COST

European Cooperation	**Brussels, 24 May 2013**
in the field of Scientific	
and Technical Research	
- COST -	

_____

COST 022/13

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IC1302: Semantic KEYword-based Search on sTructured data sOurcEs (KEYSTONE)

Delegations will find attached the Memorandum of Understanding for COST Action IC1302 as approved by the COST Committee of Senior Officials (CSO) at its 187th meeting on 15-16 May 2013.

_____

MEMORANDUM OF UNDERSTANDING
For the implementation of a European Concerted Research Action designated as

COST Action IC1302
SEMANTIC KEYWORD-BASED SEARCH ON STRUCTURED DATA SOURCES
(KEYSTONE)

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1.   The Action will be carried out in accordance with the provisions of document COST 4154/11 "Rules and Procedures for Implementing COST Actions", or in any new document amending or replacing it, the contents of which the Parties are fully aware of.

2.   The main objective of the Action is to establish a cooperative network of researchers, practitioners, and application domain specialists in order to promote research activity and technology transfer in the area of semantic keyword-based search on structured data sources.

3.   The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 48 million in 2013 prices.

4.   The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5.   The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of Chapter IV of the document referred to in Point 1 above.

_____

A. ABSTRACT AND KEYWORDS

As the demographic of Web users has shifted from tech-savvy early adopters towards the general population, keyword search has become the de-facto mechanism used to find information - as opposed to structured queries used in traditional information systems - even while more and more structured data has become available on the Web. This is mainly due to the simplicity of keyword search and the basic knowledge level required by users to get to some desired information. However, existing approaches suffer from a number of limitations when applied to multi-source scenarios requiring some form of query planning to access all of these sources, and with frequent updates making it difficult to produce an effective implementation of data indexes. Typical scenarios include open data, big data and virtual data integration systems. Therefore, building effective keyword search techniques can have a significant 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 - due to the multifaceted nature of the problem – the Action "semantic KEYword-based Search on sTructured data sOurcEs" (KEYSTONE) proposes to create synergies between several disciplines, such as semantic data management, the Semantic Web, information retrieval, artificial intelligence, machine learning, user interaction, service science, service design, and natural language processing.

Keywords: keyword-based search on structured data sources, metadata extraction and indexing, semantic database management and summarisation, analysis of keyword queries.

B. BACKGROUND

B.1 General background

There is a large amount of structured data currently available on the Internet, mainly in relational and semantic formats. Recent technologies and standards are shifting the Web from a "Web of Documents", where the information is published for consumption by people, to a "Web of Data", where information on the Web is also provided in a structured form for computers to make use of as well (similar to some of the structures you would find in databases). The European Union is contributing to this process: several EU FP7 projects are developing tools enabling the publication and interlinking of data on the web (for example, LATC, PlanetData, LOD2, etc.). The "Digital

Agenda for Europe 2010-2020" (http://ec.europa.eu/information_society/digital-agenda/index_en.htm) is promoting the publication and the reuse of public sector information in the form of open data, so that it can be publicly accessible by other Institutions and Enterprises (see "Digital Agenda for Europe 2010-2020", Action 3). To access this data, users and applications need to use the query language specific to the respective sources. This makes data access a laborious task, since a strong knowledge of the query language is required, but also because the semantics of the source data contents and its structures need to be well understood. Keyword-based search, on the other hand, has become the de-facto standard for finding information on the Web. Its simplicity is the main reason for its popularity. Given this, it seems natural to consider offering keyword-based search for querying structured data sources. Unfortunately, keyword queries are inherently ambiguous. The intended semantics of a keyword query needs to be discovered and expressed in terms of some structured query language. Existing approaches for keyword-based search, inspired by the area of information retrieval, are based on the construction of specialised content indexes that provide the database structures in which the required query keywords are found and that generate the structured queries that realise the semantics of the keyword query. Unfortunately, such techniques cannot be applied in all contexts. Databases on the Deep Web, for example, do not typically expose their full contents, but only a part of them, often through predefined web forms. The same happens for sources behind wrappers in integration systems. Furthermore, for sources that are typically under the responsibility of different owners, their content may change at any time without a notice to update an index structure, making the construction and maintenance of the index a hard task.

The research community has put a lot of effort into supporting the availability and interoperability among sources through techniques for publishing, semantically enriching, matching and integrating structured data both at the database level and at the Web level. At the database level, the research community has been investigating data matching and integration for about 20 years: many different theoretical approaches and a large number of prototypes have been proposed. These approaches typically rely on structural and lexical analysis of the sources, semantic knowledge extracted from external knowledge bases and ontologies, the application of reasoning techniques and description logics, statistics, and machine learning [RB01, BBR11]. At the Internet level, the Semantic Web initiative aims to build a Web of Data, enabling people to create data stores on the Web, build vocabularies, and write rules for handling data (http://www.w3.org/standards/semanticweb/). In this context, the term Linked Data refers to a set of best practices for publishing and interlinking structured data on the Web. Thanks to these techniques, it is possible to interconnect data from different repositories on the Web [HB11].

Recently, governments and government agencies have started to publish their data on the Internet, in the form of structured data sources (http://www.w3.org/TR/gov-data/). Moreover, social networks (e.g. Facebook, Twitter, YouTube) have allowed people to share large amounts of data. These data sources have a lot of inherent structure, but again keyword search is mainly applied to the text content (title and descriptions) rather than exploiting the structures and the context information available (user profiles, channel titles, etc.). Finally, a number of recent proposals have coupled information retrieval with data management [W07], and others have dealt with the extraction of structured data from web pages [CHM11]. All these works have made available to users large amounts of heterogeneous data stored in structured web data repositories.

Most efforts have not focused on providing techniques and tools for searching and retrieving data from these sources, which would have significantly increased their usability. Without these, access to structured web data is unavailable to a wider user base and research efforts aimed at making this data available on the Web risk becoming mere exercises of academic research, without any practical impact at a societal or technological level. The current state of research in this regard and the emerging requirements for broader access call for further research on keyword search over structured databases and create the perfect conditions for its development. The availability of such free, large open databases and the recently-achieved scientific and technological outcomes make the Action goals of paramount importance since users have started to demand the ability to search data and advanced technology that enables it is needed.

