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

Project **FP7-318652 / BioASQ**

Deliverable **D5.4**

Distribution **Public**

http://www.bioasq.org

Challenge Evaluation Report 2 and Roadmap

Prodromos Malakasiotis, Ion Androutsopoulos, Agiatis Bernadou, Nephelie Chatzidiakou, Eliza Papaki, Panos Constantopoulos, Ioannis Pavlopoulos, Anastasia Krithara, Yannis Almyrantis, Dimitris Polychronopoulos, Aris Kosmopoulos, Georgios Balikas, Ioannis Partalas, George Tsatsaronis and Norman Heino

Status: Final (Version 1.0)

October 2014

Project

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

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D5.4
Deliverable title	Challenge Evaluation Report 2 and Roadmap
Contractual date of delivery	M24 (September 2014)
Actual date of delivery	October 2014
Relevant Task(s)	WP5/Task 5.2
Partner Responsible	AUEB-RC
Other contributors	NCSR "D", ULEI, TI, UJF, UPMC
Number of pages	80
Author(s)	Prodromos Malakasiotis, Ion Androutsopoulos, Agiatis Bernadou, Nephelie Chatzidiakou, Eliza Papaki, Panos Constantopoulos, Ioannis Pavlopoulos, Anastasia Krithara, Yannis Almyrantis, Dimitris Polychronopoulos, Aris Kosmopoulos, Georgios Balikas, Ioannis Partalas, George Tsatsaronis and Norman Heino
Internal Reviewers	
Status & version	Final
Keywords	Challenge evaluation, web site and participation statistics, questionnaires, roadmap, BIOASQ inheritance, benchmark datasets, authoring and assessment tools, guidelines, evaluation infrastructure, oracles, social network, future challenges, BIOASQ3, expert interviews, user needs, porting to other domains, law, economics, social sciences, cultural heritage, digital humanities

Executive Summary

This deliverable evaluates the second cycle of the BIOASQ challenge both quantitatively and qualitatively. The evaluation was conducted by measuring the number of participants in the challenge and the workshop, the number of visitors in our websites, and the downloads of the benchmark data. We also used questionnaires distributed to the participants of the challenge and to the team of biomedical experts. The analysis showed that we managed to successfully organize the second cycle of the challenge. The participation was higher than the first cycle, and most of the participants are not only willing to participate in the third cycle of BIOASQ, but they also intend to recommend it to other research groups. We also managed to organize a successful workshop leaving a good overall impression to the participants. Finally, we had a very good cooperation with the team of biomedical experts, providing them with all the help and tools they needed for the creation of the benchmark datasets for Task 1B and the evaluation of the systems' responses, again for Task 1B.

In addition this deliverable provides a roadmap of how relevant research can be pushed further, beyond the end of the project, through new benchmarks, challenge tasks, and by exploiting the BIOASQ evaluation infrastructure. Towards that direction the third cycle of BIOASQ will run with minimum cost by exploiting the infrastructure and tools created during the project. Note that these tools are publicly available and can be adapted to other challenges. In addition, research teams can use the datasets and the Oracles to compare with the participants of the two first cycles of BIOASQ.

Apart from keeping the BIOASQ challenge running in its current form, a medium-term goal could be to modify the challenge to better reflect user needs. Towards this direction, we interviewed the members of the BIOASQ biomedical experts group to better understand how they currently search (e.g., what information they search for, where they search, how they search, what problems they encounter). The results of this study can be summarized in the following table.

Recommendations for future challenges and/or systems
Specify types (e.g., research articles, systematic reviews, clinical trial records, patents) and origin (e.g., PUBMED, trusted sites, Web) of documents to be searched.
Use more designated repositories of structured information for concepts and triples, or require the participating systems to find relevant repositories of structured information per question.
Continue to aim at generic biomedical QA systems, rather than systems targeting particular types of information (e.g., gene interactions only).

Consider assigning questions to groups of experts, possibly with complementary expertise. Extend the BIOASQ social network to allow experts to criticize or complement answers produced by systems. Move towards hybrid QA systems, combining answers provided by systems and humans. Consider a question to question matching subtask (e.g., for FAQs).
Investigate if systems that accept natural language questions actually manage to produce better answers than systems that accept keyword queries. Use previous searches, articles downloaded or shared, journal subscriptions etc. to construct user models of the experts. Address full-content access restrictions of journals.
Support filtering or ranking criteria for author, reputation, affiliation, journal name, impact factor, citations, article type, recency etc. when displaying retrieved articles in the BIOASQ authoring tool and future QA systems.
Research how biomedical experts could better organize and store retrieved relevant information and sources. Develop tools (possibly based on the BIOASQ authoring and assessment tools) that would help biomedical experts organize and store relevant information and sources per natural language question. Consider retrieving relevant images, tables, equations etc.
Use English questions and keyword queries as separate or joint inputs. Consider follow-up questions, clarification dialogues, possibly also spoken dialogues. Consider merging factoid and list questions. Consider adding list questions requiring positive and negative lists, or lists of steps. Consider adding questions requiring 'insufficient information available' or 'controversial information found' as answers. Measure the time needed to formulate and process natural language questions vs. keyword queries (e.g., in emergencies).
Continue to require relevant documents, snippets, 'exact' and 'ideal' answers per question, but measure the value of 'exact' and 'ideal' answers as opposed to having only relevant documents and snippets in different settings (e.g., clinical vs. research purposes). Consider adding relevance feedback and clustering to the authoring tool for documents and snippets, and to future QA systems. Consider structured snippets. Clarify the purpose of concepts. Author questions for which there is relevant important information in repositories of structured information. Improve the BIOASQ services that retrieve possibly relevant 'statements'. Consider improving the fluency of 'statements'.
Link more tightly the 'exact' and 'ideal' answers to supporting articles, snippets, concepts, and statements. Require bibliographic entries for the supporting sources. Consider requiring more structured 'ideal' answers for particular types of questions.
Attract more participants. Provide to the participants more information about the expected answers (e.g., types of expected 'exact' answers, length of 'ideal' answer). Consider questions requiring predictions or inference.

Finally, a longer-term goal would be to port BIOASQ to other scientific domains such as Economics and Social Sciences, and the European Law, where widely used document repositories (with a role similar to PUBMED) and concept taxonomies (with a role similar to MESH headings) also exist. Another domain that could be considered is that of Digital Humanities. However, no document repositories of universal coverage and acceptance, such as PUBMED in biomedicine, have been established yet.

Contents

List of Figures

List of Tables

1

Introduction

This deliverable evaluates the second cycle of the BIOASQ challenge. For that purpose we provide figures measuring the number of the participants both in the challenge and the workshop, the visitors in our websites, and the downloads of the benchmark data. We also distributed questionnaires to the participants of the challenge, as well as to the team of biomedical experts. The questionnaires aimed to assess the quality, appropriateness, and diversity of the challenge benchmarks and evaluation measures, the quality of the support to the participants, and the adequacy and quality of the challenge evaluation infrastructure. The results of this analysis can be found in Chapter 2. In more detail Sections 2.1 and 2.2 below present the evaluation of the second cycle of the BIOASQ challenge based on web statistics and questionnaires respectively. The questionnaires can be found in Appendix A

In addition Chapter 3 provides a roadmap of how relevant research can be pushed further, beyond the end of the project, through new benchmarks, challenge tasks, and by exploiting the BIOASQ evaluation infrastructure. In more detail, Section 3.1 provides details of what we have achieved during the 2 years of BIOASQ, Section 3.2 gives ideas of how we can keep the challenge running, Section 3.3 studies how BIOASQ could be modified to reflect user needs, and Section 3.4 gives insight on how we can port BIOASQ to other domains.

2

Challenge evaluation

2.1 Evaluation via web–site statistics and workshop participation

One of the most important features for a successful challenge is attracting participants. Table 2.1 summarizes the most important statistics concerning the evaluation of the challenge. In more detail, we had 216 users registered on the evaluation platform.[1] Moreover, several teams around the globe participated in the second cycle of the BIOASQ challenge, especially in Task 2A, including key players, like NLM, Toyota Technological Institute, and University of San Diego. In particular we managed to attract more participants than the first cycle for both tasks as shown in Tables 2.2 and 2.3. We have also organized a successful Workshop as part of the Question Answering Track of CLEF 2014 in Sheffield, with approximately 30 participants.[2]

Another important objective of the challenge is to establish BIOASQ as a reference point for the biomedical community. A look at the statistics from our websites indicates that we are moving towards the right direction. Figures 2.1 and 2.2 show that even after the end of the challenge there was high traffic on our websites indicating the community's interest in BIOASQ. In addition, Task 2A continued to run in a "non-challenge" mode helping towards that direction. Table 2.4 shows the numbers of downloads for the datasets of Tasks 2A and 2B.

[1] http://bioasq.lip6.fr/

[2] Consult http://nlp.uned.es/clef-qa/ for information on CLEF 2014 Question Answering Track. For more details about the BIOASQ workshop, visit http://www.bioasq.org/workshop.

Registered users	216 (117)
Datasets downloads	576 (584)
Teams	18 (11)
Workshop participants	30 (30)
Number of biomedical experts	10 (10)

Table 2.1: Summary of the challenge statistics. In parentheses the corresponding statistics of the first cycle of BIOASQ

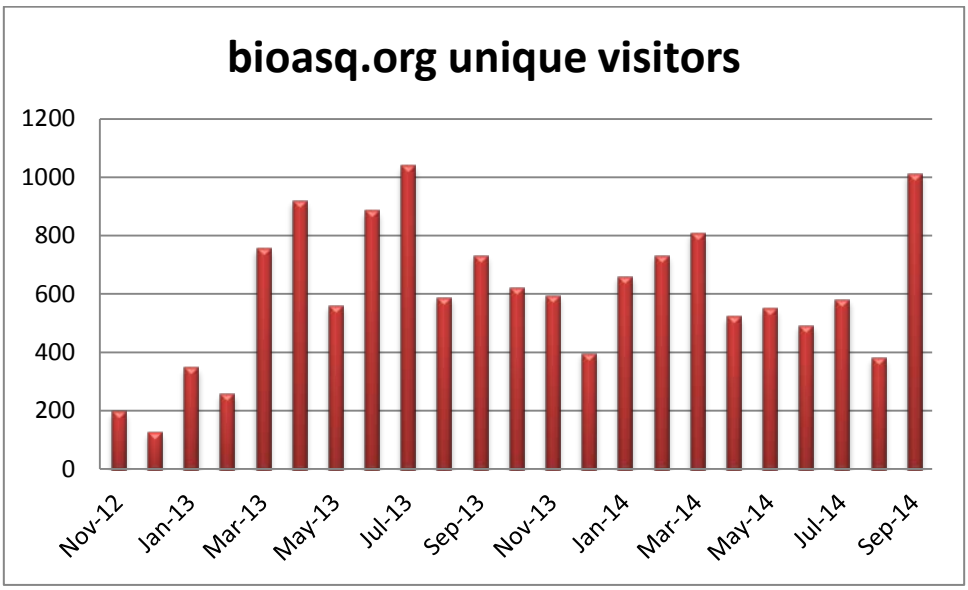


Figure 2.1: Unique visitors at BIOASQ official site from November '12 until September '14.

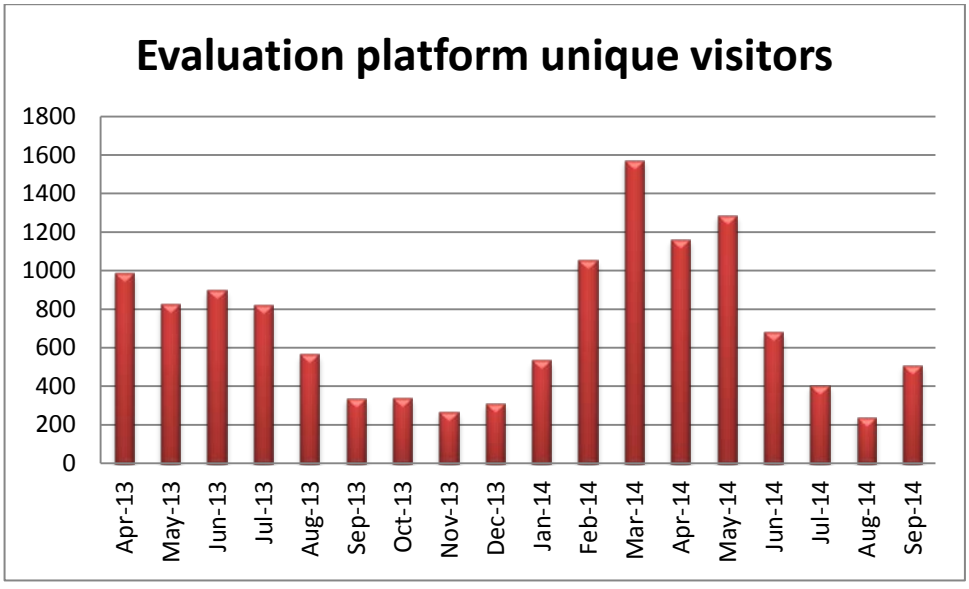


Figure 2.2: Unique visitors at the evaluation platform from April '13 until September '14.

Task 1A participation (46 systems, 11 teams)	
Mayo Clinic	North America (U.S.)
University of Alberta	North America (Canada)
Aristotle University of Thessaloniki + Atypon	Europe (Greece)
University of Vigo	Europe (Greece)
University of Colorado	North America (U.S.)
NCBI, NLM	North America (U.S.)
Universite de Rouen	Europe (France)
Fudan University	Asia (China)
UCSD	North America (U.S.)
Toyota Technological Institute	Asia (Japan)
Imran	Asia (Pakistan)
Task 1B participation	
Phase A (4 systems, 2 teams)	
Mayo Clinic	North America (U.S.)
University of Alberta	North America (Canada)
Phase B (7 systems, 2 teams)	
University of Alberta	North America (Canada)
Toyota Technological Institute	Asia (Japan)

Table 2.2: Teams participating in both tasks of the first cycle of BIOASQ challenge.

2.2 Evaluation via questionnaires

Having evaluated BIOASQ via web–site statistics, we move on to the results collected via questionnaires. We created and distributed questionnaires to the participating teams of the challenge, and the team of biomedical experts. The questionnaires aim, among other goals, to assess the quality, appropriateness, and diversity of the challenge benchmarks and evaluation measures, the quality of the support to the participants, and the tools and the support provided to the team of biomedical experts for the creation of the benchmarks. For a more detailed view of the questionnaires see Appendix A.

2.2.1 Challenge evaluation

The questionnaire distributed to the teams participating in the second cycle of BIOASQ included questions targeting several aspects of the individual tasks, like the quality of the datasets, the technical support etc. Additionally, more general questions were provided to capture the overall impression of the participants for BIOASQ. Figures 2.3–2.8 summarize the results for the most general questions for both cycles of BIOASQ. According to these figures the participants are in general satisfied with the challenge and as a consequence, not only are they willing to participate in the next cycle of the challenge, but they are willing to recommend BIOASQ to other research groups as well. Interestingly it was easier for the participants to understand the Tasks of the second cycle (Figures 2.9 and 2.10).

2.2.2 Interaction with the team of biomedical experts evaluation

One of the most difficult goals of BIOASQ was the creation of the benchmark data for Task 2B. For that purpose we had to coordinate a team of biomedical experts and provide them with tools and technical

Task 2A participation (61 systems, 18 teams)	
NCBI	North America (U.S.)
pierre curie	Europe (France)
Fudan University	Asia (China)
U.S. National Library of Medicine	North America (U.S.)
Aristotle University of Thessaloniki	Europe (Greece)
Universidade de Aveiro	Europe (Portugal)
Fudan University 1	Asia (China)
UET	Asia (Pakistan)
Fudan University 2	Asia (China)
Universidad Carlos III	Europe (Spain)
UC San Diego	North America (U.S.)
ERIAS-ISPED	Europe (France)
Seoul National University	Asia (South Korea)
Center For Spoken Language Understanding	North America (Canada)
University of Vigo	Europe (Spain)
University of St Thomas	North America (U.S.)
Holmes Semantic Solutions	Europe (France)
University of California, San Diego	North America (U.S.)
Task 2B participation	
Phase A (22 systems, 8 teams)	
University of Alberta	North America (U.S.)
Seoul National University	Asia (South Korea)
NCBI	North America (U.S.)
upmc	Europe
Hasso-Plattner Institut	Europe (Germany)
University of Massachusetts Medical School	North America (U.S.)
Fudan	Asia (China)
Toyota Technological Institute	Asia (Japan)
Phase B (18 systems, 6 teams)	
Aristotle University of Thessaloniki	Europe (Greece)
University of Alberta	North America (U.S.)
Seoul National University	Asia (South Korea)
NCBI	North America (U.S.)
upmc	Europe
Toyota Technological Institute	Asia (Japan)

Table 2.3: Teams participating in both tasks of the second cycle of BIOASQ challenge.

Dataset downloads	
Task 1A	576 (584)
Task 1B	255 (113)

Table 2.4: Tasks 2A and 2B dataset downloads. In parantheses the downloads during the first cycle.

Figure 2.3: Overall impression for the first cycle of BIOASQ by the participating teams.

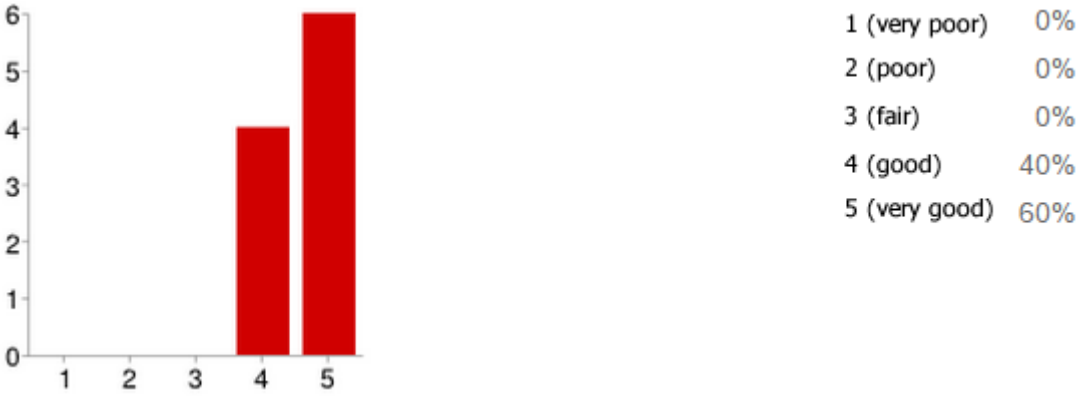


Figure 2.4: Overall impression for the second cycle BIOASQ by the participating teams.

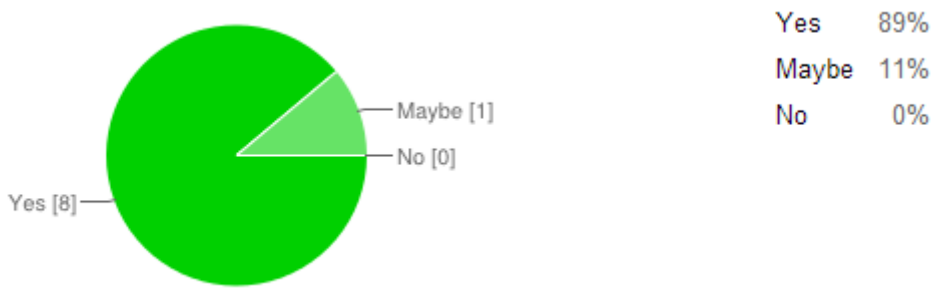


Figure 2.5: Willingness to participate in the second cycle of BIOASQ by the participating teams of the first cycle.

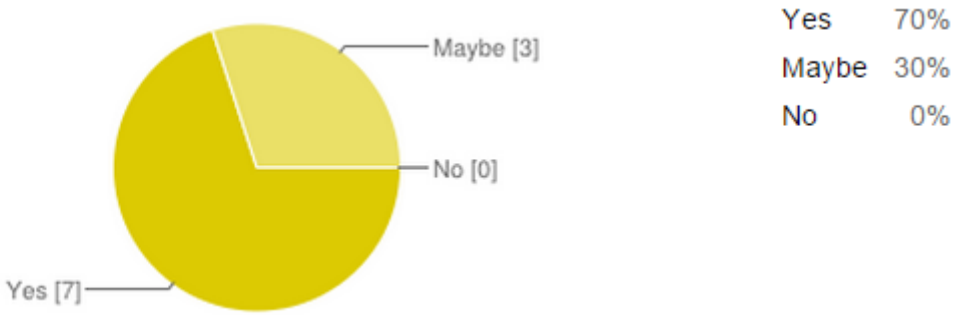


Figure 2.6: Willingness to participate in the third cycle of BIOASQ by the participating teams of the second cycle.

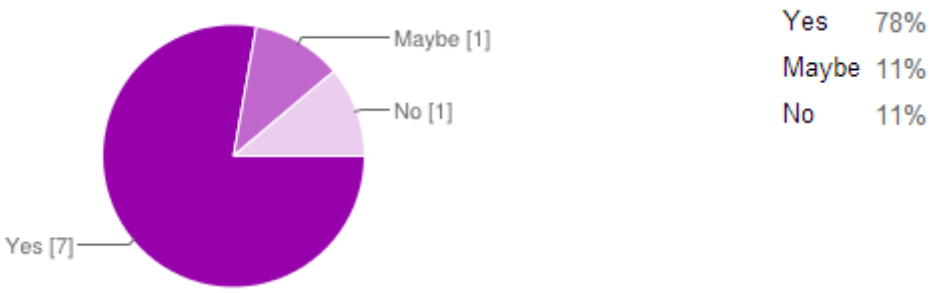


Figure 2.7: Willingness to recommend BIOASQ by the participating teams of the first cycle.

