The KEYSTONE IC1302 COST Action

Francesco Guerra[1], Yannis Velegrakis[2](✉), Jorge Cardoso[3], and John G. Breslin[4]

[1] Università di Modena e Reggio Emilia, Modena, Italy
francesco.guerra@unimore.it
[2] University of Trento, Trento, Italy
velgias@disi.unitn.eu
[3] Huawei Research Center, Munich, Germany
jorge.cardoso@huawei.com
[4] Insight Centre for Data Analytics, NUI Galway, Galway, Ireland
john.breslin@nuigalway.ie

Abstract. As more and more data becomes available on the Web, as its complexity increases and as the Web's user base shifts towards a more general non-technical population, keyword searching is becoming a valuable alternative to traditional SQL queries, mainly due to its simplicity and the lower effort/expertise it requires. Existing approaches suffer from a number of limitations when applied to multi-source scenarios requiring some form of query planning, without direct access to database instances, and with frequent updates precluding any effective implementation of data indexes. Typical scenarios include Deep Web databases, virtual data integration systems and data on the Web. Therefore, building effective keyword searching techniques can have an extensive impact since it allows non-professional users to access large amounts of information stored in structured repositories through simple keyword-based query interfaces. This revolutionises the paradigm of searching for data since users are offered access to structured data in a similar manner to the one they already use for documents. To build a successful, unified and effective solution, the action "semantic KEYword-based Search on sTructured data sOurcEs" (KEYSTONE) promoted synergies across several disciplines, such as semantic data management, the Semantic Web, information retrieval, artificial intelligence, machine learning, user interaction, interface design, and natural language processing. This paper describes the main achievements of this COST Action.

1 The Action in a Nutshell

The idea for KEYSTONE (semantic keyword-based search on structured data sources) as a COST Action was born during a joint research project involving researchers from the Universities of Modena and Reggio Emilia (Italy), Trento (Italy) and Zaragoza (Spain). The project was funded by a local foundation and was established to support research exchanges among international institutions. The goal of these exchanges was to develop a query language and associated keyword-based query engine to support users in querying data sources with complex large

J. Szymański and Y. Velegrakis (Eds.): IKC 2017, LNCS 10546, pp. 187–195, 2018.
https://doi.org/10.1007/978-3-319-74497-1_19

schemas. At the end of the project, it was decided to propose a COST action to expand the collaboration into a pan-European network, tap into the background that had been built, and produce additional techniques beyond the state of the art. COST was the right choice for many reasons: (i) It is a *flexible* scheme. It allows researchers to continue the research activities they already have in place, which means that the probabilities of success are significantly higher. (ii) It is *open*: It enables the exchange of methodologies and skills which leads to capability expansion and higher impact; and (iii) It is *oriented towards young researchers*: It has a number of instruments to support early career investigators with participation in research activities, either as trainees, or as activity coordinators.

All of the above reasons made COST the right tool for obtaining significant research experience and management skills. The partners were selected initially based on two criteria. The first was their research area. The areas that were given priority were those related to keyword search, or those that directly or indirectly could be exploited for some contribution towards the advancement of the work on keyword search. The second criterion was the scientific results that each researcher had produced that was related to the topic up until that point in time. The reason for the first criterion was to guarantee the creation of a cohesive network of researchers with common goals and interests. The reason for the second was to achieve the maximum impact and avoid starting from scratch, by building on the state-of-the-art results that had already been achieved. Based on these, researchers from eight European countries were initially selected to participate, and this set was later extended to cover almost every European country, thanks to the active promotion efforts by the participants and the management committee.

KEYSTONE [1,2] Action was approved at the beginning of 2013. The project kicked off in October 2013 and the official end date was December 2017. The principal target outcome was the coordination of collaboration amongst the fields of semantic data management, the Semantic Web, information retrieval, artificial intelligence, machine learning and natural language processing, to enable research activity and technology transfer in the areas of keyword-based search over structured data sources. The coordination effort aimed to support the development of a search paradigm that provides users with keyword-based searching capabilities for structured data sources as they currently do with documents. Furthermore, it aimed to exploit the structured nature of data sources in defining complex query execution plans by combining partial contributions from different sources.

