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

Project **FP7-318652 / BioASQ**

Deliverable **D3.7**

Distribution **Public**

http://www.bioasq.org

Tutorials and Guidelines

Prodromos Malakasiotis, Ion Androutsopoulos, Yannis Almirantis, Dimitris Polychronopoulos and Ioannis Pavlopoulos

Status: Draft (Version 2.0)

December 2013

Project

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

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D3.7
Deliverable title	Tutorials and Guidelines
Contractual date of delivery	M15 (December 2013)
Actual date of delivery	December 2013
Relevant Task(s)	WP3/Task 3.4
Partner Responsible	AUEB-RC
Other contributors	NCSR “D”
Number of pages	23
Author(s)	Prodromos Malakasiotis, Ion Androutsopoulos, Yannis Almirantis, Dimitris Polychronopoulos and Ioannis Pavlopoulos
Internal Reviewers	George Balikas, Nicolas Baskiotis
Status & version	Draft
Keywords	benchmark datasets, biomedical experts, guidelines, tutorial, annotation tool

Executive Summary

This deliverable is the second version of the tutorial and guidelines that will be provided to the team of biomedical experts to help them create the questions, reference answers, and other supportive information that will be used in the benchmark dataset of the second BioASQ challenge. The guidelines provide directions regarding the number and types of questions to be created by the experts, the information sources the experts should consider and how to use them, the types and sizes of the reference answers and the other supportive information the experts should provide etc. Taking into consideration the feedback received by the team of biomedical experts, the annotation tool of deliverable D3.3 (Ngonga Ngomo et al. (2013)) was improved to its second version to better help the biomedical experts follow the guidelines. The new tool provides access to all the necessary resources, allows the questions, reference answers, and supportive information to be edited, saved etc., within a unified and easy to use Web interface. A tutorial illustrating the usage of the annotation tool is included in this deliverable, and will be provided to the biomedical expert team, along with access to the tool and the guidelines. More technical information about the annotation tool can be found in deliverable D3.6 (Heino (2013)).

Contents

List of Figures

List of Tables

1

Introduction

This deliverable comprises the second version of the tutorial and guidelines that will be provided to the BioASQ team of biomedical experts to help them create the questions, reference answers, and other supportive information that will be used in the benchmark dataset of the second cycle of the BioASQ challenge.

The guidelines provide directions regarding the number and types of questions to be created by the experts, the information sources the experts should consider and how to use them, the types and sizes of the reference answers and the other supportive information the experts should provide, etc. Taking into consideration the feedback received by the team of biomedical experts, the annotation tool of deliverable D3.3 (Ngonga Ngomo et al. (2013)) was improved to its second version to better help the biomedical experts follow the guidelines. The new tool provides access to all the necessary resources, allows the questions, reference answers, and supportive information to be edited, saved etc., within a unified and easy to use Web interface. A tutorial illustrating the usage of the annotation tool is included in this deliverable, and will be provided to the biomedical expert team, along with access to the tool and the guidelines. More technical information about the annotation tool can be found in deliverable D3.6 (Heino (2013)).

Chapters 2 and 3 below present the guidelines and the tutorial, respectively, that will be provided to the biomedical expert team. The tutorial, which will also be available as a slide-show presentation, presupposes that the experts have studied the guidelines.

2

Benchmark Creation Guidelines

Each biomedical expert should formulate *at least* 50 English questions, reflecting real-life information needs encountered during his/her work (e.g., in research or diagnosis). Each question should be stand-alone, i.e., it should not presuppose that any other questions have been asked; for example, it should not contain any pronouns referring to entities mentioned in other questions. For each question, the expert is also expected to provide an answer and other supportive information, as explained below.

To formulate each question and to provide the corresponding answer and supportive information, the expert should follow the following steps. An *annotation tool* will be made available to help the experts follow these steps, and a tutorial showing how to use the tool is provided in Chapter 3.

Step 1: Question formulation. Formulate an English stand-alone question reflecting real-life information needs. *At least 10 questions of each one of the following four categories* should be formulated by each biomedical expert; more than 10 questions will have to be formulated for some of the four categories, since *a total of at least 50 questions* is required.

Yes/no questions: These are questions that, strictly speaking, require either a "yes" or a "no" as an answer, though of course in practice a longer answer providing additional information that supports the "yes" or "no" will often be desirable. For example, "*Do CpG islands colocalise with transcription start sites?*" is a yes/no question.

Factoid questions: These are questions that, strictly speaking, require a particular entity (e.g., a disease, drug, or gene) as an answer, though again a longer answer providing additional supportive information may be desirable in practice. For example, "*Which virus is best known as the cause of infectious mononucleosis?*" is a factoid question.

List questions: These are questions that, strictly speaking, require a *list* of entities (e.g., a list of genes) as an answer; again, in practice additional supportive information may be desirable. For example, "*Which are the Raf kinase inhibitors?*" is a list question.

Summary questions: These are questions that do not belong in any of the previous categories and can only be answered by producing a short text summarizing the most prominent relevant information. For example, "*What is the treatment of infectious mononucleosis?*" is a summary question.

When formulating summary questions, the experts should aim at questions that they can answer (possibly after consulting the literature) in a satisfactory manner by writing a one-paragraph summary intended to be read by other experts of the same field.

In all four categories of questions, the experts should aim at questions that when converted to PUBMED queries, as discussed below, retrieve approximately 10–60 articles (or abstracts). Questions for which there are controversial or no answers in the literature should be avoided. Moreover the questions should be related to the biomedical domain. Consider for example the following two questions:

Q_1: Which are the differences between Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs)?

Q_2: Which are the uses of Hidden Markov Models (HMMs) in gene prediction?

Note that although HMMs and ANNs are used in the biomedical domain, Q_1 above is not suitable for the needs of BIOASQ since there is not a direct indication that it is related to the biomedical domain. On the other hand Q_2 which links "gene prediction" with HMMs is appropriate and should be preferred.

Step 2: Relevant terms. Form a set of terms that are relevant to the question of Step 1. The set of relevant terms may include terms that are already mentioned in the question, but it may also include synonyms of the question terms, closely related broader and narrower terms etc. For the question "*Do CpG islands colocalise with transcription start sites?*", the set of relevant terms would most probably include the question terms "*CpG Island*" and "*transcription start site*", and possibly also other terms.

Step 3: Information retrieval. Facilities will be provided to formulate a query (Boolean or simple bag of terms) involving the relevant terms of Step 2, as well as to retrieve articles from PUBMED that satisfy the query (or abstracts, when only abstracts are available). The query can be enriched with the advanced search tags of PUBMED.[1] Facilities will also be provided to execute the query against biomedical terminology banks, databases, and knowledge bases, in order to obtain possibly relevant *concepts* (e.g., MESH headings) and relations (e.g., a database may show that a particular disease is known to cause a particular symptom). Relations retrieved from databases and knowledge bases will be shown in the annotation tool as pseudo-natural language statements, hereby called simply *statements*; hence, the experts do not need to be familiar with how information is actually represented in the databases and knowledge bases. Note that when retrieving concepts and statements, advanced search tags are ignored. Furthermore, when retrieving concepts, Boolean operators are also ignored, i.e., Boolean queries are turned into bag of terms queries.

Returning to the example question "*Do CpG islands colocalise with transcription start sites?*" of Step 1, a possible Boolean query involving the relevant terms of Step 2 might be "*CpG Island*" *AND* "*transcription start site*". The concepts, articles, and statements retrieved by this query are shown below; we only show the titles of the articles to save space, but the annotation tool will allow the experts to view the entire articles or their abstracts (when only abstracts are available).[2] Shown in brackets are the names of the resources the concepts come from.

[1] See http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Search_Field_Descrip for a detailed description of the tags.

[2] More concepts are actually retrieved; we only show the first 10 to save space.