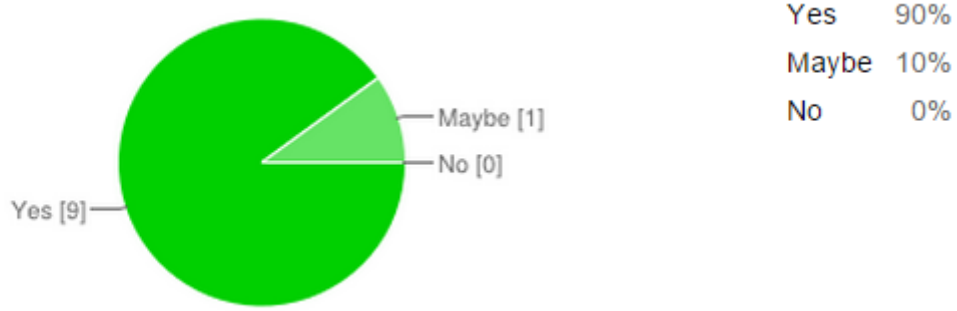


Figure 2.8: Willingness to recommend BIOASQ by the participating teams of the second cycle.

Figure 2.9: Difficulty to understand BIOASQ tasks by the participating teams of the first cycle.

Figure 2.10: Difficulty to understand BIOASQ tasks by the participating teams of the second cycle.

Figure 2.11: Overall impression for the first version of the annotation tool by the team of biomedical experts.

Figure 2.12: Overall impression for the second version of annotation tool by the team of biomedical experts.

support that would help them in the creation of the benchmarks. In addition, the team of biomedical experts manually assessed the responses of the participating teams in Task 2B, again with the assistance of a tool. After the end of the second cycle of the challenge, we distributed questionnaires to the biomedical experts, in order to assess the quality of the tools and their interaction with us. Figures 2.11 – 2.26 show that we had a very good cooperation with the team of biomedical experts. The experts were not only satisfied by the tools, but they are also willing to use them again in the future and even recommend them to others. Particularly for the annotation tool they are willing to use it for their own work, especially if it is improved. The most problematic issue concerning both tools was about the RDF triples exploitation (Figures 2.27 and 2.28). The main complaint was that pseudo-English renderings of RDF triples were difficult to make sense of. To address that we experimented with generating texts with NATURALOWL system (Androutsopoulos et al. (2013)) from Disease Ontology, one of the ontologies used in BIOASQ.

Figure 2.13: Willingness of the team of biomedical experts to use the first version of the annotation tool again.

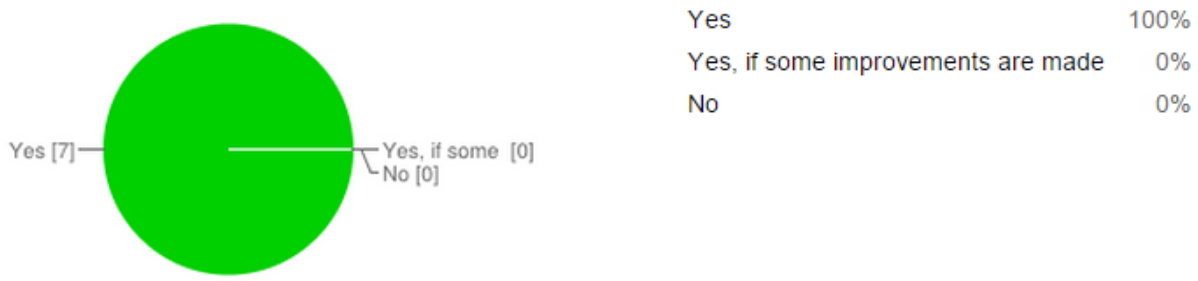


Figure 2.14: Willingness of the team of biomedical experts to use the second version of the annotation tool again.

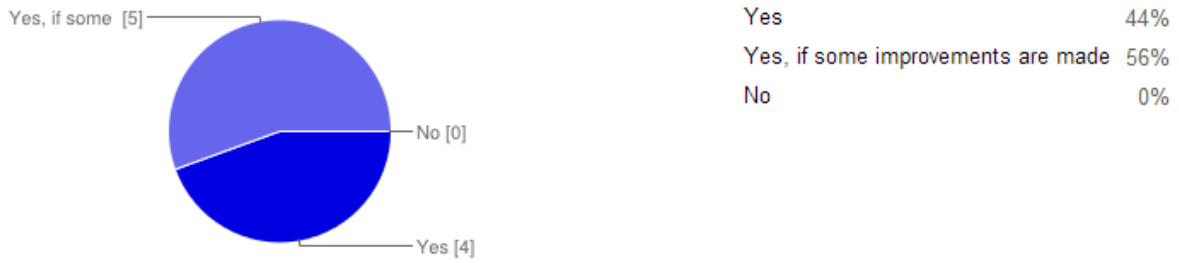


Figure 2.15: Willingness of the team of biomedical experts to use the first version of annotation tool for their work (e.g., to organize a search).

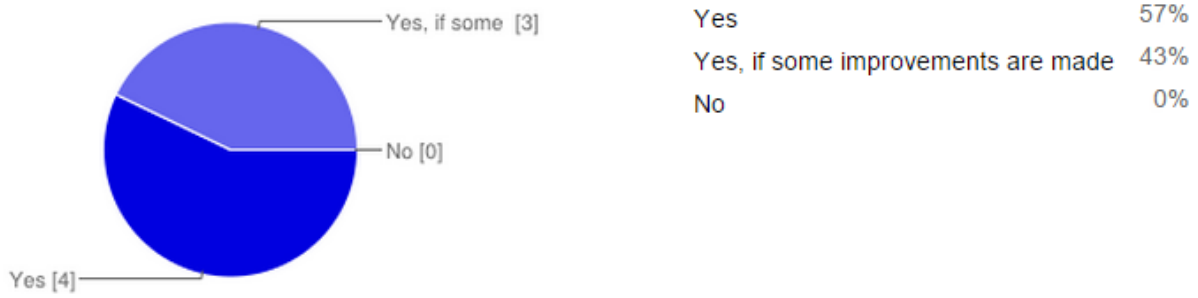


Figure 2.16: Willingness of the team of biomedical experts to use the second version of annotation tool for their work (e.g., to organize a search).

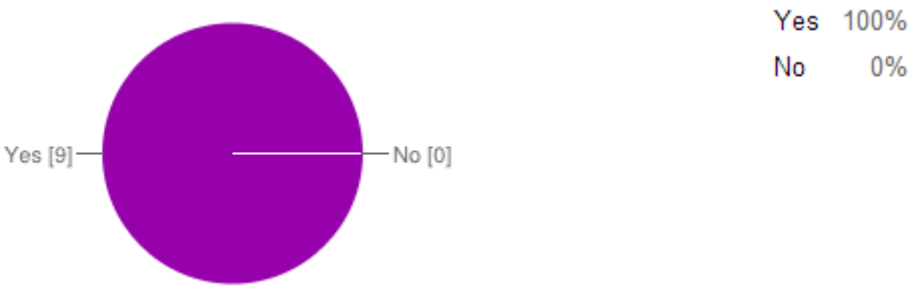


Figure 2.17: Willingness of the team of biomedical experts to recommend the first version of the annotation tool.

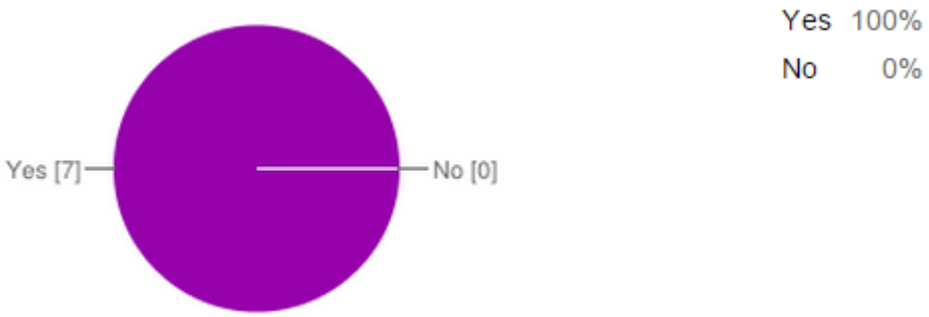


Figure 2.18: Willingness of the team of biomedical experts to recommend the second version of the annotation tool.

Figure 2.19: Overall impression for the first version of the assessment tool by the team of biomedical experts.

Figure 2.20: Overall impression for the second version of the assessment tool by the team of biomedical experts.

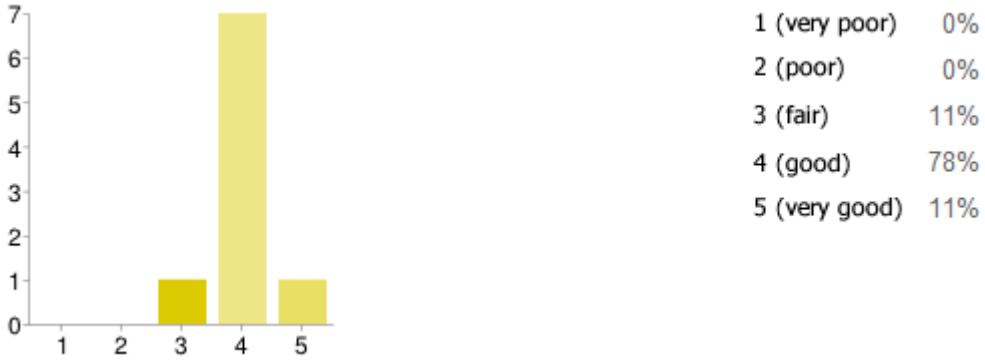
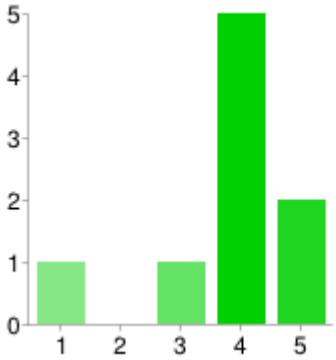


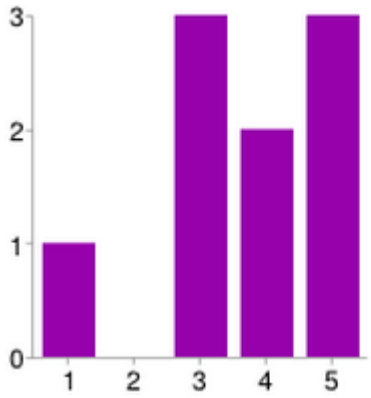
Figure 2.21: Willingness of the team of biomedical experts to use the first version of the assessment tool again.

Figure 2.22: Willingness of the team of biomedical experts to use the second version of the assessment tool again.



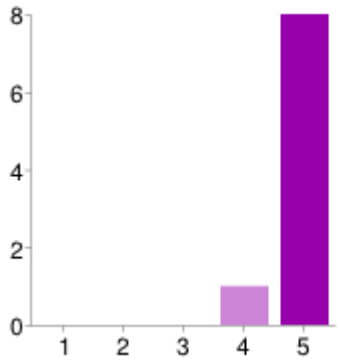
1 (very poor)	11%
2 (poor)	0%
3 (fair)	11%
4 (good)	56%
5 (very good)	22%

Figure 2.23: Willingness of the team of biomedical experts to recommend the first version of the assessment tool.



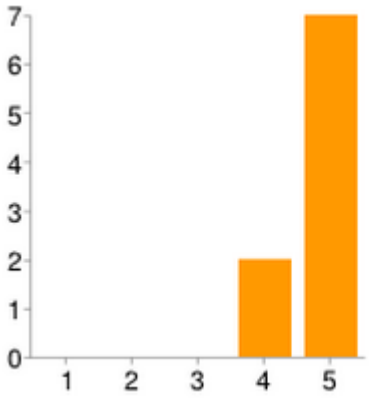
1 (very poor)	11%
2 (poor)	0%
3 (fair)	33%
4 (good)	22%
5 (very good)	33%

Figure 2.24: Willingness of the team of biomedical experts to recommend the second version of the assessment tool.



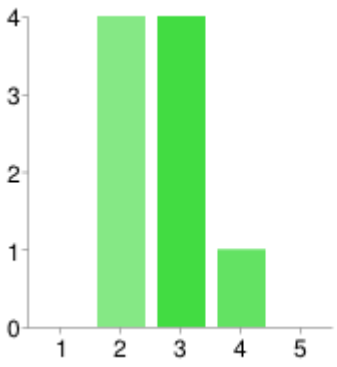
1 (very poor)	0	0%
2 (poor)	0	0%
1 (fair)	0	0%
1 (good)	1	11%
1 (very good)	8	89%

Figure 2.25: Assessment of the interaction of the team of biomedical experts with us during the first cycle.



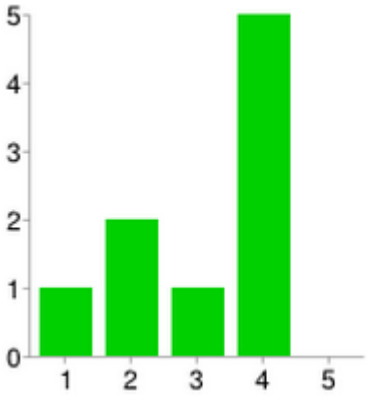
1 (very poor)	0%
2 (poor)	0%
3 (fair)	0%
4 (good)	22%
5 (very good)	78%

Figure 2.26: Assessment of the interaction of the team of biomedical experts with us during the second cycle.



1 (very poor)	0%
2 (poor)	44%
3 (fair)	44%
4 (good)	11%
5 (very good)	0%

Figure 2.27: Assessment of the RDF triples search procedure of the first version of the annotation tool.



1 (very poor)	**1**	11%
2 (poor)	**2**	22%
3 (fair)	**1**	11%
4 (good)	**5**	56%
5 (very good)	**0**	0%

Figure 2.28: Assessment of the RDF triples search procedure of the second version of the annotation tool.

Figure 2.29: Assessment of the RDF triples evaluation procedure of first version of the assessment tool.

Figure 2.30: Assessment of the RDF triples evaluation procedure of second version of the assessment tool.

Figure 2.31: Impact of improvements made for the 2nd version of the annotation tool.

Figure 2.32: Impact of improvements made for the 2nd version of the assessment tool.

3

Roadmap

Having completed successfully two cycles of BIOASQ it is highly important to push the relevant research even further. As a short-term goal towards that direction, a third BIOASQ challenge is planned for 2014–15. Task A already runs in a fully automatic manner from the organizers' point of view (batches of newly published articles continue to be released, system responses are collected and evaluated against the MESH headings provided by NLM curators). For Task B, new questions will have to be formulated by biomedical experts, but the required effort and cost will be much lower compared to previous years, since all the tools to be used by the experts and the evaluation infrastructure are now fully developed. A medium-term goal could be to modify the challenge to better reflect user needs. For that purpose, we interviewed biomedical experts to better understand how they currently search (e.g., what information they search for, where they search, how they search, what problems they encounter). Finally, a longer-term goal would be to port BIOASQ to other scientific domains such as Economics and Social Sciences, and the European Law, where widely used document repositories (with a role similar to PUBMED) and concept taxonomies (with a role similar to MESH headings) also exist.

3.1 The present: The BIOASQ inheritance

3.1.1 Datasets for Tasks A and B

In both editions of the BIOASQ challenge, i.e., 2013 and 2014, the challenge comprised two tasks:

- **Task a**: Large-scale on-line biomedical semantic indexing.
- **Task b**: Biomedical Semantic Question-Answering.

The data for Task a in both editions, i.e., Task 1a for the first and Task 2a for the second, consist of biomedical articles indexed in PUBMED, selected from a pre-defined pool of journals the papers of which experience fast annotation time. At the time of the distribution, the articles had not been assigned MESH terms by the professional indexers of *NLM*. Hence, the task for the participants was to annotate the articles using MESH headers, within 24 hours. The data were distributed in batches. The evaluation took place by comparing the assigned MESH terms given to these articles by the professional indexers at

a later time, with the system responses. The participants had the opportunity to train their systems prior to the beginning of the task. As training data, the participants were given older PUBMED indexed articles which were already assigned manually MESH terms by the professional indexers. In both training and testing, the articles were provided in their raw format (plain text) as well as in a pre-processed one (in a vectorized format), distributed as a Lucene index to the participants.

Task b of BIOASQ took place in two phases in each of the two editions, i.e., 1b for the first edition and 2b for the second edition. In the first phase (Task 1b or 2b Phase A), the participants were given a set of questions that were prepared by the BIOASQ medical experts team, and their systems had 24 hours to respond with related concepts, articles, snippets and triples. In the second phase (Task 1b or 2b Phase B) the participants were given the same questions as with Phase A, along with the golden concepts, articles, snippets and triples that the medical experts had provided. The participants had 24 hours to respond with exact answers or summaries. The data for both phases of Task b were provided in a raw text format, using JSON representation to format the fields. Instead of training data, the participants in 1b were given a "dry-run" data set to tune their systems. For 2b, the systems could have used all of the questions of 1b to train their systems on.

In the following, we provide the collected statistics from the data sets produced for both tasks, in each of the two editions. Table 3.1 summarizes the statistics collected for the training data of Task a. Table 3.2 summarizes the statistics collected for the test data of Task a. Finally, Table 3.3 provides the statistics collected for the training and the test data of Task b.

	Task 1a	**Task 2a**
Articles	10,876,004	12,628,968
Unique labels	26,563	26,831
Labels per article	12.55	12.72
Size in GB	18	20.31

Table 3.1: Statistics for the training data of Task a.

	Task 1a			Task 2a		
Week	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 1**	**Batch 2**	**Batch 3**
1	1,942 (1,553)	4,869 (3,414)	7,578 (2,616)	4,440 (3,319)	4,085 (3,422)	4,342 (3,009)
2	830 (726)	5,551 (3,802)	10,139 (3,918)	4,721 (3,734)	3,496 (2,788)	8,840 (5,883)
3	790 (761)	7,144 (3,983)	8,722 (2,969)	4,802 (3,884)	4,524 (3,274)	3,702 (2,860)
4	2,233 (586)	4,623 (2,360)	1,976 (1,318)	3,579 (2,431)	5,407 (3,923)	4,726 (3,252)
5	6,562 (5,165)	8,233 (3,310)	1,744 (1,209)	5,299 (3,693)	5,454 (3,666)	4,533 (3,252)
6	4,414 (3,530)	8,381 (3,156)	1,357 (696)	-	-	-
Total	16,763 (12,321)	38,801 (20,025)	31,570 (12,726)	22,841 (17,061)	22,966 (17,073)	26,143 (18,256)

Table 3.2: Statistics for the test data of Task a. In parentheses is the number of articles that at the time of the evaluation, had been annotated with MESH terms by the professional indexers. The total number of all articles distributed to the participants is $159,084$, out of which $97,462$ were annotated by the professional *NLM* indexers with MESH terms, by the time of the systems' evaluation.

3.1.2 Tools and guidelines to create new datasets for Task B

Datasets for Task B are created using the BIOASQ Annotation Tool (Heino (2013)) and, optionally, the BIOASQ Social Network (Heino and Ngonga Ngomo (2013)). Experts use the BIOASQ Anno-

	Task 1b				Task 2b					
	Dry Run	**Test sets**			**Training Data**	**Test sets**				
		1	**2**	**3**		**1**	**2**	**3**	**4**	**5**
Questions	29	100	100	82	310	100	100	100	100	100
Yes/No	8	25	26	26	85	32	28	36	32	24
Factoid	5	18	20	16	59	27	27	24	32	29
List	8	31	31	23	92	25	23	22	15	30
Summary	8	26	23	17	74	16	22	18	21	17
Avg #concepts	4.8	5.3	6	12.9	7.1	6.5	4.2	5.09	5.18	5.07
Avg #documents	10.3	11.4	12.1	5.4	14.2	11.4	14.8	8.66	12.25	11.07
Avg #snippets	14	17.1	17.4	15.9	18.7	17.1	14.7	10.8	14.58	13.18
Avg #triples	3.6	21.8	5.5	4.5	9.0	102.0	125.3	354.4	58.7	271.68

Table 3.3: Statistics for the training and test data for Task b. In total, 810 benchmark questions were prepared for Task b. The questions of Task 1b were given as training questions for Task 2b.

tation Tool to create and annotate questions that follow the guidelines provided in Malakasiotis et al. (2013a,b). Those questions can be published anonymously to a configured installation of the BIOASQ Social Network. The role of the BIOASQ Social Network in this process is twofold:

- It can be used to review published questions and give feedback to the experts. This allows them to iteratively improve their questions until a quality is reached that can be used in a benchmark set.
- It allows each registered user to vote for questions to appear in the next benchmark. After the voting, questions can be sorted by vote and the top k question's ID can be used to export said questions from the BIOASQ Annotation Tool.

The format for Task B challenge data sets closely resembles the format stored internally in the Annotation Tool. Therefore, only two simple steps are needed to create a new Task B dataset.

1. Exporting the Annotation Tool database in JSON format.
2. Post-processing the exported JSON file to create the desired format.

Both, BIOASQ Annotation Tool and BIOASQ Social Network have been developed as open-source software and are available from GitHub [1]. In addition, a collection of scripts has been released [2].

The script `dump_questions.js` from the scripts repository can be used to export the Annotation Tool database as a JSON file. One can pass a file with question IDs to identify the questions to be exported. After running a post-processing script on the exported JSON file, the data can be released as a challenge dataset.

Having the data export of the BIOASQ Annotation Tool a pre-processing script and a data-summary script are applied.

- The preprocessing script (`createDataFiles.py`) takes as input the BIOASQ Annotation Tool version of the data and creates the files containing the appropriate data for each phase of the task as well as the file with the golden data.
- The data summary script (`createStats.py`) takes as input the file with the golden data produced by the `createDataFiles.py` and provides statistics about the composition with respect to the 4 kinds of questions of the BIOASQ challenge and the annotations of the questions.

[1] https://github.com/BioASQ/AnnotationTool, https://github.com/BioASQ/SocialNetwork
[2] https://github.com/BioASQ/Scripts

Again, both scripts along with example data files of inputs and outputs are commented and available at the BIOASQ Github account.