Alongside these main objectives, the action also aimed towards: (i) promoting the development of novel techniques for keyword-based search over structured data sources; (ii) facilitating the transfer of knowledge and technology to the scientific community and enterprises; and (iii) building a critical mass of research activities and outcomes that would achieve sustainability of the research themes beyond the Action.

The Action Organisation

The Scientific Programme of the Action has been conceived in a manner to achieve the primary and secondary objectives. It consists of three vertical thematic areas, each one covered by a respective working group (Working Groups 1, 2, and 3) and a horizontal activity across the thematic areas that is covered by a fourth working group (Working Group 4).

Working Group 1 is composed of people studying the representation of structured data sources. It investigates possible metadata formats describing data sources and efficient structures for data retrieval from such sources. Working Group 2 works on keyword search techniques. It is the core of the Action since it puts together researchers studying techniques for matching user keywords to database data and metadata structures, and for forming the actual SQL queries that will retrieve the data. Working Group 3 focuses on user interaction and keyword query interpretation. It investigates issues related to the semantic disambiguation of queries based on context and topics related to keyword annotation with respect to some reference ontologies. It also looks at the development of languages for keyword searching and the use of users' feedback in improving the generated results. Finally, Working Group 4 is about research integration, showcases, benchmarks and evaluations, and aims to integrate the activities of the different working groups with the goal of creating a "vademecum" for developing a search engine for structured data sources.

Two Co-Chairs for each Working Group were appointed in the first meeting to guarantee the full involvement of interested researchers in the Action. The Working Group Co-Chairs, alongside the Action Chair, the Scientific Coordinator, the Training Coordinator, the STSM Coordinator and the Dissemination Coordinator, constitute the Executive Scientific Board. This board is in charge of planning and executing the activities approved by the Management Committee of the Action, composed of a maximum of two people per participating country according to COST regulations. The Action had 53 effective members, 36 substitute members and 4 observers, all of them from a total of 31 different countries.

2 The KEYSTONE People

Participation in the Action was open to everyone, but had to be approved by the Management Committee. The network of experts that participated in the various Working Groups were representative of a great number of different fields in all aspects important to the themes of the Action. These experts were renowned researchers working on many related topics that have been supported financially by national governments and the European Commission.

At the end, the Action had involved 238 Working Group Members (177 male, 76.1% – 56 female, 23.9%). These members were distributed across the groups as follows. Note that members could choose to participate in more than one group, according to their interests.

- Working Group 1: 169 Members
- Working Group 2: 170 Members
- Working Group 3: 139 Members
- Working Group 4: 118 Members

90 participants were "early career investigators", i.e. they were within eight years of the date when they obtained their PhD/doctorate.

The geographical distribution of the Working Group Members is not balanced across countries as illustrated in Fig. 1. Spain is the most represented country, with ten other countries having at least ten participants, and the remaining countries having smaller participation. This difference is due to three main reasons: (i) The ability to establish and sustain collaboration from these countries, as it is determined by the support that the country and its infrastructure provides, (ii) the nature of the Action theme, that may not be among the buzzwords of the period in which the Action took place, and (iii) the amount to which each country's research is devoted to these topics (such countries are referred to as Inclusiveness Target Countries – ITC within COST and the European Commission) Around 37% of KEYSTONE participants (88 members) were from these ITC countries.

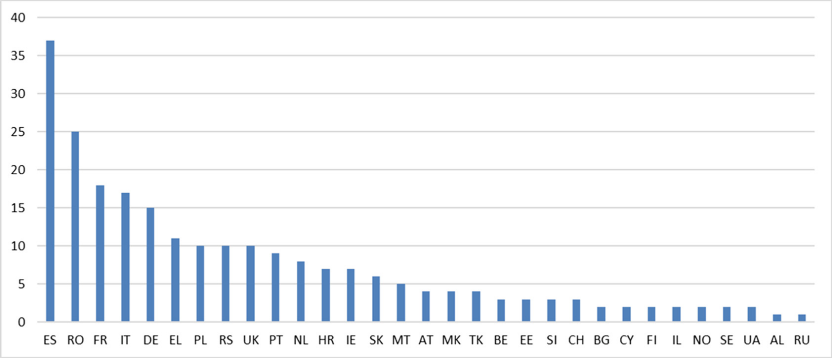

Fig. 1. Geographical distribution of the Working Group Members.

