Towards a Personalized Query Answering Framework on the Web of Data

Enayat Rajabi
Dalhousie University
Halifax, NS, Canada
rajabi@dal.ca

Christophe Debruyne
Trinity College Dublin
Dublin 2, Ireland
debruync@scss.tcd.ie

Declan O'Sullivan
Trinity College Dublin
Dublin 2, Ireland
declan.osullivan@cs.tcd.ie

ABSTRACT
In this paper, we argue that layering a question answering system on the Web of Data based on user preferences, leads to the derivation of more knowledge from external sources and customisation of query results based on user's interests. As various users may find different things relevant because of different preferences and goals, we can expect different answers to the same query. We propose a personalised question answering framework for a user to query over Linked Data, which enhances a user query with related preferences of the user stored in his/her user profile with the aim of providing personalized answers. We also propose the extension of the QALD-5 scoring system to define a relevancy metric that measures similarity of query answers to a user's preferences.

CCS Concepts
• Information systems→ Question answering.

Keywords
Personalisation; Linked Data; Question Answering.

1. INTRODUCTION
With the rapid growth of Web of Data (currently more than 154 billion triples[1]), answering users' queries and delivering the actual results have become increasingly a key issue [1]. Retrieving appropriate answers by exploring or browsing the Web content based on keyword search mostly fail to exploit the internal structures of data and reveal their underlying semantics. The search results are expected to contain information corresponding to the keywords and in most cases, the user is left with the task of sifting through these results. Question Answering is an information retrieval technique [2] that tackles this issue by retrieving exact answers to users' questions posed in natural language. On the Web of Data [3], due to its standards and schema, the question answering system is executed on a network of RDF datasets and data discovery sometimes requires integrating several datasets. Moreover, different kinds of datasets underlying Question Answering systems have been semantically improved from unstructured text to structured data [4].

One of the significant factors to be considered in a question answering system is personalisation of the query and answers contingent on the user interest and preferences, as various users may find different things relevant when searching because of different preferences, goals and interests. Thus, users may naturally expect different answers to the same query. Typically, query personalisation [5] is the process of dynamically enhancing a query with related user preferences stored in a user profile with the aim of providing personalised answers. To illustrate the question answering application, consider for example the following case:

"*Bob is a high school student and performs most of his studies and homework using search engines. However, he is tired of searching among the search engines' results, as it is a tedious work to discover the precise answer in thousands of candidate contents. He is aware of the strength of Web of Linked Data and decides to pose his queries against a personalised question answering system (PQALD). He registers in PQALD and creates his profile. He also specifies his preferences for search. For example, he is interested in reading fiction books, and romantic movies. Afterwards, he starts surfacing the Web of Data to find the answers of his questions using PQALD. The system narrows the list of results for Bob to specific answers that are close to his interests and preferences. As an example, it lists all the romantic movies (as one of Bob's interests) as the priorities of search for the following question: 'best movies of 2016?' PQALD also considers all other Bob's preferences and interests in the search. As PQALD relies on the Web of Linked Data, it links each found answer to IMDB dataset so that Bob can access more information about each movie. Bob can also specify more preferences for his search (one for his homework, another for his research, etc.) and utilise each one for the specific search.*"

Most of the studies in the area of question answering on the Web of Data [2] [6] [7] [8] present approaches to retrieve information and infer knowledge over the Semantic Web utilising a set of ontologies, reasoning capabilities, and inference engines. Some others [9] [10] investigate the issues involved in designing a query language in Semantic Web. To the best of our knowledge, query personalisation for question answering on the Web of Data has not been widely considered in studies to date. This short paper presents a personalised question answering framework with the intent of improving as well as customising the search results contingent on a user's preferences and interests. The remainder of this paper is structured as follows. In Section 2, we will outline the current studies on question answering on the Web of Data. Section 3 introduces our proposed approach, followed by conclusion and future work in Section 4.

