International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Designing Innovative Linked Open Data and Semantic Technologies for Agro-environmental Modelling

Rob Lokers[a], Stasinos Konstantopoulos[b], Armando Stellato[c], Rob Knapen[a], Sander Janssen[a]
a Alterra, Wageningen UR, The Netherlands - rob.lokers@wur.nl
b Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", Greece - konstant@iit.demokritos.gr
c University of Rome, Tor Vergata, Rome, Italy - stellato@info.uniroma2.it

Abstract: In recent years, innovative applications exploiting Linked Open Data (LOD) and the Semantic Web have opened up, combined and cross referenced high volumes of high-quality data and created tremendous new opportunities for data users as well as data providers. However, in order to serve the broadest community of users, technologies need to be developed that can manage large, constantly updated datasets and streams that are published in formats that were not designed with cross source linking in mind.
The EU-FP7 project SemaGrow aims to tackle this challenge by developing novel algorithms and methods for querying distributed triple stores, scalable and robust semantic indexing algorithms and effective ontology alignment. These innovations will be tested by applying them to data and knowledge intensive use cases from the agro-environmental domain. Aspects like the relatively large heterogeneity of datasets in this domain, their often explicit spatial and temporal dimensions resulting in relatively large volumes and their inherent nature of uncertainty provide additional challenges which are not usually dealt with till so far. This paper describes the architectural design of the SemaGrow infrastructure and how it integrates LOD concepts and a range of semantic technologies to meet these challenges. The presented SemaGrow use cases describe some concrete challenges and help understanding how applying these innovations will provide agro-environmental modellers with new opportunities to discover and combine distributed datasets for use in their models, to handle data gaps and achieve data volume reduction.

***Keywords**: Semantic Web; Linked Open Data; agro-environmental modelling; semantic technologies*

1 INTRODUCTION

The opportunities of the Semantic Web (Lee et al., 2001) and specifically the added value of semantic annotation of agro-environmental datasets and models as well as the exploitation of semantic technologies are widely recognized as a means to improve effectiveness of modelling environments. Added value can be found in efficient discovery and processing of data and knowledge required for modelling exercises as well as in the semantic coupling of models. Villa et al. (2009) describe semantically aware environmental modelling as a way of designing, implementing and deploying environmental datasets and models based on the independent, standardized formalization of the underlying environmental science. It is the result of merging the rationale of declarative modelling with modern knowledge representation theory, through the mediation of the integrative vision of a Semantic Web. Semantic technologies in practice already facilitate the discovery of agro-environmental data and metadata. One example taken from the agricultural domain is the OpenAGRIS system (Anibaldi, Jaques et al., in press), a semantic mashup application integrating knowledge by exploiting the semantics of annotated metadata of agricultural resources. Another range of possible applications evolves when not only the metadata, but also the data itself can be annotated and linked using ontologies. Linking the data semantically and exposing the spatial and temporal characteristics of data opens up possibilities to relate and combine different datasets, even to the level where integrated datasets for use in environmental modelling can be (semi-) automatically generated. Michener (2012) for instance describes a stepwise approach to ecoinformatics based on a

data life cycle. This approach includes the specific innovation of using semantic models and technologies for data integration, adopting a bottom up approach for data streamlining using the semantics of measurements as offered by for example the Extensible Observations Ontology (OBOE) and the Observations and Measurements specification.

In the meantime, also the trend to open up data and provide them freely on the Internet has intensified and has gained value by adhering to Semantic Web principles and standards for using formal, machine-readably vocabularies that enhance data interoperability. An ever-increasing collection of tools and vocabularies to re-use and extend and a continuously lowering bar for posting on the Web at a large scale has decreased the resources needed in order to cross-link data and to provide machine-readable access to it via querying engines or other Web APIs. The resulting *Linked Open Data (LOD)* cloud contains more than an estimated 50 billion facts from domains as diverse as geography, media, biology, chemistry, economy, and energy (Bauer and Kaltenböck, 2011). This creates an unprecedented opportunity for the agro-environmental modelling community to carry out innovative experiments and analyses that combine and cross-reference agricultural data with large volumes of data from different domains.