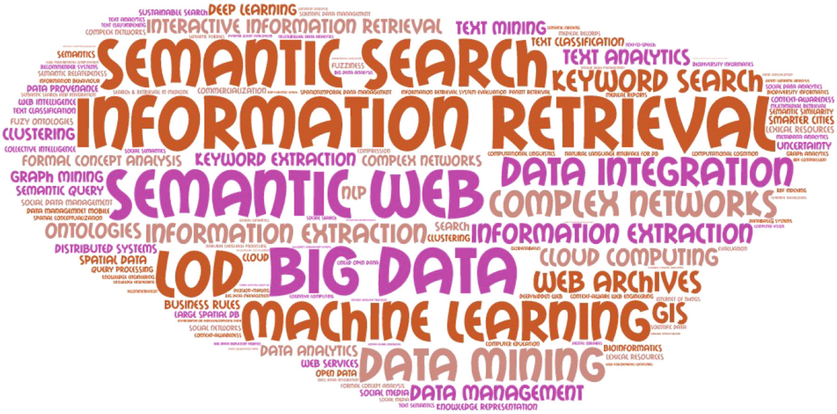

Fig. 2. Working Group Member research keywords.

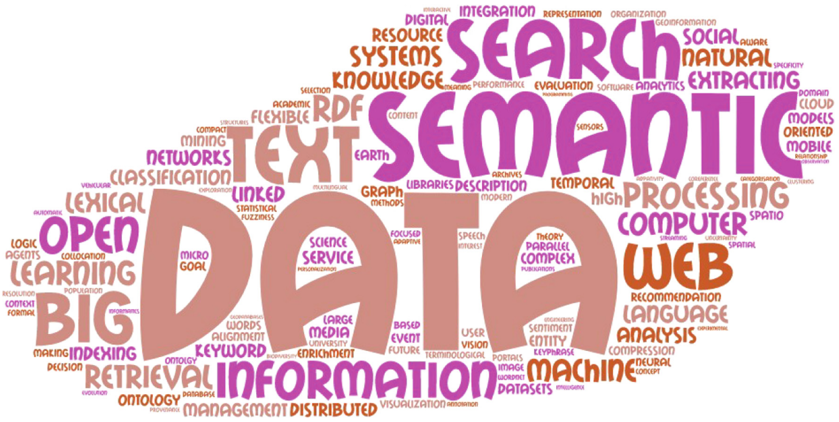

Fig. 3. Working Group Member research keywords.

In order to obtain a good idea of the actual topics on which Action participants actually work on, over the last five meetings attendees were asked to provide three keywords that they believed best characterized their work. The collected keywords (688 in total) were used to produce the word cloud illustrated in Fig. 2. As the cloud indicates, the most popular keywords were information retrieval, Semantic Web, big data, semantic search and machine learning. Figure 3 presents the same keywords but as a unigram. In this case the terms data, semantic, text, and search are those that can be clearly distinguished.

3 The KEYSTONE Activities

The COST Action's support for networking and coordination activities comes in the form of meetings, short-term scientific missions and training schools.

3.1 Meetings

KEYSTONE organized 10 meetings, with an overall participation of around 500 attendees in total, of which around 350 were financially supported by COST. From the latter, around 47% were from ITCs. In the meetings, there were brainstorming sessions where emergent techniques in keyword search research were collected, analysed, and revised. The sessions were of three types.

