W478–W483 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 3 May 2017*
doi: 10.1093/nar/gkx363

Olelo: a web application for intuitive exploration of biomedical literature

Milena Kraus*, **Julian Niedermeier, Marcel Jankrift, Sören Tietböhl, Toni Stachewicz, Hendrik Folkerts, Matthias Uflacker and Mariana Neves**

Department of Enterprise Platforms and Integration Concepts, Hasso Plattner Institute, August-Bebel-Strasse 88, Potsdam 14482, Germany

Received January 31, 2017; Revised April 12, 2017; Editorial Decision April 20, 2017; Accepted April 25, 2017

ABSTRACT

Researchers usually query the large biomedical literature in PubMed via keywords, logical operators and filters, none of which is very intuitive. Question answering systems are an alternative to keyword searches. They allow questions in natural language as input and results reflect the given type of question, such as short answers and summaries. Few of those systems are available online but they experience drawbacks in terms of long response times and they support a limited amount of question and result types. Additionally, user interfaces are usually restricted to only displaying the retrieved information. For our Olelo web application, we combined biomedical literature and terminologies in a fast in-memory database to enable real-time responses to researchers' queries. Further, we extended the built-in natural language processing features of the database with question answering and summarization procedures. Combined with a new explorative approach of document filtering and a clean user interface, Olelo enables a fast and intelligent search through the ever-growing biomedical literature. Olelo is available at http://www.hpi.de/plattner/olelo.

INTRODUCTION

Researchers all over the world regularly access the MEDLINE/PubMed database, which currently contains over 20 million scientific biomedical publications. Users usually explore this knowledge through simple keyword searches. Although advanced search options are available, these require an exact search target and proper search terms as well as knowledge on how to use the interface. Further, current search engines do not leverage the information contained in abstracts, full texts, medical vocabularies and ontologies to their full extent. Additionally, their user interfaces frequently restrict explorative search as a new browser tab needs to be opened for every relevant search result.

Lu *et al.* (1) reviewed a collection of web tools that differ in their search options to the traditional PubMed approach. Among others, the author came to the following observations: (i) most of the engines provide a list of titles and authors of the relevant documents; (ii) in systems that perform a clustering or ranking of documents, the list of documents can usually be expanded; (iii) only few approaches provide other result sets, such as tables or graphs; (iv) improved ranking and usability seem to be popular driving forces for new systems.

Question answering (QA) systems offer a user-friendly alternative to plain keyword searches and have proven to provide exact answers in the biomedical context (2,3). In particular, QA enables three advantages: (i) queries can be posed using natural language instead of keywords; (ii) results are generated according to what has been specifically requested, be it a single answer or a short summary; (iii) answers are usually based on the integration of textual documents and a variety of knowledge sources (3).

Bauer *et al.* (4) surveyed the only three available QA systems for biomedicine, namely askHERMES (5), HONQA (6) and EAGLi (7). The authors identified drawbacks in usability in terms of response time, obstacles in the web interface, as well as restrictions in the types of questions that these tools are able to process.

The Olelo Web application is derived from our previously established QA system (8) that was one of the winners in the two previous editions of the BioASQ challenges (www.bioasq.org/participate/winners). It is the entry point to an explorative search through the biomedical literature. An in-memory database (IMDB) holds all data in main memory to enable real-time exploration of the documents. Further, the IMDB provides useful built-in features for natural language processing (NLP), which we extended with advanced algorithms, such as question understanding and multi-document summarization (9).

Pang *et al.* (10) discuss desirable design principles for Web applications, in order to facilitate the explorative search in

[*] To whom correspondence should be addressed. Tel: +49 331 5509 1366; Fax: +49 5509 579; Email: Milena.Kraus@hpi.de

health-related data (Better Health Explorer). Among others, principles for design are: (i) to support information exploration, (ii) to minimize keyword search, (iii) to offer preview and summary and (iv) to leave clues for the journey. The design principles in Olelo mirror those in the Better Health Explorer prototype.