Although state-of-the-art applications in the agro-environmental continue to evolve we can at the same time also clearly observe that the immense speed of progress of developments poses new barriers which could mean that this vast potential remains to a large extent untapped. The FP7-ICT SemaGrow project attempts abating these impediments by developing and integrating information technology innovations and validating these in use cases in the agro-environmental research community through demonstrators exploiting these innovations. The targeted community, through its nature, typically consumes and produces vast amounts of data. Data-intensive analysis and modelling activities require that information from many, usually large and heterogeneous, actively maintained sources are combined. This paper describes two of the SemaGrow use cases, specifically exemplifying those issues related to agro-environmental modelling. Subsequently, we will explain the outline of the SemaGrow infrastructure as a means to overcome these barriers.

2 SEMAGROW USE CASES

2.1 Trees4Future – A Research Infrastructure for Forestry Research

Trees4Future (www.trees4future.eu) is an integrative European Research Infrastructure project that aims to integrate, develop and improve major forest genetics and forestry Research Infrastructures. It will provide the wider European forestry research community with easy and comprehensive access to currently scattered sources of information (including genetic databanks, forest modelling tools and wood technology labs) and associated expertise. To support researchers and make relevant forestry related data discoverable and accessible for among others forestry modellers, a Clearinghouse mechanism is developed as a central access point to distributed datasets. Metadata is harvested from distributed catalogues and stored in a RDF store structured according to a custom-developed ontology. The extracted concepts are semantically linked to forestry ontology and the resulting knowledge structure is exposed through a SPARQL endpoint to facilitate semantic searches.

Although first experiences in implementing this use case show good results, several bottlenecks have been observed. The first one originates from the heterogeneous character of agro-environmental data sources. Forestry research deals with a broad spectrum of knowledge domains (e.g. plant sciences, soil science, climatology and climate change) that use their own formal and non-formal vocabularies and schemata. This occurs even within these domains, where specialist subdomains like forestry genetics or forest management use their own unaligned vocabularies. Proper alignment of semantics over these domains is a tedious and costly process, while on the other hand resources of data owners in the European forestry sector are limited. Thus it is not to be expected that, even if standardized cross-domain ontologies would be available, data providers will re-organize and harmonize their data to comply with them.

A second phenomena is related to the fact that available metadata on datasets is usually not very targeted to the actual data itself. Despite the fact that most metadata standards offer a structured template that also allows specifying characteristics of the dataset like their lineage or spatial and

temporal coverage, they usually lack the required detail required to reason on the contents of the dataset itself. Even relatively basic and high level characterisations of data content, like a structured specification of available attributes and aspects like their unit and domain are not explicitly targeted in commonly used metadata standards. This obstructs some of the more data oriented approaches that are required to effectively select, combine and process data from different sources and to eventually semantically link these to the interfaces of forestry research models.

2.2 AgMIP – Comparing and linking agricultural models

The Agricultural Model Intercomparison and Improvement Project (AgMIP) is a major international effort linking the climate, crop, and economic modelling communities with information technology to produce improved crop and economic models and the next generation of climate impact projections for the agricultural sector (Rozenzweig et al., 2013). A core activity within the project is the intercomparison and improvement of different agricultural (e.g. crop and economic) models. Such exercises require model integration, intensive interchange of datasets between models and large harmonization efforts. The main integration strategy used in AgMIP is to setup a central NoSQL data repository relevant for modelling, harmonize data using a standard schema, and to interlink data and models through data translators. Semantic technologies are used as a means to harmonize data streams in AgMIP and related initiatives. Janssen et al. (2012) describe a generic data scheme that can be used to store data on agricultural systems compiled with many different purposes and scopes. The generic data schema covers aspects of soil, climate, location, crop management and crop variety characteristics. White et al. (2013) describe an implementation of the ICASA version 2.0 data standards covering crop experiment data from the AgMIP project. The AgMIP initiative has extended this ICASA data dictionary and used this extension as the basis for the AgMIP Crop Experiment (ACE) harmonized format.