Providing techniques to support keyword queries over structured databases, in multi-source scenarios, and in the absence of direct access to the instances, requires synergies between several disciplines belonging to research areas which are traditionally only partially overlapping, such as semantic data management, the Semantic Web, information retrieval, artificial intelligence, machine learning, statistics, data mining, user interaction and interface design, and natural language processing. The COST scheme offers the best opportunity for carrying out the interactions required for achieving the main objectives of the "semantic KEYword-based Search on sTructured data sOurcEs" Action (KEYSTONE). The flexibility in the nature of the activities which are scheduled year-by-year, the autonomy of participants in relation to how they carry out their research activities, and the possibility to open the Action to further participants, while it is in progress, all make this COST Action the best way to integrate and merge outcomes from different research areas. Moreover, the support for training and exchanges of researchers enables the development of a critical mass of researchers in the topic that may go beyond the Action goals and duration, enabling strong relationships to be built and consequently collaboration opportunities to be created in other research activities derived and related to the Action.

[CHM11] M. J. Cafarella, A. Y. Halevy, J. Madhavan: Structured data on the web. Commun. ACM 54(2): 72-79 (2011).

[HB11] T. Heath, C. Bizer: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.

[RB01] Erhard Rahm, Philip A. Bernstein: A survey of approaches to automatic schema matching. VLDB J. 10(4): 334-350 (2001).

[W07] G. Weikum: DB&IR: both sides now. SIGMOD 2007:25-30

B.2 Current state of knowledge

Keyword search in structured databases is particularly challenging [YQC10]; firstly, because the database instances are much, much larger, and secondly, because IR-based techniques currently and successfully adopted for search in documents cannot be directly applied in this context. Structured data sources, where the information is fragmented in several classes/tables, are intrinsically different from documents. Thus, different problems have to be addressed, the main ones being: (a) the identification of the information units in the data source; (b) the identification of the informative user's needs, i.e. which data is to be returned; (c) the creation and management of effective indexes over frequently-updating sources; and (d) the fusion of information from different sources. Nevertheless, keyword search over relational data is particularly appealing. There are already interesting proposals in the scientific literature, e.g. BANKS, DISCOVER, DBXplorer, Precis and many others presented in various surveys [CSS10, YQC10, SSS07]. The typical approach is to build a special index on the contents of the database and then to use that index to identify the appearances of the query keywords in attribute values. Once the terms are located in an index, a path discovery algorithm is used to find the different ways that these query parts are located (e.g. finding minimal joining networks or Steiner trees). Nevertheless, since these techniques rely on instance analysis, their application is limited to cases where direct access or materialisation of the source data is feasible. This is not always the case for Deep Web databases, for sources in integration systems, or for big data. Moreover, current search engines retrieve their data from multiple sources, but do not perform any kind of query analysis and planning. Complex scenarios require the combination of results from different sources for answering a user query.

KEYSTONE will study common scenarios where approaches based on indexes are not feasible or effective. In these cases, only database metadata, keyword semantics and external knowledge bases may be exploited for solving queries [BDG+11, LYM+06]. The Action will study

theories and techniques for going one step further by combining techniques exploiting syntactic and structural information based on semantic similarity, statistics and machine learning, natural language processing, query relaxation and approximate query-answering techniques. Moreover, current search engines retrieve their data from multiple sources, but do not perform any kind of query analysis and planning. Complex scenarios require the combination of results from different sources for answering a single user query. This requires the ability to select the data sources completely or partially relevant for the user query, to create a query execution plan that is able to divide the query into partial queries, to fuse the partial results obtained and to eliminate duplicate answers. Many approaches have been proposed in the literature to model data sources in terms of structures, contents and query capabilities, thus summarising their informative contents for solving a user query. These approaches are typically based on structural and lexical analysis, clustering algorithms, and matching techniques [YPS09].

The main focus of this Action, namely keyword-based searches over structured datasets, can also be complemented by a new strand of research that explores the role of data provenance in this setting. In particular, provenance provides a complementary metadata space that can be used to further qualify user searches, either explicitly, i.e. by providing users with a provenance vocabulary to be used in searches, or implicitly, as a proxy for quality and reliability. Thus, the Action provides an ideal testbed where user searches can be employed to assess the practical impact of provenance on the perceived quality of the results returned by the searches. At the same time, the architectural issues (namely, of indexing) that surround keyword searches performed on structured data extend to the causal graphs that represent their associated provenance.

Finally, the Action effort can also find applications in the field of graphical tools that assist the user in formulating queries. By finding different interpretations of a keyword query, it is possible to detect related schema structures, make suggestions and guide the user in query formulation. For example, recent systems for query construction use schema information to enable users to form expressive queries over structured data beginning with the entry of some simple keywords [DZNa12, DZNb12]. FreeQ [DZNb12] enables users to construct queries over Freebase – an important dataset within the LOD cloud. Furthermore, in the case of exploratory searches, the user can use the generated interpretations as a way to explore an (unknown) data source and better understand its semantics. Starting from this background, KEYSTONE will support the combination of research findings from different areas for enabling a new scientific direction on keyword search, where it is possible to manage sources providing only schemas for their data, and where the results will not simply be ranked lists of relevant data sources, but the combination of results from partial queries executed on multiple sources.

[BDG+11] S. Bergamaschi, E. Domnori, F. Guerra, R. Trillo Lado, Y. Velegrakis: Keyword search over relational databases: a metadata approach. SIGMOD Conference 2011: 565-576.

[CSS10] S. Chakrabarti, S. Sarawagi, and S. Sudarshan. Enhancing search with structure. IEEE Data Eng. Bull., Special Issue on Keyword Search, 33(1):3–24, 2010.

[DZNa12] E. Demidova , X. Zhou , W. Nejdl. A Probabilistic Scheme for Keyword-Based Incremental Query Construction. IEEE Trans. Knowl. Data Eng., 24(3): 426-439 (2012) .

[DZNb12] E. Demidova , X. Zhou , W. Nejdl. FreeQ: An Interactive Query Interface for Freebase. Proceedings of the 21st international conference companion on World Wide Web, Lyon, France, April 2012.