In the remainder of the article, we present details on the overall architecture of the Olelo web application including the integrated data sources, the IMDB technology, NLP algorithms and the technology stack used for our front and back end. In the 'Results' section, we describe key features and usage of Olelo in general and in specific use cases. We then discuss Olelo's contribution for an intuitive exploration of the biomedical literature.

MATERIALS AND METHODS

In this section, we describe the three main components of Olelo in detail as well as the interplay between them. Figure 1 depicts the simplified architecture behind our application. A detailed explanation of our methods is given in the Supplementary Material.

Data sources and integration

We downloaded the totality of Medline documents and PubMed Central Open Access (PMC OA) full texts from the NCBI's FTP servers. We imported approximately 16 million abstracts, for which the titles and authors are available, as well as 1.3 million full-text documents. Additionally, we integrated the Medical Subject Headings (MeSH), a thesaurus for medical terminology, and the Unified Medical Language System (UMLS), a large collection of biomedical dictionaries.

MeSH (www.nlm.nih.gov/mesh) contains 27 883 unique medical headings and their synonyms, which represent common key concepts found in medical literature. Each heading can be identified by its unique ID and contains a short definition. MeSH is organized in a tree structure. Nine high-level nodes of the tree (TREE_ID A, B, C, E01, G05.360.340.024.340, D12.776, Z, E02, C23.888) serve as key concepts in Olelo and are used for grouping of retrieved documents.

From UMLS (www.nlm.nih.gov/research/umls), we utilize the Metathesaurus and the semantic types from the Semantic Network. UMLS terms are identified by concept unique identifiers (CUI). All data sources are updated on a regular basis via a custom update script.

In-memory database

IMDB technology enables fast access of data directly from main memory. This differs from most traditional approaches in which data are processed from files with significantly higher access time. IMDB technology provides a wide range of built-in features, such as multi-core and parallelization strategies, columnar data layout and lightweight compression that can be exploited to speed up calculation-intensive processing steps. We use the SAP S/4 HANA (SPS11) database which is hosted on a machine with 120 cores and two terabytes of main memory. Our NLP algorithms consist of either built-in features of the database or custom structured query language (SQL) procedures that we implemented into the database. More specifically, we use the following built-in NLP features from the database: stemming, tokenization, part-of-speech tagging and dictionary-based named-entity recognition (NER). To support NER, we compiled custom dictionaries based on concepts from MeSH and UMLS.

Question answering

QA systems are usually composed of three main components (2,3): question processing, document/passage retrieval and answer processing. In Olelo, the question processing module is based on a system previously described in (8). Currently, it supports three question types, depending on the expected answer: (i) a definition of an MeSH term, (ii) a short list of facts (factoid) or (iii) a short summary. The expected question type is extracted through regular expressions. If neither (i) nor (ii) was identified, a short summary is generated. If the system detects a factoid question, it proceeds to identify the headword (or key concept) of the question, e.g., that diseases or treatments should be returned. All tokens (words) of the question are mapped onto the MeSH tree ID, UMLS CUI or their word stem. In the next step, Olelo uses the tokens and the matched terms to formulate a query to the database. The query also includes all synonyms and linguistic modifications of all tokens. The system retrieves abstracts and ranks them according to the occurrence and importance of the searched tokens. Finally, an answer is returned to the user depending on the type of the question.

Summary generation

Summaries of documents aim to provide a short paragraph which includes the most relevant information from the original documents. Olelo generates automatic summaries of retrieved documents through the incorporation of an extractive procedure, which is described in more detail in (9). Documents used for summary generation are either user-specified or are based on the most relevant retrieved documents (default value is 20). Relevancy is based on usual information retrieval techniques, i.e. frequency-inverse document frequency (TF-IDF) (11). TF-IDF is a numerical value that indicates the importance of a word in a document collection, and consequently, it measures how relevant is each document for a given keyword. The system calculates the similarity between each of the two sentences from the retrieved documents and constructs a graph in which the nodes are the sentences and the vertices are the similarity values between each pair. Similar sentences are then examined to find the sentences that best summarize them all. The final summary is created through multiple iterations of concatenating the best sentences, while avoiding redundancy, until the specified size of the summary is reached (default value is five sentences).

