ICT Seventh Framework Programme (ICT FP7)

Grant Agreement No: 318497

Data Intensive Techniques to Boost the Real – Time Performance of Global Agricultural Data Infrastructures

Third Annual Public Report

Deliverable Form	
Project Reference No.	ICT FP7 318497
Deliverable No.	D1.3.3
Relevant Workpackage:	WP1: Project Management
Nature:	R
Dissemination Level:	PU
Document version:	Final
Date:	10/12/2015
Authors:	UAH and NCSR-D, with input from all partners
Document description:	The present document presents the Main Activities & Achievements of the third year of the project.

PUBLISHABLE SUMMARY

1 Project Context and Objectives

As 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 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 loose data source federations at a large scale.

The SemaGrow project used the large-scale and complex *agricultural data service* ecosystem as a testbed for its technologies. During the first reporting period, the project engaged consortium members and external stakeholders to identify user needs and relevant data sources and to draft requirements and the system architecture. From the perspective of engaging stakeholders and translating user needs to requirements, the project focused on the definition of different types of use cases based on stakeholder requirements. We documented the outcomes of the stakeholder workshops, as it is from the contact with stakeholders that the project may effectively pin point how to develop and then evaluate its results. Through the definition of use cases ideas and requirements expressed in these workshops were translated into the application descriptions that form the basis for the development of service demonstrators and connect the needs of the stakeholders to the technological developments of the project. Three categories of use cases are considered in the SemaGrow project, addressing a diverse group of stakeholders and wide area of application:

- *Heterogeneous Data Collections and Streams,* focusing on data-intensive experiments in the domain of agro-environmental modelling. SemaGrow technologies were used to prepare suitable input for modelling experiments. SemaGrow supports the previously available functionality of searching metadata to identify relevant datasets, and augments it with new functionality of selecting and combining data from these datasets. This latter task was previously carried out by downloading complete datasets and creating custom scripts for slicing and combining them. The stakeholders are the *modellers* who prepare the experiments and the *IT personnel* who support them. These use cases validate the integrative semantic capabilities provided by the Semagrow Stack as well as its ability to search and retrieve large data volumes.
- *Reactive Data Analysis*, focusing on the *AGRIS portal* that serves scientific bibliography and relevant Web resources. SemaGrow technologies allow federating the diverse endpoints used to retrieve resources that are semantically relevant to a given AGRIS bibliography item. Stakeholders are the *Web developers* who maintain the portal and the *data analysis experts* who experiment to refine the relevance scoring mechanism. This use case validates the ease of experimenting with different mixtures of datasets without amending the analysis software, as well as the reactivity of the SemaGrow Stack when searching through large datasets in order to find results that are not necessarily voluminous.
- *Reactive Resource Discovery*, focusing on the *AKstem* application that serves diverse bibliographical and educational resources in the LOD cloud. The stakeholders are the *LOD professionals and enthusiasts* who put together Web applications using Web services and the AKstem users who want to publish the data inter-linked with other datasets in the community. These uses cases validate the ease of developing Web applications over the SemaGrow Stack and the ability to federate and semantically integrate a large number of heterogeneous data sources, although the contents of any individual data source do not constitute big data by itself.

The assumption is that the efficiency of the stakeholders on these use cases significantly improves when replacing current methods of data access with SemaGrow technologies *without affecting their workflow otherwise*. That is to say that, for example, AGRIS widgets that visualize data exposed by a SPARQL endpoint do not need to be re-developed in order to be able to take advantage of SemaGrow technologies to serve federations of SPARQL (and possibly other) data sources. The rigorous testing and evaluation activities foreseen in SemaGrow are meant to validate both the increase in stakeholder efficiency and the effort required to adopt SemaGrow as data access infrastructure.