3.1.3 Evaluation infrastructure and Oracles

During the first two years of the BIOASQ challenge the BIOASQ Participating Area (hereafter platform), which is available at http://bioasq.lip6.fr has been developed. Goal of the platform is to provide the necessary functionality to the participants to enter the series of the BIOASQ challenges. The functionality that the platform offers can be split in the following groups: (i) guidelines and tools, (ii) data exchange, (iii) user support, (iv) evaluation infrastructure (v) oracles. In the following paragraphs, we provide more information for the above-mentioned groups.

Guidelines and tools. Using the platform, participants can find information for the BIOASQ challenges and gain access to tools developed by the BIOASQ consortium. Detailed guidelines describing each of the two tasks along with the resources and the schedule of each task are available. We used a user-friendly template that allows users to find information fast. In addition, several supporting tools (e.g. word2vec code snippets and vectors) are available. To this direction, participants can also find code snippets to test and use the platform functionality efficiently (e.g. for exchanging using APIs).

Data exchange. Participants of the BIOASQ challenge should exchange data with the platform frequently. The training and test datasets for both tasks are available in the platform. When exchanging data, users can do it either manually i.e by following links or programmatically i.e by using web-services that are platform and language independent.

User support. A lot of effort has been spent during the challenge in order to provide adequate support to the challenge participants. To this direction, we have integrated a forum, a contact form and a frequently-asked questions (FAQ) page in the platform. Goal of the forum is to enable discussions between participants and the organisers of the challenge. The participation in the forum has increased during the second year of the challenge, reflecting the increase of interest towards the BIOASQ challenge. A contact form is also provided, for users who wish to contact the BIOASQ team in a more personal way. Finally, we keep and update a FAQ page with the most common questions of the participants to help them when looking for information.

Evaluation infrastructure. The evaluation of the submissions of participants in the tasks of the challenges is performed using automated evaluation measures. The scores are calculated periodically and tables where participants can browse over their performance are updated in the platform. The BIOASQ team has selected the official measures that decide the winners of the challenge and also provides several measures for reasons of reference and consistency with the existing literature. The calculation of the measures is performed by integrated scripts in the platform, which will become publicly available by the end of the challenge.

Oracles. In order to take full advantage of the infrastructure that has been developed during the challenge, the BIOASQ team has integrated oracles in the platform. Goal of the oracles is to give the participants the opportunity to test their systems in the following way: they can submit results for past test sets of the challenge and receive as immediate feedback of their performance the scores of the BIOASQ evaluation measures along with rankings among systems that have provided results for those test sets. For more information on the platform functionality and design details, please consult Balikas et al. (2013b,a, 2014).

Especially for the 'ideal' answers we conducted a correlation study in order to find which evaluation measures to use in the corresponding oracle. To measure the correlation we used Pearson's correlation, Spearman's Correlation and Kendal's τ. A very interesting result of this study is that the scores assigned

during the manual evaluation, namely HS_{pre}, HS_{rec}, HS_{rep}, and HS_{read},[3] seem to correlate well with each other. As a consequence we computed the average of the human scores, HS_{avg}, and measured the correlation of HS_{avg} with $BLEU$ and $ROUGE$. Table 3.4 summarizes the correlation scores. Both $BLEU$ and $ROUGE$ correlate well with HS_{avg} and as a consequence we will use them in the oracle concerning 'ideal' answers.

	Pearson's correlation	**Spearman's correlation**	**Kendal's τ**
HS_{pre} vs. HS_{rec}	0.79	0.78	0.74
HS_{pre} vs. HS_{rep}	0.80	0.79	0.76
HS_{pre} vs. HS_{read}	0.82	0.81	0.79
HS_{rec} vs. HS_{rep}	0.74	0.72	0.69
HS_{rec} vs. HS_{read}	0.77	0.74	0.71
HS_{rep} vs. HS_{read}	0.84	0.82	0.80
HS_{avg} vs. $ROUGE$	0.80	0.88	0.84
HS_{avg} vs. $BLEU$	0.85	0.91	0.87

Table 3.4: Correlation statistics of human scores.

3.1.4 Social network, communication channels and community building

As described in subsection 3.1.2, questions created using the BIOASQ Annotation Tool can be published to the BIOASQ Social Network. The BIOASQ Social Network enables a collaborative process in curating a set of benchmark questions. In order to support such a process, a vibrant community is needed. The BIOASQ Social Network supports organic community growth with the following features:

- It allows several ways for people to participate. If people do not want to express their opinion in written form they have the option to vote in favor of or against a particular question.
- People can follow and get updates from both other people and questions.
- Senior users are allowed to invite peers to the BIOASQ Social Network.
- Active users of the BIOASQ Social Network are rewarded. Rewards can be seen by everyone on the People page.

Additionally, BIOASQ enjoys the support of an **advisory board,** consisting of 33 international experts from the areas of bio-informatics, computational biology and medical informatics.[4] Different members of the board help in the successful organisation of the BIOASQ challenges in various ways, depending on their specialization. They all help also to increase the visibility of BIOASQ and to support the BIOASQ team in critical decision making.

BIOASQ has also established and organised a team of biomedical experts. The BIOASQ biomedical expert team has been formed during the first two months of the project. Several experts had been invited from a variety of institutions across Europe. The main criteria for inviting experts was: (a) seniority of the candidates, (b) complementarity of their expertise (various fields of biology, various medical specialties, bioinformaticians, etc.), (c) diversity of their occupations (scientists working for commercial organizations, university personnel, medical practitioners, etc). A team of 10 experts was

[3] For more information concerning the manual evaluation scores consult Balikas et al. (2013b,a).

[4] see http://www.bioasq.org/project/advisory-board.

finally formed. The principal task of the team was the composition of the Question/Answer benchmark datasets which were used during the two BIOASQ challenges. In addition, the members of the team have also participated in the manual evaluation of the competitors' answers, the overall challenge evaluation and the composition of this roadmap.

One particular collaboration that turned out to be very beneficial for both parties was that of BIOASQ with the US National Library of Medicine (NLM). NLM has created the Medical Text Indexer (MTI) to help MEDLINE curators in associating Mesh Terms with MEDLINE abstracts. MTI has been used in BIOASQ as a baseline system, that the challenge participants tried to outperform. We observed a significant improvement of the baseline system in the challenge, while NLM benefited from the ideas used in the participating systems to improve MTI (see the corresponding announcement of NLM: http://www.nlm.nih.gov/news/indexer_challenge.html).

3.2 Short-term future: Keeping the BIOASQ challenge running

3.2.1 Keeping the platform and oracles running

Towards meeting objective 4 of the BIOASQ project, which concerns the establishment of a framework which can be re-used for further competitions, the implemented platform along with the complete datasets of both tasks of the challenge will be available after the end of the project. Our goal is to keep the BIOASQ infrastructure functional and available online in order to allow the research teams around the world to evaluate their systems using the datasets of the BIOASQ challenge.

More specifically, the evaluation platform will keep running in order to provide all the functionalities like user registration, downloading of data, use of the Web services and uploading of results (in the oracles) by the interested users. The guidelines of the tasks will be refined using the experience of the second year of the challenge and will remain available as a reference point for the scientific community.

Additionally, the platform will migrate to a server in the premises of NCSR "Demokritos" in order to ensure the unobstructed operation of the platform as well as to facilitate its maintenance. Apart from the BIOASQ datasets (Partalas et al. (2014a,b)), which will be available the most useful functionality will be provided through the oracles.[5] Participants, using the oracles will be able to develop and evaluate their systems with respect to both tasks of the BIOASQ challenge. Until the end of the project, the complete dataset of the challenge will be integrated and will be available for submissions through the oracles.

Finally, in order to maximize the benefits from the BIOASQ challenge the code of the platform and the evaluation scripts will be provided as open source software under the GPL[6] licence. Accompanying documents describing the rationale of the implementation and instructions on how to download and install it in a local machine will also be released. Furthermore, videos describing the installation process and the use of the oracles will be provided to support anybody who will use the software after the end of the project. Those actions, along with the project deliverables will make the use of the platform and oracles sustainable and will keep them running for a long time.

3.2.2 Adding new datasets

A crucial part in the sustainability and longevity of the BIOASQ initiative is the process of adding new benchmark datasets for the future challenges at a low cost. As the challenge comprises two tasks, in the

[5] http://bioasq.lip6.fr/oracle/
[6] http://www.gnu.org/copyleft/gpl.html

following we will discuss the process of generating benchmark data sets for both of them in the future, and analyze the options and the respective costs.

Regarding Task a, already from the first edition of the challenge, the process of generating both training and test sets has been automated. More precisely, *TI* has set up services that can be called remotely, and that can automatically prepare test sets for the batches of Task a. The process can be further automated by a script that can call the services weekly, or at any given frequency. In addition, the process can be automated so that the produced data sets appear in the participants' platform automatically for distribution. The evaluation of the system submissions is also automated. As a result, the cost of maintaining the production of benchmark data sets for Task a is minimal, and is reduced to the maintenance of the servers that are running the respective services, including security and software updates, provision of fast internet connection, and utility costs.

With regards to Task b, the process of creating benchmark sets is conducted via the annotation tool. The tool uses services from *TI* to search for documents, triples, and concepts in the underlying resources. In addition to the annotation tool, the BIOASQ Social Network has been developed to aid the sustainability of the benchmark set creation for Task b. The main idea behind this tool is to serve as a platform for the experts to communicate on, publish results about and access the BIOASQ data. Beyond the end of the project, it will facilitate the compilation of novel versions of the BIOASQ benchmarks, as well as new data sets. As a summary, the social network can aid the experts towards:

- disseminating the benchmark questions,
- broadcasting the latest results around the BIOASQ core topics,
- enabling the experts to get familiarized with the BIOASQ tools and the underlying resources, and,
- tracking changes to the data sets, facilitating the curation of newly created questions and compiling novel versions of the benchmark sets.

Especially for Task b, in addition to the nature of costs that is also associated with Task a (e.g., maintenance of servers and services), there lies the additional cost of the compensation of the experts to create new questions. Since the beginning of the BIOASQ project, this cost has dropped dramatically due to the development of tools, such as the annotation tool, the social network and the associated services that aid the experts to accelerate their work. In this direction, the tutorials and guidelines that have been prepared for the experts to guide them through the process of creating questions, have provided a good basis for new experts to learn fast the task of benchmark questions creation for BIOASQ task b.

Overall, regarding figures pertaining to the sustainability costs of the two BIOASQ challenges, we summarize in the following an analysis conducted by *TI*, for a period of 8 months that covers the time span of a BIOASQ challenge edition. According to the following analysis, the cost of a BIOASQ edition at this stage is estimated at $24,400$ Euros. In more detail:

- maintenance of servers and services for both Task a and b: $8,400$ Euros; more analytically:
  - 4 servers with a total of 150 Euros needed per month for power supply, and internet connection,
  - administrative costs for security updates, and operating system updates amounting to 900 Euros for the whole period of the 8 months (i.e., 25% of a person month), and,
  - costs for data and services maintenance, i.e., upgrade to new resources'/ontologies' versions, and update *PubMed* and *PMC* indexes, estimated at $3,600$ Euros for the whole period of the 8 months (i.e., a person month).

- experts compensation for the creation of 800 benchmark questions, estimated at 16,000 Euros, i.e., 20 Euros per question.

3.3 Medium-term future: modifying the challenges to better match user needs – interviews with biomedical experts

Apart from keeping the BIOASQ challenges running in their current form, a longer term goal could be to modify the challenges to better reflect user needs. Towards this direction, we interviewed the members of the BIOASQ biomedical experts group to better understand how they currently search (e.g., what they search for, where, how, what problems they encounter). The interviewees were the same experts who had authored the questions and gold answers of the BIOASQ benchmarks for Tasks 1b and 2b (Malakasiotis et al. (2013a,b)). They had also participated in the assessment of the responses of the participating systems (Balikas et al. (2013b)). Hence, during the interviews we were also able to discuss the extent to which the experts thought that the BIOASQ challenges matched their needs (e.g., how useful it would be to be able to formulate English questions, as opposed to searching by keywords), which types of required answers (documents, snippets, triples, concepts, 'exact' and 'ideal' answers) are most useful in practice and, more generally, how the challenges could better match their needs; this part of the interviews focused on Tasks 1b and 2b, which were the tasks the experts were familiar with.

More precisely, we interviewed the twelve members of the BIOASQ biomedical experts group, plus two of their assistants who had participated in the preparation of the benchmarks and the assessment of the participating systems, 14 experts in total. Three experienced interviewers were employed, who had carried out similar studies with experts from other scientific areas (Benardou et al. (2010, 2013)).[7] Each interview involved a single biomedical expert (in two interviews, also an assistant of the expert) and one (in most cases) or two (in two initial interviews) interviewers. Each interview lasted approximately one hour. Teleconferencing (via Skype) was used in four of the interviews; for the other interviews, physical meetings took place. All the interviews were recorded and transcribed. The transcripts were checked by the experts, who also provided additional clarifications (e.g., definitions of technical terms) when the interviewers felt they were necessary.

The interviews were semi-structured. Each interview comprised an introductory section, where the experts were asked (Question Q0) to provide information about their affiliations, research areas, seniority, IT familiarity etc. Subsequently, eight questions (Q1–Q8) guided the discussion on how the experts currently search; the last two of these questions (Q7, Q8) were used only if time permitted. Four more questions (Q9–Q12) were then used to guide the discussion on how well the BIOASQ challenges matched the search needs of the experts, which types of required answers are most useful in practice, and how the challenges could better match the needs of the experts. At any point, the interviewers were allowed to ask further questions and, more generally, to diverge from the predefined list of questions (Q0–Q12), when they felt that it was worth doing so (e.g., when clarifications were needed, or when an interesting novel point was being made), but there were only rather minor diversions in practice.

Section 3.3.1 below summarizes our findings from the introductory section (Q0). Sections 3.3.2 and 3.3.3 list questions Q1–Q8 (how the experts search) and Q9–Q12 (matching BIOASQ and future challenges to user needs), respectively. For each question, we provide a summary of the most interesting points made by the experts, along with representative and/or interesting excerpts from the transcripts; some points and excerpts have been moved to different, more directly relevant questions than the ones

[7] The interviewers were Agiatis Benardou, Nephelie Chatzidiakou, and Eliza Papaki. Related work has been conducted in the projects European Holocaust Research Infrastructure (EHRI, Deliverable 16.4 – Researcher Practices and User Requirements), Preparing DARIAH (Deliverable 8.1.1), and Europeana Cloud (Task 1.3.5 – Exploring Innovative Tools in Research).

they actually originated from. For each question, we also provide recommendations for future biomedical QA systems and challenges, based on the responses of the experts. Section 3.3.4 summarizes our main recommendations.

3.3.1 Introductory section of the interviews

Q0. Name, gender, affiliation, country of residence, field of expertise, career status, years in research or profession, age, IT familiarity

Hereafter, the 14 experts are identified as "Expert 1",...,"Expert 14". The following table summarizes their genders, affiliations, countries of residence, career status, years in research or profession, and ages.

Expert ID	Gender	Affiliation	Country	Status	Years	Age
Expert 1	Male	Institute of Biology, NCSR "Demokritos"	Greece	Researcher	15	56
Expert 2	Male	Department of Biology, University of Athens	Greece	PhD Student	–	30
Expert 3	Female	Faculty of Medicine, University of Athens	Greece	Assoc. Prof.	7	–
Expert 4	Male	Institute of Clinical Physiology, NRC, Pisa	Italy	Medical Doctor, Researcher	23	48
Expert 5	Male	University of Crete	Greece	Assist. Prof.	–	36
Expert 6	Female	Biomedical Research Foundation, Academy of Athens	Greece	Researcher	19	41
Expert 7	Male	Biomedical Research Foundation, Academy of Athens	Greece	PhD student	1	27
Expert 8	Female	Cambridge University Hospital	Bulgaria, UK	Medical Doctor, Researcher	25	51
Expert 9	Female	Biomedical Sciences Research Centre Fleming	Greece	Researcher	25	47
Expert 10	Male	Institute of Biology, NCSR "Demokritos"	Greece	Researcher	20	56
Expert 11	Male	–	Greece	Med. Doctor	–	38
Expert 12	Male	Department of Genetics, Faculty of Biology, University of Barcelona	Spain	Assoc. Prof.	30	54
Expert 13	Male	Center for Genomic Regulation in Barcelona	Spain	Professor	32	55
Expert 14	Male	–	Lithuania	Researcher, Med. Doctor	–	30

The following excerpts provide an overview of the fields of expertise of the experts. Almost all of them are involved in biomedical research; Expert 8, who searches for new technologies that can be applied to hospitals, can be viewed as the only exception. Some of the experts (Experts 4, 11, 14) also work as clinical doctors (Experts 4 and 11 as cardiologists, Expert 14 as a neurologist). Six of the experts

are bioinformaticians (Experts 2, 5, 6, 7, 10 and 13), and four (Experts 1, 3, 9, 12) work in the broader area of molecular and cell biology.

Fields of expertise
"We specialize in two areas, firstly on molecular carcinogenesis and secondly we conduct research on molecular genetics for human diseases." [**Expert 1**]
"My subject was the study of DNA replication. I worked in the UK for two years and then I turned from experimental biology to computational biology and more precisely, computational sequence analysis." [**Expert 2**]
"My background is in Molecular Biology and I have specialized in genomics and pharmacogenomics. More specifically, we study cardiovascular diseases: their mechanisms of pathogenesis (i.e. mutations that lead to the disease) and consequently, the molecular players that could serve as promising therapeutic targets, and novel compounds that could serve as novel treatments against these diseases." [**Expert 3**]
"I am a cardiologist, a medical doctor and a researcher at ... and my research field is in non-invasive and non-ionizing cardiac imaging and in particular cardiac MRI. I study also the relationship between thyroid and heart and in the last year I research also in exercise physiology. More recently I started a new field about lifestyle." [**Expert 4**]
"My main research projects are related to the study of genomic DNA sequences. In particular I am mostly interested in understanding the way the DNA sequence may dictate its structure...We want to find if we can understand how the primary sequence may guide the structure, the way DNA folds in three dimensions." [**Expert 5**]
"Recently there has been a re-organization in the Systems Biology Department and our new name is 'Computational Biology and Medicine group'. [And your field of research is...] Biomedical, questions of biomedical interest." [**Experts 6 and 7**]
"By profession I'm a medical doctor and I'm specialized in General Medicine and Geriatrics Medicine...What I'm now doing is I'm looking at a lot of technology solutions that can be applied in hospitals. So we're not looking at something which is high level research, laboratory, blue sky which doesn't get into the hospital but we are trying to identify what technologies, ICT or any sensors, any monitoring devices, what can be used for our patients so that we can improve the care of the patient." [**Expert 8**]
"The broader field of research here at the Centre is biomedical sciences, and personally I am working on biochemistry and protein analysis. Generally I work with macromolecules." [**Expert 9**]
"My first degree is on chemistry, my Masters is on chemistry too and my PhD in system's self organization, which is theoretical physics and chemistry and theoretical biology. My research experience is on theoretical biology, evolutional biology, modeling in early development and later on genomics, what we call computational genomics." [**Expert 10**]
"I am a cardiologist. I worked as a researcher for many years; I conducted my PhD thesis in pharmacology, but the research object was related to cardiology...We mainly work on how the thyroidal hormone works, and how it could fix problems of the myocardium. For instance, how the thyroidal hormone could intervene in case of an ischemic episode or after a myocardial infarction, or after a heart failure." [**Expert 11**]
"I'm a developmental biologist, so I work with drosophila to understand how morphogenesis takes place in development. So mostly I am a geneticist and a developmental biologist." [**Expert 12**]

"I work with computational genomics, mostly in developing and applying programs to the analysis of genomic sequences and in particular to the signals and the processes in the genomic sequences that regulate transcription and primary and secondary production of RNA." [**Expert 13**]

"I am neurosurgery resident and researcher, I'm doing clinical research as well...I'm doing research in neuro oncology and also on behavioral medicine and endocrinology." [**Expert 14**]

All the experts have at least basic IT familiarity (e.g., they use web search engines and e-mail), many of them use specialized software (e.g., for statistical analysis, pattern matching, image processing), and some use or have used programming languages.

IT familiarity

"I conduct all analyses using software; I am also working on bioinformatics. But in any case, I am a user. On the other hand I use computers in a daily basis, so I would say that I am rather familiar with technology." [**Expert 9**]

"I use computational methods applicable to my research. I use some algorithms, both published and some custom made, which are helping me study the biology of the matter...I use pattern matching methods and these kinds of things in order to detect sequence motifs, estimate proximity to genes and to known (annotated) genomic regions, etc. For these purposes, I use mostly programming languages like Perl and awk as well as shell scripting." [**Expert 2**]

"I am familiar with Bioinformatical analysis of data that emerge from studies in genomics and pharmacogenomics that is studies that are concerned with the whole genome expression...There are tools either commercially available or freely accessible on the Internet. For example Partek, Ingenuity, microRNA analysis tools, Illumina Genetic Variant studio, GeneSpring and more." [**Expert 3**]

"Regarding Internet, I use Internet mainly for searching papers in libraries, such as PubMed. Then I use technologies to analyze images, cardio images." [**Expert 4**]

"Me, because I am a bit older, my IT knowledge is a little out of date. For programming, I am asking ... or someone else. I understand the big picture, but I am not that much "hands on" anymore. I am a Biologist and afterwards I studied bioinformatics, and software development; my post-doc was completely on software development." [**Experts 6 and 7**]

"We use computers a lot. Computers are an integral part of our work. First of all we processors are used in some machines, like the machines measuring things in the blood or the machines that connect with cameras in order to analyze images. We also use machines digitalizing what we measure. So, we use computers regularly. Furthermore, we enter our results in computers and conduct statistical analysis, so we also use statistical software. We write papers, we do image processing, we make images for our publications, and we use the Internet, e-mails..." [**Expert 11**]

"When I do research I use IT stuff all the time, I'm looking for papers and data...I'm also doing statistical analysis" [**Expert 14**]

"But my background is not computational, so somehow I am a bit behind. Especially when you start directing PhD theses or more generally, to supervise the work of younger people, you stay behind on the techniques. This is a problem... Computational techniques, –new programming languages, computational environments, it's something you always want to do but you can't find the time... Of course, when you have younger colleagues they learn those things, and you transmit other things that you know to them, it is kind of supplementary work." [**Expert 10**]

"As far as I remember myself I always used computers, as I'm growing older though I use them even less. More specifically, I spend more time in front of the computer but I learn less. The reason is that I simply don't have enough time and you don't have anyone to teach you these programs. In addition to this, since I use a number of different programs I forget them by the time I want to use them again and I have to remember them once more. This means that the complexity has increased, the variety has increased and my time has been reduced. Therefore I choose to learn something that I can remember and that will still be useful to me after some time." [**Expert 1**]

3.3.2 How the experts currently search

Q1. Are there often cases in your work where you need to search for information in the scientific literature (e.g., in journal or conference articles, books) or in structured information sources (e.g., databases, ontologies)? Can you provide one or two examples of such cases?