Back end

The back end serves static resources in the form of HTML pages, images and JavaScript, and offers a Representational

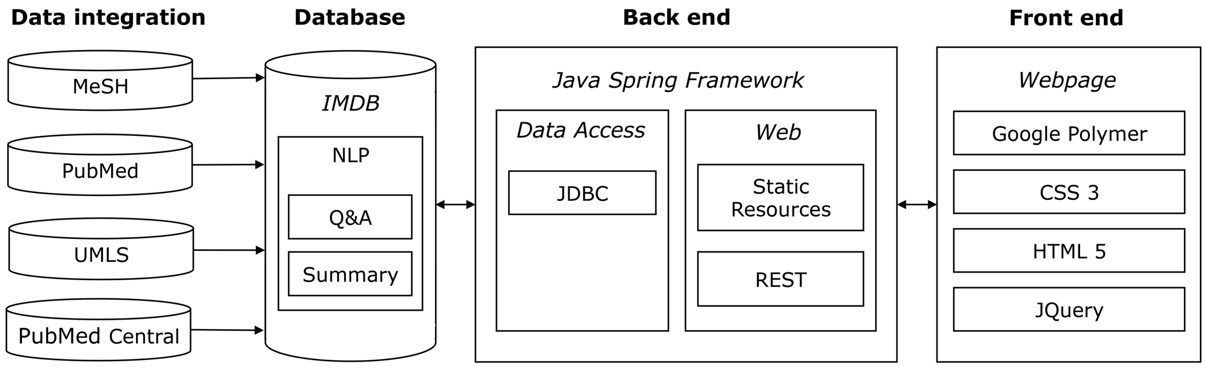


Figure 1. Architecture of the Olelo Web application. It is composed of three main components: data and database, back end and front end.

State Transfer interface to allow the front end to query information from the database procedures. The back end is built purely in Java and uses the open-source Spring application framework as this allows the back end to run on any platform while being easily extendable. Data access is provided by a Java Database Connectivity interface.

Front end

We implemented the front end mainly using Google Polymer and Google Material Design, in addition to the most recent web standards such as Cascading Style Sheets 3 and HTML 5. Polymer allows web component templating to modularize and inherit individual components, which help create a consistent and maintainable user experience. Material Design is a design language that focuses on minimalistic and clean interfaces and provides many pre-built elements such as cards, buttons and menus that can be used across different Polymer components. Several Javascript libraries, such as JQuery, were used especially for the drag-and-drop function.

RESULTS

In this section, we describe the key features in Olelo and illustrate its functionality in the scope of three uses cases.

Key features

Figure 2 gives an overview on the most relevant features and functions of Olelo. They are divided in high-level searching and filtering and a deep dive into reading and summarizing the retrieved abstracts/documents. High-level searching and filtering are displayed in case of factoid or definition questions or if a definition is requested for a term highlighted in a document or a summary. The search and filter module allows the discovery of relations between two or more concepts. We facilitate this relation detection by allowing the user to navigate along the path of biomedical concepts as they co-occur in publications. Clicking upon a key term results in a definition of the term and an automatic filtering of all documents. A definition question can also be an entry point of the system and will also lead to a list of key concepts, as well as their corresponding documents. By clicking on these key concepts, the system filters out the corresponding set of documents and displays a related list of key terms.