2. BACKGROUND AND RELATED WORK
The goal of question answering systems is "to allow users to ask questions in Natural Language (NL), using their own terminology and receive a concise answer" [11]. Recent years have witnessed the transfer of question answering techniques used for traditional

[1] http://stats.lod2.eu/ [Accessed: 28-Feb-2017]

Web or local systems to the development of the semantic query answering systems on the Web of Linked Data [6], which take queries expressed in natural languages and a given ontology as input, and returns answers drawn from one or more datasets that subscribe to the ontology [1]. Most query answering systems rely on ontology specific approaches, where the power of ontologies as a model of knowledge is directly exploited for the query analysis and translation. Aqualog [8], in particular, allows users to choose an ontology and then ask natural language queries with respect to the universe of discourse covered by the ontology. It identifies ontology mappings for all the terms and relations in the triple patterns of SPARQL query by means of string based comparison methods and WordNet. AquaLog uses generalisation rules to learn novel associations between the natural language relations used by the users and the ontology structure. Lopez et al. [1] compared several ontology-based question answering systems in a study based on a set of criteria including degree of customization, and revealed that most of the semantic question answering systems (such as QuestIO [12], FreyA [13], and Querix [14]) did not support customization in their approaches, whilst QACID [15] and ORAKEL [16] considered some levels of domain customisation that have to be performed or supervised by domain experts. For example, QACID is based on a collection of queries from a given domain that are categorised into clusters, where each cluster, containing alternative formulations of the same query, is manually associated with SPARQL queries. None of the mentioned question answering systems, however, did take the users' interests and preferences into their consideration.

With regard to query personalisation studies, Koutrika and Ioannidis [17] presented an approach on query personalisation in digital libraries over relational databases. They treated query personalisation as a query-rewriting problem and provided an algorithm that produces a personalised version of any query. They captured user preferences as query rewriting rules with assigned weights that indicate user interest. In [18], the authors formulated Constrained Query Personalisation (CQP) approach as a state-space search problem to build a set of personalised queries dynamically taking the following features into account: the queries issued, the user's interest in the results, response time, and result size. Gheorghiu et al. [19] presented a hybrid preference model that combines quantitative and qualitative preferences into a unified model using an acyclic graph, called HYPRE Graph, to personalise the query results. They implemented a framework using Neo4j graph database system and experimentally evaluated it using real data extracted from DBLP. The above-mentioned studies in this domain did not implement their approaches on the Web of Data to leverage the connectivity and availability of datasets and improve their results. However, we believe the preference model mentioned in [20] can be utilised in development of a query answering system for Linked Data. We will also leverage an extensive survey performed by Lopez et al. [1] in our implementation to precisely identify the strengths and weaknesses of other approaches with the intent of designing a robust system. Moreover, to convert the user questions to SPARQL, we will investigate the possibility of using some text to SPARQL approaches like AutoSPARQL [21], which implements an active learning approach using Query Tree Learner (QTL) algorithm.

3. PERSONALISED QUESTION ANSWERING FRAMEWORK

Searching information on the Web of Data requires user friendly approaches, similar to the ease of keyword-based search engines, but relying on the RDF. In this vein, typically Question Answering systems are proposed to retrieve the best possible answers for end users. Question Answering on Linked Data has recently been studied by researchers along with the associated challenges in the domain [12-15]. Current query personalisation systems mostly concern semi-structured or unstructured data and, to the best of our knowledge, a personalisation query approach on the Web of Data has not been considered yet. Providing an enriched knowledgebase is another step toward developing a question answering system that can be fulfilled by linking different datasets to each other or to external knowledge on the Web.

Generally speaking, query personalisation in a question answering system usually falls into two categories: a) information filtering systems wherein a stored query or set of queries comprise a user profile based on which an information filtering system collects and distributes relevant information; b) recommendation systems that produce predictions, recommendations, opinions that help a user evaluate or select a set of entities, and the system identifies other similar entities, based on which recommendations or predictions are produced regarding what the user would like.