Retrieved concepts:

1. *"Transcription Initiation Site"* (MESH)
2. *"Factor VIII intron 22 protein (Homo sapiens)"* (UNIPROT)
3. *"Factor VIII intron 22 protein (Mus musculus)"* (UNIPROT)
4. *"CpG Islands"* (MESH)
5. *"regulation of transcription, start site selection"* (GENE Ontology)
6. *"hypermethylation of CpG island"* (GENE Ontology)
7. *"hypomethylation of CpG island"* (GENE Ontology)
8. *"DNA-dependent transcriptional start site selection"* (GENE Ontology)
9. *"Cyclic 2,3-diphosphoglycerate synthetase"* (UNIPROT)
10. *"Cyclic 2,3-diphosphoglycerate synthetase"* (UNIPROT)

Retrieved articles (only titles shown here):

1. *"Putative Zinc Finger Protein Binding Sites Are Over-Represented in the Boundaries of Methylation-Resistant CpG Islands in the Human Genome"*
2. *"CpG Islands: Starting Blocks for Replication and Transcription"*
3. *"Periodicity of SNP distribution around transcription start sites"*
4. *"Comprehensive analysis of the base composition around the transcription start site in Metazoa"*
5. *"DBTSS: DataBase of Human Transcription Start Sites, progress report 2006"*
6. *"Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression"*
7. *"CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences"*
8. *"CpG islands in vertebrate genomes"*
9. *"Dynamic usage of transcription start sites within core promoters"*
10. *"Boosting with stumps for predicting transcription start sites"*

Retrieved statements:

1. *"Methyl-cpg-binding domain protein 2's specific function is binds cpg islands in promoters where the dna is methylated at position 5 of cytosine within cpg dinucleotides. binds hemi-methylated dna as well. recruits histone deacetylases and dna methyltransferases. acts as transcriptional repressor and plays a role in gene silencing. isoform 1 may enhance the activation of some unmethylated camp-responsive promoters. reports about dna demethylase activity of isoform 2 are contradictory."*

Step 4: Selection of concepts, articles, statements. *All* the retrieved concepts of Step 3 that closely correspond to the terms of Step 2 and the type of the expected exact answer (if the type of the expected answer is clear from the question) should be selected. Continuing with our example, the retrieved concepts "Transcription Initiation Site" (MESH) and "CpG Islands" (MESH) correspond to the terms "Transcription start site" and "CpG island" of Step 2 and, hence, should be selected. Note that the selected concepts should correspond closely to the terms of Step 2 and the type of the expected exact answer (if the type is clear from the question), and they should be as specialized as possible, without assuming that the exact answer is known in advance. For instance, consider the following questions:

Q_1: *Which anti-inflammatory drugs are used to treat back pain?*

Q_2: *Which drugs are used to treat back pain?*

For Q_1 the concept "Anti–inflammatory Agents (MESH)" should be considered as relevant but the more general concept "Drugs (MESH)" should not. On the other hand, for Q_2, the concept "Drugs (MESH)" should be considered as a relevant concept but the more specialized concept "Anti–inflammatory Agents (MESH)" should not as it is not directly implied by the question. Also, *all* the articles of Step 3 that are *possibly relevant* to the question should be selected. By 'possibly relevant' we mean articles that the expert would want to read more carefully in practice, to check if they contain information that is useful to answer the question. At this step, the expert is only expected to skim through the retrieved articles (or their abstracts) to figure out if they are possibly relevant. Finally, *every* statement of Step 3 that provides information that is useful to answer the question should be selected, even if the statement does not provide on its own all of the information that is needed to answer the question. In our example, the following concepts, documents, and statements might be selected:

Selected concepts:

1. *"Transcription Initiation Site"* (MESH)
4. *"CpG Islands"* (MESH)
5. *"regulation of transcription, start site selection"* (GENE Ontology)
6. *"hypermethylation of CpG island"* (GENE Ontology)
7. *"hypomethylation of CpG island"* (GENE Ontology)
8. *"DNA-dependent transcriptional start site selection"* (GENE Ontology)

Selected articles (only titles shown here):

2. *"CpG Islands: Starting Blocks for Replication and Transcription"*
4. *"Comprehensive analysis of the base composition around the transcription start site in Metazoa"*
5. *"DBTSS: DataBase of Human Transcription Start Sites, progress report 2006"*
7. *"CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences"*
8. *"CpG islands in vertebrate genomes"*
9. *"Dynamic usage of transcription start sites within core promoters"*
10. *"Boosting with stumps for predicting transcription start sites"*

Selected statements:

1. *"Methyl-cpg-binding domain protein 2's specific function is binds cpg islands in promoters where the dna is methylated at position 5 of cytosine within cpg dinucleotides. binds hemi-methylated dna as well. recruits histone deacetylases and dna methyltransferases. acts as transcriptional repressor and plays a role in gene silencing. isoform 1 may enhance the activation of some unmethylated camp-responsive promoters. reports about dna demethylase activity of isoform 2 are contradictory."*

Note that if no concept that meets the criteria mentioned above can be found, then the question is not suitable for the needs of BIOASQ and should not be preferred.

Step 5: Text snippet extraction. At this stage, the expert should read (or skim through more carefully) the set of possibly relevant articles selected during Step 4. *Every* text snippet (piece of text) that provides information that is useful to answer the question of Step 1 should be extracted, even if the snippet on its own does not provide all of the information that is needed to answer the question. Note that in the second year a text snippet should contain one or more entire sentences (consecutive sentences, if the snippet contains more than one sentences; e.g., starting after a full stop and ending with a full stop); snippets should not contain only fragments of sentences, they should contain one or more entire sentences. If there are multiple snippets that provide the same (or almost the same) useful information (in the same article or in different articles), *all* of them should be extracted, not just one of them. Snippets can be easily extracted using the annotation tool, much as one might highlight snippets that provide useful information when reading an article. In our example, the following snippets might be extracted. The numbers in square brackets point to the articles of Step 4 the snippets were extracted from.

- *"A common explanation for the G+C rise that is seen here in the mammalian profile in the proximity of the TSS is the presence of CpG islands,"* [4]
- *"Above we have made the remark that the G+C rise in mammals and maybe generally in vertebrates is probably caused by the higher number of CpG dinucleotides in the promoter region."* [4]
- *"This could mean that there is some DNA methylation and some CpG over-representation around TSS but not as much as in human."* [4]
- *"The results for Fugu (Fig. 4C,4D) show that some genes could have CpG islands (Fig. 4D) since for those the nucleotide composition is similar to the mammalian profiles."* [4]
- *"Nucleotide composition and gene expressionIt is generally known that the presence of a CpG island around the TSS is related to the expression pattern of the gene. Unmethylated DNA can have an open chromatin structure that facilitates the interaction of transcription factors with the promoter region [15]. Housekeeping genes (HK genes), which are transcribed in all somatic cells and under all circumstances (and thus should be easily activated) frequently have a CpG island in their promoter region [16,17]."* [4]
- *"CpG islands are good markers of some classes of genes because they are often linked to the promoters of those genes"* [5]
- *"In most cases, CpG islands escape DNA methylation, which suppresses gene expression in general, in almost every tissue [10] and function as part of the gene promoter [11]."* [5]
- *"In the human genome, CpG-rich promoters or CpG island promoters are dominant, occurring more than twice as often as CpG-poor promoters"* [5]
- *"Currently, the presence of CpG islands in invertebrate animals is unclear."* [5]
- *"It is well known that the enrichment of the CpG dinucleotides in CpG island promoters is maximum in TSSs [12,13], so TSSs constitute candidate regions in which CpG island promoters or CpG island-like sequences might occur in the invertebrate genome."* [5]
- *"The CpG-rich promoters can be considered to contain a CpG island."* [5]
- *"his observation led to the hypothesis that human CpG-poor promoters emerged with the deamination of methylated CpG dinucleotides in CpG island promoters"* [5]
- *"Our results confirmed that the ascidian promoters tended to have high CpG score and G+C contents around TSS, as was observed in the human promoters."* [5]

- *"Although the ascidian TSSs exhibited quite high CpG score, this fact does not necessarily mean that they have high frequency of the CpG dinucleotide"* [5]
- *"ascidian promoters tended to exhibit high CpG scores"* [5]
- *"CpG island promoters must have appeared in an early stage of vertebrate evolution"* [5]
- *"The sequences near TSSs tend to exhibit high CpG score and high G+C content, but their level and extent are actually restricted."* [5]
- *"Considering that 67.5% of responsive genes have CpG islands,"* [7]
- *"We found that more than a third (33.4-34.1%) of these tissue-specific genes had CpG islands"* [7]
- *"another observation that 24% of brain-specific promoters have CpG islands"* [7]
- *"CGIs often extend into downstream transcript regions. This provides an explanation for the observation that the exon at the 5' end of the transcript, flanked with the transcription start site, shows a remarkably higher CpG density than the downstream exons"* [8]
- *"Genes with a CGI in their promoter tended to be regulated by H3K36me3 rather than nucleosomes or CpG methylation, probably for efficient transcription elongation"* [8]
- *"CGIs and NFRs tend to coexist in some promoters, together marking an active chromatin configuration"* [8]
- *"CpG methylation is proposed to cooperate with nucleosomes and H3K36me3 to differentially regulate the elongation of pol II."* [8]
- *"These associations are consistent with the previous finding that broad tag clusters are associated with CpG islands"* [9]
- *"An interpretation of this fine-grained tissue specificity is that the differential methylation of each CpG dinucleotide affects the transcription machinery, and results in different specificities without a clear positional bias"* [89]
- *"Although there has been much success in locating the TSSs for CpG-related promoters, the performance for non-CpG-related promoters (about 25% of known genes) is still not satisfactory because of the diverse nature of vertebrate promoter sequences"* [10]

Tables 2.1 and 2.2 provide some more examples of snippets as they were extracted during the first year of BIOASQ. To measure inter–annotator agreement, the experts were coupled into five pairs with the requirement that the experts of each pair should have the same or similar research areas. The members of each pair were then asked to come up with at least five questions which they would then try to answer independently, i.e., without helping each other. The examples of Tables 2.1 and 2.2 come from such questions. "Snippet 1" in this table was extracted by the first expert of a pair, while "Snippet 2" by the second one. The snippets that should have been extracted considering the second version of the guidelines, are shown in the third column of each question.