At any time, users can open the collection of corresponding documents in a new card to take a deep dive into the scientific literature. Preview of a publication allows a glance at the title, authors and abstract. Previews can be extended to show the scrollable complete abstract or the full text, if available. At this point, it is also possible to check the primary source of the document in PubMed. For a certain document, the user can ask for similar publications or, when combined with more documents, create a custom collection. Collections can be summarized to extract the most relevant information. For any summary, the user can navigate through the corresponding collection of documents, i.e. the ones used to generate the summary. By clicking on a sentence in a summary, and thus highlighting it, the corresponding document will be opened. A click on a highlighted MeSH term leads to a new term-based search and filter loop starting with a definition of the term.

Interactive user interface

The tabs or ‘breadcrumbs’ keep track of the search flow and serve as an easy way of browsing back and forth. Every detected biomedical term is highlighted and can be selected to open the definition of that term without the need for a new tab search. Furthermore, NER-derived term highlighting allows the user to get a quick overview of the terminologies, which can help in determining the importance of the publication. All cards can be dragged and dropped and they can be collapsed to save space on the screen. Expanded abstracts/full texts combine into a collection if their cards are placed onto each other. Flicking through documents, e.g. on a tablet or with the mouse, enables the user to get a fast overview of all retrieved documents. A new question can be posed at any time without losing previous searches. Documents of a collection can be exported in the BioC format (12), a standard XML format in the biomedical NLP community. Finally, response times of Olelo are in the range of milliseconds to seconds.

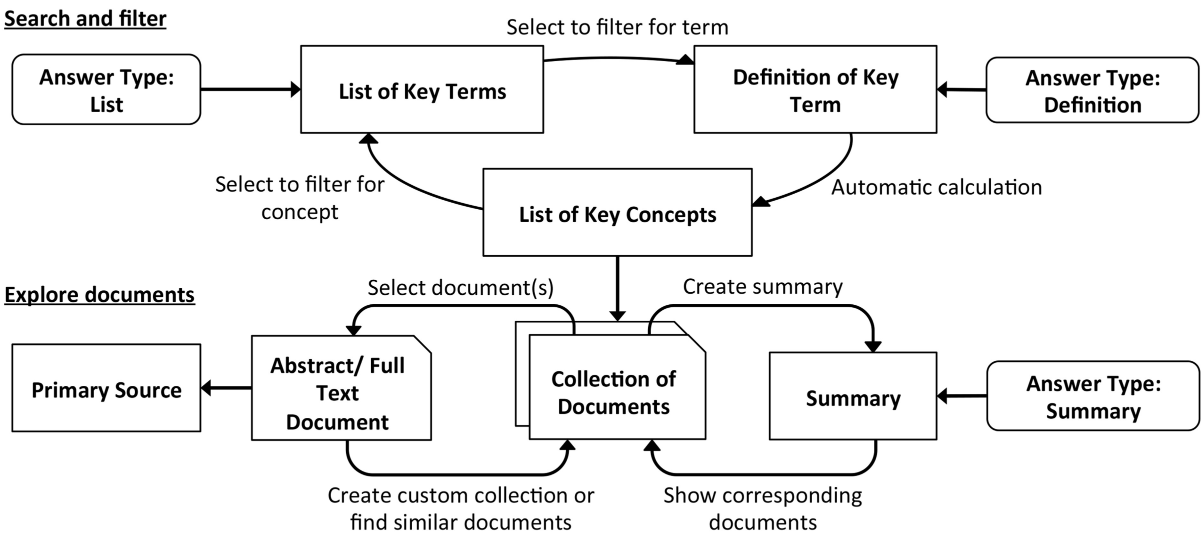


Figure 2. Summary of the main functions available in Olelo and how they are connected to each other. Rounded boxes represent answer types as identified by the QA module. They determine the entry point into either searching and filtering of key term and key concepts or into the summary generation and exploration of documents.