All the interviewees agreed that searching is a very large and important part of their work. They all search regularly in the scientific literature and many of them also search in databases and other sources of structured information. Most of the discussions for this question were interleaved with discussions that were also related to more specific subsequent questions (mostly Q2, Q3, Q4, Q6); hence, several interesting excerpts from the discussions of Q1 have been moved to the following questions. The excerpts that we list here are only intended to provide a first taste of the information needs of biomedical experts and the types of information sources they use. More detailed points and recommendations are made below, based on questions Q2–Q12.

"In the past, when I was working in a hospital I used to search for information mainly in medical databases and occasionally in more public databases to find some reports on rare conditions or for something that is not in the medical literature. But all of this has always been related to treating the patient, some reviews, reports, medical publications. Right now in my current work I search anything that I can find because one day I have questions on what monitors are there to monitor blood pressure and blood sugar, which are very technical questions. Even sometimes I search what has been commercially approved which is totally not medical or technical. I have searches that have to do with project areas where I'm writing a project, for example urology. When I was doing a proposal, it's not my specialty, I know a lot, but I don't know enough so I have to go in this specialty and start searching what is the current evidence, what are the current gold standards, so I just search everywhere!" [**Expert 8**]

"Well, I always search in the literature, mainly in PubMed. Now regarding structured information sources, we use such sources less for operational experiments... on bladder cancer, and more on databases that have to do with references to mutations, polymorphisms, in general to variations in DNA so for this kind of research there are specific databases for specific genes...We search in databases that have to do with sequences of genes or even sequences of proteins etc." [**Expert 1**]

"Yes, PubMed and all this of course, we really depend on that. We cannot work if we don't search in those. We can also use databases, like for example to check for data that have been put there from high-throughput experiments, sequence experiments" [**Expert 12**]

"First of all, we begin our search from the statistical analysis, then on a second level we conduct bioinformatics analysis, and on the third level includes data mining. For the first two steps we use bioinformatics tools such as the ones mentioned before while for data mining the tool Ingenuity and Gene Ontology classification tools, are an initial step. From then onwards the largest proportion of our work includes searching for information in PUBMED, in primary scientific articles and I would say that there lies 90% of the data mining we conduct by searching in person/ manually. This information is usually the most interesting/relevant/valuable to us. Following that, we may search for additional information in Google Scholar for scientific articles, in databases such as the Gene Expression Omnibus when we want to observe the expression of particular genes that have been observed in different gene networks. Other tools include GENECARDS, EXPASY, NCBI gene, which has collected specific information for every gene and other such similar databases. We may use MICRORNA databases for prediction or analysis of data, SWISS-PROT for the level of proteins, these are digital tools that try to present information gathered from various other sources which is presented in an organized way according to the gene or protein." [**Expert 3**]

"We are a bioinformatics group and bioinformatics combines biological data in databases and publications...For example there was a Greek endemic virus for which we lacked information; we didn't have much information about its relations. Firstly there have been some analysis on other viruses, we downloaded databases with information about other viruses in order to categorize it to the closest virus and thus to understand the use of our own virus. After we found the genus and the family of the virus then we had to consult PUBMED, which is a bibliographical database, to find more information about the other viruses that were related and to be able to make our own models. Another example is antibodies; we work a lot with them. Generally antibodies are a subject that has been growing a lot in the past ten years, and if someone wants to be up to date they have to constantly get informed. There are many publications about antibodies, so in order to be up to date someone has to search in bibliographic databases. If you want something very specific you have to go to databases like IMGT or if you want to see sequences there is SWISS-PROT. So all of us here are working with databases and bibliographic databases, both." [**Experts 6 and 7**]

"I work on proteins. So, there are specific databases only for proteins. There are also other databases containing different kind of information regarding proteins, for example their modifications. I am searching there, and of course I am looking for bibliography, for publications related to those proteins. I am looking at least into twenty different places for the same protein." [**Expert 9**]

"So well for example when I'm writing a paper and I need to read about something so I go online to search for papers related to what I'm looking for and that's probably the most common example...I haven't searched for online databases, right now we are working on the SAP enhancement Project which is basically focused on how to publish databases. So now I'm searching more on that but in general I use my own data, so papers" [**Expert 14**]

Q2. In what types of scientific literature do you search the most (e.g., journal or conference articles, books, technical reports, other)? Do you use particular repositories of scientific literature? Can you name the ones you use most often?

Most interviewees mentioned articles in scientific journals when asked about the types of scientific literature they search the most, though some of the experts also mentioned conference proceedings, technical reports, and/or books. Most interviewees tend to use specific repositories, predominantly PUBMED. Other starting points mentioned include MEDLINE (included in PUBMED), Google, Google

Scholar, SCOPUS, ISI and, perhaps surprisingly, Wikipedia. One of the interviewees pointed out that documents that have not made it through peer-review may still be useful; for example, if they refer to a trial that stopped and was never published because of an adverse effect, the adverse effect may be useful in drug repositioning.

Recommendations: The exclusive use of journal articles indexed by PUBMED as the only source of documents in BIOASQ seems reasonable, given the responses. A possibility for future systems might be to allow the users to specify if the answers should come from PUBMED articles, particular types of PUBMED articles (e.g., systematic reviews), particular types of documents on the entire Web (e.g., drug descriptions, clinical trial records, patents, conference proceedings, technical reports cited by PUBMED articles), or any Web document. For example, systematic reviews or on-line documents about particular drugs may be particularly desirable for clinical doctors; and clinical trial records may be particularly useful to experts working on drug repositioning, as already noted. For challenges like BIOASQ, however, a particular snapshot of the designated document repository has to be constructed to be able to compare the responses of the systems, which may be impractical if the repository is not confined to PUBMED. Hence, continuing to target PUBMED articles in future biomedical QA challenges may be a reasonable choice, though types of desired PUBMED articles could be added.

"We always search the literature –downloading and reading articles... For the most part it is articles in scientific journals, mostly research articles but sometimes retrospective articles too." [**Expert 10**]

"I usually search for published papers on Google Scholar, Google or others like PubMed and so on." [**Expert 4**]

"I mainly search into journals. About the technical reports, often they can be found inside the journals... Mainly on PUBMED, but I also use Google. You can get some information from Google as well, because sometimes you need to have to look for information that is not necessarily into a scientific publication, let's say in a journal. You might look for information about a drug that can be found through EOF or FDA. This information is not published in journals, but it can be found through Google in other websites." [**Expert 11**]

"I usually search ... in PUBMED... It's mostly articles (journals)... All the papers that I need can be found there. I work on Biomedicine and therefore everything I want is in PUBMED. PUBMED covers my needs." [**Expert 1**]

"I usually turn to PUBMED... Now, if there is something which interests me more, I search by keywords on SCOPUS, as well as in some specialised journals, like, in Genomics... I also use some specialised journals, as said, but I mainly search in PUBMED or in SCOPUS... The drawback with SCOPUS is that it requires subscription, it is not freely available, so you have to have institutional access to be able to use it. I usually search in journals. I search for information and articles that interest me in the reference list at the back. I look at the cited literature. These are also journals. I do look in some proceedings, like Lecture Notes in Computer Science (LNCS) and, more particularly, LNBI, Lecture Notes in Bioinformatics, which is a subseries of the former. So yes, I do use some bioinformatics-specialised conference proceedings, like RECOMB. And then, as I said, SCOPUS also makes reference to conference proceedings, which is very good. Some conferences are peer-reviewed, some others are not. I usually look at the former, the peer-reviewed ones." [**Expert2**]

"For our field, we mostly search in scientific journals... Medline and PubMed are the ones I use for articles. Google scholar helps sometimes too." [**Expert 2**]

"I do use PUBMED. It's the primary repository from which I will start searching... I read articles from scientific blogs etc. but I don't consider them part of my searches. I will always follow a link to a paper that has been peer reviewed." [**Expert 5**]

"Scientific journals. In some cases, when we are talking about younger students, they might use reviews of articles or, if they are even younger, textbooks. But in our case it is mostly scientific journals...What we mainly use is PUBMED, which surely includes all the official bibliography." [**Experts 6 and 7**]

"I also have a subscription in "Uptodate", if you are familiar with this. If I want to quickly see what's the best treatment for a disease, Uptodate is something which is updated every 6 months and I can get a quick summary from there...I wouldn't go much to Cochrane at all because the systematic review results are not something that I need that much. Often my work has to do with things that have not been researched that much and systematic reviews are done on things that have been researched so Cochrane doesn't help me ... I go to Google, I go to anything that could bring me information, not only scientific but also public, and then from the public if I find something I then try to find a scientific reference. So it's from broad to narrow. I don't start with PUBMED, I go later into PUBMED but I start from everywhere... I'm a little more open minded because for example in the UK people are conservative and they read only certain journals but in practice it seems that you actually have to read more and it is your own brain that needs to judge how good that is. For example there maybe things from the 70s which happen to have interesting information. For example, we just discussed this with a professor, a trial which was done in the 70s for a certain drug, he's old enough to remember it. And he knew that that drug had some adverse effects which were not published. This trial stopped because of the adverse effects. I wish I could have that data now because these adverse effects can be positive effects in another disease. That means that not everything that has not made it to a peer reviewed journal is rubbish, it can be very useful. In the area of drug repositioning - this is when you look at existing drugs that have new indications or even drugs that did not pass the clinical test - they can be sometimes used successfully in a different disease." [**Expert 8**]

"All of it, but mainly (journal) articles." [**Expert 9**]

"Generally I use PUBMED and Google, or actually the Wikipedia." [**Experts 12 and 13**]

"So mostly journal articles, sometimes conference papers and sometimes even books as well...Usually it's PUBMED, now I started using Google Scholar more as well and Web of Science." [**Expert 14**]

Q3. In what types of structured information sources do you search the most (e.g., databases, ontologies, terminologies, other)? Do you use particular repositories of structured information (e.g., particular databases or ontologies)? Can you name the ones you use most often?

Most of the interviewees use (or have used) repositories of structured biomedical information, but the repositories of structured information they use vary depending on their research areas, and there does not seem to be any established single point of entry (unlike PUBMED for research articles). GENE Ontology and UNIPROT (and its SWISS-PROT subset), two of the five designated ontologies for concept retrieval in BIOASQ (the other three being MESH, JOCHEM, DISEASE Ontology), were among the repositories of structured information mentioned by the experts, but several other repositories were also mentioned (e.g., Gene Expression Ombinus, microRNA.org, GENECARDS, LOVD, PDB, CATH, SCOPE) and it is unclear if their concepts are covered (and to what extent) by the five designated ontologies. It is also unclear if the additional repositories that were mentioned are included (and to what extent) in the

LINKEDLIFEDATA repository, which is the designated repository for RDF triple retrieval in BIOASQ.[8]

Recommendations: Future challenges may wish to use a larger set of designated repositories of structured information for concept and triple retrieval, or perhaps require the systems to find themselves repositories of structured information (e.g., ontologies or databases) that are relevant to each question. It should be noted that some of the repositories of structured information mentioned by the experts contain primary data (e.g., gene sequences, solved 3D structures) rather than human knowledge representations (e.g., that a particular disease is known to have a particular symptom). Hence, some of the repositories of structured information may not provide facts directly useful for the generation of the 'exact' or 'ideal' answers of BIOASQ, but it may still be useful to require systems to find repositories of primary data that could be related to a particular question (e.g., data that have been used in relevant articles or that could be used in experiments needed to answer a question).

"It depends on the project, the nature of the project. Someone could go at "Online Mendelian Inheritance in Man"" because they are interested on the mutations of a gene and on which diseases are associated to them... There is abundance of structured information... Unfortunately not all structured databases are included into one. This is why where you are going to search depends on what you are looking for. And often the structured information isn't complete, so when you use it you also have to include your own information... When, for example you are looking for solved 3d structures of a molecule you might search in one database that contains only solved 3d structures. You will find what you are looking for and if you want more information you will use this to search in bibliographic references. The same goes with Gene Ontology. You can search for what you are interested in and then go to Pub Med for bibliographic data mining." [**Experts 6 and 7**]

"I search in databases that gather information from various sources and present it in a unified structured form. We use ontologies to a large extent as well, for example GENE Ontology, but we don't visit the webpage of GENE Ontology to find information; rather the tools we use have the ability to analyze data based on Gene Ontology...Gene Ontology is a database. Gene Expression Omnibus is another database. MICRORNA database or MicroRNA.org are other databases for different type of information. Also, SWISS-PROT is database for the level of proteins, NCBI has various databases within, it functions as a main umbrella under which there exist various sources of information in the level of gene, RNA, protein. Therefore, we search a lot in NCBI while under this umbrella lies PUBMED which I mentioned earlier. From there on a researcher can find a lot of information. We also mentioned GENECARDS which is also structured information source which gathers material from other sources, there are many." [**Expert 3**]

"The last years we use GOPUBMED. This is a website based on PUBMED containing structured information supported by ontologies. For example you enter keywords and it gives you back structured results, based on the ontology. We use it sometimes. It also has the possibility to produce some statistical figures that PUBMED does not provide. For example for "ischemic episode" the system finds the bibliography and can also produce a statistical analysis about the publications, for example showing you the publications per year. This way you can see if the progress of the publications is active etc." [**Expert 11**]

[8] See http://linkedlifedata.com/sources.html.

"I use ontologies and when I conduct an analysis I download all information related to the protein I work on, I have a software that does this. The software harvests information -from around twenty databases- about ontology, pathways, interaction analysis. [...] That's commercial software. I have two or three such programs doing the same work. Ontologies are now used in the market. Of course there are also websites where you can do the same, but there are programs as well, allowing you to download information about specific proteins." [**Expert 9**]

"We use LOVD, a database curated and informed in Leiden... It has to do with variations of specific genes." [**Expert 1**]

"I don't (need them). But, for example, structural biologists looking at motifs in proteins, which are stored in databases, like CATH, SCOPE, which include, let's say, some structures, also PDB. These databases are structured, in which people can search for biological information, predominantly solved structures - which means that you have found where exactly the atoms of a molecule are located, such as the molecule of crystallin or the molecule of albumin. So, in order to understand how a medicine work you have to solve its structure. The structure is directly related to the function. So you first solve the structure and then you examine the function. So this is the use of such databases. But I don't use any." [**Expert 2**]

"Regarding structured information sources, we mostly search on things relevant to our practical work but not for the literature. That means that a large part of the research conducted goes through the analysis of data. If these data are not produced by one of your colleagues or in your own laboratory then you will be led to public data. So, we visit databases such as GEO, ArrayExpress etc. These databases contain segmented data of large scale and we take relevant information from there. But in order to reach that point this means that you had already read a paper. Usually, the paper leads you to the structured information sources. Unless you conduct research primarily based on data and in this way you may want to search in databases for all the experiments conducted in cell type for example. We don't search in ontological databases. We use standard ontologies, such as the GENE Ontology or KEGG and similar databases of biological pathway but this is done only once a few months to get the most recent updates, the information is stored and then treated locally. We have locally downloaded them in our computers and so we visit them whenever there is an update in the database and we want to get new data. But I don't use ontologies particularly that have to do with literature." [**Expert 5**]

"Hardly ever, structured information sources... In my work I don't have any particular need for interpreting data. Also, on ontologies, we created an ontology in the context of a project which wasn't very easy because it was probably too ambitious. I'm familiar with ontologies, I do look into ontologies beyond just to see what's been done. I hardly ever need to go down to trace data, actual data gathering in terms of results, long time ago I stopped doing it. I was doing it when I was sort of specializing and needed to evaluate research results but right now not at all, structured is almost out." [**Expert 8**]

"For my research, it is mainly genomic sequences. Whole genomes or parts of the genome... Also, sometimes we might use ontologies, if we want to know the nature of some genes. In our work, those are mainly ontologies of genes. For other researchers they might be ontologies of diseases or drugs... You have a vague idea on your mind; then you consult the bibliography and render it more specific. You might download some materials to run some first experiments and then revisit the bibliography. This is a continuous procedure. There are steps, forth and backwards, between textual literature or sources of text and sources of data. For me this data is genomes. For other researchers using structured databases it can be proteins or other biological material that has been digitized... I usually visit EBI, European (Molecular) Biology Laboratory. EBI is a European Bioinformatics Institute. Also I frequently use the Santa Cruise University Genome Browser and some ftp sites of Univerities or Organisations... Also, we often visit websites of several genome projects. A genome project is an organization aiming to the complete sequencing of the genome of an organism, e.g. Drosophila, or rice, or the human genome etc... About databases, a big part of our work is done with material that we download from databases. But it is not material that we use in BIOASQ. For example, say we want to study a feature of genomes. We might have to download 150 genomes of microbes. It is possible nowadays, there are hundreds of them." [**Expert 10**]

"I was actually using this structured information much more two years ago. Now I think that Google is the main entry point, because with Google you can go wherever you need. Before, I used structured databases like those at the EBI, NCBI. Now, because I am mostly supervising the work of others, I don't work directly with the data and I don't use this as much as before." [**Expert 13**]

"In my case I use a lot the database that is called FLYBASE. It's the database for drosophila. In that database you get all information you need about single genes, you can have links to PUBMED and to other paper information. So it is very useful. Many times I don't actually go to PUBMED, I go first to FLYBASE, I take the name of the gene and I find all the information through FLYBASE. In the case of drosophila this database is widely used." [**Expert 12**]

Q4. What kinds of information do you mostly search for (e.g., symptoms of diseases, appropriate medication, gene interactions)? Do you search for different kinds of information and in different repositories at different stages of your work? Can you provide one or two examples?

As one would expect, the interviewees search for different kinds of information according to their research questions, the stages of their research, and also (when applicable) their clinical and teaching needs. For example, they may search for proteins involved in a function, techniques to measure something, gene interactions, research related to a particular hypothesis, gold standards to diagnose some disease, medication. The following excerpts provide a taste of the variety of information the experts search for, along with reasons for searching particular pieces of information.

Recommendations: The variety of information needs of biomedical experts seems to justify the decision to aim at generic biomedical QA systems in BIOASQ, as opposed to systems (e.g., in information extraction competitions) aimed at extracting or retrieving specific types of information (e.g., gene interactions, protein blockers), which can nevertheless be useful components of generic QA systems. Continuing to aim at generic QA systems in future biomedical challenges seems desirable.

