", for list questions.
  - "summary", for summary questions.

Users will be able to download the JSON string using either the web interface or an API. In the second case, a GET request in a URI along with their username and password will result in downloading the test set.

### Test results upload

Users will upload their results in JSON strings. The format of the JSON string will be:

```
{ [ "questionId": { "concepts": [ "index_1", ..., "index_N" ],
  "articles": [ "PMID_1", ..., "PMID_M" ],
  "snippets": [ { "PMID": "23154875", "offset": [ "21", "68" ] }, ... ]
  "statements": [ "INDEX_1", ..., "INDEX_K" ]
}, .... ] }
```

where:

- `questionId` will be the unique identifier of each question and it will be the same given with the question when the test set was downloaded.

- `concepts` will be an array with the concepts from the indexed ontologies the system will estimate as relevant to the question.
- `articles` will be an array with the PMIDs of the articles that the system has estimated as relevant to the question.
- `snippets` will be an array. Each element will have a PMID and another array that will contain two numbers; the offsets of the beginning and the ending of the snippet in the article that the system has estimated as relevant.
- `statements` will be an array with the indices of the statements and the RDF triples that the system has estimated as relevant to the question.

The platform will store the system's answer after checking that the question's id is in the active test set and the indices exist in the ontologies which are provided for the task. What is also important in the JSON string is the ordering in the arrays with indices, as the ordering represents the decreasing confidence of the systems for their answers.

Users will be able to upload the JSON string either using the web interface and selecting a file in their computer or using an API and making a POST request to a selected URI. This process can be repeated as many times as needed before the test set expires; each time, the old upload will be deleted and the new answer will be saved.

### 4.5.3 Web services for Task 1B Phase B

#### Test set creation and download

In this phase, users will be provided with the questions as well as the golden answers of the phase A. Internally, we will select from the database the question id, the question type and the question body along with the golden annotations, which will be concepts, articles, snippets and triples that were provided from the biomedical experts during the benchmark creation. The format of the JSON string will be as follows:

```
{ "question_id": { "concepts": [ "index_1", ..., "index_N" ],
  "articles": [ "PMID_1", ..., "PMID_M" ],
  "snippets": [ { "PMID": "23154875", "offset": [ "21", "68" ] }, ... ],
  "statements": [ "INDEX_1", ..., "INDEX_K" ],
  "questionBody": "Do CpG Islands colocalize with
transcription start sites",
  "questionType": "decisive" },
.
.
. ] }
```

where:

- `questionId` will be the unique identifier of each question.
- `concepts` will be an array with the concepts from the indexed ontologies that the experts have selected.
- `articles` will be an array with the PMIDs of the articles that the experts have selected .

- `snippets` will be an array. Each element will have a PMID and another array that will contain two numbers; the offsets of the beginning and the ending of snippet in the article that the experts have selected.
- `statements` will be an array with the indexes of the statements and the RDF triples that the experts have selected.
- `questionBody` will be the question as it was formulated by the biomedical experts.
- `questionType` will be the type of the question with values:
  - `"decisive"`, for yes/no questions.
  - `"factoid"`, for factoid questions.
  - `"list"`, for list questions.
  - `"summary"`, for summary questions.

### Test results upload

The results that the users have to upload in this phase are the exact answers and the ideal answers. The format of the JSON string they have to submit is the following:

```
{[{"question_id": "0001",  
  "exact_answer": "your_exact_answer", "ideal_answer": "your_ideal_answer"}]}
```

where the `exact answer` and the `ideal answer` have been described in the previous sections. In the case of summary questions no `exact_answer` is needed.

---

## Annex

---

### Data Sources

#### MEDLINE

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

#### PubMed

PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. It comprises more than 22 million citations for biomedical literature through MEDLINE, life science journals, and on-line books. Citations may include links to full-text content from PubMed Central and publisher web sites. The United States National Library of Medicine (NLM) at the National Institutes of Health maintains the database as part of the Entrez information retrieval system. PubMed has an publically available web interface for accessing it's contents.

#### MeSH

MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. It consists of approximately 26.000 terms and new terms are added in a yearly basis. The terms are organised hierarchically in 12 trees. MEDLINE and PubMed use Medical Subject Headings (MeSH) for information retrieval. In addition, many engines (e.g. GoPubMed) are designed to access and search the MEDLINE content using MeSH terms.

#### GoPubMed

GoPubMed is a knowledge-based search engine for biomedical texts. The Gene Ontology (GO) and Medical Subject Headings (MeSH) serve as “Table of contents” in order to structure the millions of articles of the MEDLINE database. The search engine allows its users to find relevant search results significantly faster than PubMed.

---

## Bibliography

---

- Django. *Django: The web framework for perfectionists with deadlines*. URL <https://www.djangoproject.com/>.
- G. Giannakopoulos and V. Karkaletsis. Summary Evaluation: Together We Stand NPower-ed. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 436–450, Karlovassi, Samos, Greece, 2013.
- G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3):1–39, Oct. 2008. ISSN 1550-4875.
- S. Kiritchenko, S. Matwin, R. Nock, and A. Famili. Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization. In *19th Canadian Conference in Artificial Intelligence*, volume 4013, pages 395–406. Springer, 2006.
- A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. Evaluation Measures for Hierarchical Classification: a unified view through two generic frameworks. 2013.
- C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop ‘Text Summarization Branches Out’*, pages 74–81, Barcelona, Spain, 2004.
- A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, 2013.
- P. Malakasiotis, I. Androutsopoulos, Y. Almirantis, D. Polychronopoulos, and I. Pavlopoulos. Tutorials and Guidelines. Technical Report D3.4, BioASQ Deliverable, 2013.
- C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, volume 1. Cambridge University Press, 2008.
- A.-C. Ngonga Ngomo, N. Heino, R. Speck, T. Ermilov, and G. Tsatsaronis. Annotation Tool. Technical Report D3.3, BioASQ Deliverable, 2013.
- NLM. MeSH: Medical Subject Headings. Technical report, a. URL <http://www.ncbi.nlm.nih.gov/mesh>.

- NLM. MEDLINE. Technical report, b. URL <http://www.ncbi.nlm.nih.gov/pubmed>.
- S. Robertson. On GMAP: and other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 78–83, Arlington, Virginia, 2006.
- M. Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers, 2010.
- G. Tsatsaronis, M. Zschunke, M. R. Alvers, and C. Plonka. Report on existing and selected datasets. Technical Report D3.2, BioASQ Deliverable, 2013.
- G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining Multi-label Data. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.
- E. Voorhees. The TREC QA Track. *Natural Language Engineering*, 7(4):361–378, 2001.