Our approach for personalising the user queries will fall in the first category and relies on a quantitative approach which aims at an absolute formulation of user preferences, such as a user likes comedies very much and westerns to a lesser degree. This allows for total ordering of results and the straightforward selection of those answers matching user preferences. We may also use techniques in query personalisation that reveal some implicit knowledge about the user interests, when incomplete information in the user profile prevents us to retrieve appropriate knowledge for query customisation [20]. Figure 1 outlines the main components and flows of the proposed approach wherein we will analyse the questions, customise them based on users' preferences and profile, extract the answers from a set of linked datasets, and finally score the results as well as visualise them for users. Below we will explain how we implement each phase of the proposed framework.

3.1 Question Analysis

With respect to question analysis phase, several NLP techniques can be used to convert the user questions to SPARQL. In particular, the underlying idea of AutoSPARQL [21] is an interesting solution to convert a natural language expression to a SPARQL query, which can then retrieve the answers of a question from a given triple store. Our strategy for both syntactic and semantic analysis of questions is not implementing a software from scratch to convert the user question to a SPARQL query, instead we intend to apply one of the existing approaches (i.e. AutoSPARQL, GATE[2], or the approach in [22]) to select features from the question, to extract and classify them, and to support the transformation of question to SPARQL. To provide support for multiple languages, we intend to follow the approach mentioned in QALD-4 [23] by annotating the questions with a set of keywords in an XML or RDF format. The language detection step is also appropriate to identify the user's language and customise the results for him/her.

[2] https://gate.ac.uk/ [Accessed: 28-Feb-2017]

3.2 Query Personalisation

For the query personalisation phase, the idea is to design and implement a user preference model (based on current well-designed preference models e.g., [20]) and customise the query according to the user's interests stated in the user profile. Having the user preferences in one of the Linked Data vocabularies (including but not limited to FOAF or FRAP [24]), the query analyser analyses the query (which is presented in SPARQL as the output of previous step) and customises it according to the designed user preference model. The output of this phase is a new SPARQL query, which will be the input of answer extraction service.

3.3 Answer Extraction Service and Reasoning

To extract the answers from a set of linked datasets, we intend to apply a reasoning engine that uses description-logic as its basic formalism and relies one of OWL 2 flavours (e.g. OWL 2 QL) as the ontology logic. The idea is to select a reasoner that provides completeness and decidability of the reasoning problems, offers computational guarantees and has more efficient reasoning support than other formalisms. As we will follow a rule-based reasoning engine, a homogeneous approach will be applied to make a tight semantic integration for embedding rules and ontology in a common logic ground. We will also utilise either SWRL [25] or RIF [27] as the rule languages of the framework in the knowledge layer of this phase and Jena-Pellet reasoner as our reasoning engine tool.

3.4 Answer Scoring and Visualisation

One of the technologies that can be applied for Answer Scoring is using the lexical answer type. DeepQA[3], as the IBM project in NLP, includes a system that takes a candidate answer along with a lexical answer type and returns a score indicating whether the candidate answer can be interpreted as an instance of the answer type. This system utilises WordNet or DBpedia datasets to search for a link of hyponymy, instance-of or synonymy between answer and its lexical type. We will extend this approach to discover a link between the answers and the user's profile or preferences.

This phase has also a visualisation service to visualise the final candidate answers (the most matched answers to the user's profile) to the user.