Step 6: Query revision. If the expert believes that the articles (or abstracts), snippets, and statements gathered during Steps 2–5 do not provide enough information to answer the question, or if the expert believes that additional relevant concepts, articles, snippets, or statements could have been retrieved, the terms of Step 2 and the query of Step 3 should be revised. The revised query will be used to perform a new search, which may produce different concepts, articles, and statements; the expert will again select (in Step 4) concepts, articles, and statements among those retrieved, and then snippets (in Step 5). It is important to try to retrieve *all* the relevant concepts, articles,

What is the mode of inheritance of nemaline myopathy?		
Snippet 1	Snippet 2	Correct snippet
A missense mutation, Glu41Lys, in the beta-tropomyosin gene TPM2 was identified in both patients but was absent in their healthy relatives. CONCLUSIONS: The results indicate that mutations in TPM2 may cause nemaline myopathy as well as cap disease with a dominant mode of inheritance.	The results indicate that mutations in TPM2 may cause nemaline myopathy as well as cap disease with a dominant mode of inheritance.	The results indicate that mutations in TPM2 may cause nemaline myopathy as well as cap disease with a dominant mode of inheritance.
the mode of inheritance appears to be recessive. Apart from a few instances of dominant inheritance, most cases published also seem compatible with recessive inheritance.	We conclude that in the Finnish CNM patients, the mode of inheritance appears to be recessive. Apart from a few instances of dominant inheritance, most cases published also seem compatible with recessive inheritance.	We conclude that in the Finnish CNM patients, the mode of inheritance appears to be recessive. Apart from a few instances of dominant inheritance, most cases published also seem compatible with recessive inheritance.
These may represent heterozygous manifestations of recessive gene.	–	These may represent heterozygous manifestations of recessive gene.
investigate the inheritance in congenital nemaline myopathy (CNM)	–	In order to investigate the inheritance in congenital nemaline myopathy (CNM), we studied the family histories and pedigrees of 13 patients with CNM from 10 families, and the 20 patients, by physical examination, single fibre electromyography, ultrasonography of muscles, measurement of serum creatine kinase, muscle biopsy, and electrophoresis of muscle proteins.

Table 2.1: Example of snippets extracted during the first cycle of BIOASQ.

Does Serca2a bind PLN in the heart?		
Moreover, PLN-R14Del did not co-immunoprecipitate with SERCA2a (as did WT-PLN),	Consistent with the lack of inhibition on SR Ca-transport and contractility, confocal microscopy indicated that the PLN-R14Del failed to co-localize with SERCA2a. Moreover, PLN-R14Del did not co-immunoprecipitate with SERCA2a (as did WT-PLN), but rather co-immunoprecipitated with the sarcolemmal Na/K-ATPase (NKA) and stimulated NKA activity.	Moreover, PLN-R14Del did not co-immunoprecipitate with SERCA2a (as did WT-PLN), but rather co-immunoprecipitated with the sarcolemmal Na/K-ATPase (NKA) and stimulated NKA activity.
The human phospholamban Arg14-deletion mutant localizes to plasma membrane and interacts with the Na/K-ATPase	–	–

Table 2.2: Example of snippets extracted during the first cycle of BIOASQ.

snippets, and statements, using more than one queries if necessary. The annotation tool provides facilities that allow the concepts, articles, and statements that the expert has already selected (before performing a new search) to be saved, along with the snippets the expert has already extracted. The query can be revised several times, until the expert feels that the gathered information is sufficient to answer the question and no relevant concepts, articles, snippets, and statements have been missed. If despite revising the query the expert feels that the gathered information is insufficient, or if there seem to be controversial answers, the question should be discarded.

Step 7: Exact answer. At this step, the expert should provide what we call an *exact answer* for the question of Step 1. For a yes/no question, the exact answer should be simply "yes" or "no". For a factoid question, the exact answer should be the name of the entity (e.g., gene, disease) seeked by the question; if the entity has several names, the expert should provide, to the extent possible, all of its names, as explained in the tutorial of Chapter 3. For a list question, the exact answer should be a list containing the entities seeked by the question; if a member of the list has several names, the expert should provide, to the extent possible, all of the member's names, again as explained in the tutorial of Chapter 3. For a summary question, the exact answer should be left blank. The exact answers of yes/no, factoid, and list questions should be based on the information of the statements and text snippets that the expert has selected and extracted in Steps 4 and 5, respectively, rather than, for example, personal experience.

Step 8: Ideal answer. At this step, the expert should formulate what we call an *ideal answer* for the question of Step 1. The ideal answer should be a one-paragraph text that answers the question of Step 1 in a manner that the expert finds satisfactory. The ideal answer should be written in English, and it should be intended to be read by other experts of the same field. For the example yes/no question "*Do CpG islands colocalise with transcription start sites?*", an ideal answer might be the following:

"Yes. It is generally known that the presence of a CpG island around the TSS is related to the expression pattern of the gene. CGIs (CpG islands) often extend into downstream transcript regions. This provides an explanation for the observation that the exon at the 5' end of the transcript, flanked with the transcription start site, shows a remarkably higher CpG density than the downstream exons."

The ideal answer should be based on the information of the statements and text snippets that the expert has selected and extracted in Steps 4 and 5, respectively, rather than, for example, personal experience. The experts, however, are allowed (and should) rephrase or shorten the statements and snippets, order or combine them etc., in order to make the ideal answer more concise and easier to read etc.

Notice that in the example above, the ideal answer is longer than the exact one ("yes"), and that the ideal answer provides additional information supporting the exact answer. If the expert feels that the exact answer of a yes/no, factoid, or list question is sufficient and no additional information needs to be reported, the ideal answer can be the same as the exact answer. For summary questions, an ideal answer must always be provided.

3

Annotation Tool Tutorial

The biomedical experts will be assisted in creating the benchmark sets (questions, answers, and supportive information) by the second version of the annotation tool. The annotation tool can be used via a Web interface, which is available at: `http://at.bioasq.org/`. This chapter demonstrates the usage of the annotation tool, assuming that the reader has already studied the guidelines of Chapter 2. More technical information about the annotation tool can be found in deliverable D3.6 (Heino (2013)).

3.1 Registration and log-in

Since the biomedical experts had registered in order to use the first version of the annotation tool, they do not need to register again. Each expert can log into the annotation tool by filling in his/her e-mail address and password (the ones entered during registration) and clicking on the "Login" button of the annotation tool initial page (Figure 3.1). Experts who have forgotten their passwords should click on the "Forgot your password?" button (Figure 3.1) to receive further instructions.

3.2 Question formulation

Having logged in, the expert can view all the questions he/she has created so far. A tick mark appears next to each finalized question (Figure 3.2). The expert can create a new question by clicking on the "+ Add question" button (Figure 3.2). A form will then appear (Figure 3.3), where the expert can fill in the question (in English) and select its type ("yes/no" question, factoid question, list question, or summary question). Consult Step 1 of the guidelines of Chapter 2 for more information on the types of questions.

After filling in the question and selecting its type, the expert should click on the "Save" button (Figure 3.3) to save the question. By clicking on the question the expert can view the details of the selected question. He/she can then edit the question (change the phrasing, the question type, etc.) or delete it. By clicking on the "Proceed" button the expert can move on to the next step (Figure 3.4).

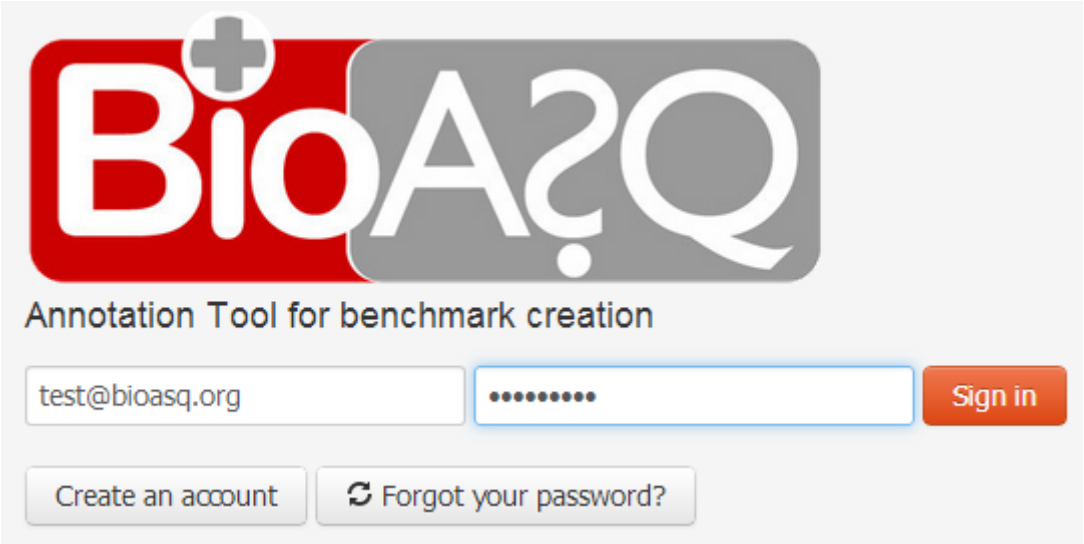

Figure 3.1: Logging into the annotation tool.