[LYM+06] F. Liu, C. T. Yu, W. Meng, and A. Chowdhury. Effective keyword search in relational databases. In SIGMOD, pages 563–574, 2006.

[SSS07] Semantic Search Survey:

http://swuiwiki.webscience.org/index.php/Semantic_Search_Survey

[YPS09] X. Yang, C. M. Procopiuc, D. Srivastava: Summarizing Relational Databases. PVLDB 2(1): 634-645 (2009).

[YQC10] J. Xu Yu, L. Qin, L. Chang: Keyword Search in Databases M&C Publishers 2010.

B.3 Reasons for the Action

The development of a keyword search engine for structured sources that is able to work even without direct access to data instances has multiple benefits. From the technological perspective, building a search engine with these requirements is a challenge since no industrial-grade tool has yet been implemented. One of the main challenges is to achieve high performance in terms of response time, precision and recall. From the scientific perspective, challenges include the development of novel methods that span across areas such as keyword query semantics discovery, entity recognition, management of heterogeneous databases, query formulation and data fusion. From the social perspective, the Action achieves two important results. First, any implementation of the scientific results allows non-professional users to easily access structured repositories as they currently do with documents. Since several studies have estimated that the largest part of the data available on the Web is actually stored in structured databases, which are not completely indexed by current search engines, making these contents accessible to users has a paramount importance. Second, the query result may be provided at different granularity levels. It may be a ranked list of data sources (as happens for documents retrieved by search engines), where portions of contents of those data sources or data contents are extracted from different data sources and combined by

means of data fusion techniques. In this way, the current paradigm to search for data is revolutionised. The KEYSTONE Action can support the dissemination of scientific, technological and social results to the broader research and industrial community.

The themes addressed in KEYSTONE are still very topical and relevant in Horizon 2020. The development of techniques and tools for searching across the data available on the Web represents an important building block for the development of our knowledge-based society in Europe. This knowledge-based society is the backbone for many ICT goals that have been set by the European Commission towards improving the quality of life of European Citizens and the competitiveness of European Businesses, as they are described in the Innovation Union initiative. Innovation Union is supporting over thirty action points listed in the Communication on Innovation Union published on 6 October 2010. Among them, those most interesting to KEYSTONE are:

a. Promoting excellence in education and skill development: The skills to be acquired by the researchers in the Action are consistent with marked needs as business sector analyses show. A recent report from Strategic Business Insight (Rob Edmonds, Knowledge-Management Tools, Strategic Business Insight, May 2010) shows new emerging opportunities in the areas of Enterprise-Content Management and Enterprise Resource Discovery and Search, where, in the next few years, the sectors related to intelligent content delivery and context-aware enterprise search will grow. The report states that new technologies will provide access to "hidden" corporate contents that are currently unavailable, with tools that require less configuration and programming. KEYSTONE will contribute to the same direction. Moreover, an analysis promoted by the European Agenda for Skills and Jobs shows that there is a large number of vacancies in the ICT sector (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0682:FIN:EN:PDF) and, at the same time, there exists a clear gap between ICT-related qualifications provided by the education and training sector and those required by enterprises (http://www.cedefop.europa.eu/EN/articles/17459.aspx). In KEYSTONE, the interconnection between Academy and Industry and the particular research activities will bridge this gap and guarantee professional opportunities to the researchers.

b. Delivering the European Research Area (ERA): KEYSTONE contributes to innovation in the ERA, with international, intersectoral, and interdisciplinary research and training activities in a sector that has been considered strategic for the Digital Agenda for Europe. The KEYSTONE research goals are also related with the NESSI ICT European Technology Platforms (ETP) where the lack of interoperability is addressed by open standards, open platforms and open sources. KEYSTONE can support this ETP by providing the techniques needed for access to hidden data sources.

c. Promoting openness and capitalising on Europe's creative potential: The research results in terms of academic publications, datasets and benchmarks, code libraries, prototypes will be freely available through the Action website.

d. Spreading the benefits of innovation across the Union: The development of techniques and tools for keyword search over structured data sources is at an early stage. There is the opportunity now for making Europe a central reference point. This will be possible only by joining synergies among different countries and by joining efforts both from the academic and from the industrial side. The research outcomes, the researchers trained, and the events organised are foundation stones for building towards this direction.

e. **Exploring new forms of innovation, including social innovation and creativity:** The EC is fostering the publication of data on the Web by the Public Sector (Open data - An engine for innovation, growth and transparent governance 2011/INFSO/037). Providing easy access to open data available on the Web will be essential to promote innovation in efficient, open and citizen-centric public services (eGovernment). KEYSTONE will help to create the basis for these activities, by enabling access to and analysis of the data in open data sources. Finally, the impact of the Action will be further strengthened by a recent report from Cheskin Added Value (Industry Signals on the Future of Search, September 2010, Cheskin Added Value) showing that in the next few years the performance of search engines will improve through the exploitation of data structures. The report claims that winner applications will be those that can 1) provide good access to data from different sources (e.g. by using structures extracted from sources like Amazon, Facebook, and other social data); 2) do computations that cut across data types to figure out the meaning and join the information; 3) show and build structure in the data.

B.4 Complementarity with other research programmes

To the best of the Action's knowledge there is no project funded by any European Framework that is addressing the same issues addressed in the Action. However, there are other research projects that face issues complementary to the Action. This is an advantage for the Action, as it may directly benefit by involving researchers from these projects in the Action (as experts in the open working group meetings, as teachers in training schools, or as members in the Action) or, indirectly, by exploiting the techniques and the prototypes developed. These projects may support KEYSTONE by providing techniques to be evaluated and extended, and datasets to be used for the internal experiments. In particular, the main sibling projects are:

COST ICT 1002 MUMIA: MUltilingual and Multifaceted interactive Information Access. This is a

COST Action providing coordination in themes related to search technology, and in particular approaches to search, filter, extract, combine, integrate, and process multiple and distributed sources of multilingual content, delivered to an even wider global audience and variety of population, are studied. In contrast to this Action, MUMIA's targets are mainly web pages, patents search, and other next-generation web applications.