3.5 Evaluation

To evaluate the proposed query answering system, we intend to utilise the QALD evaluation approach [6], which provides a common evaluation benchmark and allows for an in-depth analysis of a question answering system and its progress over time. In this benchmark, the task for our system would be to return, for a given natural language question and an RDF data source, a list of entities that answer the question, where entities are either individuals identified by URIs or labels, or literals such as strings, numbers, dates, and Booleans. We will extend the benchmark to evaluate the closeness of the query results to the user preferences or profile. Particularly, multilingual questions are provided in seven different languages in QALD-4 [23] that helps us to cover the users' language in their

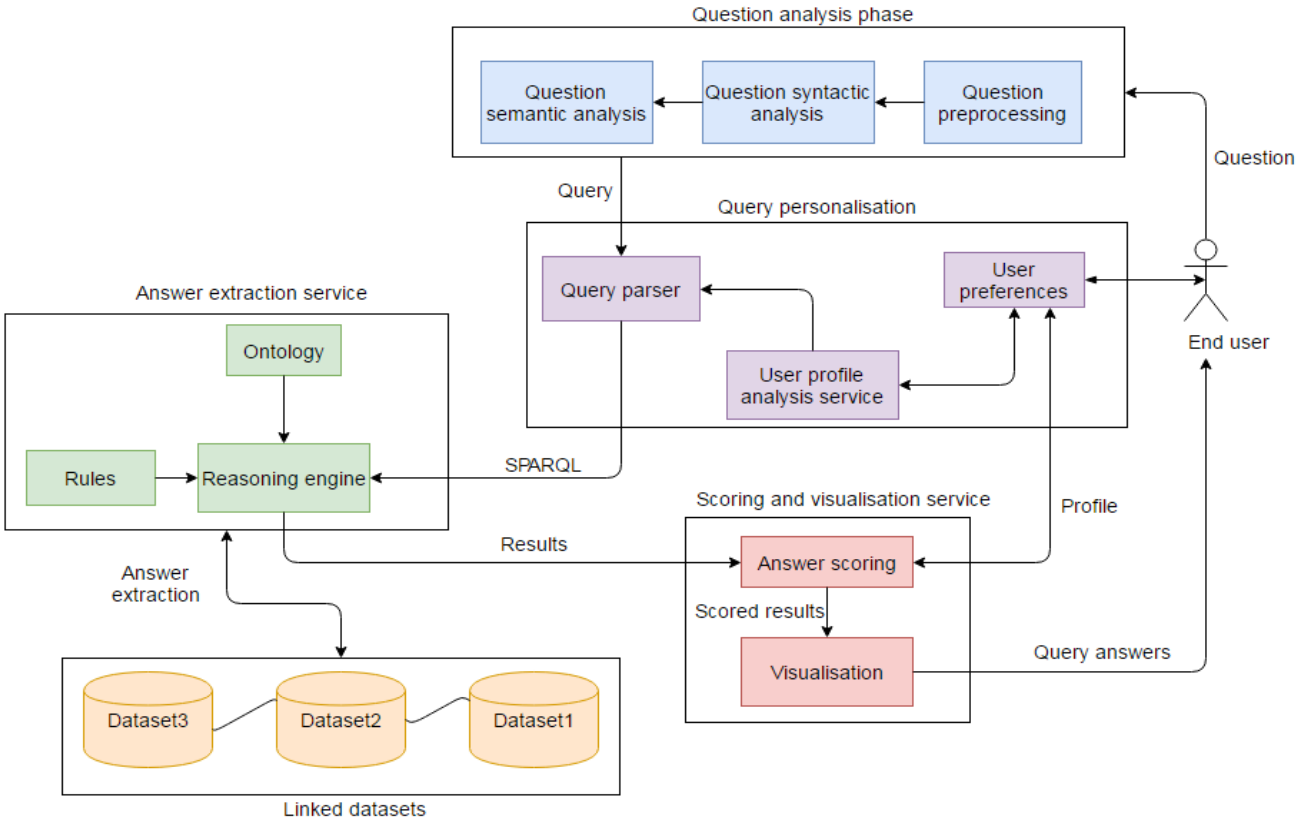

Figure 1. Personalised question answering framework

[3] https://www.research.ibm.com/deepqa/deepqa.shtml [Accessed: 28-Feb-2017]