Figure 3.2: Creating a new question or selecting a previously created one.

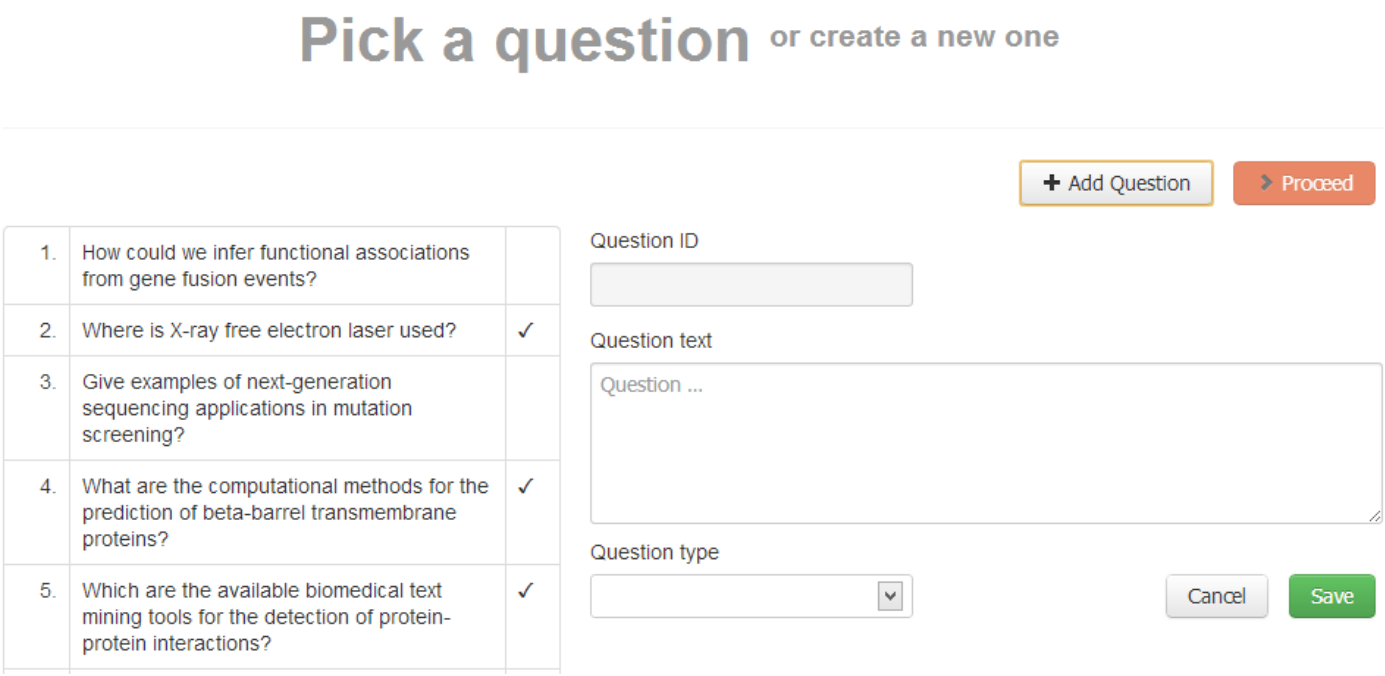

Figure 3.3: Creating a new question.

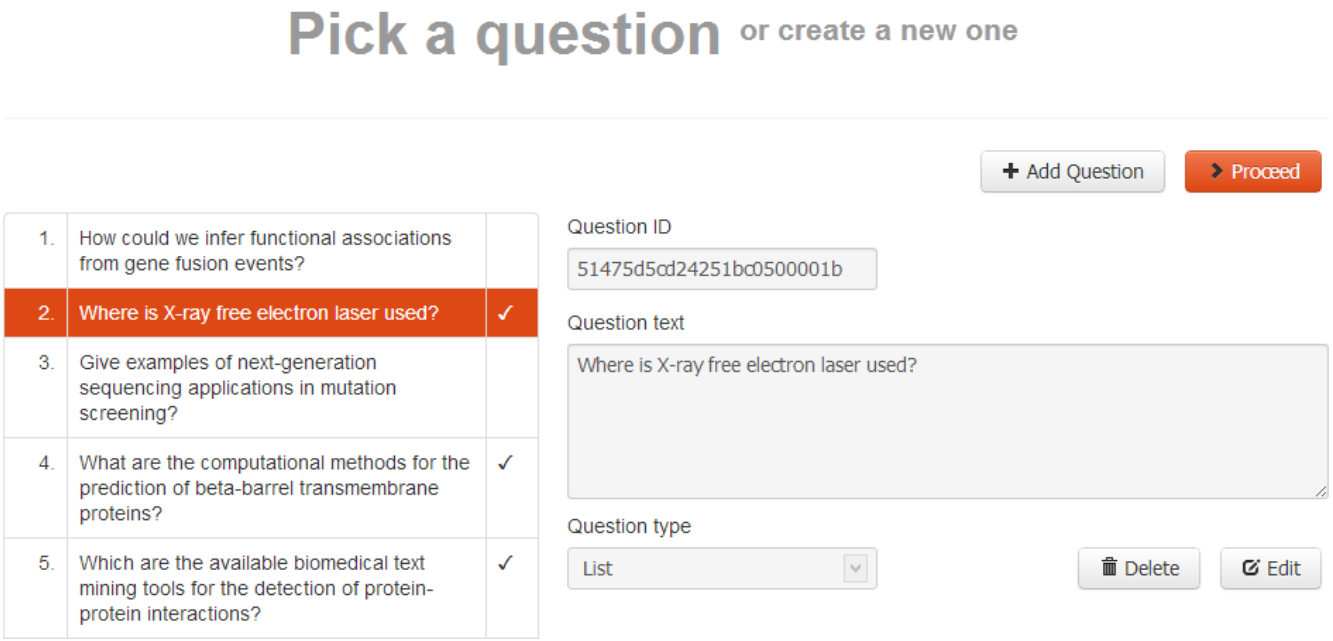

Figure 3.4: Selecting a question

3.3 Relevant terms and information retrieval

Having selected a question to work with, the expert can proceed to formulate a query involving terms that are relevant to the question, as discussed in Steps 2 and 3 of the guidelines of Chapter 2. The query has to be entered in the "Query..." text box of Figure 3.5. It can be a "bag-of-words" query or a Boolean query. A bag-of-words query is simply a set of terms, as in the following example:

```
"di-glycine signature" Trypsin human
```

Multi-word terms, like "di-glycine signature", should be enclosed in quotation marks, as in the example above. The annotation tool attempts to retrieve concepts, articles, and statements that contain as many as possible of the specified terms. Recall that statements are entity relations retrieved from databases and knowledge bases, shown as pseudo-natural language sentences.

In Boolean queries, the terms are connected with AND and OR operators; brackets can also be used to clarify the scope of the operators.[1] Multi-word terms are again enclosed in quotation marks. For example, the following Boolean query retrieves articles that contain the term "disease" and (at the same time) at least one (or both) of the terms "quantitative trait loci" or "splicing".

```
disease AND ("quantitative trait loci" OR "splicing")
```

Once the query has been entered, clicking on the "Search" button (Figure 3.5) executes the query.

3.4 Selection of concepts, articles, and statements

When the search specified by the query is completed, three lists containing concepts, articles (shown as "documents"), and statements appear (Figure 3.6). The contents of these lists can be viewed by clicking on the them. The expert should select *all* the retrieved concepts that closely correspond to the terms of the query, *all* the possibly relevant articles (all the articles that the expert feels he/she should read or skim through more carefully), and *all* the statements that provide information that is useful to answer the question, as discussed in Step 4 of the guidelines of Chapter 2.

When a list is expanded, the expert can select items (concepts, documents, or statements) from the list by clicking on the corresponding "+" icons (Figures 3.7, 3.8 and 3.9). When an item is selected, its "+" icon turns into a "–" icon. If an item has been accidentally selected, clicking on its "–" icon will remove it from the set of selected items. Figures 3.7,3.8 and 3.9 show examples of selecting concepts, documents and statements respectively. In order to decide whether a document (article) is possibly relevant or not, the expert can view (inspect) it by clicking on the "⇝" icon (Figure 3.8). An "⇝" icon is also available for each concept and by clicking it some additional information concerning the concepts is displayed. Recall that the concepts come from biomedical terminology banks, databases, and knowledge bases (Chapter 2) and not all of them are appropriate for every query. For that reason, five buttons appear above the retrieved concepts (Figure 3.7). Each button corresponds to a resource from which concepts are retrieved. By clicking on these buttons, the expert can see the retrieved concepts of the corresponding resource. A deeper orange colour of the button indicates the resource of the corresponding concepts.

