PRESS RELEASE

Launch of the *SemaGrow Project* in the context of the FP7 Intelligent Information Management

Current Situation

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 analyse 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 schema, size, and update rate*; accepting that some schemas might be better suited to a given dataset and application and that there is no consensus about a "universal" schema or vocabulary for any given application, let alone for the Semantic Web and related initiatives such as the LOD cloud. In other words, we need infrastructure that besides being efficient, real-time responsive and scalable is also *flexible and robust* enough to *allow data providers to publish in the manner and form that best suits their processes and purposes and data consumers to query in the manner and form that best suits theirs*.

Our Vision

This will be a decisive factor in maintaining the momentum of the linked open data movement by including in the cloud *large, live, constantly updated datasets and streams* that are published in formats that were not designed with linking across sources in mind. This will not only increase the value of all public data, but can also provide both the incentive and the opportunity to follow Semantic Web standards and linked data best practices for publishers that will not or cannot directly and immediately make this transition.

SemaGrow is partially funded by the Seventh Framework Programme (FP7) of the European Commision (FP7-ICT-2011.4.4a Intelligent Information Management) under Grant Agreement No. 318497

SEVENTH FRAMEWORK PROGRAMME

Our Challenges

In order to achieve this ambitious vision and solve a difficult data management problem, we aim to address 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.
- Since it is, in the general case, not possible to align schemas and vocabularies so perfectly that there is no loss of information, investigate how to *minimize losses* and how to *not accumulate* them over successive schema translations.

To *address these challenges,* SemaGrow carries out fundamental databases research and develops methods and infrastructure that will be *rigorously tested* on *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.

Our Use Case

Agricultural resource management is a good example of a real-world situation where data-intensive analysis needs to combine information from different, large-scale sources that are actively maintained in incompatible schemata: the agricultural domain includes various different topics with subjects varying from plant science and horticulture, to agricultural engineering, to agricultural economics. These different subjects are extensively researched by scientists all over the world, consuming as well as producing an enormous volume of data; agricultural scientists are inundated by an abundance of data as well as reported results relevant to their research as much as their colleagues from different disciplines. SemaGrow aims to reinforce the ability of a wide range of innovators working over agricultural data infrastructures to engage intensive data analysis and create value beyond the original purpose of the data. Beyond the rather obvious goal of increasing the ability to find, reuse and exploit relevant data resources, from diverse sources and stakeholders into different and unforeseen contexts, SemaGrow wants to facilitate new scientific investigations over large, interconnected agricultural data resources. Furthermore, SemaGrow wants to develop and deploy the necessary infrastructural components that will support the development

SemaGrow is partially funded by the Seventh Framework Programme (FP7) of the European Commision (FP7-ICT-2011.4.4a Intelligent Information Management) under Grant Agreement No. 318497

and rigorous testing of technologies that will scale analytic capabilities in step with the witnessed or predicted growth of data.

Our Partners

Experts from different, yet complementary disciples from 8 different organizations in 6 different countries form the SemaGrow consortium, combining both the academia and the industry: the Information Engineering research unit of the Universidad de Alcala from Spain (coordinator), the Institute of Informatics & Telecommunications of the Greek National Center for Scientific Research "Demokritos" from Greece, the Artificial Intelligence Research group of the University of Tor Vergata from Italy, the Semantic Web Company from Austria, the Institute of Physics of the University of Belgrade from Serbia, the Alterra Institute of the University of Wageningen from the Netherlands, the Food and Agriculture Organization of the United Nations from Italy and the Agro-Know Technologies from Greece.

Project Coordinator:
Professor Miguel A. Sicilia
Universidad de Alcalá (UAH)
msicilia@uah.es

For more information: www.semagrow.eu

SemaGrow is partially funded by the Seventh Framework Programme (FP7) of the European Commision (FP7-ICT-2011.4.4a Intelligent Information Management) under Grant Agreement No. 318497