By addressing these use cases, the project evaluates its approach to the following major technical challenges:

- *Finding small results in big data:* in this “needle in a haystack” situation we are joining results from different big datasets in order to retrieve a result set that does not constitute big data by itself. The challenge here is to have

an intelligent *query execution planner* that guides the *query execution engine* along a query execution plan that never retrieves large quantities of results that do not contribute to the end-result because they do not join with subsequent query patterns. Such a planner relies heavily on the availability of accurate instance-level metadata.

- *Big results from big data:* in this situation the result set constitutes big data, and no amount of query plan optimization can avoid this, since this is what has been requested by the client. Besides the challenges for the query execution engine, handling this situation is also relevant to the ability of the histogram maintenance mechanism to efficiently handle query feedback that is big data.
- *Integrating heterogeneous data:* 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, we are developing technologies that 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.

We proceed to present progress during the third period along these major technical objectives, as well as progress in developing the overall SemaGrow system.

2 Work Performed and Main Results

Federated Query Processing

The central product of the project is the *SemaGrow Stack* that federates remote SPARQL end-points and offers a single endpoint to its clients. Earlier in the project, we developed the *Query Decomposition* component that calculates the optimal *query execution plan* for a given query. Query execution plans break down a query into fragments and specify the federated end-point that will execute it and the optimal execution *ordering* for merging the partial results returned for each fragment. This calculation is based on the estimated *cost* of executing a query fragment on a given endpoint. Cost is estimated from *metadata* about the federated data sources: schema-level information but also instance-level statistics, served from the *Resource Discovery* component.

Cost estimation was originally based on metadata expressed using the *Vocabulary of Interlinked Datasets (VoID)*. During the final year, we formalized *Sevod*, a VoID extension that is expressive enough to represent detailed instance-level metadata akin to relational database *histograms*. This improves cost estimation, but *Sevod* models are practically impossible to maintain manually. To address this, we developed a tool that extracts histograms from RDF dumps and we have also extended Y2 work on *adaptive query processing*, i.e., into a method for maintaining histograms by analysing the results to the user-requested query workload. In this manner, SemaGrow builds and maintains its *Sevod* descriptions both with and without access to complete data dumps.

Although the research above has given us considerable advances towards handling the needle-in-a-haystack problem of identifying the best strategy to retrieve data out of a large-scale repository, intelligent execution planning does not improve retrieving large amounts of data. To efficiently combine query results at a large scale and to handle the dynamic nature of the LOD cloud, we developed an execution engine as efficient as that used by FedX, but with the added benefit of being based on state-of-the-art stream processing technologies to achieve asynchronous and non-blocking operation. The benefits of the SemaGrow system are already visible in the relatively mid-scale FedBench suite, where SemaGrow outperforms FedX and SPLENDID, the two systems that constituted the state of the art prior to SemaGrow. The difference becomes more pronounced in the LargeRDFBench suite, where SemaGrow even more clearly outperforms FedX and where SPLENDID cannot complete due to timeouts.

Managing Heterogeneity

Besides integrating large-scale data sources, SemaGrow also tackles integrating *heterogeneous* data sources. This pertains to efficiently applying resource mappings within the query execution engine of the SemaGrow Stack, but also to a system of tools that complement the SemaGrow Stack and that provide the mappings that should be applying. While the first period of the project focused on developing technologies that support robust alignment over heterogeneous knowledge models and languages (MAPLE, Lime) and on refining the SYNTHESIS system for automatic alignment, the second period focused on robustness and integration. Specifically, SYNTHESIS has been provided with an appropriate data access layer, moved to a thread-based architecture, and provided with ontology modularization methods to improve scalability. Furthermore, the architecture of the VocBech environment for manual alignment and of its backing framework, Semantic Turkey, has been reworked to allow for multi-project and multi-graph management, dynamic context injection inside the services, seamless navigation of local and web data laying out a common ground for a user-centric ontology alignment experience. Semantic Turkey also features a new Linked Data explorer.

Finally, during its second period the project experimented with the CODA system for knowledge extraction and transformation, applying it to a datasheet import scenario. The objective is to strike a balance between enforcing

conventions (affording ease of use) and capacity to deal with complexity. This effort resulted in Sheet2RDF, a *datasheet to RDF* import and transformation system, which by following a "convention over configuration" approach makes so that the "evident" and trivial imports can be dealt in an almost automatic way, whereas more complex transformations can be still managed through the more powerful capabilities of CODA.