3.5 Text snippet extraction and answer formulation

Having selected concepts, documents, and statements, the expert should now read (or skim through more carefully) the possibly relevant articles he/she selected. By clicking on the "Answer" tab of the

[1] Other operators are also available, but AND and OR should suffice in most cases.

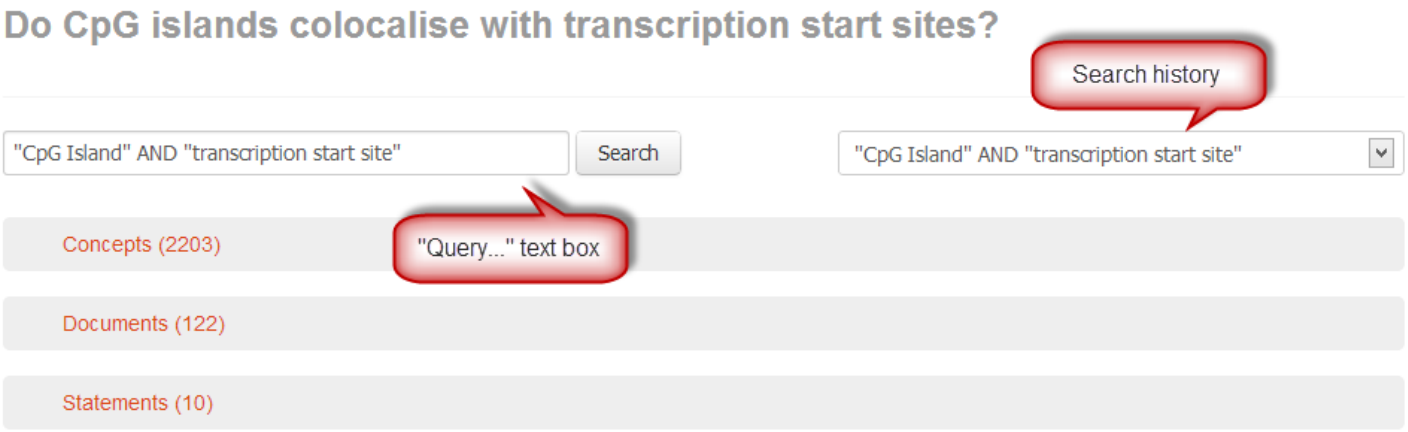

Figure 3.5: Performing a search.

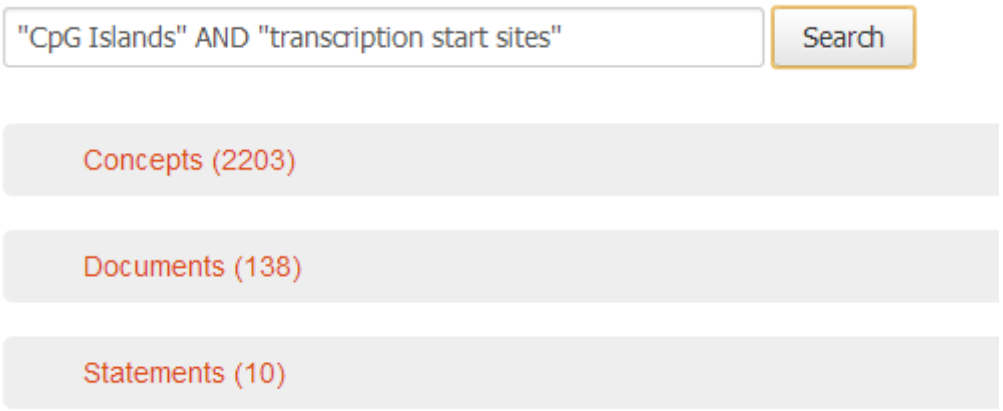

Figure 3.6: Search results.

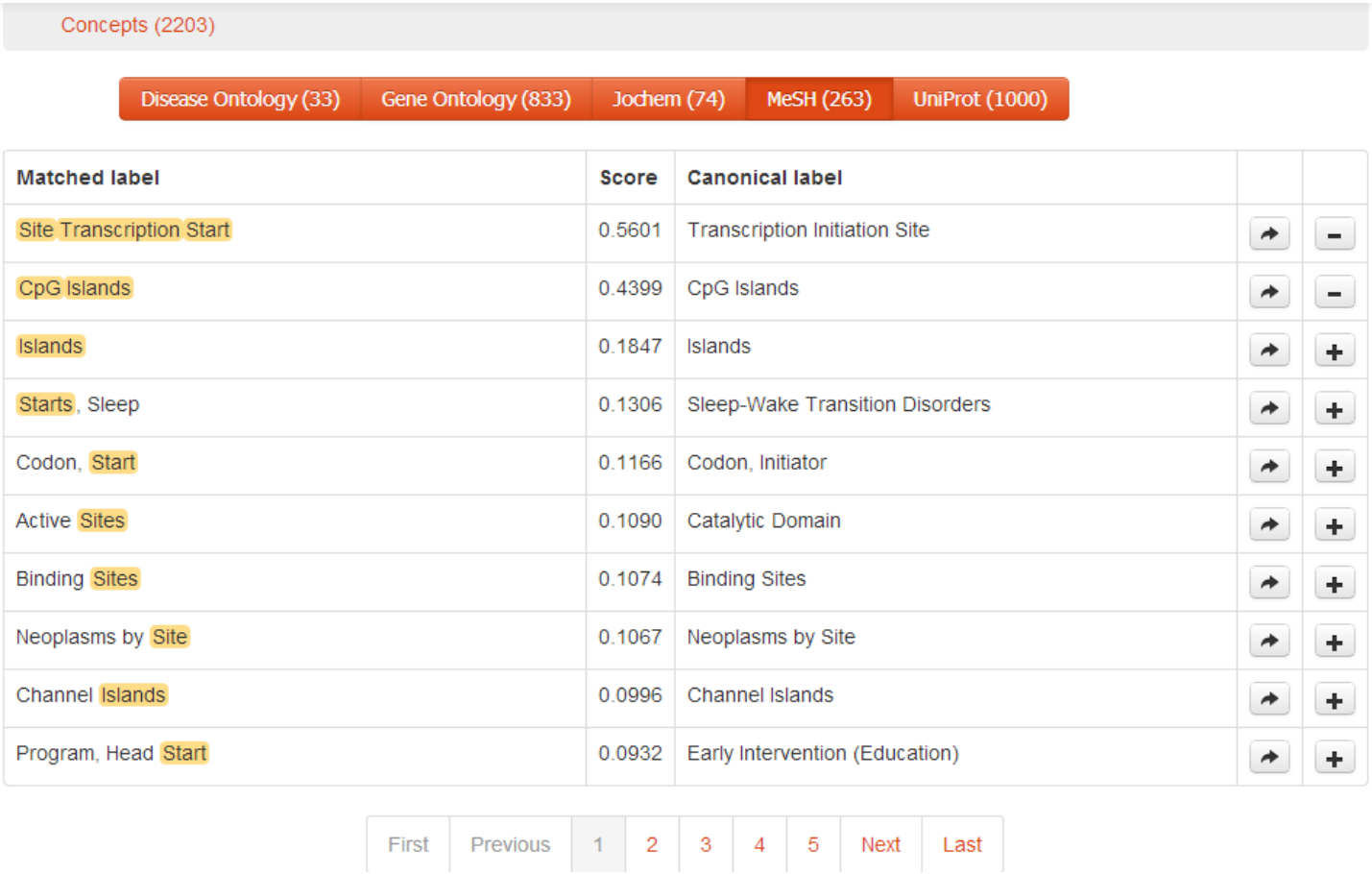

Figure 3.7: Concept selection.