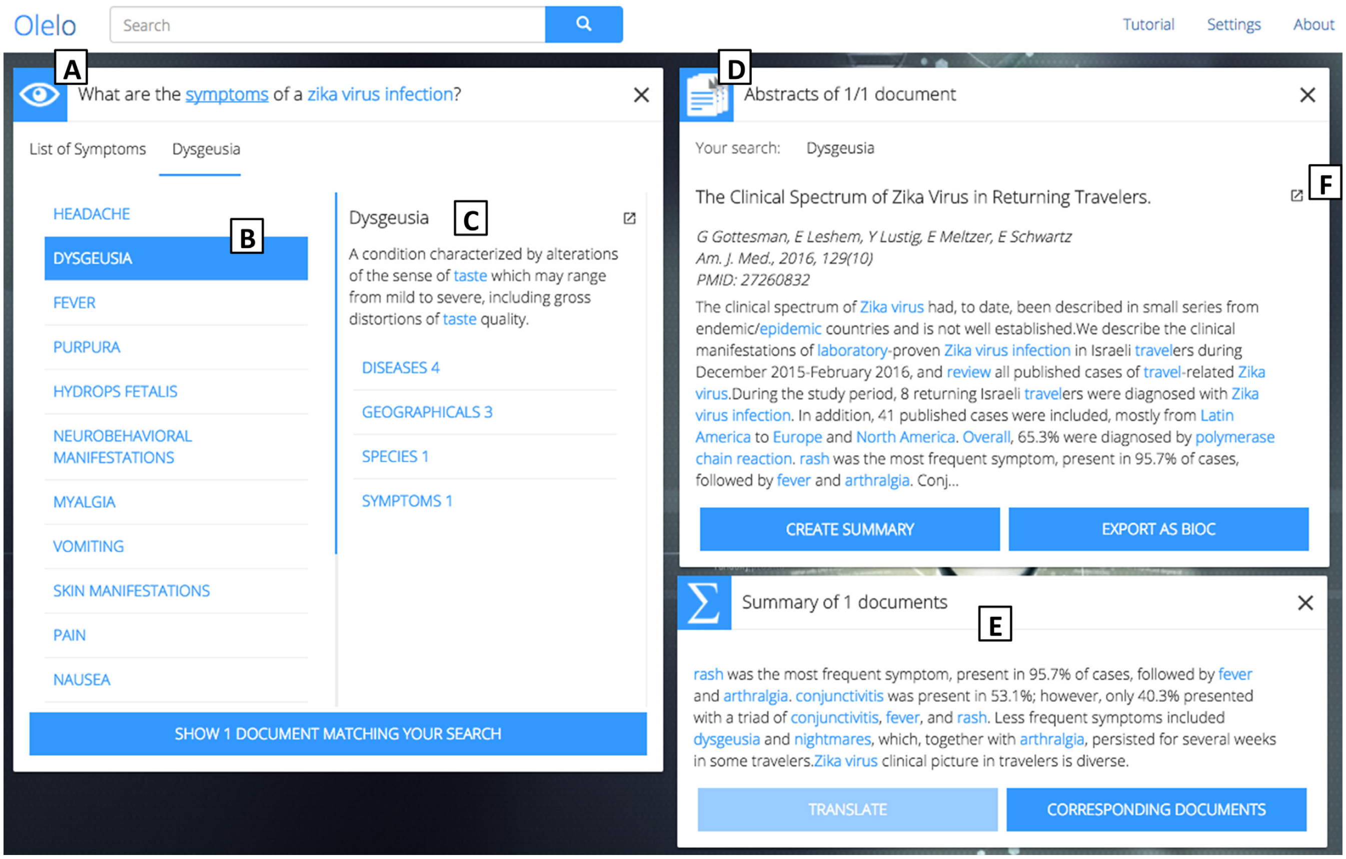
Use cases

Olelo was designed to support explorative searches. The presented user stories are limited to specific examples but can be extended to a complete exploration of the topic's key concepts and key terms and the underlying literature (Physician and Researcher). The BioASQ use case is based on the benchmark of questions for biomedical QA. We do not provide a systematic evaluation but the presented BioASQ questions can be answered by up to a couple of navigation steps. Please note that regular database updates may result in different retrieved documents or navigational steps.