COST IS 1004 WEBDATANET: web-based data collection. This is a COST action providing research coordination to address web-based data collection (surveys, experimenting, testing, non-reactive data collection, and mobile Internet research), its methodologies, scientific validity, and use in the social sciences. The web sources defined in this Action will be evaluated in KEYSTONE as possible benchmarks for keyword search engines.

ERC SECO: Search Computing. This project aims at producing the answers to complex search queries by interacting with cooperating search services, using ranking and joining of results as the dominant factors for service composition. The project is developing new foundational theories, new language and description paradigms for expressing queries and for discovering services, new interfaces and protocols capturing ranking preferences and enabling their refinement. The project also studies the economical and legal implications of search computing. The techniques developed in SECO will be evaluated with the Action's benchmarks.

FP7 LOD2: creating knowledge out of interlinked data. The project aims to develop: adaptive tools for searching, browsing, and authoring of Linked Data; enterprise-ready tools and methodologies for exposing and managing very large amounts of structured information on the Data Web; algorithms based on machine learning for automatically interlinking and fusing data from the Web; and standards and methods for reliably tracking provenance, ensuring privacy and data security, as well as for assessing the quality of information. The project provides interesting techniques to be evaluated in the Action.

FP7 TATOO: Tagging Tool based on a Semantic Discovery Framework. This will set up a semantic web solution to close the discovery gap that prevents full and easy access to environmental resources on the web. In particular, the tools project is developing tools to easily discover environmental resources (data and/or services residing on different information nodes) on the Web and to add valuable information in the form of semantic annotations to these resources, thus facilitating future usage and discovery, and kicking off a beneficial cycle of information enrichment. The techniques developed in the project will be evaluated in KEYSTONE.

FP7 I-SEARCH: A unified framework for multimodal content SEARCH. The project aims to provide a novel unified framework for multimodal content indexing, sharing, search and retrieval, oriented to handle specific types of multimedia and multimodal content. Moreover, I-SEARCH

introduces the use of advanced visual analytic technologies for search results presentation in order to facilitate their fast and easy interpretation and also to support optimal results presentation under various contexts (i.e. user profile, end-user terminal, available network bandwidth, interaction modality preference, etc.). The techniques and datasets developed in I-SEARCH will be evaluated in the Action.

FP7 LarKC: The Large Knowledge Collider. This is a project that ended in 2011 and that built a platform for massive distributed incomplete reasoning, removing the scalability barriers of currently existing reasoning systems for the Semantic Web. The LarKC platform is open source and the functionality of the platform will be studied and potentially adopted/extended for querying structured databases.

FP7 PlanetData. This project aims to establish a sustainable European community of researchers that supports organisations in exposing their data in new and useful ways. The project aims to efficiently make sense out of the enormous amounts of data continuously published online, including data streams, (micro)blog posts, digital archives, eScience resources, public sector data sets, and the Linked Open Data Cloud. The interaction with PlanetData is crucial for KEYSTONE, since the datasets provided by this Network of Excellence will be evaluated as possible benchmarks for keyword search tools.

FP7 NOE Promise. PROMISE will provide a virtual laboratory for conducting participative research and experimentation to carry out, advance and bring automation into the evaluation and benchmarking of complex information systems, by facilitating management and offering access, preservation, re-use, analysis, visualisation, and mining of the collected experimental data. PROMISE aims to foster the adoption of regular experimental evaluation activities, bring automation into the experimental evaluation process, promote collaboration and re-use over the acquired knowledge base and stimulate knowledge transfer and uptake. The collaboration with members of this network will be really useful for WP4, where benchmarks for evaluating keyword search engines will be defined.

SoFWIRed, Southampton Fraunhofer Web Science, Internet Research & Development, is a common project of the University of Southampton together with Fraunhofer. The goal is to leverage the value of open data for e-business by exploring the value of open data in supply chains and logistics, and to deploy mechanisms for public engagement in collecting and leveraging open data related to e-government. The open data exploited in this project can be used as datasets for the tools / techniques developed in the Action.

C. OBJECTIVES AND BENEFITS

C.1 Aim

The main objective of the Action is to launch and establish a cooperative network of researchers, practitioners, and application domain specialists working in fields related to semantic data management, the Semantic Web, information retrieval, artificial intelligence, machine learning and natural language processing, that coordinates collaboration among them to enable research activity and technology transfer in the area of keyword-based search over structured data sources. The coordination effort will promote the development of a new revolutionary paradigm that provides users with keyword-based search capabilities for structured data sources as they currently do with documents. Furthermore, it will exploit the structured nature of data sources in defining complex query execution plans by combining partial contributions from different sources.

C.2 Objectives

The main objective of the Action is complemented by the following secondary objectives:

1) Promote the development of novel techniques for keyword-based search over structured data sources. In particular, the Action aims to:

   a) promote a critical analysis of the emergent techniques for keyword-based search in structured data sources, identifying and critically reviewing the most promising approaches, such as

      - describing and summarizing data source contents

      - discovering the intended meaning of the user keyword queries

      - matching the user keywords with data structures and domains

      - fusing, ranking , and visualizing the results

   b) enable the creation of networks/discussion forums/mailing lists with the support of new web technologies, e.g. semantic wikis and semantic microblogs, for sharing ideas and promoting joint and multidisciplinary ideas, and promoting joint and multidisciplinary research

c) enable the development of a conceptual map of the open and closed issues related to the Action themes

d) support the discussions and idea sharing among early-stage and senior researchers by means of training schools and symposia, and by promoting visiting activities and exchanges through Short-Term Scientific Missions

e) support the development of scenarios, showcases and benchmarks for evaluating and comparing the approaches and the prototypes

2) Facilitate the transfer of knowledge and technology to the scientific community, practitioners and the enterprises, through:

a) the dissemination of the results through high quality conferences and journals

b) the development of targeted publications, newsletters and discussion groups

c) communications exploiting the capabilities offered by the most common social networks (e.g. by means of the creation of a Twitter channel, Facebook Page, LinkedIn group, YouTube channel, etc.)

d) the participation and organisation of meetings and events with people from industry or government promoting the reuse of the Action outcomes in professional applications

e) the development of scenarios, use cases, benchmarks, prototypes and libraries of source code

3) Build a critical mass of research activities and outcomes that achieve the sustainability of the research themes beyond the Action through:

a) the collection of the most relevant material (slides, tutorials and other training materials, scientific papers, prototypes, datasets, scenarios and benchmarks) developed within the Action in an open annotated bibliography for improving their dissemination, reuse and extension

b) the enlargement of the network of academic and professional researchers beyond the ones initially involved for the preparation of the Action

c) the development of lectures for Master Thesis and PhD courses, guaranteeing the persistence and circulation of the methods and applications relevant for the Action

Other broader objectives addressed in the Action are:

- to promote the publication, use and sharing of open structured data through the web

- the involvement of internationally acknowledged research centers working in the Action themes

- the development of cooperation relations among European research institutions to perform joint activities beyond the ones proposed in the Action (the participation in other research proposals, the development of joint doctorate courses, …)

C.3 How networking within the Action will yield the objectives?

