1. PUBLISHABLE SUMMARY

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 318497

Project Context and Objectives

During the last years, the trend to open up data and provide them freely on the Internet has intensified in volume as well as quality and value of the data made available. The linked data community has grasped the opportunity to combine, cross-reference, and analyze unprecedented volumes of high-quality data and to build innovative applications. This effort has caused a tremendous network effect, adding value and creating new opportunities for everybody, including the original data providers.

But most of the low-hanging fruit has been picked and it is time to move on to the next step, combining, cross-indexing and, in general, making the best out of all public data, regardless of their size, update rate, and schema; accepting that centrally-managed repositories (even distributed) are not able to meet the challenges ahead and that we need to develop the infrastructure for the efficient querying of large-scale federations of independently-managed sources.

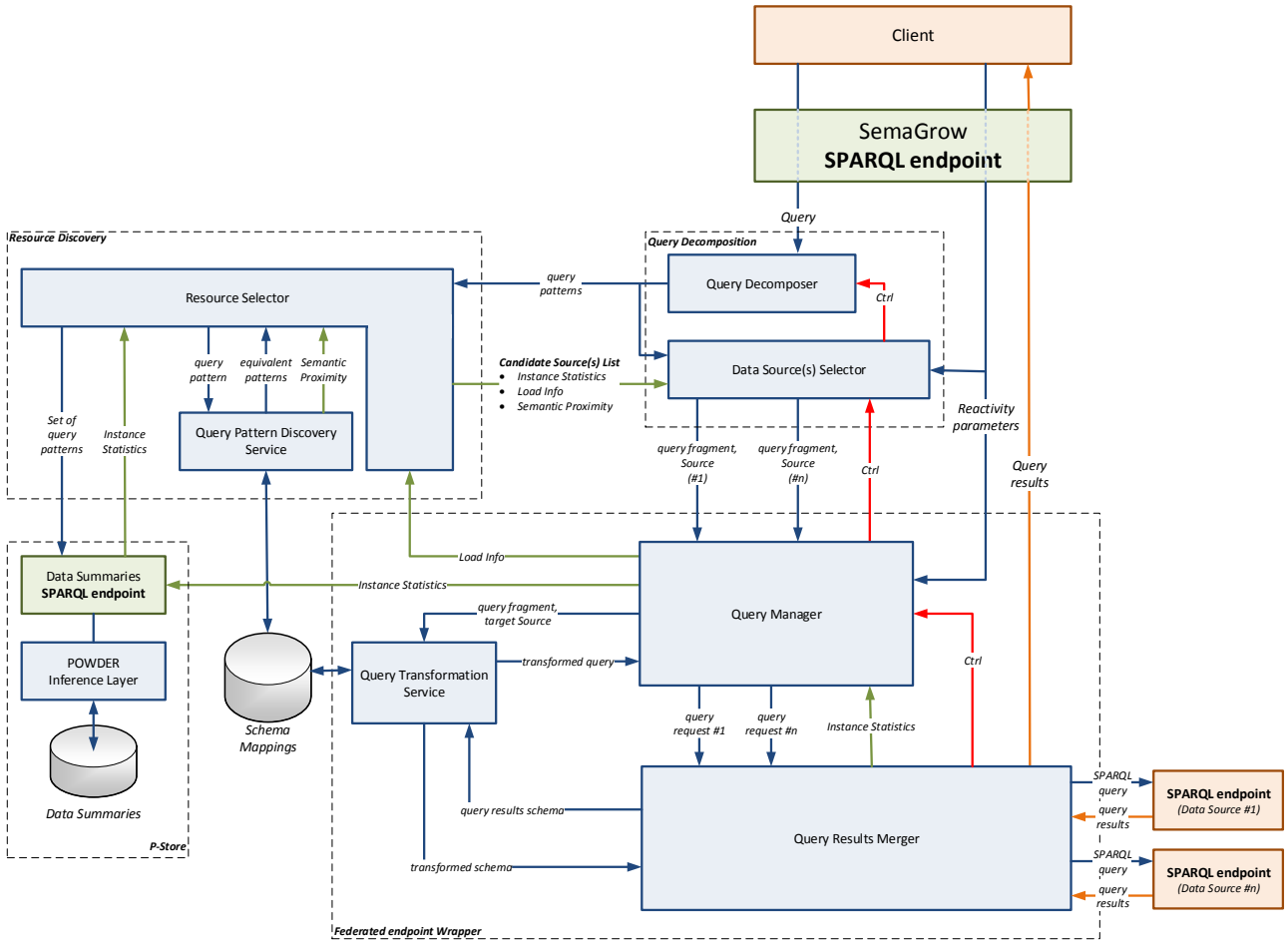
SemaGrow carries out fundamental databases research and develops methods and infrastructure that will be rigorously tested on three large-scale current use cases as well as on their projected data growth beyond project's end: we are laying the foundations for the scalable, efficient, and robust data services needed to take full advantage of the data-intensive and inter-disciplinary Science of 2020, by addressing the following key challenges:

- Develop novel algorithms and methods for querying distributed triple stores that can overcome the problems stemming from heterogeneity and from the fact that the distribution of data over nodes is not determined by the needs of better load balancing and more efficient resource discovery, but by data providers.
- Develop scalable and robust semantic indexing algorithms that can serve detailed and accurate data summaries and other data source annotations about extremely large datasets. Such annotations are crucial for distributed querying, as they support the decomposition of queries and the selection of the data sources which each query component will be directed to.

Progress So Far

SemaGrow Stack

The SemaGrow Stack integrates the components needed in order to offer a single SPARQL endpoint that federates a number of heterogeneous data sources, also exposed as SPARQL endpoints. The main difference between the SemaGrow Stack and most existing distributed querying solutions is that SemaGrow targets the federation of heterogeneous and independently provided data sources. In other words, SemaGrow aims to offer the most efficient distributed querying solution that can be achieved without controlling the way data is distributed between sources and, in general, without having the responsibility to centrally manage the data sources of the federation.



The Core Components that comprise the SemaGrow Stack are:

- The SemaGrow SPARQL Endpoint
- The Query Decomposition Component
- The Resource Discovery Component
- The Data Summaries Endpoint
- The Federated Endpoint Wrapper

The Maintenance Components that are activated in SemaGrow are the following:

- The SemaGrow Authoring Tool
- The SemaGrow Ontology Alignment Tool
- The Content Classification and Ontology Evolution Tool

Experimentation Activities

Resource Discovery

The Resource Discovery component identifies the federated sources that are most likely to hold triples matching a given query and for breaking up queries in the fragments to be dispatched to each source for execution.

During the first year we focused on the development of a triple store that is capable of inferring triples from POWDER statements, as these provide a convenient (but hard to reason over) formalism for succinctly storing data summaries of what data is stored where in the federation. We have chosen PostgreSQL as the physical layer for our POWDER store, due to its stability, efficiency, and extensive usage as triple store backend. Furthermore, it provides a built-in mechanism for customized indexing. Since POWDER inference is based on URIs' matching regular exceptions, we – effectively – need an efficient way of retrieving all tuples holding a value that matches a regular expression. We started with

implementing list of tuple pointers for all tuples that match a single, fixed regular exception. Besides gaining familiarity with PostgreSQL indexing internals, this exercise has also shown us the maximum gain that we can ever hope to achieve: this is the perfect index for this single regular expression. As a second step, we extended this code so that instead of hard-wiring the regular expression, we now cache substrings vs. tuple pointers. In this manner we restrict the number of values that needs to be checked against a regular expression. What is critical in such an approach is the strategy for dropping elements from the cache when it is full, where we weigh how frequent a substring is in queries and how selective it is with the data, in order to retain the most valuable cache elements that both appear often in queries and drastically restrict the volume of values that needs to be checked against the regular expression.