Physician: latest developments in Zika outbreak. The complete result screen of this user story is shown in Figure 3 and letters in brackets refer to the figure labels. This practical example of how Olelo helps in searching the literature can be given by the recent outbreak of the Zika virus. When searching in PubMed for the keyword 'Zika virus infection' (ZVI), only 66 articles were published between 2008 and 2015, while 1007 articles were published in 2016. A physician in Brazil, who might be confronted with patients suffering from a ZVI, needs to regularly update his knowledge on the disease. As he is not able to read the whole amount of articles available, a possible question to Olelo could be the following: 'What are symptoms of a Zika virus infection?' (Figure 3A). The system outputs a list of symptoms that were found in the relevant documents, for instance, 'fever' and 'headache' are common answers that are supported by more than ten documents. The physician might have been aware of these. In contrast, 'Dysgeusia' (distorted taste) does not sound familiar and is only supported by one document (Figure 3B). In Olelo, a click on the term filters all documents found for the term and also displays a short definition (Figure 3C). To explore the relationship of ZVI and dysgeusia the physician opens the corresponding document, which displays a preview of the abstract in a document collection card (Figure 3D). The highlighted MeSH terms help to get an overview and show that 'dysgeusia' is not part of this short preview. But this can be solved by opening the preview of the full abstract by clicking on the link-out icon (Figure 3F). When generating a summary for this document, which includes the most important sentences extracted from the abstract, it reveals that dysgeusia is in fact a rarely reported symptom of ZVI in returning travelers (13) (Figure 3E). The physician could continue his search by exploring other hits of his search, e.g. 'purpura'.

List items are grouped into key concepts and are found in abstracts relevant to the given question. In this use case, dysgeusia can be identified as a direct symptom of ZVI. However, for example 'Hydrops fetalis' is only speculated to be a consequence affecting the fetus when the infection occurred during pregnancy (14). Thus, the nature of the found relationship, e.g. a speculation, direct/indirect correlation or just a co-occurrence of terms, is not provided by Olelo and needs to be inferred by the user.

Researcher: connection of S1P and diabetes. Olelo supports researchers in finding insights and relationships which might be new to them. For instance, a particular publication could have just been published or the researcher, e.g. a PhD student, might not yet be familiar with a finding. Alternatively, the researcher might not be aware of a relationship because it spans diverse disciplines of research. The PhD student might have heard of 'S1P' in her surroundings of diabetes research and could ask Olelo for some insights on the term. By posing a question such as 'What is S1P?', she finds that it is short for 'Sphingosine-1 phosphate' in the generated summary. A click on the found MeSH definitions for 'S1P receptors' within the summary reveals that S1P binds to lysosphingolipid receptors, a subfamily of G-protein-coupled receptors. But, 'What is the connection of S1P and diabetes?'. This question leads to another summary, which contains the sentence 'We show that eleva-



What are the symptoms of a zika virus infection?

Dysgeusia

A condition characterized by alterations of the sense of taste which may range from mild to severe, including gross distortions of taste quality.

Abstracts of 1/1 document

Your search: Dysgeusia

The Clinical Spectrum of Zika Virus in Returning Travelers.

G Gottesman, E Leshem, Y Lustig, E Meltzer, E Schwartz
Am. J. Med., 2016, 129(10)
PMID: 27260832

The clinical spectrum of Zika virus had, to date, been described in small series from endemic/epidemic countries and is not well established.We describe the clinical manifestations of laboratory-proven Zika virus infection in Israeli travelers during December 2015-February 2016, and review all published cases of travel-related Zika virus.During the study period, 8 returning Israeli travelers were diagnosed with Zika virus infection. In addition, 41 published cases were included, mostly from Latin America to Europe and North America. Overall, 65.3% were diagnosed by polymerase chain reaction. rash was the most frequent symptom, present in 95.7% of cases, followed by fever and arthralgia. Conj...