- Brain writing
- Round table discussions with academic and industry participants
- Theoretic hackathons

Brain writing is a way to take advantage of group priming effects through writing and reading interaction that reduces barriers with traditional brainstorming due to inhibitions inherent in face-to-face interactions [5]. The participants write down their ideas on a piece of paper, pass them on to a second participant, who reads and further develops them by adding his or her own ideas and comments, and in the sequence passes the paper to a third participant, and so on. The ideas are only passed forward, screened and further developed by different participants, without returning to the original source. Although based on the same principles, group Brain writing has been proven to be more effective than both individual brain writing and traditional brainstorming, specifically when it comes to heterogeneous groups whose members have different levels of knowledge about the issue at hand. KEYSTONE experimented with a mix of brain writing and plenary discussion sessions. In particular, a number of questions are typically defined before the meetings and revised in an initial plenary session. After that the brain writing session starts. After the session, the answers are analysed both in small groups and in plenary sessions. The results of these activities have been materialised as content published on the meeting web pages[1], and a white paper on keyword search in Big Data [3].

Round table discussions and keynotes delivered by various people from industry have been organized in almost all the meetings. The goal of these sessions was to report and analyze some interesting use cases and scenarios. Some examples include the talk delivered by Yahoo's Edgar Meij on "Web-scale semantic search at Yahoo", by Djoerd Hiemstra on "Federated search for real: combining 150 search engines, and counting" in the 1st WG Meeting[2], by Veli Bicer from IBM Ireland on "Handling city data deluge challenges and Applications" at the KEYSTONE Conference 2015[3], Radu Tudoran from the HUAWEI's European Research Centre at the KEYSTONE Conference 2016[4], and Jacek Kawalec from Voicelab at the KEYSTONE Conference 2016[5]). The Action also had calls for discussion sessions on major industrial technology needs (e.g., a panel was held in the Autumn WG Meeting 2014[6] involving four companies, and another session on "Bringing Big Data Analytics to Small & Medium Enterprises" was delivered at the KEYSTONE Conference 2015).

Last but not least, a "theoretical hackathon" was realized in another Working Group Meeting. The attendants were divided into six working groups, and were asked to develop a theoretical proposal for addressing a particular problem. All the solutions were evaluated in a plenary session at the end of the meeting. The topic of the hackathon was proposed by Mauro Dragoni from FBK (Trento, Italy). It consisted of a small dataset containing 331 documents (blog posts) and 35 queries. The documents were enriched with four metadata/semantic layers. Specifically, there was a URI Layer (links to entities detected in the text and mapped to DBpedia entities), a TYPE Layer (conceptual classification of the named entities detected in the text and mapped to both DBpedia and Yago

[1] http://www.keystone-cost.eu/keystone/outreach/meetings/.

[2] http://www.keystone-cost.eu/keystone/outreach/meetings/spring-2014-mc-and-wg-meetings/.

[3] http://www.keystone-cost.eu/ikc2015/program.php.

[4] http://www.keystone-cost.eu/ikc2016/program.php.

[5] http://www.keystone-cost.eu/ikc2017/program.php.

[6] http://www.keystone-cost.eu/keystone/outreach/meetings/autumn-2014-mc-and-wg-meetings/.

knowledge bases), a TIME Layer (metadata related to temporal mentions found in the text by using a temporal expression recogniser), and a FRAME Layer (output of an application of semantic role labelling techniques). The final goal of the theoretical hackathon was to develop a functional architecture for a system addressing the problem in hand. In particular, each component of the system had to be described in terms of inputs, outputs, and processes, i.e., algorithms to be implemented. The overall system had to be described in terms of external resources needed (if any), and in terms of the experimental evaluation process that was to be followed alongside the data (datasets, measures, etc.), and format. The results produced by the different groups have been published on the KEYSTONE website[7] and one of them has since been completed, refined and implemented in a paper accepted in the KEYSTONE Conference 2016 [4].

3.2 Short-Term Scientific Missions and Training Schools

Short-term scientific missions are institutional visits aimed at supporting individual mobility, and fostering collaboration between individuals. The Management Committee decided to assign the grants through calls issued periodically during the Action. In four years, KEYSTONE was able to fund 64 missions through 11 calls.

KEYSTONE has also organised three training schools[8], that required an investment of around 70K Euros for supporting the various missions, that was made available to different institutions.