In SemaGrow the objective for this use case is to further exploit the possibilities of the Semantic Web by semantically linking data sources. Challenges in this use case are in many ways similar to those in the Trees4Future use case. The heterogeneity of data and the different schemes used over the different domains will be important barriers to be passed, although the initiatives to define harmonized data structures mentioned above will help facilitating this. However, it is presumable that these standards do not fully suffice to adequately characterize the data level. Spatial and geographical querying and scaling over datasets are required to support regional and local applications. Existing data gaps require proximity approaches, e.g. for finding data in regions that have similar climatic and biophysical conditions or on crop varieties with similar characteristics as a target variety. Thus, the needs to intelligently align and combine datasets for modelling are possibly even more pressing in this use case.

3 SEMAGROW ARCHITECTURE

3.1 Problem Statement

As illustrated by the SemaGrow use cases, application of semantic technologies is hindered by several aspects. However, most can be traced back to a single cause: the highly decentralized nature of the LOD cloud that makes it so attractive for data publishers. Minimizing the requirements for publishing data is very important for having publicly available data in the first place, but at the same time this poses data interoperability issues to data consumers. In other words, LOD might be in principle interoperable and machine readable but in practice, and due to its de-centralized nature:

- Not all publicly available data sources are reliably available. Although not a concern for the core data sets in our use cases, it becomes a concern when linking these with data from the LOD cloud;
- There will never be consensus over how data is structured and annotated, since different communities, or even individual providers, will follow the schemas that are better suited for their own workflows and applications. This is certainly the case for the agro-environmental domain, with its cross-domain orientation and high data heterogeneity.

- Data-intensive scientific methods and sensor deployment for scientific measurements are reaching scales where data volumes have reached (or are rapidly approaching) the point where copying, transforming, and curating all the needed data is not a viable approach to data consumption. Data integration needs to become more flexible and dynamic and should not rely on centralized repositories maintaining integrated and curated copies of all relevant datasets.

Moreover, it is clear that the complexity of the required semantic support increases as the use case shifts from solutions requiring only metadata towards solutions requiring insight in and handling of the actual data content. Problems then tend to shift towards the big data domain requiring massive processing capacity. Moreover, it assumes detailed and less commonly available knowledge about the semantics of data over different domains.

3.2 The SemaGrow Stack

The *FP7-ICT SemaGrow* project assumes the given problem statement as a starting point and attempts to solve these barriers through information technologies that facilitate dynamic *data integration* and *distributed querying*, and are specifically designed for this cloud of large, distributed, heterogeneous, live and constantly updated datasets. The focal point of this infrastructure is the *SemaGrow Stack*, an integrated infrastructure which provides to data consumers a SPARQL endpoint that *federates* multiple SPARQL endpoints independently maintained by data providers (Prud'hommeaux and Buil-Aranda, 2013).

There are several key features of the SemaGrow Stack that address the use cases described above: it provides a querying interface that uses the result of ontology alignment to completely hide schema heterogeneity and also applies methods from artificial intelligence in order to take into account complex endpoint selection considerations involving data contents and querying efficiency. The ontology alignment and dynamic vocabulary transformation facilities address the required harmonization of the broad and fragmented vocabularies used to annotate agro-environmental data sources. At the same time, the intelligent endpoint selection is based on methods that automatically extract detailed metadata about the content of the federated endpoints, overcoming the lack of detail in the manually provided annotations. As an added benefit, the SemaGrow Stack foresees mechanisms that facilitate falling back to feasible alternatives in the face of unavailability of an endpoint and, furthermore, does not require any modification of the federated endpoints or any other obtrusion of existing workflows.