"My main research questions have to do with molecular mechanisms. That is how proteins and protein complexes interact inside a cell nucleus in a mechanistic spectrum. From this point of view one might say that we are interested in three things. Firstly, new players in the game, which means a new gene, a new protein that was found to be involved in a function. In another level, new techniques with which you can measure something or get a redout which you could not previously get ... Thirdly, the actual result of a research, like as you said gene interactions or what changes in a molecular level in a disease. I do search for them at different stages of my work. For example, when I would like to think of an algorithm application that we are developing, having as aim to be able to predict the locations/positions of some proteins on DNA, there are proteins that are attached to the DNA in non-random positions, reading the sequence of DNA again to observe if a protein will be attached or not. The first thing that we will examine is if there are experimental data saying where this protein is attached to as we need a gold standard to test the algorithm and our predictions. So this will lead us to find experimental structures defining this thing. On the other hand, it would interest us if a new protein attached to the DNA emerges, which we didn't previously know of, to see if we can build another model for this protein which would predict its position or to change an existing model. Therefore, new players interest us as they can strengthen our research tools or to give as the motivation for new tools. The third example is if the proteins interact with each other they may change the total number of positions on which these proteins are attached to the DNA." [**Expert 5**]

"Basically it's collecting the information what is there in the world in this area. So one type of information is well the gold standard for, let's say a diagnosis, what is the gold standard for establishing a diagnosis. Then the next question is what new developments are there that might become better than the gold standards but right now they are not, so what are the new developments there so I read through those and I see how they compare. For example, for one project proposal I did literature search on PSA which is a gold standard for prostate cancer. The research world has been trying to develop a better test for years. So I go and try to find out what other tests have been developed or are in the pipeline, how do they function. Then I go and check how complicated it is to do the new test, how expensive it is to do it, how realistic it is to do it, because for my new project idea I want to only pursue tests which you can apply in everyday practice. It is not of interest to me if an academic team has discovered something which is terribly expensive and that has limited application in practice, I care which is the test which is practically more applicable. Thus I go a lot into cost issues, often it's outside the scientific literature, often it's some reports, presentations, anything from the websites. So how much does it cost, is it registered, is it available in this country." [**Expert 8**]

"For example, there are cases when you need to find help about methodological issues, when you try to set up a new technique or to make a measurement that you haven't tried up to now and you want to see if it has been done by anyone else and if yes if it has been methodologically described so as to follow the same steps and possibly get faster to your result. This is one kind of information; another kind of information has to do with the results and the design of a research in the sense that when you start designing a research protocol you set a hypothesis. This hypothesis could be "Can the thyroid hormone restore the function of the myocardium after an ischemic episode?" Firstly you need to see if anyone else has done the same thing, because if it has already been done by many there is nothing new to add. So you have to see if this has been already done, you are going to look for "thyroid hormone" and "ischemic episode" and see what other publications there are, what studies have been conducted. Secondly you will have to look for information that could be indirectly related to this. Perhaps nobody has done the same thing, conducted this experiment and you check if there is any information that could support your hypothesis. What actions of the thyroid hormone could be related to the ischemic episode indirectly? Another thing is the dose. You may want to give thyroid hormone to animals with ischemic episode. Is there, in the bibliography, information about the dose? There are many kinds of information." [**Expert 11**]

"If I'm writing an article I read anything that will help me write this specific article. If I'm just searching around, which happens often, then you start searching for something and you might be led to a different area because you found something interesting there, and so you also save that source. But in general I search for something really specific... The main kind of information that I search for is gene variations." [**Expert 1**]

"In our research, we make long lists of genes. You ask, for example, what are the differences between normal and pathogenic samples and you concentrate on 200-300 genes that behave differently. Then you have to study those 500 genes, what is their biological role, what do they do inside a cell. Therefore, you can either search manually for each one of these genes in PubMed where for each gene the system will return to you 50-100-1000 articles, which is practically impossible or you can begin with tools that make functional categorization, they organize the genes in networks based on information gathered from various sources of data. Therefore, the part of the bioinformatics analysis you search aiming to gather as much information and to organize your genes into groups according to their function. Following that, after creating groups and having a clearer view, you decide in which of these networks/groups you will focus according to your subject of research. For example, if I study heart disease and a group of genes are relevant to cancer, this does not have anything to do with my subject so it's a parallel result on which I don't want to focus on. Therefore, I choose the groups of genes that I want to focus on, and afterwards, for these genes I will go and read for each one, to see what their role is, what do they do etc." [**Expert 3**]

"I mainly search for clinical and physiopathological papers, because my main research is in Physiology and Clinics, like heart failure or cardiac function... I search continuously. When I started writing the hypothesis of the work, of this study, so at the beginning to see if there are references in the literature and then my data, my results in my research paper and then to compare my data with others, already published." [**Expert 4**]

"It's the combination, always. I look also for genes, because the information you get about its form or its expressions until the protein, which is the final product... – In the course of my work, I also deal with drugs, blockers of some proteins. So I try to face the issue from all points of view. I am trying to see all sides: the gene, and the protein, and where it is placed, in which tissue, where inside the cell it is, what it does, with what other proteins it is combined and collaborates." [**Expert 9**]

"I teach biology of cancer, so I may look for genes and diseases. If I look for something related to my work I might look for protein modifications, for example. So in my case I look for both, research and teaching." [**Expert 12**]

"I was looking for a gene that we found linked to hypertension. So actually it has not been related to hypertension before. But it has been related to glaucoma. Then we found that people with hypertension have glaucoma. [So if the question is] is there a relationship between glaucoma and hypertension, of course I type "glaucoma" and "hypertension" and there is a paper that relates glaucoma to hypertension." [**Expert 13**]

"So for clinical purposes yes I look for clinical stuff like medication and surgery symptoms and imaging data. For research purposes it very much depends on the topic of the paper that I'm working on." [**Expert 14**]

Q5. Do you cooperate with colleagues or librarians when searching for information? If yes, at what stages, why, how?

Cooperating with librarians when searching for information seems to be something that the biomedical experts do not do, or at least not any longer. On the contrary, they all collaborate with colleagues, including supervisors, students, and colleagues with different areas of expertise. Collaboration can take place during each stage of research, particularly on projects, and some of the interviewees (Expert 2 being a counter-example) also collaborate when searching for information (e.g., to divide the search load, or to confirm their findings).

Recommendations: The responses of the interviewees seem to justify the decision not to involve librarians or curators in the construction of the BIOASQ benchmark datasets for Tasks 1b and 2b, which focus on questions of biomedical experts.[9] It seems reasonable to adopt the same policy in similar tasks of future challenges. Furthermore, given that biomedical experts are used to collaborating, often even when searching, it may be worth assigning questions to groups of experts (e.g., 2-3 experts in each group). By contrast, most of the questions (and gold answers) of BIOASQ were authored by a single expert each, and only a few questions were assigned to pairs of experts to measure inter-annotator agreement. Assigning questions to groups of experts may help improve inter-annotator agreement (see Paliouras (2014)). It may also allow the experts to formulate more challenging questions that require combining expertise from different biomedical areas, if groups of experts with complementary, but related expertise are formed. The social network of BIOASQ (Heino and Ngonga Ngomo (2013)), which allows experts to follow and comment upon questions (and gold answers) prepared by their peers (e.g., suggest missing related articles) can be seen as a step towards this direction, and it is in line with the fact that biomedical experts are used to seeking the advice of their colleagues. The social network of BIOASQ could also be extended to allow biomedical experts to criticize or complement answers produced by systems. This could lead to hybrid QA systems that would integrate answers provided by systems with answers provided by humans, though mechanisms to detect spam or disguised advertising (e.g., of drugs) would also be needed. Future challenges may also include tasks to match questions to other questions that have already been answered (e.g., in FAQs or discussion fora).

[9] By contrast, Tasks 1a and 1b attempt to automate part of the work of NLM biomedical curators (assigning MESH terms to newly published articles); they use the terms assigned to articles by biomedical curators as gold answers.

"Bioinformatics favours collaboration because it is an interdisciplinary field anyway, as it combines biology and informatics, so I collaborate with various computer scientists, or computer engineers, and I use various methods which have been developed, classification methods on biological data, so collaboration is useful...Searching is solitary." [**Expert 2**]

"In the biomedicine field nobody is working alone...There is a library, but the librarians are not very specialized in the kind of questions we are asking. It is mostly the young people, those who are conducting their PhD thesis, that are doing the specialized questions in bibliography" [**Expert 6**]

"For me the best way is to read the literature and be informed, even if you don't have a specific problem to answer. It is important to read without always having a set research question, because you will be led to a research question while reading. So the best way is the journal clubs, to read in groups. What we do in my group is that once a week or once every two weeks we share journals among us, which the leader of the group chooses, and it is a good way to scan what's out there. When I say reading I mean scanning, reading the abstracts etc, go in depth to parts that seem more interesting to me, and the day that we meet, which might take 2 or 3 hours, each of us summarizes what he has found interesting. In this way, we are constantly kept up to date on what has been published in journals of our interest. On the other hand, we need someone that has organizational responsibilities as it might, for example, a subscription of a journal to expire. We then need to contact the library, talk to the appointed person, find in which repository this journal belongs to etc. I don't go to the library to search for journals anymore, I'll only go there for books. The same thing can be done virtually." [**Expert 5**]

"But the reference search which changes, I say look I found this and the other one says I found this so we put them in one shared space like Dropbox, we put all the information we found. This is during the submission and then somebody takes charge of selecting the information and compiling it. If a submission is successful and we start a project then, the collaborative work continues - we have certain tasks called deliverables and there are, let's say, 5 or 6 partners who are involved in this. In the same way I give tasks to each partner and I say "you collect this information, I collect that". But if I find something in a new area, that is not of my interest but I know another partner will need it for his reports, I'll give to him to help with his search. So in my job at the moment, in my new project we have eleven partners writing and we have collected about 200 pieces of information. Then I give a task to one partner to summarize a specific part because he might have the best experience or I will give a task to another partner "can you make a diagram out of this", so we work together in collecting, reading and in writing." [**Expert 8**]

"When I started my research, when I was a student I was collaborating (with librarians). Back then there was PUBMED, which was very difficult, we were downloading journals, carrier content... – And when I was using the PC I always had help from librarians. Of course now that the technology has advanced and I also have gathered experience I do it by myself. I don't have help from a librarian." [**Expert 9**]

"Yes, in the sense that often we double-check our opinions. Each one of us has a question in mind, something that we investigate and we search on our own. Afterwards we double-check our information, our findings, we discuss. Because there is a lot of information and you can get lost, each one finds something different and we double-check our findings and discuss about them." [**Expert 10**]

"Before, sometimes, when you could not find a paper or you were looking for a journal that wasn't in your library you would go to them, but now I ask the people in my group if they could search something about a relationship of something with something else, or I do it myself. But I'm not really collaborating with librarians." [**Expert 13**]

"I try to cooperate so again if I'm looking for some clinical stuff the easiest way is just to ask somebody who has more experience about that stuff, so to hear from people from their experiences or from what they have read. For research, and if I'm doing interdisciplinary research like Genetics stuff I always try to consult geneticists because I'm not an expert in those fields. Likewise If I'm doing cardiology related stuff so then I try to speak to cardiologists or to people who have done more research, like senior colleagues, more experienced researchers so I try to ask their opinion. I don't usually ask librarians to do my search for me although we have this option in my university but I try to do it by myself or ask somebody that I trust." [**Expert 14**]

Q6. Do you use search engines or other similar tools to search for information in repositories of scientific literature and structured information? What kinds of queries do you formulate most often (e.g., sets of terms, Boolean queries, other)? How useful are these search engines or tools? What are the main problems you face when using them?

All the interviewees regularly use search engines and consider them very useful and necessary. It should be noted, however, that the interviewees seem to refer mostly to search engines for article repositories (e.g., PUBMED) or Web search engines, rather than search engines for structured information. The interviewees use mainly sets of keywords describing the topic of the information they seek. They occasionally also use Boolean queries and/or include in the queries author names or publication dates, but overall they report that they use mostly simple keyword queries. They often try several queries, replacing keywords by more general or specific terms, depending on how many relevant documents seem to have been retrieved. Some queries are navigational, e.g., intended to quickly lead to a particular Web page ("For instance, if I look for a function of a gene Google always directs me to this database called 'Gene Cards'. I can also go there directly and type the name of the gene in Gene Cards, but actually it is easier because you have the Google window there, so you type directly.") In most cases, the interviewees manage to find what they are looking for and they seem satisfied with current search engines, though they also mention problems one might expect: having to study large sets of retrieved documents to identify the information they seek, missing results (false negatives), some of which can be spotted by using multiple search engines, irrelevant results (false positives), ambiguous terms (especially acronyms), difficulties in specifying particular semantic relations that should connect the terms of the query. Another common practical difficulty is that search engines often return pointers to papers whose full content is not available to the experts, due to journal subscription restrictions. Some of the interviewees also use notifications from particular journals they follow, article recommendations from search engines (e.g., papers that refer or are similar to their own papers), and social filtering (e.g., sharing interesting references via Mendeley).

Recommendations: The interviewees seem used to, and reasonably satisfied with specifying their information needs via keywords. Hence, one may wonder if systems that would accept natural language questions (like the ones authored by the BIOASQ experts) instead of keyword queries are really necessary. The interviewees seem to agree that formulating natural language questions is more natural, closer to the way they think and, hence, desirable (see Q9 below), but it would be worth investigating in future challenges if systems that accept natural language questions actually manage to produce better answers than systems that accept keyword queries. For example, future challenges may first distribute

the keyword queries that the experts used for each natural language question instead of the question itself. Subsequently, they may also release the natural language questions, and check if the results of the systems improve (possibly because the natural language questions allow the systems to better disambiguate terms, or because they more clearly specify desired relations between entities, or because they allow SPARQL queries to be generated).[10] The responses of the interviewees also remind us that queries or questions are not the only ways to specify information needs. The previous searches of the experts, the articles they have stored locally or shared with colleagues in social filtering platforms, the journal notifications they have subscribed to are examples of additional, indirect ways of specifying user needs and preferences, which could be incorporated in future systems and challenges (McCreadie et al. (2014)). Finally, future challenges need to be aware of content access restrictions (this was already an issue in BIOASQ) and possibly negotiate with publishers special licenses for the challenge participants.

"We wouldn't be able to do anything without them. We rely on them, even if sometimes they are not as good as we wish." [**Expert 12**]

"By using keywords you can determine what you are looking for easily and fast...Given the results they present us up to now, they are very useful. But you can never know if they are working well and if the results are always right. What you get is always what you can see. If you have the time to separately put the same question in more than one search engines and then compare the results... Personally I have done this once, when I was looking for genes related with the metabolic syndrome and I have spotted some problems." [**Experts 6 and 7**]

"If I am trying to find a function of a particular gene I just type the name of the gene. Usually Google always drags you to the same places. For instance, if I look for a function of a gene Google always directs me to this database called GENECARDS. I can also go there directly and type the name of the gene in Gene Cards, but actually it is easier because you have the Google window there, so you type directly... Of course if you have a complex query that is more conceptual, then it is sometimes difficult to find the information." [**Expert 13**]

[10] The keyword queries that the experts used for the BIOASQ questions have been stored and could be included in the BIOASQ benchmark datasets.

"PUBMED also helps you conduct more specialized searches. For example, like Google Scholar, it gives you the opportunity to get feeds from newly published papers which mention either your own work or relevant work. In this way a large part of information is filtered and you receive a notification that these papers of the latest month might be relevant to your research etc... Regarding search engines, I mainly use PUBMED and Google Scholar but I often use all these hybrid social media type search engines such as Mendeley or an older engine named Conoteia. For example, in my old laboratory... we had a group in Connotea in which we were sharing papers among us, sending articles to colleagues relevant to their research, or links from conferences, this is how it worked. Now we are doing the same using Mendeley as I'm collaborating with various persons specialized in different domains inside our laboratory and outside... We have several groups in Mendeley, we add there references to articles that are interesting and thus a more social way of searching is conducted led by either your colleagues or by the search engines which are really helpful...The bulk of information, that's the main problem. For example, if someone has some extra time and starts reading the results of a search then this might never end! ... The bulk of information is a blessing and a curse. You want to reach results but on the other hand you have to set some limits. Another main problem that we face is of financial nature, we don't have open access in several journals and so we must find several ways to access that information... What I personally do is that there are about ten standard journals in which relevant to my research articles are published, apart from the more general journals that I believe are read by most of the researchers of various specializations such as Nature, Science etc, or in Biosciences the journal Cell which is of wider interest. In these journals you can find articles of wider interest and therefore, even if I don't read something on Bioinformatics I would like to be up to date with the latest developments in the field of Molecular Biology in particular which is the research subject that I'm most interested in." [**Expert 5**]

"I really use Google...Keywords mainly, very seldom do I put in a sentence so it's usually keywords which take me somewhere and from there I search or occasionally if I find an article or some sort of source I put the quote and see if I can get into the article from a different library...One of the main problems is that the material is locked and I cannot read it, sometimes you have an abstract and you can't reach the whole material and it's annoying but I just learned to go around." [**Expert 8**]

"We search mainly through keywords. It can be, for example, 'thyroid hormone and myocardial fraction'... I mainly use PubMed, but I also use Google... Sometimes you get irrelevant results. That's the main problem. Sometimes. In other cases you get very good results. It mainly depends on the query. Sometimes the results have no relation with what you are looking for." [**Expert 11**]

"As for other search machines, I use Google Scholar... I make simple queries, nothing too complicated. And then I search on my own. I study the CNEs, and search for related elements. That's it. And then sometimes I use some filters, like articles which have been published from 2004 onwards, etc. But my queries are simple...I use various queries such as "CNEs *and* development". This was I combine some keywords and so you can... For example you may look for p53, which is an oncogene, and colon cancer, so "p53 *and* colon cancer"; this oncogene has been proven to be related to colon cancer. Or try and look for some side effects of P53. So you can put "p53 *and* alopecia". So this synthesis is very common and I use it a lot to find out about genetic correlations/interactions and so on." [**Expert 2**]

"I usually search with keywords in PUBMED. If it's something familiar then you can go on searching by the name of the author or by the name of a certain technique... I mainly use PUBMED, sometimes I search through Google but in these cases I'm not interested in the amount or quality of information that much, I'm rather interested in finding an article that has to do with a specific topic. Or even sometimes in PUBMED you might find a really old paper which is listed in the 35th page of results so in these cases I search through Google because I know that I will trace it faster...I search by keywords mostly. I might use the name of the author sometimes... In the level that I'm searching for information, I almost always find what I want. I don't have any particular problem. When using PubMed or Google I can find what I'm searching for... My research area is really focused and because I know what I'm looking for I don't have so many results. If the number of retrieved information is really big then I can search by more specific keywords, and sometimes they are too specific and I even have no results!" [**Expert 1**]

"Keywords basically... Two of the most common problems are when an abbreviation exists in the literature and corresponds to two or three different terms and therefore the search engine can return a large amount of irrelevant information while on the other hand if you search by using the full name the search engine might not find any results because publications might use mostly the abbreviation or the word with different punctuation/characteristics. The second problem is the extent to which the reply can be comprehensive and focused. For example, when we search using more than two key words, usually the system returns results less targeted/focused because it might find the specific keyword or it might find those key words without necessarily having any relation between them. Because we might not want to find those words in a row but we would like them to be connected. That's another problem. And the third problem is that these search engines, return to you the primary information which you then have to link, after reading all of it, by yourself. They do not conduct mining of the relevant elements of information for you." [**Expert 3**]

"I had problems searching some kinds [when formulating keyword queries in the BIOASQ authoring tool], to have an accurate selection of documents. Probably I was mistaken in the selection of keywords but I did not find all the papers already published... [In search engines, I use] keywords... When I try to make a question, the selection is not accurate. The selection of the papers is not so great." [**Expert 4**]

"Sometimes I do something very specific. I am looking for a protocol. In that case I will enter a keyword for exactly what I want. And in most cases I get my results. If I ask in a very broad way I might ask the search engine to put some date restrictions...It is very rare not to get a result. Especially in our field there is a lot of information, thousands of publications, many things that are similar... We always have an alert when a new issue of a journal that interests us is published. So we are following specific journals, but also specific molecules..." [**Expert 9**]

"Google, yes...Mainly bag of words. And also simple Boolean queries, very often. Simple ones. It has never been necessary for me to use filters in the abstract, the title etc. Simple Boolean queries using associations." [**Expert 10**]

"For example, now I'm working on a paper about evaluating clinical impact of metabolic syndrome, association of metabolic syndrome on mortality so I'm looking for papers on that topic so I type "metabolic syndrome" and "mortality" or something like that, I try to type more specific. So first I type some broader terms and then if I get too many articles then I try to narrower my query, instead of "mortality" let's say "cardiovascular mortality", or just describe the study population like primary patients, that's a paper that I'm working on right now...When I'm using PubMed sometimes I don't have full access to full paper or you know, my university has lots of those papers subscribed but still not all of them so sometimes I'm not able to see the full paper. For Google I might get too many, all this random association... Usually what's a good option in PubMed is that you can see what's related... that usually helps me because if I see a paper that's something that I'm looking for and that I can see obvious related articles, that's helpful. For Google it's more difficult as I said so usually I try to work on Google Scholar but I don't know how this works" [**Expert 14**]

Q7. [If time permits] What are the main criteria you use to evaluate the results (e.g., retrieved articles, snippets, database records) returned by the search engines or tools (e.g., relevance to the query, publication year, well-known journal or repository, review article, other)?

Most interviewees had retrieved journal articles in mind, when answering this question. The criteria mentioned included the names of the author(s), their reputation, their affiliations, the name of the journal, its impact factor, the citations pointing to the article, the type of the article (e.g., review or research article), its recency etc.

Recommendations: It would be useful to add support for filtering or ranking criteria like the above in the BIOASQ authoring tool (Heino (2013)), in the functions that allow the experts to retrieve possibly relevant articles, which were often too many. This could reduce the time needed to prepare new questions and gold answers in future challenges. Support for the same criteria might also be useful in future deployed biomedical QA systems.

"Who wrote the article, in which journal it has been published. Those are the main criteria." [**Expert 9**]

"It depends on how familiar you are with the topic. If you are totally unfamiliar it is really difficult to judge how reliable the information is... You need to have some knowledge. Of course if it is in a scientific journal you may trust it more than if it was in some sort of obscure page you do not know the author of, but it's not always clear how you would judge the quality of information." [**Expert 13**]

"How recent the articles are, or from what journals. We know which journals from our field are of a better quality, let's say. Or someone might want to use more objective criteria, like the impact factor. Sometimes we also know the authors. We also look where the authors come from, from which labs. We can also see if it is a review article or a research article. There are many criteria. In addition, we look at how many times an article has been cited" [**Expert 6**]

"If a subject is really "sexy" and researchers are working on it then in that case you would like to see the more recent results as there would have been changes. PUBMED and Google Scholar give you the most recent results in any case... Due to new methodologies, something that is 10 years old and more may be considered outdated. This has to do with the research field." [**Expert 5**]

"First, what I do is I read the title of whatever I get, then if I say that the title is something that I'm interested in then I open an abstract if it's like from PubMed and then I read the abstract." [**Expert 14**]

Q8. [If time permits.] What do you do with the search results you consider relevant to your information needs? For example, do you keep notes on hard copies of the articles as you study them? Do you highlight relevant snippets? Are there any particular steps you follow to synthesize and organize information from multiple articles, books, ontologies, databases etc.?

