LarKC Annual Report

http://www.larkc.eu

The objective of LarKC is to go beyond the limited storage and inference solutions currently available for semantic computing, and to fulfil needs in sectors that are dependent on massive heterogeneous data sources such as telecommunication services or bio-medical research. For this purpose, LarKC develops an infrastructure that extends the current reasoning paradigms which are strictly based on logics by fusing reasoning with techniques from information retrieval, machine learning or even economics. The resulting platform, the Large Knowledge Collider, is a pluggable framework for reasoning at Web-scale, implemented on a distributed computational platform, and engineered to scale to very large distributed settings. LarKC trades quality for computational cost by embracing incompleteness and unsoundness.

Summary of Activities

Around a stable LarKC platform, the consortium is building a marketplace of plug-ins of various kinds. Leveraging cognitive science, statistical semantics and information retrieval theories, novel selection strategies are developed which reduce the problem space for reasoning tasks. Interleaving reasoning and selection is an important aspect of LarKC, and investigations continue to seek better performance at lower cost – without too much loss in quality. In terms of reasoning, LarKC is working on inductive methods based on machine learning, and various deductive methods including parallelized engines and heuristic methods. As a whole, LarKC has reached a state of maturity in which all these methods can be plugged via the LarKC platform to serve the project's use cases that have in common their demand for large scale reasoning over dynamic, inconsistent and incomplete data. While so far, the listed efforts were conducted in partial isolation, the project is now in the position to provide comprehensive reasoning experiments, and is ready to target new domains and problems. The project welcomes new early-adopters of the LarKC technology!

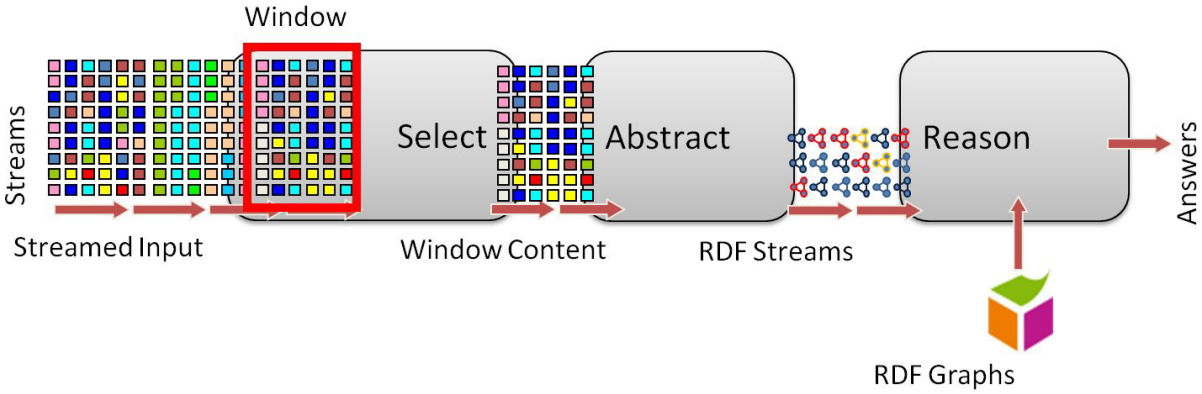
Data Selection Strategies

The work on selection strategies is driven by the requirement to reduce the problem space for reasoning by applying cognitively inspired models and information retrieval theories to large RDF graphs. LarKC offers real-time interest-based selection from Web streams to determine the most relevant data items for a particular task or query. Statistical semantics, in a related approach, is applied for random indexing and the generation of virtual documents from RDF graphs. This approach was evaluated with domain experts from the consortium member AstraZeneca. Selection is also addressed from a more human-centred perspective. LarKC produces an infrastructure to use human judgment as a gold standard for the rating of answers to queries. The idea is to add a 'star rating' to query result in order to allow users to report good or bad answers. This is a major redesign for linked data querying, and very important to make informed decisions on how satisfying the current state-of-the art is for users.