During the final period, and besides experimentation and improvement of the individual tools, significant integration results have been achieved: SemaGrow has adopted the INRIA Alignment API as a specification followed by all alignment tools, allowing the project to integrate its automatic alignment and manual editing tools into a system for semi-automatic extraction, validation, and editing of vocabulary mappings. CODA has also been integrated to Semantic Turkey and, thus, to VocBench that uses Semantic Turkey as its underlying RDF management framework.

Prototyping and Rigorous Testing

In order to be able to deploy and test the SemaGrow system, the project prepared and maintained hardware infrastructure, integrated the research prototypes described previously in a *SemaGrow Stack* prototype, and developed the software that comprises the SemaGrow ecosystem of tools, including some peripheral and domain-specific tools that are not part of the core SemaGrow system. These tools have been produced by factoring out of the tools developed for piloting SemaGrow of the parts that could be generally useful. SemaGrow also develops an automated testing and system health monitoring component, based on the experimental setups used for system evaluation. For this purpose, we executed the following experiments on query execution and managing large scale:

- We have measured the optimality of the query execution plans calculated by our Query Decomposition component and the efficiency of our query execution engine using the FedBench benchmark to compare SemaGrow against FedX and SPLENDID. This experiment is done for evaluation purposes only and no automated testing component will be derived.
- Based on the FedBench experiments, we prepared software that automates testing. During the final year, we re-worked the logging facilities of the SemaGrow Stack so that it can trace query execution across different threads. That is to say, we can identify the query that is served by each executor thread, although such executor threads are re-used from a thread pool and not spawned specifically for processing new queries. This allows measuring CPU time consumed and not only wall time.
- The ability to trace query execution across different threads is not only used in experimentation, but also by the Automated Rigorous Testing Web app developed during the final period. This Web app is integrated into the SemaGrow endpoint Web app and plots run-time graphs that monitor system health.

Furthermore, we established and ran experiments to evaluate our methods for managing heterogeneity:

- We used the heterogeneous data sources from the Reactive Resource Discovery use cases to measure the validity of ontology alignment, by comparing automatically calculated mappings against prior manual mappings.
- We used the FAO document base to evaluate the appropriateness of the PEARL language and the CODA architecture to cover a broad variety of extraction/classification and triplification scenarios.

Deployment

The project developed three demonstrators, that is, three SemaGrow Stack client applications, which were used to validate the three use case categories presented above:

The *Trees4Future/AgMIP Demonstrator* validates the *Semagrow* system on the *Heterogeneous Data Collections & Streams* use cases. This demonstrator focuses on the support of scientific communities in the domain of agricultural and forestry modelling. Specifically, it supports end-users in preparing suitable input dataset for their modelling exercises from the wealth of heterogeneous big data collections and streams available, validating the integrative semantic capabilities provided by the *Semagrow Stack* as well as its ability to search and retrieve large data volumes. The software and the larger of the datasets used in this pilot have been integrated in the agINFRA research data hub and can be accessed by searching for "semagrow".[1]

The *AGRIS Demonstrator* validates the *SemaGrow* system on the *Reactive Data Analysis* use cases. The *AGRIS Web Portal* serves bibliographical data from the agricultural domain, combined with widgets that retrieve and present Web resources that are relevant to the currently viewed bibliography item. The AGRIS Demonstrator validates that using the *SemaGrow Stack* as a backend component helps IT personnel to more easily extend the range of data sources federated under the AGRIS Web Portal; and that the *SemaGrow Stack* is able to reactively search through large

[1] Direct links: http://ring.ciard.info/software-all?search_api_views_fulltext=semagrow (software) and http://ring.ciard.info/datasets?search_api_views_fulltext=semagrow (datasets)

datasets in order to find results that are not voluminous. The AGRIS Demonstrator was used to calculate the recommendations served by the "Activities from the Web" button on the AGRIS site.[2]