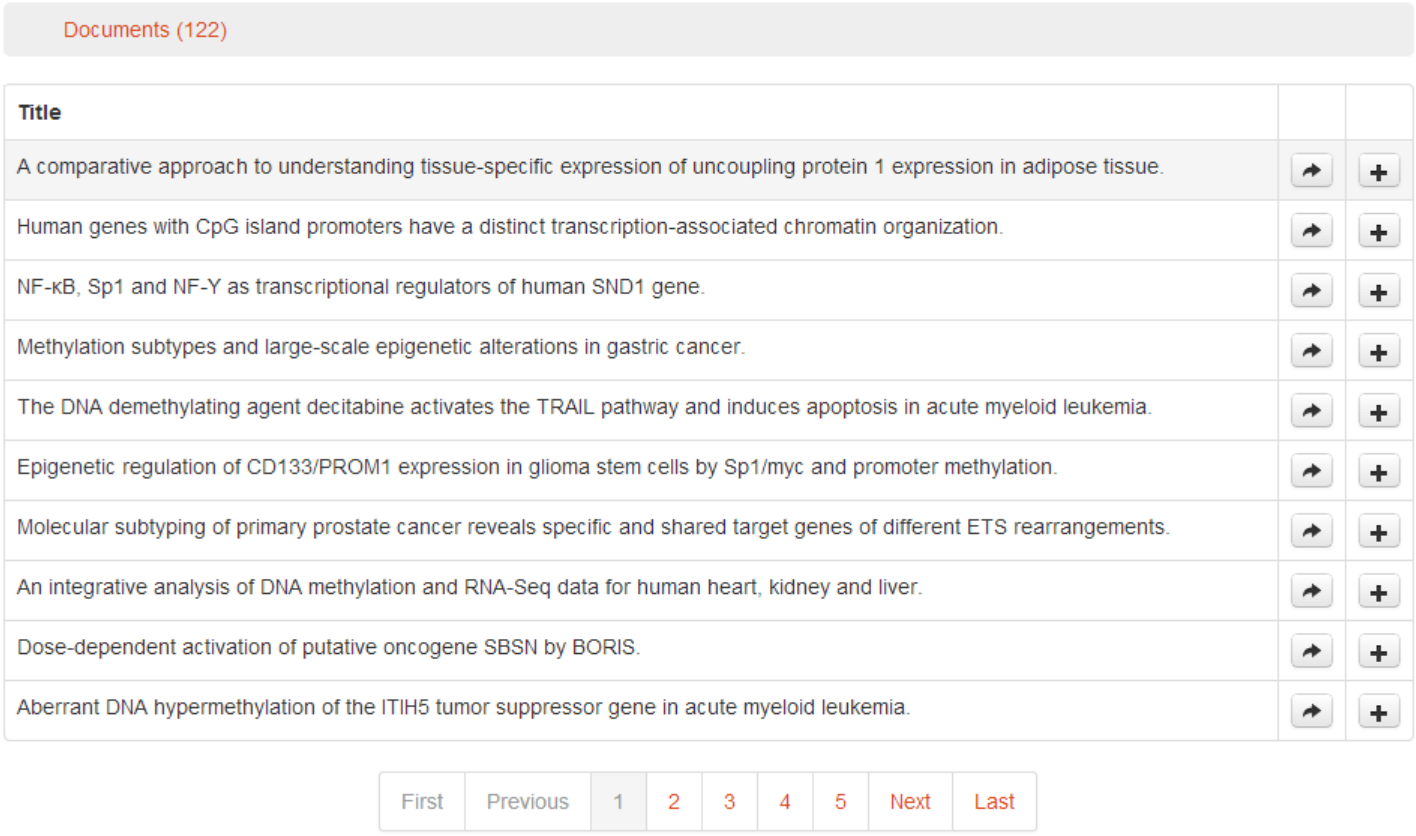

Figure 3.8: Document selection.

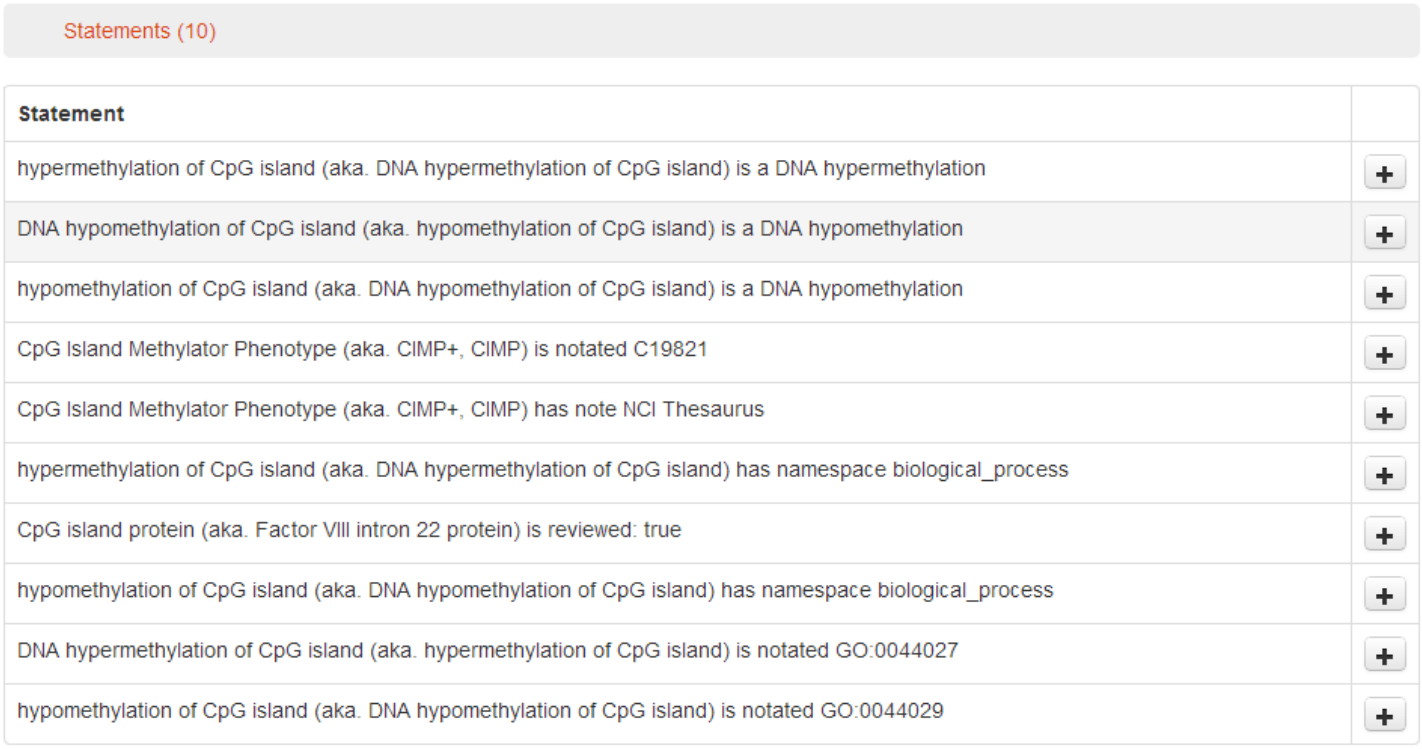

Figure 3.9: Statement selection.

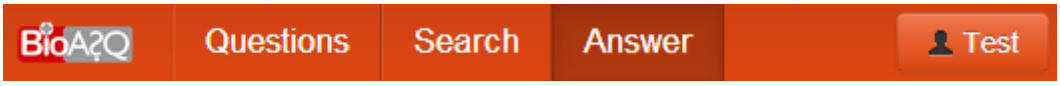

Figure 3.10: Selecting the "Answer" tab of the upper navigation menu.

upper navigation menu (Figure 3.10) the expert can formulate the answer to the question. On the upper part of the screen he/she can formulate the 'exact' and 'ideal' answer of the question (Figure 3.11). Concerning the exact answer different options are provided depending on the type of the question. For yes/no questions the expert can click on the appropriate radio button to select either "Yes" or "No" (Figure 3.14). For factoid questions the expert can use the ">" button to add synonyms (Figure 3.12). Finally, for list questions the expert can use the "∨" button to add an item to the list and the ">" button to add a synonym of the corresponding item to the list (Figure 3.13).

On the lower part of the screen the expert can see the items (concepts, documents, statements) that have been selected during the search, by clicking on their type (Figure 3.14). To remove an item (e.g., to remove a document that turned out not to be relevant), click on its "X" icon. Again, clicking on the "⇝" icon of an item displays the original source of the item (e.g., the corresponding PUBMED page for articles).

Clicking on the title of an article (document) displays the article (or its abstract, if only the abstract is available) and allows snippets to be extracted from the article (Figure 3.15), as discussed in Step 5 of the guidelines of Chapter 2. To extract a snippet, highlight it with the mouse and click on the "Annotate with selected snippet" button (Figure 3.15). The extracted snippet then appears highlighted in gold. Clicking on the "X" button at the end of the snippet cancels the extraction (selection) of the corresponding snippet (Figure 3.16).