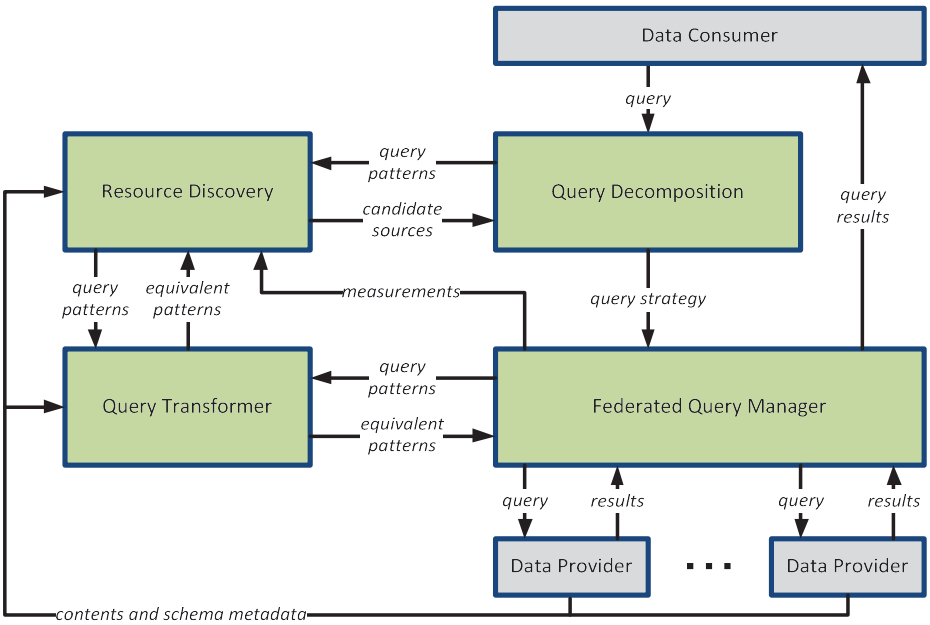

Figure 1: The architecture of the *SemaGrow Stack*.

In the architecture of the *SemaGrow Stack* (Figure 1), this transparent data integration is performed by the *Query Decomposition* component that builds *query strategies* that inform the *Query Manager* about which patterns of the overall query should be used with each of the endpoints known to the federation. The Query Manager is then responsible for implementing the query strategy and for requesting alternative strategies in the face of possible endpoint unavailability. Query strategies are optimized based on metadata regarding the availability and reactivity track record of each endpoint, the types of data exposed by it, the schemas this data follows and known mappings between different schemas. This metadata is served by the *Resource Discovery* component that provides candidate data sources for a given query pattern, exploiting novel representations for succinctly describing Web resources (Konstantopoulos et al., 2012). *Resource Discovery* is based on *content and schema metadata* provided by human annotators, including third-party data curators so that the data consumers themselves can add sources to the federation and annotate them. It is, furthermore, based on characteristics regarding performance, reliability, and other relevant metadata extracted by statistically analysing the *measurements* that the Query Manager forwards to the Resource Discovery component.

The integration of semantically heterogeneous data is performed by the *Query Transformer* that uses *schema metadata* and knowledge extracted by *ontology alignment* in order carry out transformations between equivalent entities in different schemas. Thus, effective methodologies and tools to perform ontology alignment over heterogeneous schemas are an essential asset to support the *SemaGrow Stack.*

3.3 SemaGrow Ontology Alignment

Alignment of different data sources has the clear objective of enabling information sharing across different peers, creating a common basis for understanding among agents. However, the context and specific objective of an alignment is also dependent on the nature of the linked resources.

In general, OWL Vocabularies provide schemes for modelling and structuring data. Consequently, proper alignments between two vocabularies are always difficult to obtain, as the correspondences between them not only have to qualify the nature of the related elements, but also to model data transformations. Two classes from the aligned vocabularies could in fact denote the same set of domain objects, (thus needing a “class equivalence” mapping), still the mutual representation of these objects could strongly differ, making the mapped data almost unusable if seen “with the eyes” of the other vocabulary On the other hand, when dealing with Knowledge Organization Systems (KOS), like thesauri and other controlled vocabularies, the scenario is quite different: Organizations usually adopt KOS to internally organize their document bases and facilitate selection and discovery of material. In such a scenario, the porting to a web scale environment consists in a plethora of peers providing pairs of documents bases/KOS, with KOS acting as indexes for retrieving documents from their bases. The differences with OWL vocabularies are that (1) there is actually no data to remodel/transform, and the sole objective is to enable a shared Information Retrieval environment and .(2) the shallow semantics of the languages used to represent KOS (e.g. SKOS (World Wide Web Consortium (W3C), 2009) and SKOS-XL (World Wide Web Consortium (W3C), 2009) allow for very flexible mappings without any risk of causing inconsistencies or undesired behaviour.