The *AKstem-Linked Open Data Hubs* application[3] replaced the Agricultural Discovery Space (ADS) application based on the recommendations of the 2nd Review Meeting and the results arising from the R&D department of Agro-Know. It validates the SemaGrow system on the *Reactive Resource Discovery* use cases. The SemaGrow powered linked open data interface of each hub allows the retrieval of resources that are related with the resources of the hub but they are from other diverse data sources. This demonstrator validates that using SemaGrow as a backend supports developers in re-using code for apps that facilitate the discovery and the linking of relevant data sources.
In order to achieve these deployments, the project prepared and developed hardware infrastructure and software for accessing it, server-side tools, clients and client-side tools, seeding an *ecosystem* of software tools, semantic vocabularies, and public Web services developed around the main SemaGrow system.

These SemaGrow deployments have validated the ability to deploy SemaGrow with minimal disruption of existing workflows, while also gauging end-user satisfaction with the new features and capabilities. Furthermore, SemaGrow pilots have aimed to gauge ease of use for IT personnel who install and maintain it and for programmers who develop client software. Through these pilots, besides the SemaGrow software tools we also improved our *guidelines* and *getting started* document distributed together with the SemaGrow Stack and WebApp.

3 Results and Potential Impact

SemaGrow carried out fundamental databases and Semantic Web research and developed and validated methods and infrastructure laying the foundations for the scalable, efficient, and robust data services needed by the data-intensive Science of 2020. The tangible outcomes of the project comprise:

- The *SemaGrow Stack*, which integrates the project's research prototypes under the Sesame architecture (http://rdf4j.org). The SemaGrow Stack is developed on Github, http://github.com/semagrow and is also distributed in binary packages for Ubuntu Lucid from the SWC repository http://semagrow.semantic-web.at/deb/ As a further dissemination platform and focal point for potential users, we have used the Web hosting facility offered to Github projects to collect technical information and guides (http://semagrow.github.io)
- A series of experimental setups and the *Automated Rigorous Tester (ART)* which automates the execution of these experiments. Besides facilitating experimentation during the project, ART also plots real-time query processing time to allow the monitoring of system health.
- The alignment tools used to provide the vocabulary mappings upon which the SemaGrow Stack relies to handle heterogeneity. Alternative alignment tools can be used, as long as a list of mappings of URI resources is produced. An import facility for third-party mappings is under development.
- Sevod, a vocabulary for serializing histograms of RDF data, and a series of tools for creating such descriptions.
- Datasets and services from one of the SemaGrow pilots have been integrated in the agINFRA data hub for agriculture, food, and the environment.
- The other two pilots that have been integrated in actual production systems.

Further targeted project outcomes, with a view on sustainability beyond project's end, include the transition of the piloting prototypes into production systems for the three piloting partners as well as to engage relevant developer and user communities into building on SemaGrow technology beyond project's end. Especially with respect to the latter, the project organizes a series of Hackathons to attract the attention of these communities to project outcomes.

4 Contact Details

Project website: http://www.semagrow.eu
Project Coordinator: Prof. Miguel A. Sicilia, Universidad de Alcalá
E: msicilia@uah.es W: http://www.cc.uah.es/msicilia
Scientific Manager: Dr Vangelis Karkaletsis, National Centre for Scientific Research "Demokritos"
E: vangelis@iit.demokritos.gr W: http://users.iit.demokritos.gr/~vangelis
Technical Manager: Dr Stasinos Konstantopoulos, National Centre for Scientific Research "Demokritos"
E: konstant@iit.demokritos.gr W: http://www.iit.demokritos.gr/people/konstantopoulos-stasinos

[2] Please visit http://agris.fao.org and select a paper to see its bibliographic entry and related information widgets, including the SemaGrow "Activities from the Web" widget.
[3] Please visit http://www.akstem.com and in particular the "Share" service.