preferences and answer the correspondent question accordingly. Moreover, QALD-5 [28] allows users to annotate hybrid questions with several attributes including answer type and aggregation. Extending the question annotations by adding more attribute associated with the users profile in a query allows the question answering system to consider the user preferences in the process. According to this approach, to measure the overall precision of a question q, we consider three following metrics (precision, recall, and relevance):

$$\text{Recall(q)} = \frac{\text{number of correct system answers for q}}{\text{number of gold standard answers for q}}$$

$$\text{Precision(q)} = \frac{\text{number of correct system answers for q}}{\text{number of system answers for q}}$$

$$\text{Relevance(q,u)} = \frac{\sum_i^n \text{Relevance(q},a_i)}{\text{number of correct system answers for q}}$$

Where Relevance(q,u) (0<=value<=1) is the total similarity (according to the user profile) of all the correct answers (a_i) to question q rated by user u.

$$\text{Relevance(q)} = \frac{\sum_i^n \text{Relevance(q},u_i)}{\text{number of rated users}}$$

Relevance(q) is total relevance of all users $(u_i, \ldots, u_n)$ to question q.

Overall F-measure in our approach is computed as follows:

$$\text{F} - \text{Measure(q)} = \frac{2 * \text{Precision(q)} \times \text{Recall(q)}}{\text{Precision(q)} + \text{Recall(q)}} \times \text{Relevance(q)}$$

The gold standard answers in our system are defined as most matched answers with the user preferences.

3.6 Evaluation Scenario

To select a set of linked datasets for evaluation and test of the proposed question answering system, we formulated a set of criteria to assess the abilities and robustness of system. Containing large-scale data, multilinguality, ontological structure, and linkability were part of these criteria. Our knowledgebase will be chosen from one or more of the following datasets for the evaluation:

- DBpedia[4] which is the central interlinking hub for the emerging linked data cloud [29]. The English version of DBpedia includes around 4.6 million things. This dataset has been linked to 41.2 million entities to YAGO categories[5].
- MusicBrainz[6] as a collaborative open-content music dataset, contains all of MusicBrainz' artists and albums as well as a subset of its tracks, leading to a total of around 15 million RDF triples.
- British National Bibliography[7] (BNB) dataset that publishes books and digital objects as Linked Data by British Library linked to external sources including GeoNames[8]. Currently, BNB includes around 3.1 million descriptions (more than 109 million triples) of books and serials published in the UK over the last 60 years.
- WordNet [30] dataset with hundreds of thousands of facts that provides concepts (called synsets), each representing the sense of a set of synonymous words.

The presented framework is a domain independent system. However, to evaluate the functionality of the system, we intend to apply the mentioned dataset(s) in an educational system wherein students are willing to do their homework/research by answering a set of questions. First of all, we will provide a set of in-scope and out-scope test questions (e.g., 30 questions). To assess the efficiency and robustness of proposed system, some in-scope questions will require linking datasets to be answered and some others will not be in the scope of knowledgebase (out-scope). For example, for the question of "list of most sold books in 2013", the system will explore more than one datasets to discover the answers for users. On the other hand, students will set their profile and specify a set of preferences that can be applied for their research purposes. For example, information such as student's grade, language, and field of study will be specified in his/her profile. Also, student's interests such as his/her favourite subjects, music, and books will be provided in the system. Figure 2 illustrates a prototype that the final system will look like, wherein the answers have been personalised based on user' interests and preferences in the right side of picture. Students will select some case questions and the system will provide them a set of candidate answers based on their preferences and profile. Eventually, the students will be asked to rate the results, that it, we will specify the similarity of generated answers and what the students expect to see as the output of system. This metric will be used for the evaluation of the accuracy of proposed system.