Most of the interviewees try to organize the results of their searches. The methods they use include: storing retrieved articles into folders (sometimes shared across a lab) using indicative file and folder names; highlighting snippets and writing notes on electronic or, less frequently, hard copies of articles; extracting snippets and copying them to separate files or bibliographic databases (sometimes also shared across a group) possibly along with notes; tagging articles with tags reflecting the purpose of the search (not necessarily the terms provided by the authors and journals); and organizing the retrieved information as slide presentations. There does not seem to be any clearly dominant approach. Most of the interviewees seem to have adopted their own personal approach, and at least some of them do not seem entirely satisfied with the approaches they use ("I am really chaotic", "This is a problem"). One interviewee (Expert 13) has completely given up trying to organize the retrieved information, saying that "the effort that is required to organize the information doesn't pay off the fact that you can actually search again". Interestingly, Expert 1 reports that (at least in his/her field) reading an entire article is rare, and that particular emphasis is placed on studying images and captions ("You rarely read an entire article, particularly in our work we don't read the introduction for example, we rarely read the findings, we simply observe the images and their captions.")[11]

Recommendations: The responses of the interviewees suggest that organizing and storing the retrieved relevant information and its sources may be as challenging as, if not more challenging than, searching for relevant information. A combination of the BIOASQ authoring and assessment tools (Heino (2013)) that would allow experts to inspect, edit, and store for each English question lists of relevant articles, snippets, concepts, triples, along with 'exact' and 'ideal' answers, all initially suggested by a system or found by the experts themselves (or colleagues, see Q5), might be a useful tool to organize and store relevant information and sources per question.[12] In effect, a tool of this kind would subsume several of the approaches the interviewees have adopted. For example, grouping and storing retrieved items (articles, snippets, concepts, triples) per natural language question is similar to tagging them with tags reflecting the purpose of the search that retrieved them, and to some extent also similar to storing them into folders corresponding to particular information needs; storing relevant snippets corresponds to highlighting or clipping them; editing 'exact' and 'ideal' answers is similar to keeping notes summarizing the findings of a search, especially if each 'exact' answer and each part (e.g., sentence) of an 'ideal' answer is linked to the sources (e.g., article snippets) that support it (see also Q11 below). Future research could study more extensively (e.g., with a larger sample of experts) the approaches that biomedical experts (and possibly experts from other scientific areas) use to organize the results of their searches, in order to propose best practices and propose future challenges based on them. The fact that most of the interviewees store (and some also share) search results (e.g., articles, snippets) they have found useful in the past also points to the possibility of using these previous results as indirect ways of specifying user needs and preferences, as already noted in Q6. Finally, future challenges should pay more attention to retrieving relevant images, tables, and perhaps other non-textual elements (e.g., equations) of articles (or other

[11] This was also pointed out by J. Sack in a blog post titled "Helping Researchers See Farther Faster", along with the point that nowadays "Finding is easy... but reading is hard"; see http://googlescholar.blogspot.gr/2014/09/10th-anniversary-series-helping.html.

[12] Some BIOASQ experts also pointed out this possibility while using the authoring and assessment tools.

sources), given the importance that they have in at least some biomedical areas. BIOASQ was not concerned with non-textual elements of articles, but it may be possible to exploit results of VISCERAL in future challenges.[13] IMAGECLEF tasks may also be relevant.[14]

"I am really chaotic with this. It depends on what period of the year. If I am doing research then I compile the information in a very organized way. But many times I have just one folder with the topic and the name and the recent papers there. Sometimes I print them. If I print them I highlight on the paper. I'm not very used to highlight them on the computer. But I must tell you that I am completely chaotic on that. And then I have some different folders in my computer, I know how to look for them, or some piles on my desk and then I know which pile corresponds to what topic. But I'm not very well organized in this." [**Expert 12**]

"I gave up trying to organize the information because I believe that the effort that is required to organize the information doesn't pay off the fact that you can actually search again" [**Expert 13**]

"I download the PDFs of published articles that I have used for a project. My notes are already there, since I have read them. You can also add text, for example a small resumé about the content of a paragraph...another way for keeping notes is in a small database. It is the database where I keep my bibliographic references. I have created my own database, like the one we also have here at the Lab, where we have all our references. You can search in many ways and look for articles published from this Lab or articles that we had used in the past. This is how all of them are stored. If you want to have a better look inside the contents of the article then you will have to search for it elsewhere." [**Experts 6 and 7**]

"We try not to print a lot any longer. This is probably a new trend. But I store them in folders and sometimes I print them... It's rather mixed up...I would place them in a PowerPoint altogether, parts of them. Like a presentation." [**Expert 9**]

"I download the .pdf documents, I have them organized in a Dropbox folder which I have named CNEs, for example, and then I open the .pdf documents and I do exactly what you said, I highlight snippets, with the tool that Acrobat Reader for example offers, which makes parts of the text yellow. I then copy those snippets to a "Notes" folder and then I add this too on Dropbox as a text file. For every article I read I use its title as an identifier. And then I put the snippets. And then I just read them. And then I am done with the article. I do it like I was a machine. Of course they do not make much sense as such, right? These may include different things, but at least the important things for me are there, out of each article." [**Expert 2**]

[13] See http://www.visceral.eu/.
[14] See http://www.imageclef.org/.

"I don't use snippets. What I do is that I do detailed tagging. By using tools such as Mendeley, previously this was done more manually by writing keywords on the article in a row below the title. The author does not always provide this or in some cases we are not interested on what the author or the journal gives as keywords as we might have different views. So, the papers were organized according to their manual hashtags. Now I do the same digitally but it follows the same reasoning. For example, I'm currently reading a survey that has to do with the analysis of the RNA as it is produced. What I do is that I scan all the papers that I have downloaded, their title and abstract, and I tag them based on this...What I have done is that I concentrated all of my published papers, my PhD etc, I gathered all the literature digitally, I scanned a database and I inserted all this information into Mendeley. Now I have in there almost every paper that has attracted my interest in the past while anything new that I read I add it there. You now have the possibility to search inside this corpus even with a simple text search as it is a narrower bulk of information, you don't search in PUBMED or in a database. So you gather there the most relevant papers, you are more flexible in the kind of literature that you gather and then you start narrowing the results when selecting what to read etc." [**Expert 5**]

"Either if there is a PDF I save it and then I mark it. I do have Evernote in my computer and I transfer if there's a webpage or something I transfer it so I work a lot with Evernote, I clip things, I transfer and then I start cleaning from there. So this is my fastest way but sometimes it's just cut and paste and put it together because there are so many, I go through hundreds of webpages and documents and I do remember what I found where but if I rely on my memory there's no way. So whatever I find I put it in one place basically and then I check it." [**Expert 8**]

"I usually save the papers in directories according to their subjects and I have recently stopped printing, as you can see my office is full of paper! Because of that, even if I have some papers printed I don't remember where to find them. For example, I'm editing a paper, which we want to send for publication to a journal, that some researchers sent to me recently and I wanted to take this paper to edit it at home over weekend. So I printed it. But sometimes, before I start writing a paper I have downloaded all the relevant literature. From what I have saved to my computer I print those that I consider more important. Because the articles that I will print are more possibly those that I will entirely read. You rarely read an entire article, particularly in our work we don't read the introduction for example, we rarely read the findings, we simply observe the images and their captions. In our work this is really important. This will tell us more than the writings in an article, because usually authors exaggerate when presenting their findings and therefore the image clearly shows you the findings. I often print these to think more on them, I keep notes in parallel regarding what is useful to use, what I can discuss more in our case, common findings with our work that can be discussed.... We also use Endnote to download, to create a library of sources so that we can easily write the references for a paper, again with keywords. You might also use date there. In this case the topic is not as important, the dates, the authors, any other irrelevant information or a very characteristic key word which is not that common might help better... " [**Expert 1**]

"I would possibly do this in my PowerPoint presentation. Most of my results are in PowerPoint presentations, on which I also compile the things related to a particular research project and the references. All those things are in my presentations." [**Expert 13**]

"This is a problem, because the quantity of the information gathered is huge and the problem is how to organize all this and how to find again something that you know you have already found and you know that you stored it and you wonder where it has been stored. This is often a problem. We try to catalogue them by ourselves with some kind of logic, often giving them names related to the content of the article... Saving a PDF, usually the articles are PDFs, with names like "thyroid hormone and ischemic episode". Another solution is to save them with the name of the author, often the name of the first author." [**Expert 11**]

"" [**Expert**]

"" [**Expert**]

"" [**Expert**]

"" [**Expert**]

3.3.3 Matching BIOASQ and future challenges to user needs

Q9. Is it worth trying to develop (or improve) systems that will allow queries to be formulated as natural language questions, as opposed to sets of terms, Boolean queries etc.? Are the types of questions used in BIOASQ (yes/no, factoid, list, summary) enough? Are there any additional types of questions that you believe should be considered?

Most of the interviewees were positive (some even enthusiastic) about the prospect of using natural language questions instead of keyword queries. The expected advantages of using natural language questions include: being able to express more naturally, directly (closer to how experts think), and perhaps faster (e.g., in emergencies) information needs; being able to specify more precisely information needs (e.g., by specifying particular relations between entities), thus hopefully also obtaining more specific answers; being able to use morphological variants of words, synonyms, or alternative phrasings of questions with the system hopefully still being able to retrieve the right information. Some of the experts, however, were aware of the difficulties that systems face when attempting to understand natural language and, hence, were skeptical about the practical value of using natural language questions instead of keyword queries, given that they are familiar with formulating keyword queries and they seem to work reasonably well (see also Q6). One of the interviewees (Expert 5) also pointed out that formulating the right natural language question may not be as direct or easy as one might think ("What I want to say is that the problem with the natural language question is that you must be careful how you will formulate the question. While it seems to be simple, it leaves great flexibility to the user as well as responsibility at the same time. Doctors who would like to use natural language questions as queries must think about the question really carefully."). We interpret (rather freely) this comment as saying that keyword queries (presumably less precise than natural language questions) may be easier to formulate when the users do not know exactly what they are searching for and they may still return interesting results, i.e., they may be more suitable to exploratory search.

Regarding the types of English questions of BIOASQ (yes/no, factoid, list, summary questions), the interviewees seemed overall satisfied. Two interviewees proposed a variant of list questions, where separate positive and negative responses would be required (e.g., pros and cons of a treatment). It was also pointed out that what may initially look like a factoid question (e.g., seeking for a single entity) may turn out to be a list question (multiple entities may satisfy the constraints of the question), and there is currently no easy way in the authoring tool to change the type of a question. The interviewees also pointed out that in reality there are often questions for which there are no clear-cut answers ("there are answers in research that are not 'yes-no', they are 'maybe', they are 'yes' in this case and 'no'

in the other case"), or questions for which there are contradictory or no answers in the literature. We note, however, that the guidelines that were given to the BIOASQ experts (Malakasiotis et al. (2013a)) instructed them to avoid questions for which there were controversial or no answers in the literature.

Recommendations: It may be worth using both English questions and keyword queries as separate or joint inputs to participating systems in future challenges (see also Q6). Another possibility would be to add follow-up questions (e.g., gradually more specific questions, possibly including pronouns or other expressions referring to previously mentioned entities) or to support clarification dialogues (as suggested by Experts 6 and 7). It might also be useful to consider spoken dialogues, though this was not mentioned during the interviews.[15] Factoid questions can perhaps be merged with list questions, since they are, in effect, a special case of list questions (a list of one element is required). List questions requiring separate lists of positive and negative answers (e.g., pros and cons) could also be added, though pairs of list questions (requiring positive and negative answers, respectively) could be used instead. Questions requiring lists of steps (e.g., to peform a medical procedure) could perhaps also be added, though they were not mentioned in the interviews.[16] Although questions with no clear-cut answers may be important in practice, it would probably be particularly difficult to evaluate answers to these questions and, hence, it is probably best to continue avoiding them in future challenges. It may be possible, however, to add questions for which the correct response would be that there is insufficient information in the literature to answer them, or that the literature contains controversial information (possibly with pointers to contradicting articles, snippets etc.). in future work it may also be worth measuring the time taken for the users and systems to formulate and process, respectively, natural language questions vs. keyword queries, along with the corresponding evaluation scores of the retrieved results (MAP, accuracy, F-measure etc.), especially for questions that reflect urgent information needs (e.g., in emergencies); by contrast, users submitting questions for research purposes may be more interested in the quality of the results, rather than fast responses (see also Q11).

"Sometimes it is difficult to say that I'm going to put AND, OR. This combination is not like a full sentence. You have to combine words. Sometimes the way you combine them can mean something different. So that would be great if you can write like a sentence." [**Expert 12**]

"Natural language helps. Yes, I believe it is worth trying. I reckon that a scientist would do both: use natural language and keywords. But [natural language] might render the answer more specific." [**Expert 9**]

[15] See https://www.stonetemple.com/great-knowledge-box-showdown/ for a comparison of spoken QA in Apple Siri, Microsoft Cortana, and Google Now.

[16] Google already provides some support for questions requiring step-by-step instructions; again, see https://www.stonetemple.com/great-knowledge-box-showdown/.

"I believe that it is worth it, but we are still far away from the ideal outcome. Being able to do such a thing would be perfect, because there is so much information; we use most of our time looking for information... If we could formulate a question and get the exact answer back, that would be perfect... [Possible other type of questions:] Questions for which the answer was not a list of things but a list of things that are recommended and a list of things that are prohibited... It is a list, but with plus and minus. This is not included in the types of questions that we have...When you are using keywords you are obviously making an abstraction regarding your question and you let the system do the data mining for you. Asking a whole question is something more specific. Of course there could be a smart system which, in case it does not understand your question, would come back and ask for explaining the question until the question is fully understood. While with keywords I believe that you have to search inside the answers in order to find the one you are looking for. Whereas if the system fully understands the question –It would be much better to have such a smart systems that would ask again and again and offer a full answer." [**Experts 6 and 7**]

"I believe that this would be a general goal, for all specialties in this field. I think, as with all systems, that these should at first function in parallel. What I want to say is that the problem with the natural language question is that you must be careful how you will formulate the question. While it seems to be simple, it leaves great flexibility to the user as well as responsibility at the same time. Doctors who would like to use natural language questions as queries must think about the question really carefully... We just have to learn to do it automatically... Summary [the type of summary questions] is really comprehensive, it covers all questions that are not really specific. I can't think of anything else." [**Expert 5**]

"It would be really useful and I consider this to be the next absolutely necessary step, especially in the area of Omics, referring not only to Genomics but also to Proteomics, Metabolomics, etc, because technology allows us now to perform high throughput and massive scale analyses. As a result we collect many terabytes of information in really small periods of time which makes it impossible to analyse and utilize this information. At this moment, the main bottleneck is our limited brain power." [**Expert 3**]

"I don't really care. I'm used to searching by keywords, I search like this since 1994. Even before that when I was conducting research in the National Documentation Center I searched by keywords. So I don't have problems, I'm used to this kind of searching. I'm not sure if it can be done otherwise. In any case, keywords are the important parts of a sentence. So if you add "and" or "because" etc this can return to you more "noise" than relevant results. Therefore I believe that keywords are convenient... Keywords are really convenient and they return to you a large number of results. If you search by snippets for example, this might be useful but I'm not clear about it because this may filter results to a higher extent and thus you will receive few or no relevant documents back. If you search something really specific then you can formulate a snippet of your choice to receive more focused results, while this snippet shouldn't include many words of general content... Factoid questions are a special type of list questions which has sometimes made things more difficult for me or on which I met some problems in the software. This is because while I had a specific answer in mind and I considered it to be factoid, at the end there was a second answer which I haven't thought about it earlier and then I had to change the factoid question to list questions and so I was confused!... In Biology there is rarely only one answer and that's why list questions are really convenient." [**Expert 1**]

"Yes of course... This is predominantly time-saving, especially when dealing with emergencies. OK, for us, maybe not as much. But I have seen it happen in ... hospital, I saw this doctor who would enter such a system at 3 am and read about a certain medicine's side effects." [**Expert 2**]

"Yes I would use it (natural language questions). I mainly use factoid and yes/no questions." [**Expert 4**]

"I tried to put questions in natural language and even the way I'm asking is kind of very difficult for a machine to exactly get what I want. So things as they are at the moment I think it's better if you just put the keywords and find the connection. For example, if I ask what is the association between disease A and disease B or what is the cause of something, the way you ask matters because you can get very contradicting answers, the causality in a question is a problem. "disease A causes disease B" or "B causes disease A" or "do they have a common cause". I don't get what I want at present trying to use natural language. But what I do is I put the keywords and then I find the article in the direction of association I want. I prefer keywords, it's easier... What I'm saying is that it might be a waste of effort to try to process the language and you still may not get what you want... [Regarding the types of questions:] I find it more difficult to ask a factoid question because you are looking for one word or one thing and the system doesn't get too much. So if I compare the quality of making questions and answering questions probably the factoid is the hardest one to do. Yes and no is easy, although there you have a problem that half of the literature says yes and the other half says no and I'm evaluating what comes up. For example, within the project if I say "is there an association" and if an article says yes and the other article says no I'm the one who chooses yes but there is an article which says no. So, we're not exactly representing what comes out because I'm deciding, my brain and I'm ignoring "no". So there is a bit of a catch there, the list and the summary questions are easiest." [**Expert 8**]

"Of course... Natural thought is imprinted in natural language. And this thought that you have in mind you try to transform it into keywords. This is not always possible. Natural language is of course better. You cannot always express what you think in keywords... [The types of questions] I believe they cover the whole spectrum. Maybe the only that is not exactly covered by these categories is that sometimes, rather often, there are answers in research that are not "yes-no", they are "maybe", they are "yes in this case and no in the other case". This is not exactly covered. For example there is a "yes-no" category. There are sub-cases, very frequently. Or the answer could be "we do not know", "I do not know", because there is no information." [**Expert 11**]

"We are used to speak and to write in natural language. Not many people are able to use coded or structured language or a typed language. It is more difficult. For people it's easier to ask a question the way they speak that lets you know the specific domain of the language that is being used, the code, and the ontology that is being used in a particular domain. It's difficult to know how to ask a question. I go for instance in this ontology in a database. If I want to look at a function unit, to know the name of a function, it has been typed into the ontology. It can be "embryology" or "embryologists". Sometimes it is a small spelling difference that makes the difference... In terms of types of answers I think it was O.K., these different kinds of answers. What else would you ask? I think in terms of possible questions it was not bad. Yes-no, factoid, list and summary would cover almost everything." [**Expert 13**]

"Yes, because If there could be a software that would interact this way, accepting questions in natural language and providing answers in natural language, of course with links, that would be very comprehensive. And also very satisfying for the one who is asking... At next stages of this kind of work one could also envisage to implement more complex forms of questions, but now the important thing to improve is the correctness of the answers returned by the systems...[Regarding the types of questions:] Those were questions for which the answer was not a list of things but a list of things that are recommended and a list of things that are prohibited. This is something that has been realized. It is a list, but with plus and minus. This is not included in the types of questions that we have." [**Expert 10**]

Q10. For each English question, BIOASQ requires the systems that participate in the challenge to return (i) relevant documents (or abstracts), (ii) relevant snippets of documents (or abstracts), (iii) relevant concepts (from ontologies, terminologies etc.), (iv) relevant statements (English-like renderings of facts from ontologies, databases etc.), (v) an "exact" answer (e.g., name of a disease, list of symptoms), and (vi) an "ideal" answer (summary of the most important relevant retrieved information). Would all of (i)–(v) be useful in practice, assuming for the moment that systems that reliably return (i)–(v) can be constructed? Which of (i)–(v) would be most useful and why? Given the outputs of the participating systems that you have assessed, which of (i)–(v) are more likely to be returned reliably by systems in the near future?

Most of the interviewees agreed that documents (or abstracts) and snippets were the most useful among the answers of Task b – Phase A (documents, snippets, concepts, statements), though of course going through a large list of returned documents or snippets can be tedious. By contrast, statements were considered to be the least useful and most problematic among the answers of Phase A, mostly because they often did not convey useful information (e.g., they expressed obvious is-a relations), they did not "make sense" (e.g., they were difficult to read and understand), and/or they were too many. For example, one of the experts showed us the following top statements that were retrieved by the BIOASQ authoring tool for the Boolean query "CpG islands" AND "plant genomes"; note that 'Genomics' and 'Genome Res.' are presumably journal articles.

CpG island protein (aka. Factor VIII intron 22 protein) is referenced in Genomics

CpG island protein (aka. Factor VIII intron 22 protein) is referenced in Genome Res.

hypermethylation of CpG island (aka. DNA hypermethylation of CpG island) is a DNA hypermethylation

DNA hypomethylation of CpG island (aka. hypomethylation of CpG island) is a DNA hypomethylation

hypomethylation of CpG island (aka. DNA hypomethylation of CpG island) is a DNA hypomethylation

CpG Island Methylator Phenotype (aka. CIMP+, CIMP) is notated C19821

CpG Island Methylator Phenotype (aka. CIMP+, CIMP) has note NCI Thesaurus

hypermethylation of CpG island (aka. DNA hypermethylation of CpG island) has namespace biological_process

Concepts were overall considered less problematic and more often useful, compared to statements, but still less useful than documents and snippets. Furthermore, the purpose of concepts and statements was unclear to at least some of the interviewees. For example, it was unclear if the statements would in practice provide useful information to the users (experts) submitting the questions, or if they were intended only to help (or evaluate) the participating systems. Similarly, it was unclear if the concepts were to be considered parts of the answers sought by the users (and if yes, how they would help them in practice), or if they were intended to help the systems by enhancing the questions with additional

concepts the users knew were relevant. Also, the quality of (possibly relevant) concepts and statements that were shown to the experts (during the authoring of the questions, gold answers, and the assessment of the system responses) varied a lot from question to question ("sometimes I could use the concepts and the statements well, because they were quite clear and sometimes I couldn't find anything because there were pages and pages of statements that had nothing to do with that").

The perceived value of the answers of Task b – Phase B ('exact' and 'ideal' answers) varied across the interviewees. Some considered 'exact' and 'ideal' answers less useful than documents and snippets in practice ("the systems cannot give the right answers, even when gathering the right snippets"), but others considered them more useful ("Personally, the 'exact' and 'ideal' answers are the types that I consider most useful"), perhaps more useful for clinical purposes and less for research (Expert 14). One expert (Expert 12) said that it was unclear how long an 'ideal' answer should be; we interpret this point as saying that it was not always clear what should or should not be included in a gold 'ideal' answer. Another expert (Expert 11) said that evaluating 'ideal' answers was difficult, because they were incoherent texts ("The only thing about the ideal answers, as a general comment, is that the answers were like a patchwork of sentences found inside the documents that sometimes didn't make sense. It wasn't a structured text."). Finally, we note that some interviewees did not discuss 'exact' and 'ideal' answers, which may be an indication that they were more interested in the results of Phase A.