The Scientific Programme, which divides the Action in three vertical thematic areas, each one covered by a Working Group (WG1 to WG3) and a horizontal activity (WG4) extending across the three areas, has been conceived for achieving the main and secondary objectives and for responding to the specific needs of the Action. The network of experts participating in these Working Groups represents a vast collection of expertise in all the aspects important for the themes of the Action, and is composed of renowned researchers working in relevant research projects funded at the national and European level, that ensures that the Action is based on substantial and sustainably

funded scientific activity during and beyond the life of the Action. Practitioners from industry are asked to participate in the Action, thus ensuring the availability of real scenarios, the exploitation of the outcomes in real applications and the technology transfer of the achievements.

KEYSTONE supports the following actions for the achievement of its objectives:

Activity 1: Setup of a common repository of knowledge, supporting the development of an open annotated bibliography of the relevant literature and a collection of the materials produced by the Action

Activity 2: Organization of intra - Working Group meetings for sharing the research achievements involving, when required, external audience from related disciplines. These meetings may take the form of virtual meetings keeping an internal update on the current activities at different groups

Activity 3: Organization of open workshops for the dissemination and the discussion of the outcomes achieved within the Working Groups to a larger audience of researchers, practitioners and experts belonging to other Working Groups and / or external to the Action

Activity 4: Organization of training and disseminations activities including tutorials, summer schools and industrial events

Activity 5: Exchanges of early stage researchers for visiting other partners, performing joint researches and sharing methodologies. This action may lead to Ph.D. courses within joint programs and to the co-supervisions of Ph.D. students by members of the Action in different countries

Activity 6: Visiting activities for senior researchers to coordinate KEYSTONE, identifying the relevant open issues in the field, the promising research directions, and the methodologies to be exploited

C.4 Potential impact of the Action

This is intended to be a high-impact Action both at the academic and at the industrial level. Its objectives have a high potential, enabling:

- Networking in complementary disciplines leveraging innovation in an upcoming research area, by

    o the creation of a critical mass of research in all the phases of keyword search, i.e., 1. query meaning discovery, 2. selection of the sources and query formulation, and 3. visualization of the results

    o the development of techniques and tools that span across several research areas such as keyword query semantics discovery, entity recognition, management of heterogeneous databases, query formulation and data fusion

- Bridging the gap between the academic research and the industry, by

    o the support of research groups promoting research, training and technology transfer (by means of tutorials, PhD courses, Master Thesis, academic and training lessons) in hot challenging topics

    o enabling the development of professional tools and services for keyword searching over relational databases providing advanced functionality for the end users to explore and search in unknown, frequently updated databases

    o promoting the creation of new professional opportunities for early stage researchers belonging to the Action, since 1. the cross disciplinary nature of the Action programme equips them for careers in working sectors related to the content management, discovery and search; 2 the Action helps fellow developing a solid scientific basis with in-depth knowledge of structured data sources. This allows the researchers to be flexible and the skills to be transferred in related sectors.

- Enabling the use, reuse and sharing of the information available on the web, by

   o the development of a new paradigm for searching for data that allows users to combine results from different sources and to include in the results data sources that do not directly expose their contents

   o developing tools for searching in the large mass of structured and semantically rich data that the researches in the areas of Semantic Web and Linked Open Data made available

C.5 Target groups/end users

The effects of the Action are mainly concerned with two kinds of users:

1. **Academic and industry researchers** that benefit from the techniques, prototypes, use cases, scenarios and benchmarks that are developed and made available in the Action. For this purpose, a particular attention is devoted to the development and the constantly updating of an open and annotated bibliography with the relevant publications in the field. Research communities different from the ones directly involved in the Action may also reuse the achieved outcomes in different contexts

2. **End users and applications** that exploit the results for having access to the large mass of data made available by databases that do not provide direct access to their contents (sources from the deep web and database results of a virtual data integration process), structured data sources where the schema and contents are unknown (tables and open data made available from the websites of statistical, research and governmental Institutions and newspapers), and databases which are subject to frequent updates, where the costs for maintaining search applications based on data analysis updates may be too high to be effective.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

KEYSTONE focuses on reviewing, classifying, coupling, designing, developing, and evaluating techniques dealing with keyword-based searches over structured data sources. Despite the effort that the research community has put on the field, several issues with a large impact at the scientific, technological and societal level are still open. The most relevant ones concern the requirement for current approaches to interpret a query language (typically SQL or SPARQL) and understand how information is represented in the data sources, i.e. the data source schemas and contents (societal impact); the inapplicability of current techniques in several scenarios that do not allow the application of instance analysis, e.g., the deep web, virtual integrated sources, and frequently updated databases (scientific/technological impact); the absence of standardized benchmarks for evaluations and comparisons (scientific/technological impact); the need of more expressive techniques for formulating keyword queries and visualizing results (societal/scientific/technological impact); the exploitation of query execution plans for solving queries in multi-source scenarios collecting and fusing the results from different sources (scientific/technological impact).

The Action goal is to study algorithms and techniques that are applicable for keyword search over structured data sources in different contexts with different preconditions, thus providing new solutions where current techniques cannot be applied. This requires synergies from several disciplines, such as, e.g., data management, semantic web, information retrieval, artificial intelligence, machine learning and natural language processing.