Ontology Alignment

The Ontology Alignment Tool is a semi-automatic component that will employ / integrate different alignment methods in order to provide the vocabulary mappings needed for querying heterogeneous sources.

During the first year, we have reviewed the state of the art and carried out preliminary experiments, focusing on synthesis approaches and collaborative, semi-automatic alignment methods and the comparison of automatic alignment tools. A first version of SYNTHESIS, the automatic alignment platform, incorporating four individual alignment methods, was implemented and participated in the OAEI 2013 Campaign. The platform synthesizes the results of the underlying methods, dynamically allocating a subset based on the characteristics of a given alignment task and aiming at maximizing the social welfare of the interacting parties. Furthermore, a first prototype of a GUI for human-assisted alignment was developed, and work on producing a standardized API for the semi-automatic environment has begun. Finally, we are carrying out an initial investigation of methods for improving on the scalability of semi-automatic alignment systems. Additionally, we have interacted with the Ontology-Lexica Community Group (Ontolex) of the W3C and contributed linguistic annotations-based alignment as a use case for Ontolex, as well as an API for editing and creating linguistic metadata in the LIME format.

Investigation Activities

Content Classification & Ontology Evolution

Content Classification and Ontology Evolution Tools will incorporate (a) content classification methods for automatically annotating content with a finer schema than its current annotations, in situations where the granularity difference between two schemas makes them un-alignable; (b) ontology evolution methods for recommending refinements for a coarser schema, such that it will align better with schemas it is often queried in conjunction with.

During the first two months of the relevant task we have carried out an initial investigation of suitable methods

Heterogeneous Distributed Semantic Querying

The Query Decomposition and the Federated Endpoint Wrapper components of the SemaGrow Stack will extend and adapt distributed querying methods and systems, so that they can exploit the results of resource discovery and implement a querying strategy that dispatches query fragments to those sources that are most likely to yield results. During the first two months of this task we have carried out an initial investigation of suitable methods.

Large-scale Experimentation

Semantic Store Infrastructure

During the first year of the project, the PARADOX cluster where the SemaGrow large-scale experiments will take place underwent a major update, leading in significant increases in computational power and storage capacity. Access to the infrastructure has been provided via various interfaces that were appropriately set up (batch, gLite, gUSE, RESTful interface).

Scalability & Robustness Experimental Methodology

The Scalability & Robustness Experimental Methodology will provide the guidelines for the development of automatic rigorous testing components that allow the project to reliably measure and compare the efficiency of the developed research components system under realistic conditions. The methodology will take into account individualities that occur due to distinct properties of the system such as the heterogeneity and the distributed nature of the repositories.

The methodology will define the measures for evaluating both the distinct components that realize the core SemaGrow research outcomes, as well as, the overall performance of the system. The SemaGrow POWDER store will be evaluated against both current POWDER implementations and against state of-the-art non-POWDER stores where POWDER-inferred triples have been made explicit. With respect to the latter, we will test against the best large-scale stores that are freely available or for which academic research licenses can be obtained, such as Virtuoso, 4store, or bigdata. In this case, the evaluation will comprise four metrics: (a) compression in number of triples, (b) compression in disk volume, (c) responsiveness, measured as the time to retrieve the first query result and time between successive query results, and (d) throughput, measured as the time to retrieve all query results.

Finally, the reactivity and scalability of the overall distributed system will be measured in terms of (a) the size of the data summaries needed by the source selection algorithm as a function of the total size of the federated repositories; (b) the overhead of the method as a function of the time it would take to query all repositories; and (c) the accuracy of the source selection in predicting which sources hold data that satisfy a given query.