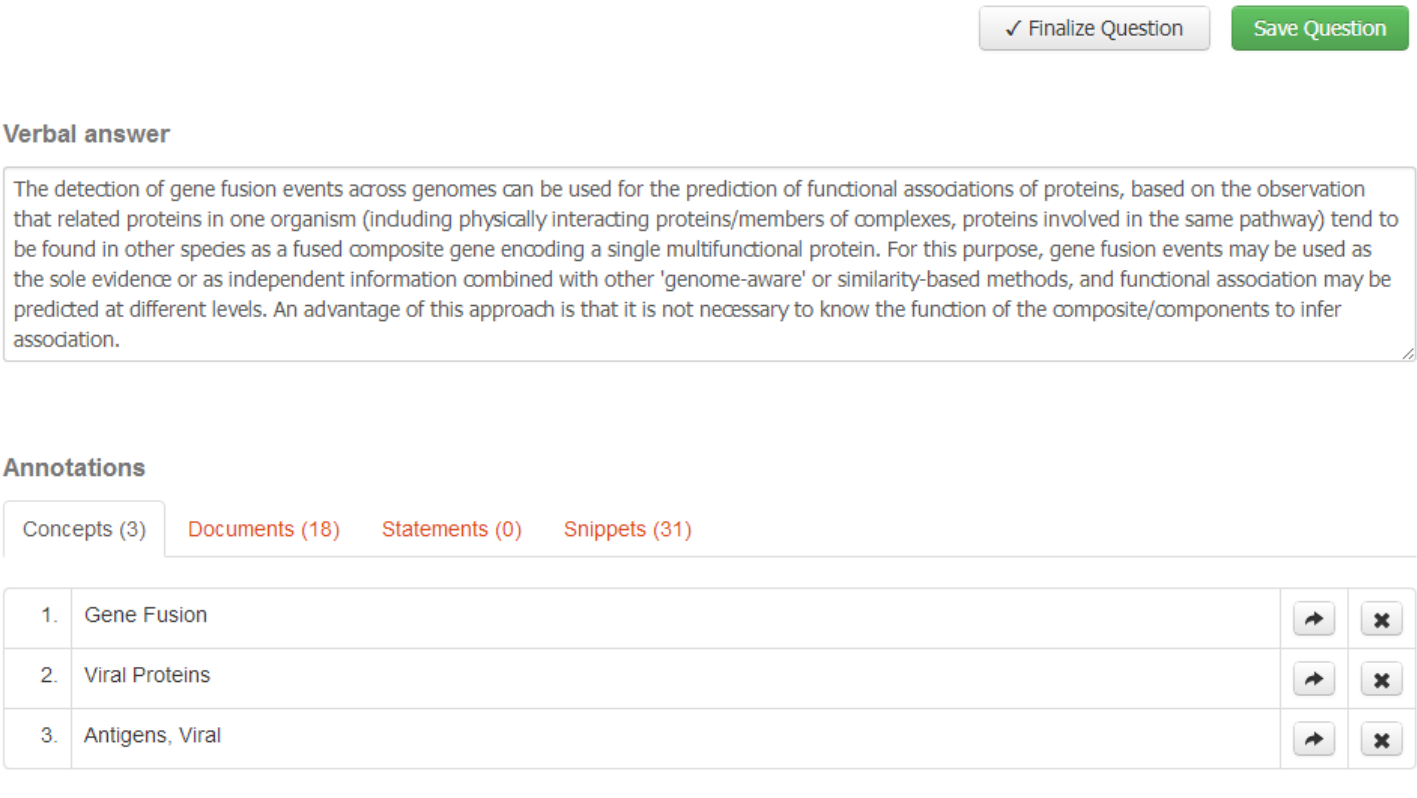

Figure 3.11: The answer formulation screen. On the upper part the expert can formulate the answer; on the lower part the selected items appear. This example corresponds to a summary question.

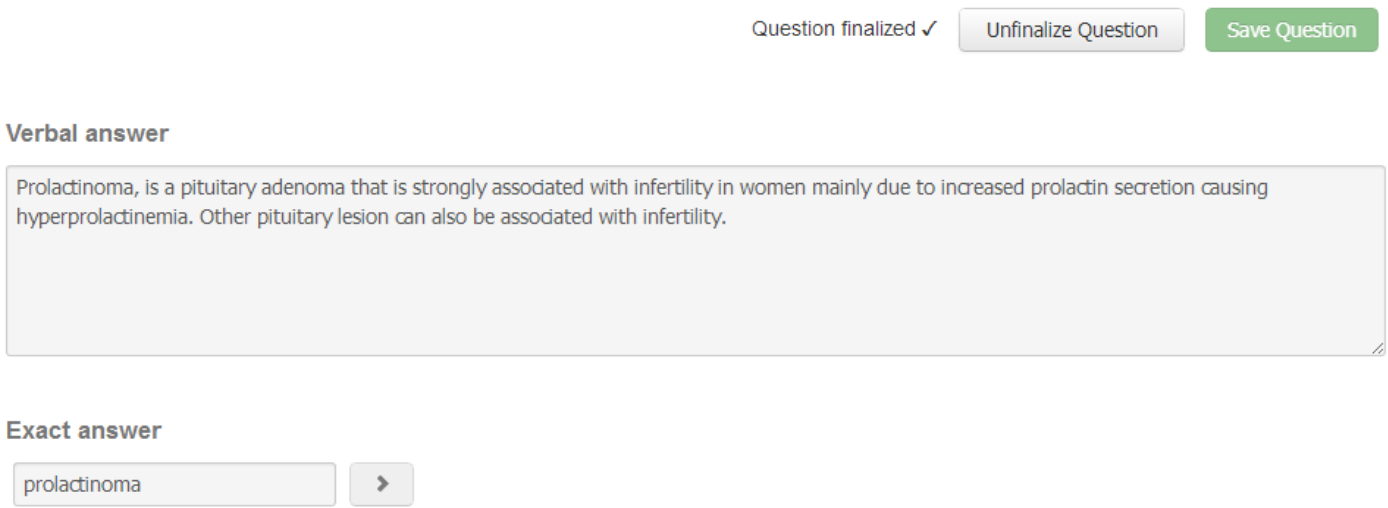

Figure 3.12: The answer formulation screen for a factoid question.

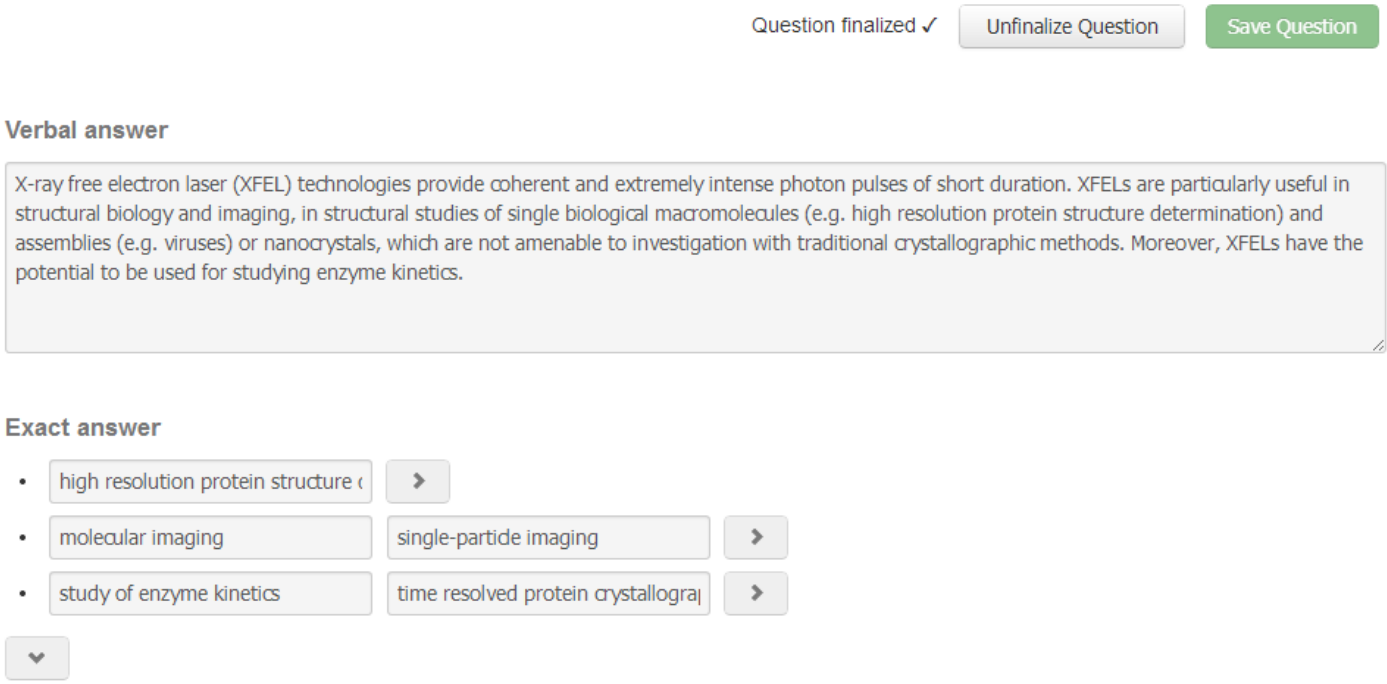

Figure 3.13: The answer formulation screen for a list question.