It is shown (van Ossenbruggen, Hildebrand, & de Boer, 2011) that shared test cases (often including strong preprocessing of the considered resources) often do not help very much in predicating the behavior of alignment approaches. It also appears that in many real-world scenarios even simple mediators based on label matching may lead to effective (semi-automatic) solutions (Caracciolo, et al., 2012), providing that the necessary verification/validation is supported by appropriate tooling. Furthermore, even state-of-the-art techniques are useless if their implementations are unable to understand modeling patterns used in the input ontologies. For these reasons, and since SemaGrow tackles real use cases dealing with large heterogeneous datasets, the project focuses on improving the robustness of alignment systems, on guaranteeing system-awareness about the alignment scenarios and on widening system’s capabilities to deal with them.

SemaGrow uses the alignment platform MAPLE to (1) provide a proxy between data to be aligned and the specific mediators, disburdening them from understanding specific data idiosyncrasies or simply different modeling choices and to (2) support coordination of the alignment activity as a whole, by exploiting metadata about the input ontologies, the linguistic resources and the mediators. To support coordination, MAPLE relies on a combination of traditional metadata standards (Alexander, Cyganiak, Hausenblas, & Zhao, 2011) and information about the linguistic expressivity of datasets and RDF linguistic resources. A linguistic metadata vocabulary: LIME (Fiorelli, Pazienza, & Stellato, 2013) is used to support this. The focus on language is fundamental, as it allows retrieving linguistic resources useful to reduce the linguistic distance between the datasets to be aligned, by different means (translation or lexical expansion).The meta mediator system SYNTHESIS selects the proper mediator by integrating MAPLE information with specific information about available mediators, in order to get the best performing system, case by case. Finally, VocBench (Stellato, et al., 2011) is a collaborative editing framework for SKOSXL thesauri. Its main features are a collaborative maintenance-to-publish workflow and the capability to use different triples store technologies, in order to balance complexity of installation and licensing costs with the scalability needs of each user.

4 CONCLUSIONS

The SemaGrow project develops novel algorithms and methods for querying distributed triple stores, scalable and robust semantic indexing algorithms and effective ontology alignment. It focuses on dynamic data integration and distributed querying, specifically designed for the cloud of large, distributed, heterogeneous, live and constantly updated datasets. The design of the SemaGrow architecture presented in this paper consists of the SemaGrow Stack, composed of integrated components allowing effective querying of federated Linked Open Datasets through a single entry point. Moreover, a suite of methods and tools for ontology alignment are presented that will help supporting the required cross-domain integration of knowledge.

The presented use cases show a variety of challenges associated with environmental modelling that range from metadata oriented information retrieval issues to heavily data-oriented problems related to big data mining and data integration. SemaGrow will verify its developed technologies against these use cases. Applications that are currently already operational in the use cases will be adapted and deployed on top of the SemaGrow Stack and will be evaluated with stakeholders. This will provide benchmarks regarding the performance and correctness of the infrastructure, as well as its ability to scale and be useful in a research environment where data-intensive science is accumulating data at an ever-increasing pace. It will also provide insights regarding the usability of the infrastructure and the ease of integration in currently used tools in real-world workflows and applications. Although it is shown through the use cases that there is considerable complexity in semantic alignment, especially towards the data level and its automation, it is expected that considerable ground can also be covered in the more data-oriented use cases. Recent relevant achievements within the Semantic Web community, including the publication of the *RDFCube* vocabulary (Cyganiak & Reynolds, 2014) go to show that extensive support for numerical conditions and operations is an immediate goal for Semantic Web stores laying the foundations for future extensions of SemaGrow technology covering large-scale numerical data integration.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 318497. For more details please see http://www.semagrow.eu.