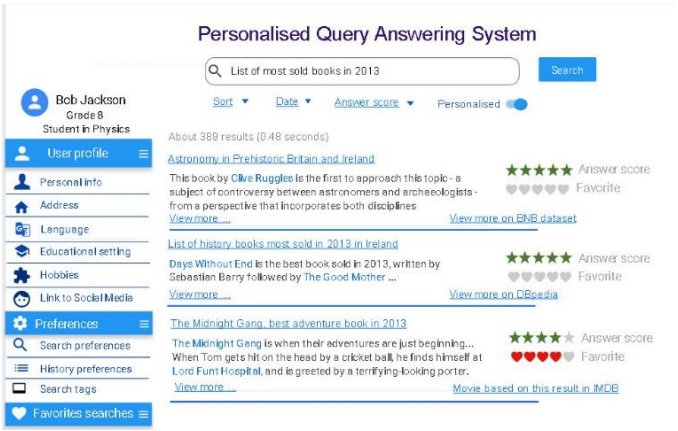

Figure 2. A prototype for the personalised query answering framework on the Web of Data

4. CONCLUSION AND FUTURE WORK

This paper described a personalised question answering framework to improve the results of a question answering system based on a user's preferences and interests. We also proposed a relevancy metric to measure the similarity between the answers and the user profile by extending the QALD-5 scoring system. The proposed framework will be implemented on the Web of Data, where the question answering system uses a set of linked datasets, an API for converting questions to SPARQL queries, and a robust answer scoring system to obtain the most interested results for users.

[4] http://dbpedia.org [Accessed: 28-Feb-2017]

[5] http://www.mpi-inf.mpg.de/yago-naga/yago/ [Accessed: 28-Feb-2017]

[6] http://musicbrainz.org/ [Accessed: 28-Feb-2017]

[7] http://bnb.bl.uk/ [Accessed: 28-Feb-2017]

[8] http://www.geonames.org/ [Accessed: 28-Feb-2017]

ACKNOWLEDGEMENT

This study is partially supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre for Digital Content Technology Platform Research (http://www.adaptcentre.ie/) at Trinity College Dublin.

REFERENCES

[1] V. Lopez, V. Uren, M. Sabou, and E. Motta, "Is Question Answering Fit for the Semantic Web?: A Survey," *Semantic web*, vol. 2, no. 2, pp. 125–155, 2011.

[2] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng, "Web Question Answering: Is More Always Better?," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, pp. 291–298, 2002.

[3] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked Data on the Web (LDOW2008)," in *Proceedings of the 17th International Conference on World Wide Web*, New York, NY, USA, 2008, pp. 1265–1266.

[4] S. Shekarpour, K.M. Endris, A. Jaya Kumar, D. Lukovnikov, K. Singh, H. Thakkar, and C. Lange, "Question answering on linked data: Challenges and future directions". In Proceedings of the 25th International Conference Companion on World Wide Web, pp. 693-698, 2016.

[5] G. Koutrika and Y. Ioannidis, "Personalization of queries in database systems," in *20th International Conference on Data Engineering, 2004. Proceedings*, 2004, pp. 597–608.

[6] V. Lopez, C. Unger, P. Cimiano, and E. Motta, "Evaluating question answering over linked data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 21, pp. 3–13, Aug. 2013.

[7] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Matfield, "Information Retrieval on the Semantic Web," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, New York, NY, USA, pp. 461–468, 2002.

[8] V. Lopez, M. Pasin, and E. Motta, "AquaLog: An Ontology-Portable Question Answering System for the Semantic Web," in *The Semantic Web: Research and Applications*, A. Gómez-Pérez and J. Euzenat, Eds. Springer Berlin Heidelberg, pp. 546–562, 2005.

[9] R. Fikes, P. Hayes, and I. Horrocks, "OWL-QL—a language for deductive query answering on the Semantic Web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, pp. 19–29, Dec. 2004.

[10] I. Horrocks and S. Tessaris, "Querying the Semantic Web: A Formal Approach," in *International Semantic Web Conference*, Springer Berlin Heidelberg, Springer Berlin, pp. 177–191, 2002.