Novel Reasoning Methods

Reasoning at Web-scale is the objective of LarKC, and various novel means of reasoning are developed within the project, including inductive and deductive methods. SUNS, for example,

is a scalable machine learning and reasoning framework based on prediction methods. Inductive reasoning is applied to discover potential truth relationships between instances in very large and complex RDF graphs. Other methods apply heuristics to achieve Web-scale reasoning. Knowledge summarization techniques, for example, are a means to achieve scalable inference via the interleaving of selection and reasoning. Entirely novel, and founded by LarKC, is the work on querying RDF data streams. LarKC has developed an engine that is able to efficiently interpret C-SPARQL queries (deductive reasoning), and to present query results as linked data. Although very different in their nature and approach, all of these reasoning algorithms can be integrated via the LarKC platform.



Parallelization of Reasoning Algorithms

WebPIE (Web-scale Parallel Inference Engine) is a MapReduce distributed RDFS/OWL inference engine written using the Hadoop framework. The engine applies RDFS and OWL Horst rules and materializes the derived statements from RDF graphs. WebPIE is the first engine that can inference over 100 billion triples. WebPIE encodes the reasoning task as a set of MapReduce operations for performance, fault-tolerance and simplified administration. As such, it aims at high scalability in terms of the number of processing nodes and data size, which is achieved by optimizing the required joins. WebPIE won the first prize in the SCALE Challenge at the 10th Int'l Symposium on Cluster, Cloud and Grid Computing. With similar goals in mind, LarKC is now working on parallelizing IRIS, the first open-source Java-based Datalog engine. While WebPIE focused on very large datasets for the fixed rules of RDFS and OWL, the work on ParIRIS addresses potentially arbitrary RIF rule bases.

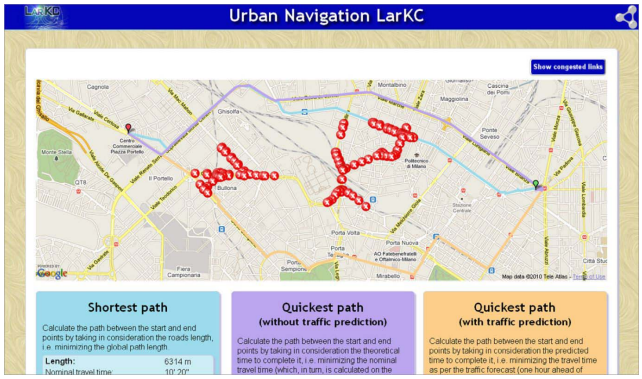
LarKC Platform 2.0

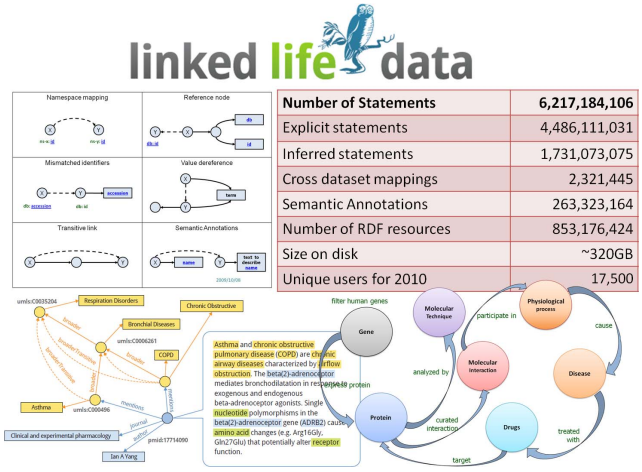
The goal of the LarKC platform is to provide an open, flexible, lightweight and scalable infrastructure for semantic computing and advanced reasoning. Since the beginning, the LarKC platform has evolved from a simple linear pipeline to a flexible workflow execution environment. Thanks to the clear separation of workflow specification and execution (workflows are described in RDF, as are the plug-ins constituting a workflow), and new features such as remote execution and distribution, data streaming and caching, as well as splits and merges of data flow for automatic parallelisation, fairly complex scenarios can now be easily designed and executed at maximum performance and scalability. Last but not least, technologies such as instrumentation and event processing have been directly integrated into the platform in order to provide a full-featured experimental environment for deploying, running and testing new applications and reasoning concepts.