Recommendations: Future challenges should definitely continue to require relevant documents and snippets per question. Relevance feedback or clustering could perhaps be added to the BIOASQ authoring tool (and future QA systems) to help the experts filter and organize more efficiently the possibly relevant documents and snippets that their queries retrieve. Structured snippets, meaning snippets accompanied by important, easy to read facts extracted from the snippets (e.g., for side-effects, dosage) might also help.[17] We also recommend continuing to require 'exact' and 'ideal' answers, given that at least some experts considered them particularly useful (see also Q11). However, it would be worth measuring the value of 'exact' and 'ideal' answers in different settings (e.g., clinical purposes vs. research), as opposed to having systems that return only relevant documents and snippets. The purpose of concepts should be clarified in future challenges; the guidelines that were provided to the experts in the second year of the challenge (Malakasiotis et al. (2013b)) made it clearer that they should be concepts closely related to the terms mentioned in the questions (e.g., synonyms, near hyponyms, near hypernyms, mostly to be used for query expansion), not parts of the answers, but this point may have to be made clearer. Before including tasks requiring statements (in effect, RDF triples) to be returned, future challenges should ensure that there are indeed relevant interesting statements to be returned per question in the designated repositories of structured information (the Linked Life Data repository, in the case of BIOASQ). One possibility might be to ask the experts to formulate questions for which the answers can be found partly in the designated repositories of structured information and partly in the designated repositories of documents, i.e., questions that require combined use of both types of repositories. It may not be easy, however, for the experts to understand exactly what information is available in repositories of structured information (e.g., LINKEDLIFEDATA datasets) and how repositories of this type can be queried for relevant information. Improvements are also needed in the BIOASQ services that retrieve possibly relevant RDF triples and convert them to pseudo-English statements. Future mechanisms should produce more statements that are relevant to the questions, and much fewer statements that are irrelevant. Finally, it would be worth investigating if the fluency of the pseudo-English statements can be improved, and if this also improves the perceived (by the experts) value of the statements.

[17] Google already returns structured snippets for some queries; see https://www.stonetemple.com/great-knowledge-box-showdown/.

"To me "documents" is the one that is more useful. For what we are working on in this Lab it is important to find the right information. So the most useful is the fact that the system gives us documents..."Snippets" also. The fact that a snippet defines exactly where the information that you need is inside the article is also very important... So it is like saving the effort to read the whole article and find what is really interesting for you inside it. So that's very important. Now, the other element, "concepts", that was about ontologies –its nature was kind of weird. On one hand I could search to find something that was related to something else or e.g. look for the genes related to a disease, but I could not add possible answers inside the concept. The concept had to be closed. Although it was a concept and it was included in the results provided by the system I could not include e.g. the information that this protein is related to what I am looking for. I had to pretend that I don't know this and wait to find it out from the three other fields...I believe that "documents" where more reliable. Although it depended on the kind of information I was looking for... For example when I was looking for the "cardasial syndrome", I had to read all the articles that came out from the system, all the 200 articles, to understand which of them contain the information. So I believe that some more work needs to be done for "documents" too. Now, if you were looking for information that is more close to bioinformatics or biomedicine you would need to read even more in order to filter the information returned. The rest of it, we were not using them much in our Lab." [**Experts 6 and 7**]

"They are useful, but it depends on the biological question you ask. [...] Since we are a Lab of bioinformatics and medical informatics, most of the things we were asking had mainly to do with methods, with medical informatics and medical biology. So we could not always have statements back" [**Expert 7**]

"I would say that, for me, the snippets are the most important type of answer. And then the documents are the second most important... As an idea, all of them sounded useful. In practice, however, we realised that the statements do not work that well. The idea behind them is very good, to get an answer which comprises of a subject, a verb and an object... Now, as for the concepts, they only bring back some very general information. They are the - so to say - most primitive kind of answer... And, of course, in this set of results there should also be an answer which could be readable. There is no point in unreadable answers. You should be able to make computers understand natural language but you should also be able to receive answers in natural language." [**Expert 2**]

"I'm sure you have already heard it from others as well, statements did not work well. The statements derived from repositories and databases in which you couldn't filter the results. For example, you wrote a question about if a protein is involved with rheumatoid arthritis and the statements returned results that mentioned that protein without giving the relevant answer or if some of the statements included the right answer it was difficult for you to trace it. So from this point of view, statements did not work. It has to do with the sources of information...Regarding the rest types of answers I would rank them having snippets on top, which derive from documents, and then documents as for me this is the base, all my work is based on documents. Concepts are useful most of the times but not always so I would rank them next." [**Expert 5**]

"Well, relevant documents or abstracts are useful, snippets are also useful because you must find among all this information what is useful, I'm not sure about relevant concepts... For example, this type of answer doesn't help me at all, but I don't know if these are useful to programs/software. So concepts are not useful to me, statements were not that accurate so I don't consider them useful. I don't think that statements would be useful in any case. Maybe in another discipline this type might be useful, but statements in Biology is not accurate, nothing is certain in Biology. Even what is definitely certain and you see it, it is that certain for some reasons, under other circumstances things would be different. Therefore, statements shouldn't exist in Biology, or they will be only few... I would like to add that the snippets returned so far by the system which we had to assess, only a third of them were relevant while the rest were either nonsense or long abstracts... Now regarding the exact' and ideal' answer...An ideal' answer is a sentence in which you can include varied information; you don't have to exactly say what is required. Exact' answer is a useful type. These are both useful! But with second thoughts, maybe exact' answer could be skipped? For example, the ideal' answer is not that exact' but in reality it can't be that exact' in any case...In the evaluation phase, the ideal' answers have made my life more difficult. Talking specifically about the exact' and ideal' answers, the exact' answers were often good but sometimes they were incomplete. There were one or two rare cases in which the system had found something more than the answers that I have thought to include in the exact' answer. Usually the exact' answers were fine, some incomplete but in general fine. Regarding now the ideal' answers the system had more flexibility to provide a better answer. The problem lies in the fact that I don't know how they searched for information and how they formulated the final answer. I had to read a number of staples of texts which included correct elements of information but also wrong or irrelevant irrelevant information. Therefore, I don't believe that the system has operated really well...So far, because I haven't completed the evaluation phase of the answers, I have assessed something more than half of them. So far, this is my opinion. That the answers which I received as ideal' answers were difficult to be assessed. I have marked them from 1 to 5 if they are readable etc, however it was not easy for me to decide how to mark them. Sometimes, due to the fact that it was so time consuming, I assessed them fast because I had to move on to the next questions. Sometimes I just had to have a break because I couldn't assess any more!... I believe that to a great extent the systems traced the correct/relevant abstracts. I think that the problem lies in the fact that the programs/softwares that these colleagues have developed don't know well the human language, so the problem is in another level. In this way, the systems cannot give the right answers, even when gathering the right snippets. I can imagine though that this is difficult. What could be done is if staples of snippets were gathered in order to form sentences, which again is not certain that these would be correct or at least correct for a person to read them. An expert might understand, in any case we search by keywords, we scan documents fast, see if they are interesting and then we decide if we will read it or not. These machines operate like this. What these machines do not offer to the user is that they don't give you a real answer which if compared to its question they could be matched." [**Expert 1**]

"I think that documents and snippets are always more important than statements and concepts to obtain an exact' answer." [**Expert 4**]

"Personally, the "exact" and "ideal" answers are the types that I consider most useful... The other types of answers have not helped me that much because, depending on the type of question, there may or may not be concepts or the answers might be more general, same applies for the statements... Yes, all of the answers that I have seen so far were reliable, some of them complete, some of them less complete, well-written or not as well-written, but they were reliable. Honestly, very few times I found totally irrelevant information. It varied on the quality... Snippets are useful, snippets would be the type following (the exact' and ideal' answer) because it marks if the information that has been returned to the user has been copied correctly. Following that, between concepts and statements I suppose I would rank concepts higher than statements." [**Expert 3**]

"The statements I haven't seen much value in there, the concepts to some point but the statements I haven't found anything that really helps me in there at all...Documents and snippets I think they are actually perfect. And the way you formulate the "ideal" and the "exact" answer is appropriate." [**Expert 8**]

"I imagine that concepts and documents are useful. We had a problem with statements in a recent challenge. It was difficult for us to define –What a computer does is different of what a human does. The statements were not making much sense to us humans. Because we were told that for computers they are necessary. But the rest is useful. "Documents" is what I would use as a tool...Snippets, as they are derived from documents, are basically what we want. Often the snippet itself is a phrase containing the information we were looking for. So, if snippets are actually good you don't need to read the whole document." [**Expert 9**]

"In the stage where the databases are, I believe that the statements are not mature yet in order to be included in –. There is an ambiguity. It is not clear to me if the statements and the concepts will be added in the answer when the final product is ready. Let's say it becomes commercial, a software that provides answers... It is normal to provide the biomedical user with the ideal answer and the exact answer, and to connect them with texts, sources. It will also probably provide them with the most correct, the most relevant snippets. My question is if the final users also get triples and statements or if those exist only for the assessment of the questions. Because what we did was an assessment platform. This is not clear to me. So the answer that I am going to give you is that those are mainly useful because of the importance they have for the objective answer, the ideal answer, the exact answer and of course as sources of text, either documents or snippets. The other two kinds of objects it is not clear to me if they are going to be useful for the one who asks the question... There are questions for which the ideal answer is enough. There are other questions for which it is enough, but what is important is inside the exact answer. And there are other cases where what you are looking for cannot be provided by the software, at least not in its current form. So you are going to use it as a means to get to the literature. I believe that all three – summary, textual answers and exact answers, lists, and bibliography are needed all together as well as in turn." [**Expert 10**]

"(Statements) they are to be thrown away. Not one in a thousand was useful. If they are improved, I don't know. It is clear that they are useless. The other ones are useful, the ideal answers, the snippets, of course they are useful. I think that the ideal answer should [be improved]. They could be very useful and they reach a satisfactory level researchers would save a lot of time. They [researchers] download articles, they read them and finally what they need to do is to write down a summary of what they have read, a resum of this information. This often takes many hours. The ideal answer is largely doing the same thing, providing a summary of the information. I believe that this would be very useful... The exact and the ideal answer were generally good. Of course there was some useless information included, but in general it was useful. The only thing about the ideal answers, as a general comment, is that the answers were like a patchwork of sentences found inside the documents that sometimes didn't make sense. It wasn't a structured text. The researcher gets this information from the documents, structures it in a different way and makes a text of his own, not a collage of sentences. This is a basic difference... the snippets were o.k., I mean sometimes they were relevant while sometimes not. It was the statements that were completely irrelevant." [**Expert 11**]

"The articles and the snippets are fine. The statements and the concepts –It depends very much on the question and how many there were. Sometimes they were reliable and sometimes they had nothing to do. It was very different between one question and another. So, sometimes I could use the concepts and the statements well, because they were quite clear and sometimes I couldn't find anything because there were pages and pages of statements that had nothing to do with that... Sometimes it is difficult. To me it was not clear how long the ideal answer should be. It could be very straight or sometimes it could be a summary. Sometimes it was difficult to me to write the ideal answer depending on the question. If it is "yes" or "no" that is clear but if it's like a summary this might be –It's your opinion, no? So it might be –And now, when we were evaluating the results of different programs I have seen that the ones that appeared in the top of the list often fitted much better with my answers, probably because they were much shorter and straight to the point. Other types of programs that write long paragraphs may repeat things. My idea was that the real answer should be as short as possible. I don't know if this was the idea or if this was my criterion when I used that." [**Expert 12**]

"The articles or the snippets are good. The statements sometimes were useful but not very useful, and the concepts were not good at all." [**Expert 13**]

"It's too much information I think, usually. It depends on your purpose so if you are trying to write a new grant and you want to see as much new information as possible, then probably yes. But for clinical purposes, for writing a paper and you want to find a reference to support whatever statement so then it can be too much information. So for those purposes I think, for clinical purposes I think that ideal' or exact' answers might be the most important. But for research purposes those are not so important, it's more important to find the data sources then documents probably would be the most important because then you'll need to cite those...From my experience in BIOASQ, statements were the least useful." [**Expert 14**]

Q11. Imagine a system that provides a machine-generated "exact" and an "ideal" answer for each English question, along with hyperlinks leading to the particular articles, snippets, concepts, and statements that it used to produce each "exact" and "ideal" answer. Would a system of this kind be useful in practice? How could it be made more useful (apart from reducing its errors)?

Overall, the interviewees agreed that a system of the kind described in Q11 would be useful, but they stressed the importance of providing links to the sources (e.g., articles, structured data) that would have been used to produce the 'exact' and 'ideal' answers. The 'exact' and 'ideal' answers might be more useful for clinical doctors (see also Q10), but again links to the information sources would be required, to ensure that the answers would be reliable, to obtain more details when necessary, and to fully understand the 'ideal' answers (the original sources can be easier to understand, especially when the 'ideal' answers look like a patchwork of extracted sentences, rather than a coherent text). The 'exact' and 'ideal' answers might also be useful as a first step towards understanding if a research question has already been answered in the literature, but otherwise researchers (as opposed to clinical doctors) might be more interested in the links to the information sources rather than the 'exact' and 'ideal' answers themselves. Providing bibliographic references for the information sources (ideally in a form directly usable in new publications) would also be important for researchers. Possible improvements to the system described in Q11 included: reporting additional relevant information that has not been explicitly requested (e.g., "yes this is related to glaucoma and by the way glaucoma is linked to this as well"), providing more than one 'ideal' answers (possibly generated by different systems), reporting relevant biological pathways (networks), and producing structured answers (e.g., trials with a particular drug a question is about, population used in the trials, safety profile of the drug).

Recommendations: In future challenges, the 'exact' and (especially) 'ideal' answers should be more tightly linked to the information sources that support them. For example, each sentence of the 'ideal' answer could be hyper-linked to a list of articles, snippets, concepts, and statements that support it; by contrast, in BIOASQ it is currently difficult to figure out exactly which articles, snippets, concepts, and statements of Task b – Phase A have been used to produce each part of the 'ideal' answer. Similarly, each 'exact' answer should be directly linked to the information sources that were used to derive it; these may not be all of the articles, snippets, concepts, and statements of Task b Phase A (e.g., some of them may support parts of the lengthier 'ideal' answer that are not included in the 'exact' answer). Future challenges could also require more structured 'ideal' answers for particular types of questions (e.g., with special sections or tables for dosages, side-effects, trials in questions about particular medicines). Future biomedical QA systems should also make it easier for researchers to obtain bibliographic references (e.g., BibTeX entries) for the information sources that support the 'exact' and 'ideal' answers; this could also be a subtask of future challenges.

> "Maybe some sort of more elaborated summary from the article and maybe to point to relationships that are not included in the original question... For example [if the question is] "is this related to glaucoma" to give information also about hypertension. I have not asked about hypertension, so this would not have appeared because it is in some other paragraph in the article. But if [the answer says] "yes this is related to glaucoma and by the way glaucoma is linked to this as well" it would actually be useful. New relationships that were not in the original query." ..A new abstract that would summarize everything that is relevant to my question"" [**Expert 13**]

"I'd get the hyperlinks... This is one step later though. I'd get the exact and ideal answers and then I would have to go through the hyperlinks. OK, that would be useful. But then again, as for me, that would be ok, especially in the initial stages of the research it would be like "ok, don't go looking for that, it has already been answered". But for a clinical researcher who would like direct results, it would also be useful... There is this issue there, because some medicines are still under clinical testing, and the average doctor would not know about them yet. He has to keep on searching. Some medicines which have been developed for a certain type of tumour are proven to be effective on another type of tumour instead. That means that this kind of thing would even save a patient's life... The reliability of the information is the most crucial, and then, yes, the speed of retrieval." [**Expert 2**]

"I'm not sure. Would that mean that for each question there would be only one answer? If yes, then that wouldn't be good. But if there was flexibility and the system returned to me 5 ideal' answers for example. Let me think about it... I believe that it would be useful yes. I believe that someone working in research wouldn't even read the answer, he would go straight to the links. It would be useful in regards to the fact that the answer has gathered relevant information, links and data underneath which the researcher can access." [**Expert 5**]

"I'm very happy with the search. Apart from some functional things that sometimes don't work, which is kind of normal especially in these systems, I'm extremely happy with the search. I mean if I could use this system for my own work it's something that I would use and I would recommend it to others as well." ... "It would help me a lot for my work if I wanted to have something super special I would check. But what I've been guessing out of the system is that various searches have been super enough for my work. For my everyday general summary that system does exactly what I need." [**Expert 8**]

"Yes, that could cover the need which I described before. As the answer cannot be given in a fine sentence of human language, we could be referred to the text which has been also written in human language but it's maybe more understandable...the snippets returned so far by the system which we had to asses, only a third of them were relevant while the rest were either nonsense or long abstracts." [**Expert 1**]

"Yes, that would be useful. Because I might need more details than the ones given in the "exact" answer or you might want to use the bibliographic reference for a publication that you are working on. We always write bibliographic references, so if these are incorporated in the "exact" or "ideal" answer we should be able not only to check them but also to document them to others that this is the source of our information. It's absolutely necessary not just desirable." [**Expert 3**]

"No, I prefer to know the essential information also, not only the ideal answer. Because I believe that no system, like BIOASQ, can replace human about the process of knowledge. So if you use BIOASQ it is a case that BIOASQ can produce an ideal' answer. But I think at the end you have to make your searching, you have to know the sources of the information for your knowledge process. I think that technology cannot replace human." [**Expert 4**]

"It would be really very useful. Having such a search engine which understands the field –It is not just googling something. This is something very specialized. You need to study a lot in order to find such information...If it is done correctly and it includes all those things that we say it will include, concepts, snippets, statements and all those, it will be even more useful because it will provide a broader and a full idea about what you are looking for. Now, if networks also can be included that would be even more useful." [**Experts 6 and 7**]

"Extremely useful. Especially if it was doing it in a sufficient extent... It could lead to databases, of those which already exist, depending on the question. There could be another functionality to lead to links containing structured data." [**Expert 10**]

"Of course, if the error is minimized and if the system can restructure the information in a way that it is unified, not a patchwork of sentences. The information needs to be structured and unified... This would be useful because it is something that the researcher is doing by him/herself. We take the information, we restructure it and we add the bibliography we used, in order to make a retrospective or a chapter in a book. This is what we do, we gather the information from works that have been published and we restructure it." [**Expert 11**]

"Yes that would be useful but I think you know that information should be very reliable. Because if you are doing clinical research and you're asking to give medication about a patient so you should trust that engine very well. Because you know if you give the wrong medication to a patient he might die. So, that system should be very reliable so I think... But usually you don't have time to read all these papers, all that stuff. So I think yes that would be good to have so ideal', exact' answers would be the most important and alongside that there should be a link to the most important clinical trials at least which have that particular medication, that particular symptom, the population of patients. So if you want to read what's safe or whatever drugs you just go and read. My issue with that would be if you trust a machine to do the thinking for you that can be hard and it's your problem and your patient's problem. That's not so much for research purposes, you can't kill somebody when doing research. Then it would be good for research stuff to have hyperlinks to original datasets first but for clinical purposes there should be a link at least, it depends on the disorder as well. You could have at least separate lists of clinical trials that test the medication in that or in different population, in that population there are overall trials of that drug, a safety profile etc. But sometimes you don't want to see all these cross sectional studies but that again depends on the disorder." [**Expert 14**]

Q12. Having participated in the preparation of the BIOASQ datasets and challenges, can you think of any modifications to the challenges that would make them more realistic or useful?

Regarding possible improvements to future challenges, two of the experts (Experts 5, 6) asked for more and better participating systems; presumably this might also make the manual assessment of the system responses more interesting for the experts. Two experts (Experts 9, 11) pointed out that different questions require different kinds of answers (e.g., different types of entities, more or less detailed answers). The expected (or desired) types of answers may be known to the experts submitting the questions, even if the actual questions are unknown to them. Hence, it might be worth providing more information about the expected kinds of answers to the participating systems. A more ambitious suggestion was to include questions that would require the systems to make predictions or, more generally, infer and report new information, as opposed to only retrieving information. Several other interesting ideas from the discussion of this question have been moved to the discussions of previous related questions.

Recommendations: Future challenges should aim to attract more participants. More information about the expected answers could be provided to the participating systems, along with the natural language questions. For example, the expected (or desired) length of the 'ideal' answers could be provided to the participants per question, as opposed to simply providing a maximum allowed length (the same for all the questions). The types of the expected 'exact' answers (e.g., disease, gene, symptom) could also be provided to the participants as concepts from designated ontologies. These concepts would differ from the relevant concepts of Phase A of BIOASQ Task 2b, in that they would be provided by the ex-

perts formulating the questions, instead of requiring the systems to guess them, and they would refer to the expected 'exact' answers, rather than being concepts closely related to terms of the questions. Adding questions requiring predictions or inferences might help attract participants working on inferencing methods and may also lead to more useful systems, though it would also increase the difficulty of the challenges.

"Maybe there is space for filtering that would distinguish between asking for general information like Wikipedia, and asking specific questions, like what's the temperature at which this enzyme reacts. That's very specific and I don't want to get needless information. There could be a variant, a graduation of the filtering. And also some sorting of what is double or triple" [**Expert 9**]

"Maybe in the future if it could link out with more databases like the ones we mentioned before, aiming to integrate all these different types of information This could possibly allow even more specialized questions to be phrased... (and by specialized questions I mean) to be able to link to prediction tools, for example could this microRNA control the behavior/expression of these mRNAs? Even if there is no proof in the literature, the system could link to prediction tools which will carry out this analysis resulting to a prediction that this might happen with x per cent possibility. This could be the next step as this would be a more complicated system." [**Expert 3**]

"I believe that if the participation in BIOASQ was bigger, from better systems or more developed systems, which means that there would be greater competition so as to reach golden answers that would have great possibility to be right then that would be really interesting." [**Expert 5**]

"I would say that it should be a little more open. To have more machines providing answers." [**Expert 6**]

"(...) through BIOASQ I realized that when searching for something, we search while partially knowing the answer. And this is very important. When you formulate a question you know roughly what the answer is going to be. You do not know the answer itself, but you know the type of answer. Let me provide an example. You have the question "what is the incidence of this disease on the population?". You know that you are looking for a number. So, when you are looking into a document that does not have any numbers, you know that you cannot get this information. You know that it is a number; it could be "one in a thousand" or "one in a million". You know that. You know the kind of answer, the category. This is very helpful when searching. But the machine does not know it. If it traces the words "incidence" and "ischemic episode" in a document it will tell you that this document is relevant. But it isn't. This might be useful for those developing the systems. For example you might look at what are the manifestations of the ischemic episode, what are the symptoms. You know that by "symptoms" you expect something specific. You have in mind a list of specific things that are the symptoms, and it is them that you are looking for in the document. So, you have broadly in mind what you are looking for. For example you ask what the most known mutations of a gene are that cause a specific disease. You know that you are looking for mutations and mutations are expressed in the form of codes in the documents, for example methyonine 232. So you look into the document to find those codes. If those codes are not there it means that this information is not there." [**Expert 11**]

3.3.4 Summary of recommendations

The following table summarizes our recommendations for future biomedical QA challenges and systems, based on the responses of the experts to questions Q2–Q12; there were no particular recommendations from questions Q1 and Q2.