Summary of 1 documents

rash was the most frequent symptom, present in 95.7% of cases, followed by fever and arthralgia. conjunctivitis was present in 53.1%; however, only 40.3% presented with a triad of conjunctivitis, fever, and rash. Less frequent symptoms included dysgeusia and nightmares, which, together with arthralgia, persisted for several weeks in some travelers.Zika virus clinical picture in travelers is diverse.

Figure 3. Result screen showing three different result cards. The explorative search card (**A**) shows the answers for a factoid question. The MeSH definition (**C**) is given for the selected term 'Dysgeusia' (**B**) and corresponding documents were grouped into four categories, as shown below the definition. In this specific case, a single document was opened in a collection card (**D**) and a summary (**E**) was created. A link to the primary source can be opened via the symbol shown in (**F**).

tions in Plasma S1P [...] correlate with metabolic abnormalities such as adiposity and insulin resistance.' which is given in (15). The found document is the starting point for the deeper exploration of the underlying literature.

BioASQ benchmark. BioASQ organizes challenges on biomedical semantic indexing and QA and provides a corpus of questions and answers created by experts (16). Below we present the results for two questions (identifiers 54db7217c4c6ce8e1d000003 and 56c3327c50c68dd41600000c):

i) How many and which are the different isoforms for the ryanodine receptor?
The answers to the question—three ryanodine receptor isoforms(RyR1-RyR3)—can be found in one of the sentences in the generated summary: 'Generally, three ryanodine receptor isoforms (RyR1-RyR3) are known;' (17)
ii) Which DNA sequences are more prone for the formation of R-loops?
The answer—guanine-rich sequences—is given in the first sentence of the summary: 'The possible formation of three-stranded RNA and DNA hybrid structures (R-loops) in [...] guanine-rich genic and inter-genic regions [...]' (18).

DISCUSSION

QA systems aim to replace the cumbersome keyword search by using methods of NLP to understand questions and to extract and generate specific answers based on a collection of texts. In contrast to the display of simple lists of relevant documents, QA systems provide an answer according the user's request. To the best of our knowledge, there are only three QA Web systems (5–7). In our work, we addressed shortcomings of these systems, such as long response time and limited usability. There is still room for improvements in our methods, especially for document retrieval and extractive summarization, whose current approach cause nonsensical summaries in some occasions. However, in contrast to other systems, Olelo always returns an answer to a question (e.g. a summary) and thus allows the user to further explore the corresponding documents.

Olelo presents a new approach, as answers are the entry point to the explorative search module, which provides a new way of filtering and narrowing down relevant documents. Users can take a deep dive into these documents and are offered a fast way of selecting documents of interest. The summarization procedure extracts relevant sentences into a short paragraph. All of the NLP procedures are executed inside the IMDB and data are held in main memory, which allows responses in real time. Additionally, Olelo provides a clean and intuitive user interface that supports the fast and intelligent search through the ever growing literature.

In the future, Olelo will be extended by optional user accounts, in order to save searches and document collections. Further, we are also working on the integration of translation procedures. We are constantly working on all NLP methods to provide more answer types, for instance, yes/no questions, improved ranking and better answers to currently supported question types.

AVAILABILITY

Olelo is available at http://www.hpi.de/plattner/olelo. We tested Olelo in four modern browsers, namely Chrome, Internet Explorer, Safari, and Firefox. As we use the Google Polymer Framework, we recommend Chrome for best visualization and intuitive handling.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Harry Freitas da Cruz for his reviews, Akshay Bhushan and Guenter Hesse for their help on the graphical abstract and Matthias Herzog for the technical support. We would also like to thank Dr Christian Gerecke (University of Potsdam) for his helpful insights on our user stories.

FUNDING

German Federal Ministry of Education and Research [031A427B]. Funding for open access charge: Hasso Plattner Institute for IT Systems Engineering; Enterprise Platforms and Integration Concepts.

Conflict of interest statement. None declared.