KEYSTONE aims to bridge the gap between unstructured strings of keywords used for search and the fixed structure that organizes data instances in a database. The Action will study how to reduce the gap size by creating a higher level of abstraction over database schemas using metadata. Data mining for knowledge extraction, statistics for summarization, and natural language processing can be used to create the metadata. For example, clustering algorithms (e.g. K-means, distribution- and density-based clustering) from the data mining field can be used to segment database instances into predefined groups which will be later labeled with metadata. These labels enable to create a conceptual layer which is semantically richer and closer to keyword-based query strings. To further close the gap, techniques for automatically annotating keywords of query strings using domain and upper ontologies (such as DOLCE) will be investigated. These two activities, bottom-up metadata generation and top-down keyword annotation dramatically reduces the gap between unstructured and structured forms of data representation. Ontology alignment (e.g. FOAM framework) and hybrid schema matching algorithms (e.g. COMA++) can provide relationships between metadata and ontologies. Semantic connections between annotated keywords and instances, sets of instances, or database attributes will enable to trigger SQL queries based on keyword-based queries.

Adopted approach

In KEYSTONE, it is adopted a modular approach where the issues related to keyword-based searching are grouped into three main thematic areas, each one analyzed by a working group (WG1-WG3). The same tasks will be addressed in all the areas and will concern the development of a solid state of the art with the aim of defining the existing open and closed issues in the field, the design of a roadmap for the achievements of the outcomes established in the Working Group, the coordination of research activities and the dissemination of results.

The experts participating in the Action belong to different areas of the research and industry community. This is definitely an advantage, since it allows researchers to share the knowledge and the methodologies typical to one discipline and to promote cross-fertilization among multiple disciplines. Nevertheless, an initial effort is required for the background harmonization among partners and the formalization of the goals and issues in order to establish a common research agenda. Moreover, the Action will create a website where an annotated bibliography of the main results available in the literature along with the results achieved in the Action will be published. The goal is to make this knowledge base an important reference point for future research activities in the field.

Thematic areas

The Action will review, design, develop, implement and evaluate techniques in all the typical three phases of keyword-based search: a) user keyword input; b) result computation; and c) result output. For each phase, the most promising approaches will be analyzed and new approaches will be proposed for (i) analyzing the user keywords by identifying the concepts related to them and generate lexical alternatives that may make sense for each data source; (ii) matching user keywords with the underlying data structures in the sources; (iii) formulating queries in the native data source languages corresponding to the user keywords; and (iv) performing the fusion, cleaning and ranking of the results of the possible queries generated during the previous step.

These aspects will be addressed in the Action within three vertical thematic Working Groups (WG1 to WG3) and a horizontal Working Group (WG4) extending across the three areas. In particular, WG1 refers to the definition of metadata for describing data sources, WG2 is about the keyword search process, WG3 analyzes issues related to user interactions, and WG4 concerns the coupling of the techniques developed in each single Working Group for the creation of an effective framework and the identification of techniques for the evaluation of the approaches.

Research tasks addressed

The tasks that the thematic areas address are:

1. Critical review of the existing emerging techniques. Common criteria for evaluating the approaches developed in the literature are studied. The goal is the creation of an open annotated bibliography of the most important approaches and techniques that have been developed

2. Definition of open/closed issues in the field and definition of a roadmap for achieving the Action ouctomes. Nowadays, keyword-based search over structured data sources is a hot and challenging research topic. Some approaches have been already proposed in the literature, but some interesting and critical issues are still open. For a high quality research, the important outcomes have to be defined and a roadmap for proposing solutions and approaches has to be established

3. Coordination of research activities. This task refers to the coordination of the research activities that will provide techniques and prototypes in the three vertical thematic Working Groups of the Action and the design of scenarios and evaluation benchmarks for the horizontal Working Group

4. Coordination of short term scientific missions. Short term missions are an indisputable opportunity offered by the action for early stage researchers of sharing research activities and methodologies with other groups. The missions have to be balanced among the members of the actions and the research directions established by the Management Committee

5. Dissemination of the outcomes (results, tools, showcases, benchmarks). The aim of the Action is to provide a complete reference study about the issues related to keyword search over structured data sources such as relational databases. This implies to be able to disseminate the outcomes to the research community. Results of the Action include publications in the most important international journals and conferences and a freely accessible website where an annotated bibliography, reference datasets and query sets, scenarios, benchmarks, prototypes and software libraries will be available. The website will be designed in order to enable communications and discussions among the Action partners and external researchers interested in the Action fields.

D.2 Scientific work plan – methods and means

The Action will address three vertical thematic areas, each one covered by a Working Group (WG1

to WG3) and a horizontal activity (WG4) extending across the three areas.

WG1 Representation of structured data sources. This WG investigates possible metadata describing data sources and efficient structures and indexes for their retrieval. Metadata available a priori at the data source, available at run time and extracted a posteriori will be taken into account. A lot of effort in providing techniques for publishing structured data sources on the web in a semantically rich form, making them sharable by different users has been provided. The approaches proposed in this area will be carefully and critically reviewed in the Action and exploited and/or extended.

WG2 Keyword search. This WG studies techniques for matching the user keywords with data structures and domains of selected sources and formulating the corresponding queries. This roughly implies to address three main issues: 1. Selection of the most relevant sources for each specific user query; 2. Discovering the structures containing the keywords and how these structures may be joined in order to formulate a set of queries, expressed in the native structured query language of the source (e.g., SQL, SPARQL) approximating the user intended meaning of the user query; 3. Fusing, cleaning and ranking the results obtained by the execution of the queries computed in the previous step.

WG3 User Interaction and keyword query interpretation. This WG investigates issues related to the semantic disambiguation of the queries based on the context and on the keyword annotations with respect to some reference ontologies, the development of languages for keyword searching and the use of users' feedbacks for improving results. Moreover, the WG studies techniques for identifying the "scope" of a keyword query, i.e. determining what are the data source elements to be returned to the user and in which form (e.g., in a graphical way).

WG4 Research integration, showcases, benchmarks and evaluations. This WG aims to integrate the activities of the other WGs with the goal of creating a "vademecum" for developing a search engine for structured data sources. The idea is to develop a set of scenarios and, for each scenario, when available, a list of the functionalities required, solutions proposed in the literature, prototypes and software libraries, datasets and query sets. Besides this task, the Working Group will study benchmarks and criteria for evaluating prototypes. Moreover, data coming from real scenarios, e.g. provided by a Portuguese telecommunication company that demonstrated its interest on the Action, will be used for testing the prototypes.