Questions	Recommendations for future challenges and/or systems
Q2	Specify types (e.g., research articles, systematic reviews, clinical trial records, patents) and origin (e.g., PUBMED, trusted sites, Web) of documents to be searched.
Q3	Use more designated repositories of structured information for concepts and triples, or require the participating systems to find relevant repositories of structured information per question.
Q4	Continue to aim at generic biomedical QA systems, rather than systems targeting particular types of information (e.g., gene interactions only).
Q5	Consider assigning questions to groups of experts, possibly with complementary expertise. Extend the BIOASQ social network to allow experts to criticize or complement answers produced by systems. Move towards hybrid QA systems, combining answers provided by systems and humans. Consider a question to question matching subtask (e.g., for FAQs).
Q6	Investigate if systems that accept natural language questions actually manage to produce better answers than systems that accept keyword queries. Use previous searches, articles downloaded or shared, journal subscriptions etc. to construct user models of the experts. Address full-content access restrictions of journals.
Q7	Support filtering or ranking criteria for author, reputation, affiliation, journal name, impact factor, citations, article type, recency etc. when displaying retrieved articles in the BIOASQ authoring tool and future QA systems.
Q8	Research how biomedical experts could better organize and store retrieved relevant information and sources. Develop tools (possibly based on the BIOASQ authoring and assessment tools) that would help biomedical experts organize and store relevant information and sources per natural language question. Consider retrieving relevant images, tables, equations etc.
Q9	Use English questions and keyword queries as separate or joint inputs. Consider follow-up questions, clarification dialogues, possibly also spoken dialogues. Consider merging factoid and list questions. Consider adding list questions requiring positive and negative lists, or lists of steps. Consider adding questions requiring 'insufficient information available' or 'controversial information found' as answers. Measure the time needed to formulate and process natural language questions vs. keyword queries (e.g., in emergencies).
Q10	Continue to require relevant documents, snippets, 'exact' and 'ideal' answers per question, but measure the value of 'exact' and 'ideal' answers as opposed to having only relevant documents and snippets in different settings (e.g., clinical vs. research purposes). Consider adding relevance feedback and clustering to the authoring tool for documents and snippets, and to future QA systems. Consider structured snippets. Clarify the purpose of concepts. Author questions for which there is relevant important information in repositories of structured information. Improve the BIOASQ services that retrieve possibly relevant 'statements'. Consider improving the fluency of 'statements'.
Q11	Link more tightly the 'exact' and 'ideal' answers to supporting articles, snippets, concepts, and statements. Require bibliographic entries for the supporting sources. Consider requiring more structured 'ideal' answers for particular types of questions.
Q12	Attract more participants. Provide to the participants more information about the expected answers (e.g., types of expected 'exact' answers, length of 'ideal' answer). Consider questions requiring predictions or inference.

3.4 Long-term future: Porting to other domains

BIOASQ is currently focused on the biomedical domain. But can we expand the know how of BIOASQ to other domains? In order to be able to answer this question, we must first consider the two different axes towards which BIOASQ could be expanded, namely, the domain of the data and the end users (table 3.21).

	BioMedical Domain	**Other Domain**
Expert	BIOASQ	Sciences (e.g. material science)
Non-expert	MedWhat++	HARD

Table 3.21: Possible extensions

Different Domain. In order to be able to run challenges like BIOASQ to a new domain, the latter should satisfy specific criteria concerning the available resources. In particular, there should exist the following types of data on this domain:

- document repository (like Pubmed)
- knowledge bases and ontologies, in order to annotate documents with terms and concepts.

Given the above resources, another requirement is that it should be *possible* and *useful* to form natural questions on this domain, which could be answered based on the above resources. Finally, a team of experts on the particular domain who could create annotated data, should be available. A specialized domain like *sciences* (e.g. *material sciences*) could fit well on this profile. For example, the arxiv[18] could be consider as the science equivalent to PubMed. In the same vein, a different domain could be *digital humanities*, *legal texts* or *ecomonics*. In the following subsections (3.4.1,3.4.2 and 3.4.3), the available resources and infrastructures are discussed, as well as ideas on how a challenge could be adapted on the above domains.

Different end-users. Concerning the second axe, in the case of BIOASQ, end users are biomedical experts, and the whole challenge has been built on this direction. If we would like to focus on a different audience, like non-experts, we could consider a system like MedWhat[19]. The latter tries to answers medical questions of simple users. BIOASQ could be expanded to support challenges on this direction. To achieve this goal, annotated data would be needed in order to train the existing system on such kind of questions. In addition, we should consider that non-experts can ask different kind of questions, which may not be able to be answered by the current resources, so an expansion of the used resources should be also considered.

If we consider the possibility of moving towards both directions (new domain and non-expert users), that would radically change the initial focus of BIOASQ. An example on this direction would be a natural language search system, like Yahoo answers. The latter is actually out of scope of BIOASQ. One of the reasons is that in such systems, there are questions that are answered by humans and the resources for these answers are not available.

[18] http://arxiv.org/
[19] http://www.medwhat.com

3.4.1 Legal documents

In the domain of law, there are several available resources that could be considered as equilevant to PubMed and MeSH.

In particular, EUR-Lex[20] provides direct free access to European Union law (e.g. directives, regulations, decisions), international agreements, preparatory acts (e.g. legislative proposals, reports, white papers), official EU journal etc. It is a multilingual resource, supporting 24 languages. It offers extensive search facilities, such as keyword search and/or search via EuroVoc descriptors.

On the other site, Eurovoc[21], which can be considered as the MeSH equivalent, is a multilingual, multidisciplinary thesaurus covering the activities of the EU, the European Parliament (EP), parts of the European Commission and many national and regional parliaments or other organisations in the European Union. EuroVoc consists of 6797 descriptors (also referred to as classes or categories) which are organised into a hierarchical structure of up to 8 levels. Human indexing professionals are typically librarians or linguists, employed or freelance, with a developed conceptual understanding of the themes dealt with in the thesaurus.

Based on EuroVoc, the European Commission's Joint Research Centre (JRC)[22] has developed the JRC EuroVoc Indexer[23] (JEX) aiming to provide end users with a tool for the automatic annotation of documents with descriptors from the EuroVoc thesaurus. The automatic classification is less accurate than that carried out by human professionals, but it has the advantage that it is extremely fast and perfectly consistent. It can be considered as the MTI equivalent of BIOASQ.

Tasks equivalent. If we would like to apply a challenge like BIOASQ on legal documents, and given the above resources, we could consider the following taks equivalents:

- **Task a:** *Assign EuroVoc descriptors to EUR-lex documents.* Distribute documents before human curators have assigned descriptors, and use human descriptors as gold answers, as in BIOASQ. The above Requires cooperation/interest of EU Parliament, Commission and JRC.
- **Task b - Phase A:** *Retrieve relevant items.* Given a natural language question (we could also consider multiple languages), return relevant EuroVoc descriptors, documents and snippets. Variant: systems also given keyword queries, EuroVoc descriptors. Possible addition: resolve the (often very complex) references.
- **Task b – Phase B:** *Formulate 'exact' and 'ideal' answers.* Given a NL question, relevant documents, snippets, and EuroVoc descriptors, return an 'exact' and 'ideal' answer (we could also consider multiple languages).

A team of legal experts is needed in order to help with the initial user studies to fine-tune task requirements. Also, the team wil be responsible for authoring and evaluating questions/answers of Task b.

3.4.2 Economics/ social sciences

Other possible domains for porting BIOASQ are the domains of economics and/or social sciences. In both domains there are resources that could be considered equivalent of PubMed. For example, RePEc

[20] http://eur-lex.europa.eu/
[21] http://eurovoc.europa.eu/drupal/
[22] https://ec.europa.eu/jrc/
[23] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer

(Research Papers in Economics)[24] is a collaborative effort of hundreds of volunteers in 82 countries to enhance the dissemination of research in Economics and related sciences. It consists of a decentralized bibliographic database of working papers, journal articles, books, books chapters and software components, all maintained by volunteers. It aggregates 1,600 archives. The resource includes abstracts, often also downloadable full text for approximately 1.4 million documents from 1,800 journals and 3,800 working paper series.

For the domain of social sciences the SSRN is available. The latter includes approximately 570K abstracts and 470K full-text documents.

Both resources provide keyword search, as well as JEL codes. The JEL classification codes can be consider as the equivalent of MeSH. It includes 20 top-level codes, up to 3 levels of sub-codes, with approximately 1,200 leaves. It is used in AEA's EconLit to index more than 120 years of Economics literature.

Tasks equivalent. Given the above resources, the tasks of the challenge could be modified as follows:

- **Task a:** *Assign JEL codes to RePEc/SSRN documents*. A cooperation with AEA Econ/Lit curators is possibly needed, as in the case of BIOASQ.
- **Task b - Phase A:** *Retrieve relevant items*. Given an NL question, return relevant JEL codes, documents and snippets.
- **Task b - Phase B:** *Formulate 'exact' and 'ideal' answers*. Given an NL question, relevant documents, snippets and JEL codes, return an 'exact' and 'ideal' answer.

As in the case with legal texts, a team of economic/Sociologist experts group need to be formulated, in order to perform the initial user studies to fine-tune task requirements, as well as author and evaluate the questions/answers of Task b.

3.4.3 Digital humanities

Digital Humanities (DH) are concerned with the intersection of computing and the various humanities disciplines. As such, DH enable new kinds of research in the humanities, but also in computer science. Humanities-related collections of digital resources, textual as well as in other media and datasets, are being developed at an increasing rate. However, no document repositories of universal coverage and acceptance, such as PubMed in biomedicine, have been established yet.

On the other hand, resources like The European Library and Europeana operate as union catalogues to libraries and cultural content across Europe. Although these do not provide access to the content itself, they offer rich metadata based on a widely applicable data model (the Europeana Data Model, EDM) and they expose these resources to exploitation through an API. The key-value structures supported by Europeana are gaining acceptance through use, however there is no vocabulary normalization activity in the humanities at the scale encountered in biomedicine. This is naturally related to the absence of an intense annotation activity of scientific publications, as is the case in biomedicine. Thus, in terms of the availability of a suitable testbed on which to carry out competitions like BIOASQ, the field of digital humanities lags behind. However, the decisive factors are (a) the trend in developing digital collections and knowledge organization systems and (b) the information practices of researchers in the humanities.

[24] www.repec.org

The trend is clearly to develop widely accepted knowledge organization systems and we expect to witness accelerated progress in this direction due not only to Europeana and The European Library, but also to the emergence of research infrastructures, such as the Digital Research Infrastructure for the Arts and Humanities (DARIAH-EU[25]), the ARIADNE network on archaeology[26]), the European Holocaust Research Infrastructure (EHRI[27]), and the Research Data Alliance (RDA[28]).

Humanities information practices are attracting continued interest due to the proliferation of the digital, e.g., work by the Oxford Internet Institute (Humanities Information Practices[29]), or the studies on scholarly practices undertaken by the Digital Curation Unit in the projects DARIAH, EHRI, Europeana Cloud and DYAS/DARIAH-GR. As a result, research infrastructures need to address the particular requirements and information practices of humanities scholars. In what concerns information seeking, knowledge organization systems, annotation services and information retrieval services should address the tendency towards associative search observed in the humanities.

Tasks equivalent. A challenge on large scale indexing and question answering in the humanities could be run using some emerging wide-scope access platform, such as The European Library or Europeana in conjunction with knowledge organization resources (e.g. taxonomies) developed in infrastructure projects like DARIAH or ARIADNE, and adopting the structures of BIOASQ tasks A and B so as to cater for the information seeking biases of the humanists. Such a challenge could equally serve as a booster of the normalization of knowledge organization structures, as a competition proper.

[25] https://dariah.eu
[26] http://www.ariadne-infrastructure.eu/
[27] http://www.ehri-project.eu/
[28] https://rd-alliance.org/
[29] http://www.oii.ox.ac.uk/research/projects/?id=58

A

Questionnaires used

In the following pages we include the questionnaires used for the qualitative evaluation of the first cycle of BIOASQ.

BioASQ evaluation form for challenge participants

Short survey on the quality of the challenge, from the participant point of view. For more information about BioASQ, please visit http://bioasq.org.

* Required

Learning about BioASQ

1. **How did you hear about BioASQ?** *
   Mark only one oval.

   - ◯ Web site
   - ◯ Mailing list
   - ◯ Personal contact
   - ◯ Twitter
   - ◯ LinkedIn
   - ◯ Other: .....

2. **How easy was it to understand the tasks overall?** *
   Mark only one oval.

	1	2	3	4	5	
Very difficult	◯	◯	◯	◯	◯	Very easy

Registration

3. **Please rate the registration process according to the following.** *
   Mark only one oval per row.

	Very bad	Bad	Fair	Good	Very good
Registration guidelines	◯	◯	◯	◯	◯
Registration platform	◯	◯	◯	◯	◯

Task 2A

Please answer the following questions only if you participated in Task 2A.

4. **Please rate Task 2A according to the following.**
   Mark only one oval per row.

	Very bad	Bad	Fair	Good	Very good
Guidelines					
Data format					
Data quality					
Downloading and uploading procedures					
Scheduling (e.g., release of test sets, time to submit results, etc.)					
Evaluation methods					
Notifications (e.g., release of test sets, evaluation results, etc.)					
Technical support					
Overall impression					

5. **Additional comments**
   Anything not covered by the questions above.

.....

.....

.....

.....

.....

Task 2B Phase A

Please answer the following questions only if you participated in Task 2B Phase A.

6. **Please rate Task 2B Phase A according to the following.**
   Mark only one oval per row.

	Very bad	Bad	Fair	Good	Very good
Guidelines					
Data format					
Data quality					
Number of questions					
Diversity of questions					
Difficulty of questions (how difficult were the questions)					
Downloading and uploading procedures					
Other supporting software and web-services (e.g., searching for concepts, articles, etc.)					
Scheduling (e.g., release of test sets, time to submit results, etc.)					
Evaluation methods					
Notifications (e.g., release of test sets, evaluation results, etc.)					
Technical support					
Overall impression					

7. **Additional comments**
   Anything not covered by the questions above.

.....

.....

.....

.....

.....

Task 2B Phase B

Please answer the following questions only if you participated in Task 2B Phase B.

8. **Please rate Task 2B Phase B according to the following.**
Mark only one oval per row.

	Very bad	Bad	Fair	Good	Very good
Guidelines					
Data format					
Data quality					
Number of questions					
Diversity of questions					
Difficulty of questions (how difficult were the questions)					
Downloading and uploading procedures					
Scheduling (e.g., release of test sets, time to submit results, etc.)					
Evaluation methods					
Notifications (e.g., release of test sets, evaluation results, etc.)					
Technical support					
Overall impression					

9. **Additional comments**
Anything not covered by the questions above.

.....

.....

.....

.....

.....

Overall impression

10. **What is your overall impression of BioASQ** *
Mark only one oval.

	1	2	3	4	5	
Very bad						Very good

11. **Will you participate in the next year's BioASQ Challenge?** *
Mark only one oval.

- Yes
- Maybe
- No

12. **Will you recommend the BioASQ Challenge to others?** *
Mark only one oval.

- ◯ Yes
- ◯ Maybe
- ◯ No

Contact details

13. **Name** *

.....

14. **Email** *

.....

15. **May we contact you in due course?** *
Mark only one oval.

- ◯ Yes
- ◯ No

BioASQ Biomedical experts questionnaire

Please rate your experience as a member of the BioASQ team of biomedical experts

* Required

Annotation tool

Please rate the following functionalities of the annotation tool on a scale of 1 - 5

1. Very poor: I was unable to use it.
2. Poor: Needs substantial improvements.
3. Fair: I could use it but still needs some improvements
4. Good: Almost everything worked smoothly. With a few improvements it would be excellent
5. Very good: Everything worked smoothly. No improvement is needed.

1. **Registration** *
   Mark only one oval.

	1	2	3	4	5	
Very poor	◯	◯	◯	◯	◯	Very good

2. **Question creation** *
   Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

3. **Concepts search** *
   Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

4. **Documents search** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

5. **Statements search** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

6. **Snippets annotation** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

7. **Exact answer creation** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

8. **Ideal answer creation** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

9. **Saving your work** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

10. **Overall impression** *
Mark only one oval.

	1	2	3	4	5	
Very poor	◯	◯	◯	◯	◯	Very good

11. **What did you like most in the tool?**

12. **What did you like less in the tool?**

13. **Would you use this tool again in the future?** *
Mark only one oval.

- Yes
- Yes, if some improvements are made
- No

14. Please give a short justification

15. **Would you recommend this tool to others?** *
Mark only one oval.

- Yes
- No

16. Please give a short justification

17. **Do you think you could use the tool in your own work (e.g., to organize a search)?** *
Mark only one oval.

- ◯ Yes
- ◯ Yes, if some improvements are made
- ◯ No

18. Please give a short justification

.....

.....

.....

.....

.....

19. **Do the changes of the 2nd version of the tool solve the issues of the 1st version?**
Mark only one oval.

	1	2	3	4	5	
None of the issues were solved	◯	◯	◯	◯	◯	All of the issues were solved

20. Please give a short justification (e.g., issues solved, issues not solved, new issues raised, etc.)

.....

.....

.....

.....

.....

Assessment Tool

Please rate the following functionalities of the assessment tool on a scale of 1 - 5

1. Very bad: I was unable to use it.
2. Bad: Needs substantial improvements.
3. Fair: I could use it but still needs some improvements
4. Good: Almost everything worked smoothly. With a few improvements it would be excellent
5. Very good: Everything worked smoothly. No improvement is needed.

21. **Ideal answers assessment** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

22. **Exact answers assessment** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

23. **Snippets assessment** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

24. **Concepts assessment** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

25. **Documents assessment** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

26. **Statements assessment** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

27. **Saving your work** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

28. **Overall impression** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

29. **What did you like most in the tool?**

30. **What did you like less in the tool?**

31. **Would you use this tool again in the future?** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

32. Please give a short justification

33. **Would you recommend this tool to others?** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

34. Please give a short justification

35. **Do the changes of the 2nd version of the tool solve the issues of the 1st version?**
Mark only one oval.

	1	2	3	4	5	
None of the issues were solved	⬭	⬭	⬭	⬭	⬭	All of the issues were solved

36. Please give a short justification (e.g., issues solved, issues not solved, new issues raised, etc.)

.....

.....

.....

.....

.....

Overall impression

37. **Please rate your interaction with the BioASQ team.** *
Mark only one oval.

	1	2	3	4	5	
Very bad	⬭	⬭	⬭	⬭	⬭	Very good

38. **Please add ideas/comments of possible ways of improving BioASQ.**

.....

.....

.....

.....

.....

Contact details

39. **Name** *

.....

40. **Email** *

.....

Bibliography

I. Androutsopoulos, G. Lampouras, and D. Galanis. Generating natural language descriptions from owl ontologies: the naturalowl system. *Journal of Artificial Intelligence Research (JAIR)*, 48:671–715, 2013. URL http://dblp.uni-trier.de/db/journals/jair/jair48.html#AndroutsopoulosLG13.

G. Balikas, I. Partalas, N. Baskiotis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, T. Artieres, E. Gaussier, and P. Gallinari. Evaluation Framework Specification – 2nd version. BioASQ Deliverable D4.5, 2013a.

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

G. Balikas, I. Partalas, N. Baskiotis, T. Artieres, E. Gaussier, and P. Gallinari. Evaluation infrastructure software for the challenges 2nd version. BioASQ Deliverable D4.7, 2014.

A. Benardou, P. Constantopoulos, C. Dallas, and D. Gavrilis. Understanding the information requirements of arts and humanities scholarship. *International Journal of Digital Curation*, 5(1), 2010.

A. Benardou, P. Constantopoulos, and C. Dallas. An approach to analyzing working practices of research communities in the humanities. *International Journal of Humanities and Arts Computing*, 7(1-2): 105–127, 2013.

N. Heino. Annotation Tool, 2nd Version. BioASQ Deliverable D3.6, 2013.

N. Heino and A.-C. Ngonga Ngomo. Social Network. BioASQ Deliverable D3.3, 2013.

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

P. Malakasiotis, I. Androutsopoulos, Y. Almirantis, D. Polychronopoulos, and I. Pavlopoulos. Tutorials and Guidelines. BioASQ Deliverable D3.7, 2013b.

R. McCreadie, C. Macdonald, I. Ounis, and J. Brassey. A study of personalised medical literature search. In *Proceedings of CLEF 2014*, pages 74–85, Sheffield, UK, 2014. Springer.

G. Paliouras. Project Periodic Report. BioASQ Deliverable D1.4, 2014.

I. Partalas, G. Balikas, N. Baskiotis, D. Polychronopoulos, Y. Almirantis, E. Gaussier, T. Artieres, and P. Gallinari. Pre-processed benchmark set 1. BioASQ Deliverable D3.5 and D4.2, 2014a.

I. Partalas, G. Balikas, N. Baskiotis, D. Polychronopoulos, Y. Almirantis, E. Gaussier, T. Artieres, and P. Gallinari. Pre-processed benchmark set 2. BioASQ Deliverable D3.8 and D4.6, 2014b.