REFERENCES

Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J., 2011. Describing Linked Datasets with the VoID Vocabulary (W3C Interest Group Note). Retrieved May 16, 2012, from World Wide Web Consortium (W3C): http://www.w3.org/TR/void/

Anibaldi, S., Jaques, Y., Celli, F., Stellato, A., & Keizer, J., (in press). Migrating bibliographic datasets to the Semantic Web: the AGRIS case. Semantic Web. doi: 10.3233/sw-130128.

Bauer, F., Kaltenböck M., 2011. Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers. Vienna.

Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Keizer, J., & Jacques, Y., 2012. Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. International Journal of Metadata, Semantics and Ontologies (IJMSO), 7(1), 65-75. doi:10.1504/IJMSO.2012.048511

Cyganiak, R. & Reynolds, 2014 D. *The RDF Data Cube Vocabulary*. W3C Recommendation, 16 Jan. 2014. URL http://www.w3.org/TR/vocab-data-cube

Fiorelli, M., Pazienza, M., & Stellato, A., 2013. LIME: Towards a Metadata Module for Ontolex. 2nd Workshop on Linked Data in Linguistics: Representing and Linking lexicons, terminologies and other language data. Pisa, Italy.

Janssen, S. J. C., Kraalingen, D. W. G. v., Boogaard, H. L., Wit, A. J. W. d., Franke, G. J., Porter, C., & Athanasiadis, L. N., 2012. *A generic data schema for crop experiment data in food security research.* Paper presented at the Proceedings of the sixth biannual meeting of the International Environmental Modelling and Software Society, Leipzig.

Konstantopoulos, S., Archer, P., Karampiperis, P., Karkaletsis, V., 2012. The POWDER protocol as infrastructure to serving and compressing semantic data. International Journal of Metadata, Semantics and Ontologies 7(1):1-15.

Lee, T.B., Hendler, J., Lassila, O., 2001. The semantic web. Scientific American 284, 28–37.

Michener, W. K., Jones, M. B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution (Personal edition), 27(2), 85-93.

Ossenbruggen van, J., Hildebrand, M., & de Boer, V., 2011. Interactive vocabulary alignment. Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries (pp. 296-307). Berlin, Germany: Springer-Verlag.

Prud'hommeaux, E., Buil-Aranda (eds), 2013. SPARQL 1.1 Federated Query. W3C Recommendation 21 March 2013. URL http://www.w3.org/TR/sparql11-federated-query.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., … Winter, J. M. (2013). The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agricultural and Forest Meteorology, 170(0), 166-182. doi: http://dx.doi.org/10.1016/j.agrformet.2012.09.011

Stellato, A., Morshed, A., Johannsen, G., Jacques, Y., Caracciolo, C., Rajbhandari, S., . . . Keizer, J., 2011. A Collaborative Framework for Managing and Publishing KOS. The 10th European Networked Knowledge Organisation Systems (NKOS) Workshop. Berlin, Germany.

Villa, F., Athanasiadis, I. N., & Rizzoli, A. E., 2009. Modelling with knowledge: a review of emerging semantic approaches to environmental modelling. Environmental Modelling & Software, 24(5), 577-587. doi: http://dx.doi.org/10.1016/j.envsoft.2008.09.009

White, J. W., Hunt, L. A., Boote, K. J., Jones, J. W., Koo, J., Kim, S.,... Hoogenboom, G., 2013. Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. Computers and Electronics in Agriculture, 96(0), 1-12. doi: http://dx.doi.org/10.1016/j.compag.2013.04.003

World Wide Web Consortium (W3C), 2009. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). (A. Miles, & S. Bechhofer, Eds.) Retrieved March 22, 2011, from World Wide Web Consortium (W3C): http://www.w3.org/TR/skos-reference/skos-xl.html

World Wide Web Consortium (W3C), 2009. SKOS Simple Knowledge Organization System Reference. (A. Miles, & S. Bechhofer, Eds.) Retrieved March 22, 2011, from World Wide Web Consortium (W3C): http://www.w3.org/TR/skos-reference/