E. ORGANISATION

E.1 Coordination and organisation

The management and the organizational strategy of the Action exploit the typical COST Action tools and, in particular, is defined by the following organizational frameworks:

- **Management Committee (MC) and Executive Scientific Board (ESB).** The MC is in charge of supervising and coordinating the Action. According to the COST 4154/11 document, it is composed of not more than two representatives of each participating Country. A Chair and a co-Chair of the Action are elected by the MC. Moreover, the MC has to nominate a Scientific Coordinator of the Action and a Coordinator of the Dissemination activities. The responsibility for planning and executing of the activities can be delegated by the MC to the ESB, composed of the Chair and the co-Chair, the Scientific and Dissemination Coordinators of the Action plus the Working Group Leaders.
- **Editorial Board.** It is composed of a sub-set of the MC members and is in charge of planning the publication strategy, reviewing the Action publications and deliverables, and defining the website contents.

The following Committees have the main role of implementing the main tools of the Action, which are:

- **Working Groups**, where interesting outcomes resulting from research carried out in the participating countries are shared, and the knowledge transfer and the research exchange is undertaken. The research coordination is implemented through meetings (at least one per year, see Section F.), where activities and reports are decided and organized. The MC selects a Coordinator for each WG with the role of assuring the achievements of the goals within the established deadlines and the integration of the activities with the other WGs.
- **Short Term Scientific Missions (STSMs) and Training Schools.** STSMs allow young researchers to share expertise and adopted methodologies with other research groups, and senior researchers to evaluate the research carried out in the other groups and plan complementary and integrated research activities with them. The Training Schools organized in the Action have a twofold goal. First, some sessions provide fundamental knowledge about the Action topics with the goal of harmonizing the skills of the participant in the Action. For this reason, they are typically oriented to Young Researchers. Second, other more technical sessions may allow researchers to gain detailed and updated knowledge about some specific Action topics. The Training Schools include seminarian sessions where

the researchers can share their perspectives, integrate their vision and depict new scenarios and directions for the Action. A STSM Coordinator and a Training Coordinator appointed by the MC are responsible for planning these activities.

- **Conferences and Workshops**.   A final conference will be held jointly among all the WGs, aiming to summarize and disseminate the Action outcomes. Workshops and tutorials will be proposed in conjunction with the most important conferences in the field in order to increase the relevance of the Action, create a critical mass of researchers in the field and disseminate the intermediate results.

- **Website.** It represents the most important source of information about the Action and thus will continuously maintained updated throughout the lifetime of the Action. Information about results, innovations and applications generated are published on the website in order to disseminate the results to experts and professionals. The website also implements tools from the Web 2.0, enabling the discussion of the outcomes among the partners and external interested experts and researchers. A website restricted area implements a collaborative work platform for the communication within the WGs, the publication of internal reports, minutes, draft of publications, schedules and calendars of the activities.

The Action encourages and promotes the participation of young researchers not only as researchers and participants to the activities (as WG members, participants in the Training schools, …), but also as Action coordinators (WG, STSM, Training, Scientific Coordinator,…). This is motivated by two main aspects: first, most of the researchers involved in the Action proposal are young researchers and, second, the passion and enthusiasm of young researchers may, in case, supply their minor experience. Nevertheless, it will not constitute a risk for the success of the Action, since Senior researchers will participate in all the activities of the Action and can redress any possible inexperience.

Milestones

The milestones of the Action include:

1. The kick-off meeting, where the organizational structure is defined, as well as the research agenda, work plan and task distribution
2. WG annual meetings, where the state of the art of the research and the activities carried out by the participating members are evaluated, and the forthcoming tasks are planned
3. Final conference, that will provide a comprehensive overview of the results of the Action. A

particular attention will be paid in the design of activities (participation in other call,) for continuing the research in the field beyond the end of the Action

4. Final publication, i.e., a book summarizing the main achievements reached in the Action

E.2 Working Groups

The Action will address three vertical thematic areas each one describing a specific aspect of the keyword-based search and that is covered by a Working Group (WG1 to WG3) and a horizontal activity (WG4) extending across the three areas.

- WG1 Representation of structured data sources.
- WG2 Keyword search.
- WG3 User Interaction and keyword query interpretation.
- WG4 Research integration, showcases, benchmarks and evaluations

In Section D, the specific tasks that will be carried out by the Working Groups are detailed. From the organizational perspective, the MC will appoint a coordinator for each WG. The Action members will decide to freely participate in a WG on the basis of their research interests. The main WG task is to coordinate the research activities and organize the meetings, i.e. the place where ideas and proposals may be exchanged and discussed, in case, in interaction and coordination with the other WGs. Since several aspects of the Action may require a multi-disciplinary approach, external experts, training activities and invited talks may be foreseen in some meetings. Finally, some virtual meetings, typically concerning small specific parts of the WG actions, are organized.

E.3 Liaison and interaction with other research programmes

In section B.4 the main EU FP projects of interest for the Action are listed. It is really important for the success of the Action to establish mutual relationships with researchers involved in these projects. Some of them have already expressed their interest in participate in the Action. The MC has the responsibility to manage the relationships with other FP research projects and to proactively search new possible collaborations with other existing or future initiatives and programs.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance in all its activities and the Management Committee will place this as a standard item on all its MC agendas. It is well-known

the gender imbalance in IT, mainly due to stereotypes related to IT jobs and lack of flexibility (http://www.usnews.com/news/blogs/stem-education/2012/01/03/wanted-technical-women). All actions foreseen in the project will be in accordance to the Council Resolution (20 May 1999) on women and science, The Council Resolution on Science and Society and on Women and Science" (OJ C 199 of 14.7.2001). KEYSTONE aims to increase the participation of women in the variety of the different activities that concerns the various researchers, stakeholder and targets' groups. Having in mind that the share of the women across Europe who are employed in/ dealing with engineering occupations (e.g. teaching, laboratory work, research and industry), or are studying in engineering and informatics-related specialities/classes is much lower than men's presence, the Action will aim at purposeful and proper addressing the gender dimension issues evoked by the specific nature of the Action thematic.