REFERENCES

1. Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, **2011**, baq036.
2. Athenikos,S.J. and Han,H. (2010) Biomedical question answering: a survey. *Comput. Methods Programs Biomed.*, **99**, 1–24.
3. Neves,M. and Leser,U. (2015) Question answering for biology. *Methods*, **74**, 36–46.
4. Bauer,M.A. and Berleant,D. (2012) Usability survey of biomedical question answering systems. *Hum. Genomics*, **6**, 17.
5. Cao,Y., Liu,F., Simpson,P., Antieau,L., Bennett,A., Cimino,J.J., Ely,J. and Yu,H. (2011) AskHERMES: an online question answering system for complex clinical questions. *J. Biomed. Inform.*, **44**, 277–288.
6. Cruchet,S., Gaudinat,A., Rindflesch,T. and Boyer,C. (2009) What about trust in the Question Answering world. In: *Proceedings of theAMIA 2009 Annual Symposium*. AMIA, San Francisco.
7. Gobeill,J., Patsche,E., Theodoro,D., Veuthey,A.L., Lovis,C. and Ruch,P. (2009) Question answering for biology and medicine. In: *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine*. IEEE, Larnaca, Cyprus, pp. 1–5.
8. Schulze,F., Schueler,R., Draeger,T., Dummer,D., Ernst,A., Flemming,P., Cindy,P. and Neves,M. (2016) HPI Question Answering System in BioASQ 2016. In: *Proceedings of the Fourth BioASQ Workshop at the Conference of the Association for Computational Linguistics*. ACL, PA. pp. 38–44.
9. Schulze,F. and Neves,M. (2016) Entity-supported summarization of biomedical abstracts. In: *Proceedings of theFifth Workshop on Building and Evaluating Resources for Biomedical Text Mining*. Association for Computational Linguistics, PA, pp. 40–49.
10. Pang,P.C.-I., Verspoor,K., Pearce,J. and Chang,S. (2015) Better Health Explorer: designing for health information seekers. In: Plauderer,B (ed). *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. ACM, NY, pp. 588–597.
11. Jurafsky,D. and Martin,J.H. (2013) *Speech and Language Processing*. 2nd edn. Prentice Hall International, New Jersey.
12. Comeau,D.C., Islamaj Doan,R., Ciccarese,P., Cohen,K.B., Krallinger,M., Leitner,F., Lu,Z., Peng,Y., Rinaldi,F., Torii,M. *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, **2013**, bat064.
13. Meltzer,E., Leshem,E., Lustig,Y., Gottesman,G. and Schwartz,E. (2016) The clinical spectrum of Zika virus in returning travelers. *Am. J. Med.*, **129**, 1126–1130.
14. Sarno,M., Sacramento,G.A., Khouri,R., do Rosário,M.S., Costa,F., Archanjo,G., Santos,L.A., Nery Jr,N., Vasilakis,N., Ko,A.I. *et al.* (2016) Zika virus infection and stillbirths: a case of hydrops fetalis, hydranencephaly and fetal demise. *PLoS Negl. Trop. Dis.*, **10**, e0004517.
15. Kowalski,G.M., Carey,A.L., Selathurai,A., Kingwell,B.A. and Bruce,C.R. (2013) Plasma sphingosine-1-phosphate is elevated in obesity. *PLoS One*, **8**, e72449.
16. Tsatsaronis,G., Balikas,G., Malakasiotis,P., Partalas,I., Zschunke,M., Alvers,M.R., Weissenborn,D., Krithara,A., Petridis,S., Polychronopoulos,D. *et al.* (2015) An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, **16**, 138.
17. Prinz,G. and Diener,M. (2008) Characterization of ryanodine receptors in rat colonic epithelium. *Acta Physiol.*, **193**, 151–162.
18. Jenjaroenpun,P., Wongsurawat,T., Yenamandra,S.P. and Kuznetsov,V.A. (2015) QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res.*, **43**, W527–W534.