Real-life Experimentation

Envisaged Applications & Use Cases

The Use Cases foreseen in SemaGrow are classified under three categories, covering the different aspects of real-life experimentation:

- *Heterogeneous Data Collections & Streams*: The perspective from which extremely large and very complex agriculture-related data sets are considered is the one of research activities, during which the users need to cope with heterogeneous data collections & streams in order to achieve new scientific investigations that may help forecast and address societal challenges such as food production in changing climate conditions.
- *Reactive Data Analysis*: The perspective from which extremely large and very complex agriculture-related data sets are considered is the one of information management, during which the users need to cope with reactive analysis of the data within the time scale and processes that they need to support in order to create value through extensive data collection and analysis that may help timely and better decision making related to societal challenges like food security.
- *Reactive Resource Discovery*: The perspective from which extremely large and very complex agriculture-related data sets are considered is the one of education, during which the users need to cope with reactive resource discovery in order to be able to find, reuse and exploit data resources created in one environment in very different contexts.

Real-life Deployment & User Evaluation Plan

In order to realize an evaluation of the system in real-life situations, a detailed plan for the implementation of three (3) service demonstrators on top of the Semantic Store will be designed. The plan will describe how the demonstrators should be developed and deployed and include a methodology for evaluating user satisfaction.

Expected Results and Potential Impact

SemaGrow is designed to contribute to significantly scaling up data analysis so that it may keep pace with the rate of growth of data streams and collections, as well as to enable novel forms of real time intelligence that only become possible on extremely large data volumes such as the agricultural one to be considered for rigorous testing. More specifically, SemaGrow is strongly focused to:

- Carry out and deliver fundamental research related to the development of novel indexing and reactive algorithms and the rigorous analysis of their complexity;
- Develop infrastructural components: this is based on the POWDER framework and aims to create a significant breakthrough in semantic infrastructures;
- Rigorous testing in a realistic environment: this is taking place over currently existing and realistically projected volumes of data for 2015 and beyond;

– Realistic ideas for possible deployments: this is actually taking the form of operational service prototypes that real users will test during controlled trials, together with prototype integration with an already deployed data infrastructure.

SemaGrow is expected to impact the general socioeconomic status quo by trying to help participating institutions that are working towards achieving major societal challenges related to agriculture. As it was clearly stated and declared in the World Summit on Food Security that took place at the FAO Headquarters in Rome (November 2009), all countries committed to the following mandate:

> "We will promote research for food and agriculture, including research to adapt to, and mitigate climate change, and access to research results and technologies at national, regional and international levels.
>
> We will reinvigorate national research systems and will share information and best practices.
>
> We will improve access to knowledge."

Towards this vision, SemaGrow aims to play a crucial role in demonstrating that it is possible to overcome the barriers of the currently available technologies, even in the face of huge amounts of unstructured and unrelated data.

A more detailed summary of the project activities during its first year is available via the SemaGrow website, as the first annual public report of SemaGrow (deliverable D1.3.1 – Annual Public Report). You can obtain the First Annual Public Report by following the link below.

http://www.semagrow.eu/sites/default/files/D1.3.1-SemaGrow%20Public%20Annual%20Report.pdf

Contact Information

Project website: www.semagrow.eu

Project Coordinator: Prof. Miguel A. Sicilia, Universidad de Alcalá (UAH)
msicilia@uah.es
http://www.cc.uah.es/msicilia/

Scientific Manager: Dr. Vangelis Karkaletsis
National Centre for Scientific Research "Demokritos" (NCSR-D)
vangelis@iit.demokritos.gr
http://users.iit.demokritos.gr/~vangelis/

Technical Manager: Dr. Pythagoras P. Karampiperis
National Centre for Scientific Research "Demokritos" (NCSR-D)
pythk@iit.demokritos.gr
http://users.iit.demokritos.gr/~pythk/

Universidad de Alcalá

NCSR "Demokritos"

Università di Tor Vergata

Semantic Web Company

Institute of Physics Belgrade

Stichting Dienst Landbouwkundig Onderzoek

Food and Agriculture Organization of the United Nations

Agro-Know Technologies