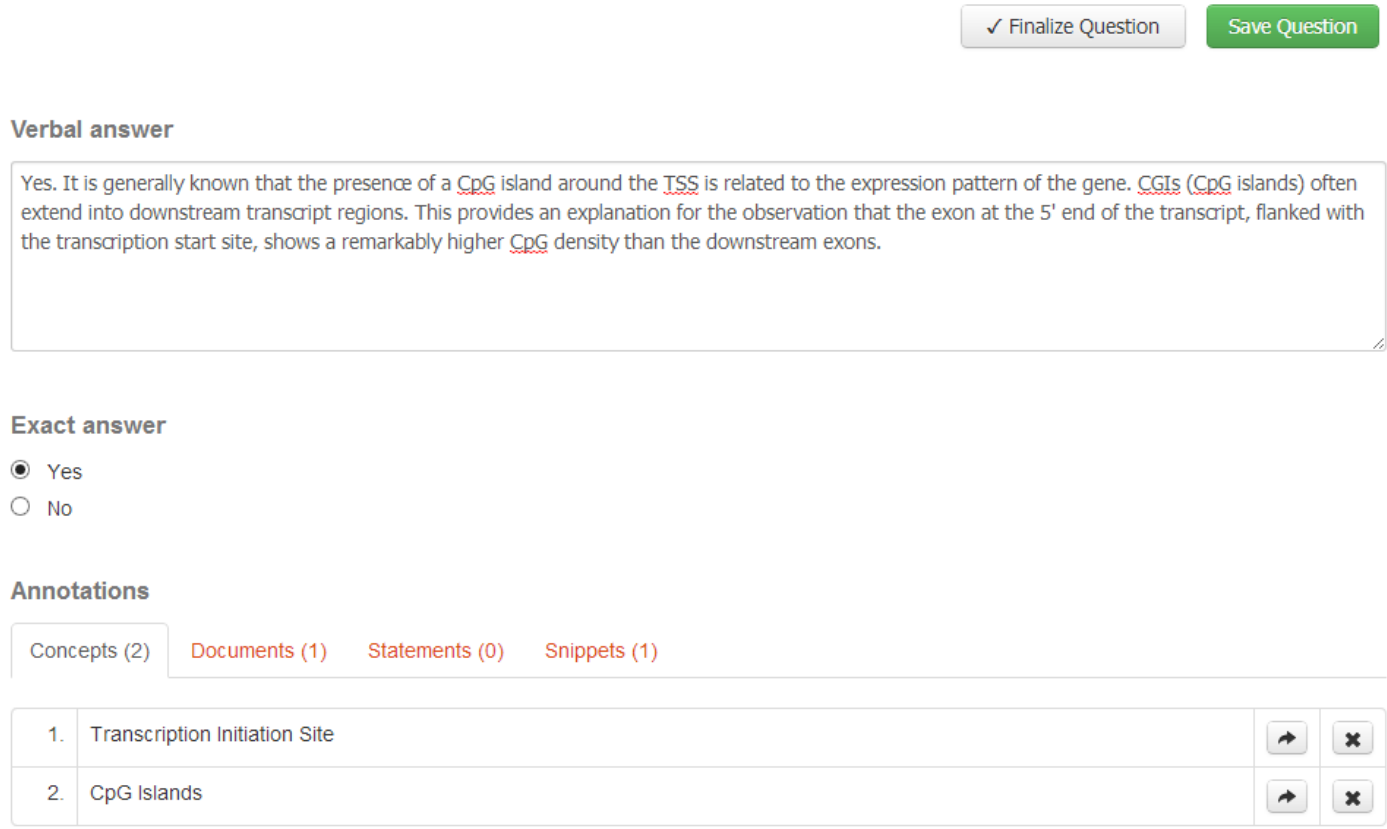

Figure 3.14: The answer formulation screen for a yes/no question.

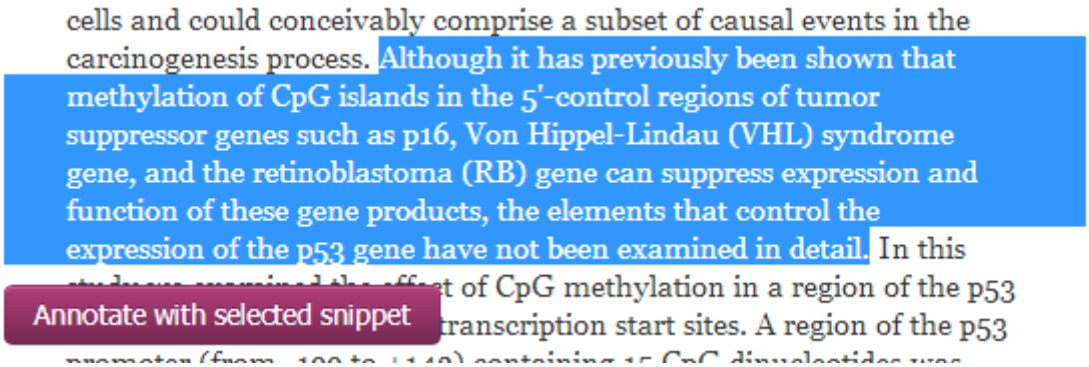

Figure 3.15: Extracting a snippet.

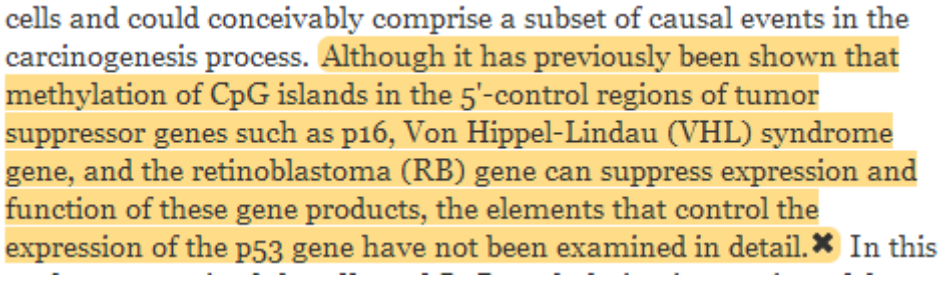

Figure 3.16: A selected snippet.

3.6 Query revision

If at this stage the expert feels that the selected statements and the extracted snippets do not provide enough information to answer the question, or if the expert believes that additional relevant concepts, articles, snippets, or statements could have been retrieved, he/she should modify the search query, as discussed in Step 6 of the guidelines of Chapter 2. Clicking on the "Search" tab of the upper navigation menu (Figure 3.10) allows a new query to be entered, as discussed in Section 3.3. When the new query is executed, three new lists with the items (concepts, documents, and statements) retrieved by the new query appear (Figure 3.5), and the expert can again select the items that are appropriate. The items that had been retrieved by previous queries (for the same question) and had been selected by the expert (before executing the new query) are retained. They are also shown on the lower part of the screen (Figure 3.14) that appears when the "Answer" tab of the upper navigation menu (Figure 3.10) is active.

Clicking on the "Save" button (Figure 3.14) saves the ideal and exact answer that have been entered. A message will appear confirming that the ideal and exact answers have been saved. To finalize a question, i.e., to signal that work on the particular question has been completed, click on the "Finalize Question" button (Figure 3.14). Note that if a question is finalized it can still be edited by clicking on the "Unfinalize Question" button (Figure 3.13).

3.7 Other useful functions of the annotation tool

To log out or to change password at any time, the person-like button of Figure 3.10 can be used. Clicking on that button leads to the form of Figure 3.17, where the expert can either log out or change his/her password.

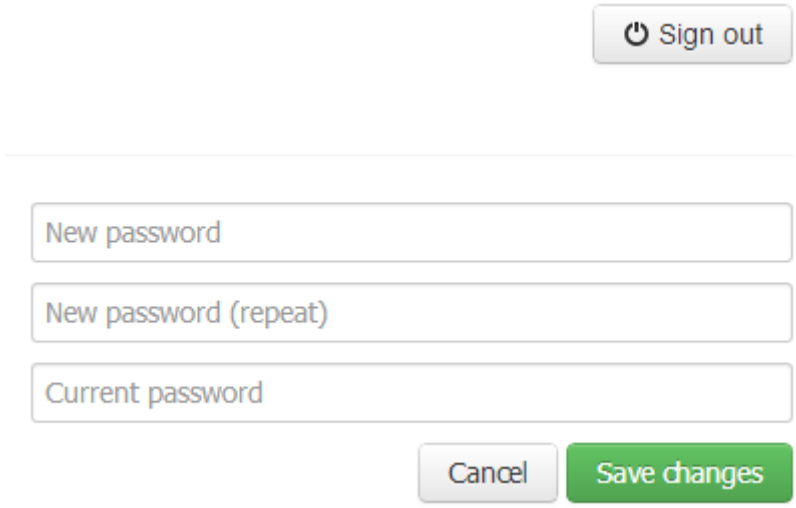

Figure 3.17: Logout or change password form.

Bibliography

N. Heino. Annotation Tool, Second Version. Technical Report D3.6, BioASQ Deliverable, 2013.

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