Application Demonstrators

Within two application domains that are urban computing and life science research the project is developing interactive online demonstrators for traffic-aware path finding, road sign management, linked life data and genome wide association studies. The traffic-aware path finding application uses traffic sensor and topology data from the city of Milano together with weather and calendar information in order to optimize routing in the city. The applied reasoning technologies rely on neural networks, operational research algorithms and semantic

querying for input data selection. The road sign management use case focuses on the validity checking of road sign information in public data such as OpenStreetMap through validation rule reasoning and data cleansing. Linked life data provides unrestricted SPARQL access to a collection of over six billion statements offered by the integration of more than 20 popular biomedical data sources. The linked life data application offers graphical interfaces to enable users to describe and control the internal data source update process. Genome Wide Association Studies attempt to find genetic markers of diseases, and to use these to predict what genes might be involved in the disease. LarKC has built a service to help decide if a marker is significant, by taking into account prior knowledge about genes, and combining it with the experimental data in a statistical model. The LarKC GWAS service has been made available to epidemiologists at the World Health Organisation and at universities around the world, where it has proved to be a useable and useful service for assisting in the prediction of genetic markers of disease.





User Involvement, Promotion and Awareness

LarKC is organizing workshops on particular scientific topics with the aim to provide a home to the research community relevant for LarKC. These workshops cover topics such as new forms of dynamic and scalable reasoning for the Semantic Web or stream reasoning. Moreover, LarKC is very active in building an early-adopter community and organizes dedicated workshop and tutorials on a regular basis; the last one in Beijing, China with almost 100 participants from universities and industry. These early-adopter workshops offer technical presentations, training and hands-on sessions on using the LarKC platform and on developing plug-ins. A further highlight in the project's promotion activities is the newly produced LarKC movie, which had its world premier at the ICT 2010 event in Brussels. LarKC continues to make a significant effort in enlarging the visibility and impact, and in building a sustainable user group that reaches beyond the list of consortium members; further early-adopters tutorials are already on the way.

Future Work and Exploitation Prospects

The LarKC research and development is maturing in all aspects of its work. The plan foresees for the next year that much effort will be put into the consolidation of results and the integration of complex workflows spanning over many different research results and plug-ins. The latest platform release that is published in December 2010 is essential in this respect.

LarKC expects workflows which apply novel reasoning techniques and selection algorithms in large and parallelized scenarios that respond to the needs of dynamic and large scale Web data sets such as LarKC's own linked life data. Offering comprehensive LarKC-based solutions includes as well the creation of tools and techniques that facilitate the implementation of workflows and plug-ins and that make the purpose and benefits of LarKC understandable in the larger context of Web-scale reasoning. For this purpose, and to ensure leadership and sustainability, the consortium is publishing a whitepaper on "Vision, Engineering and Science on Web-enabled Reasoning".

Further Information

Contact

Project Coordinator	Dieter Fensel, dieter.fensel@sti2.at
Scientific Coordinator	Frank van Harmelen, frank.van.harmelen@cs.vu.nl
Technical Director	Michael Witbrock, witbrock@cycorp.eu
Project Manager	Alice Carpentier, alice.carpentier@sti2.at

Links

LarKC Getting Started	http://www.larkc.eu/getting-started/
LarKC Platform	http://larkc.sourceforge.net/
LarKC Movie	http://www.youtube.com/watch?v=hjUbbl4cnAE
FactForge	http://factforge.net/
Linked Life Data	http://linkedlifedata.com/
WebPIE	http://www.few.vu.nl/~jui200/webpie.html

Partners

STI · INNSBRUCK
vrije Universiteit amsterdam
cycorp
AstraZeneca INTERNATIONAL
CEFRIEL FORGING INNOVATION
HLRS
IRF
MAX-PLANCK-GESELLSCHAFT
ontotext
SIEMENS
Saltlux
The University Of Sheffield.
Softgress