The Action will also be committed to considerably involve early-stage researchers. This item will also be placed as a standard item on all MC agendas. The Action promotes the participation of young researchers not only as researchers and participants to the activities, but also as Action coordinators. The goal is to give young researchers the possibility of acquiring not only research but also managerial skills.

Keywords-based search research is a fairly emergent field and in common with many IT related fields currently has an unequal gender distribution, particularly in academia. The Action will aim to address this, and encourage the involvement of early-career researchers, by:

- Ensuring that key Action positions are shared as equally as possible by men and women
- Reserving positions of WG leaders to Early Stage Researchers
- Holding Young Researcher Forums in tandem with WG meetings
- Scheduling meetings outside weekends and evenings where possible to mitigate childcare issues
- Ring-fencing places at the Actions Conferences
- Focusing STSMs and Training Schools on PhD students and embryonic stage researchers

The Action makes special provision for supporting early stage researchers and PhD students (as stated previously, through training schools, competitions, and STSMs). This is essential for this Action, and for the build-up of research into keyword search, database management, open data and big data analysis, with its long term, ambitious goals. While training schools will be the primary focus for the assembly of a community of early stage researchers in ICT, the Action will make provision for dedicated web sites, mail lists and/or social networks to help develop a supportive relationship among younger researchers in these fields throughout Europe. Additionally, workshops

will encourage the involvement of early stage researchers by the inclusion of "student paper" categories in their call for papers.

F. TIMETABLE

The Action lasts four years. MC meetings will take place at least once a year with a minimum of six meetings during the Action. During the Kick-Off meeting the Executive Scientific Board and the responsibility roles will be appointed. By the end of the first four months, an initial website will be made available and updated during the Action. Each Working Group will meet at least once a year (6 times during the Action). STSMs are planned after the first WG meetings, once the participants have identified common research interests. Three Training Schools are planned in the Action. WG and MC meetings are organised at the same time and place to reduce travel costs and organisation costs, and to promote interactions between the WGs. The main activities in the Action are shown in the following table.

	Year 1			Year 2			Year 3			Year 4		
MC Meetings	*		*	*			*			*		*
WG Meetings	*		*	*			*			*		*
Training Schools					*			*			*	
Industrial Days / WG Open Meetings			*			*			*			
Final Conference												*
Annual / Final Reports		*			*			*				*
STSMs for Early Stage Researchers	At least 12											
STSMs for Senior Researchers	At least 12											

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: CH, DE, EL, ES, FI, FR, IE, IT, NL, PL, PT, UK. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has

been estimated at €48 Million for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

H. DISSEMINATION PLAN
H.1 Who?

The Action aims to develop effective keyword search techniques for querying relational databases even when the database instances cannot be directly accessed. The Action participants agree with the European Charter for Researchers recommendations: in particular, all partners will place emphasis on the outreach activities by firmly sustaining "dissemination, exploitation of results". Horizon 2020 shares the same goal: "A key added value of research and innovation funded at the Union level is the possibility to disseminate and communicate results on a continent wide scale to enhance their impact"- COM(2011) 811 final 2011/0402(CNS), "Promoting innovation through, inter alia, open access to scientific publications, a database for knowledge management and dissemination and promoting technology topics in educational programmes" - COM(2011) 812 final 2011/0400 (NLE). For this reason, the Action will put a high effort on supporting impact through dissemination. The outcomes will be interesting for the international research community and for practitioners in industry who may reuse and extend the techniques, prototypes, showcases and benchmarks produced. Outcomes will also be useful for end users who will derive benefit from new tools for easily accessing data that has thus far been difficult or even impossible to obtain. The dissemination of the activities and goals achieved is one of the main Action tasks. A Dissemination Coordinator (as described in section E) is in charge of designing and monitoring an effective strategy, maximising impact on the following target audiences:

1.   **The International research community** carrying out research on structured databases, the Semantic Web, the Deep Web, user interfaces, data mining, statistics, natural language processing will be interested in the Action topics. Dissemination can enable the reuse of the outcomes in different scenarios and / or encourage new participants in the Action.

2.   **Industry professionals** may be interested in the Action for implementation of Action results in commercial applications. The scenarios, benchmarks and showcases developed in the network may also be interesting for evaluating relevant industrial applications.

3.   **End users (practitioners and decision makers)** will be interested in the prototypes and tools developed in the Action for obtaining access to information thus far unavailable. Several organisations, public institutions and enterprises are starting to collect and publish data on the Web in a structured form.

4.   **Education Institutions** may benefit in their Masters and PhD degree courses from the knowledge accumulated in the Action and the training materials made available.

H.2 What?

The Action outcomes will be disseminated through events / publications oriented to different targets (academics, industrial researchers, end users) and with different levels of technical detail, as summarised in the following lists:

Publications:

▪   Scientific publications (to be sent to prestigious international conferences such as SIGMOD, VLDB, CIKM, ICDE, EDBT, ISWC, ESWC, WWW, and others)

▪   Open bibliographies detailing the main publications in the field

▪   Website, including social tools for enabling discussions among the researchers

▪   Showcases, use cases, benchmarks

▪   Training materials for the schools organised by the Action, for PhD and Masters courses taught by people involved in the Action or who are interested in the Action contents

▪   Mailing lists / newsletters

▪   Flyers

- Final report (book)

Events:

- Organisation of workshops and tutorials at the main international conferences

- Training schools

- Industry days

- Open WG meetings

H.3 How?

The Action website will be the main tool for dissemination. The website is to be built and kept updated under the responsibility of the Dissemination Coordinator. It shall contain all information about the Action, including goals, outcomes, researchers involved, research activities under development, work plan, events, etc. All the dissemination material produced by the Action (papers, technical reports, training materials, newsletters, etc.) will be directly accessible through the website. Tools based on social web technologies are to be deployed, enabling discussions among the partners and external / interested researchers. An open bibliography, built during the Action and including the main publications in the field, is to be hosted in the website. Mailing lists will be created for supporting communications among the MC and WG members, and also for interested people that want to receive information about the Action results. The website provides indications about how to participate in the Action and become an active member.