3.3 Dissemination and Scientific Results

KEYSTONE organized a number of dissemination events. In particular, it organized three conferences (IKC2015, IKC2016, and IKC2017). It promoted these conferences via the main mailing list, alongside other specialised channels. Furthermore, the Action proposed the organization of workshops at International Conferences. The PROFILES workshop series (2014–2017) was started within the KEYSTONE Action. These workshops aimed at gathering innovative query and search approaches for large-scale, distributed and heterogeneous linked datasets in line with dedicated approaches to analyze, describe and discover endpoints, as an inherent task of query distribution and dataset recommendation. The PROFILES workshops aimed to become a highly interactive research forum, bringing together researchers and practitioners in the fields of Semantic Web and Linked Data, Databases, Semantic Search, Text Mining, NLP as well as Information Retrieval. Finally, two editions of the SDSW workshop (Surfacing the Deep and the Social Web) was also organised in 2014 and 2015. The Action has also promoted two special issues, one on Keyword Search and Big

[7] http://www.keystone-cost.eu/keystone/outreach/meetings/5th-mc-meeting-and-winter-wg-meeting-2016/.

[8] http://www.keystone-cost.eu/keystone/training-schools/.

Data published in the Springer LNCS Transactions on Computational Collective Intelligence (TCCI) Journal[9], and the second as a special issue on Dataset Profiling and Federated Search for Linked Data published with the International Journal on Semantic Web and Information Systems (IJSWIS), 12(3), 2016[10].

COST does not fund research activities, only networking tools. For this reason the results of an Action should not be evaluated based on the scientific products generated by Action Members. Nevertheless, to provide a complete overview of the activities, KEYSTONE Members were very active in publishing articles and papers in international conferences and journals. In the KEYSTONE repository, a set of around 80 papers has been collected, written by at least two people from different countries belonging to the COST Action (53 with an acknowledge to the KEYSTONE Action). Half of them have been authored by people who have met thanks to the Action tools. Finally, 17 applications to international projects, involving at least 2 KEYSTONE Members, have been submitted.

4 Conclusions

The KEYSTONE COST action was completed in December 2017. The Management Committee Members were proud and delighted with the results achieved during the Action. From a networking perspective, the Action organised a large number of events (meetings, training schools, and calls for short-term scientific missions), and a significant number of people were able to participate in these activities. Moreover, new research teams, new connections and new projects were started thanks to the Action. From a scientific perspective, the Action was able to organise a large number of dissemination events, and members were able to publish in important venues. What is interesting to note is that many networks that were created during the Action, became long-term collaborations beyond the end of the Action, which is a strong indication of its success.

Acknowledgements. This paper was supported by COST (European Cooperation in Science and Technology) under Action IC1302 (KEYSTONE), and Science Foundation Ireland under Grant Number SFI/12/RC/2289 (INSIGHT).

References

1. ICT COST Action IC1302. http://www.cost.eu/COST_Actions/ict/IC1302. Accessed 11 Mar 2017
2. KEYSTONE COST Action IC1302. http://www.keystone-cost.eu. Accessed 11 Mar 2017

[9] Available at http://link.springer.com/book/10.1007/978-3-662-49521-6 and http://www.springer.com/it/book/9783319592671.

[10] https://www.igi-global.com/journal/international-journal-semantic-web-information/1092.

3. Amaro, R., Breslin, J.G., Cardoso, J., Guerra, F., Trillo-Lado, R., Velegrakis, Y.: KEYSTONE - collecting and generating new ideas. Technical report, KEYSTONE COST Action (2015)
4. Azzopardi, J., Benedetti, F., Guerra, F., Lupu, M.: Back to the sketch-board: integrating keyword search, semantics, and information retrieval. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) KEYSTONE 2016. LNCS, vol. 10151, pp. 49–61. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_5
5. Brown, V.R., Paulus, P.B.: Making group brainstorming more effective: recommendations from an associative memory perspective. Curr. Dir. Psychol. Sci. **11**(6), 208–212 (2002)