[11] L. Hirschman and R. Gaizauskas, "Natural Language Question Answering: The View from Here," *Natural Language Engineering*, vol. 7, no. 4, pp. 275–300, Dec. 2001.

[12] V. Tablan, D. Damljanovic, and K. Bontcheva, "A Natural Language Query Interface to Structured Information," in *European Semantic Web Conference (ESWC)*, Springer Berlin Heidelberg, pp. 361–375, 2008.

[13] V. T. Danica Damljanovic and K. Bontcheva, "A Text-based Query Interface to OWL Ontologies," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 28-30, 2008.

[14] E. Kaufmann, A. Bernstein, and R. Zumstein, "Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs," *In: 5th ISWC*, pp. 980–981, 2006.

[15] Ó. Ferrández, R. Izquierdo, S. Ferrández, and J. L. Vicedo, "Addressing ontology-based question answering with collections of user queries," *Information Processing & Management*, vol. 45, no. 2, pp. 175–188, Mar. 2009.

[16] P. Cimiano and M. Minock, "Natural Language Interfaces: What Is the Problem? – A Data-Driven Quantitative Analysis," in *proceeding of International Conference on Application of Natural Language to Information Systems*, Springer Berlin Heidelberg, pp. 192–206, 2009.

[17] G. Koutrika and Y. Ioannidis, "Rule-based query personalization in digital libraries," *International Journal of Digital Library*, vol. 4, no. 1, pp. 60–63, Aug. 2004.

[18] G. Koutrika and Y. Ioannidis, "Constrained Optimalities in Query Personalization," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, pp. 73–84, 2005.

[19] R. Gheorghiu, A. Labrinidis, and P. K. Chrysanthis, "Unifying Qualitative and Quantitative Database Preferences to Enhance Query Personalization," in *Proceedings of the Second International Workshop on Exploratory Search in Databases and the Web*, New York, NY, USA, pp. 6–8, 2015.

[20] G. Koutrika, E. Pitoura, and K. Stefanidis, "Preference-Based Query Personalization," in *Advanced Query Processing*, B. Catania and L. C. Jain, Eds. Springer Berlin Heidelberg, pp. 57–81, 2013.

[21] J. Lehmann and L. Bühmann, "AutoSPARQL: Let Users Query Your Knowledge Base," in *Proceedings of ESWC*, 2011.

[22] D. Song, F. Schilder, C. Smiley, and C. Brew, "Natural language question answering and analytics for diverse and interlinked datasets," in *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 101–105, 2015.

[23] P. Cimiano, V. Lopez, C. Unger, E. Cabrio, A.-C. N. Ngomo, and S. Walter, "Multilingual Question Answering over Linked Data (QALD-3): Lab Overview," in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pp. 321–332, 2013.

[24] L. Polo, I. Mínguez, D. Berrueta, C. Ruiz, and J. M. Gómez, "User preferences in the web of data," *Semantic Web*, vol. 5, no. 1, pp. 67–75, Jan. 2014.

[25] "SWRL: A Semantic Web Rule Language Combining OWL and RuleML." [Online]. Available: https://www.w3.org/Submission/SWRL/. [Accessed: 22-Aug-2016].

[26] "RIF Overview (Second Edition)." [Online]. Available: https://www.w3.org/TR/rif-overview/. [Accessed: 03-Dec-2016].

[27] C. Unger *et al.*, "Question Answering over Linked Data (QALD-4)," presented at the Working Notes for CLEF 2014 Conference, 2014.

[28] C. Unger *et al.*, "Question Answering over Linked Data (QALD-5)," in *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, 2015, vol. 1391.

[29] E. Rajabi, S. Sanchez-Alonso, and M.-A. Sicilia, "Analyzing broken links on the web of data: An experiment with DBpedia," *J Assn Inf Sci Tec*, vol. 65, no. 8, pp. 1721–1727, Aug. 2014.

[30] "RDF/OWL Representation of WordNet." [Online]. Available: https://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/. [Accessed: 18-Jan-2017].