A DATAHUB FOR SEMANTIC INTEROPERABILITY IN DATA-DRIVEN INTEGRATED GREENHOUSE SYSTEMS

Jack Verhoosel[1], Barry Nouwt[1], Roos Bakker[1], Athanasios Sapounas[2] and Bart Slager[2]

[1]TNO, Department of Data Science, Soesterberg, The Netherlands

[2]TNO, Construction Innovation Centre, Delft, The Netherlands

jack.verhoosel@tno.nl, barry.nouwt@tno.nl, roos.bakker@tno.nl, athanasios.sapounas@tno.nl, bart.slager@tno.nl

ABSTRACT

This paper deals with the challenge of semantic alignment of different data sources in the horticultural sector. In this sector, greenhouses are used to grow vegetables and plants and the main goal for a greenhouse grower is to control the climate such that crop is optimally cultivated against the lowest cost. Combining available data sources to extract trends and patterns via data analysis, it is important to better support growing decisions. A Common Greenhouse Ontology (CGO) has been developed and used in a Datahub to make data sources accessible via Resource Description Framework (RDF) and a SPARQL interface on top of an Apache Jena Fuseki triplestore. The Datahub was applied in a trial use case in which three data sources where made accessible for a linear regression component that derived patterns between nutrients used and crop growth. One of the lessons learned is that the use of a common ontology very well supports the aligned use of data in analysis and thus better supports decision making.

Keywords*:* semantic alignment, ontologies, greenhouse, data analysis, decision support

1. INTRODUCTION

In the horticultural sector, greenhouses are used to grow vegetables and plants. The main goal for a greenhouse grower is to control the climate such that crop is optimally cultivated against the lowest cost. Sufficient expertise of the grower about the crop in relation to the greenhouse climate is an important prerequisite for achieving this goal. On the other hand, a greenhouse climate computer is one of the most important systems to support the grower in his decisions about defining the climate and growing strategy. However, there are also other data sources around the greenhouse that can be used to further support growers, such as weather, configuration and performance of greenhouse systems, crop growth, yield figures, fertigation strategies and labour planning. The challenge is to semantically align the data from different data sources, such that it can be jointly used to extract trends and patterns via data analysis to better support decisions taken by the grower or even to be used for automatic guidance of control systems. In a national project DDINGS (Data-Driven INtegrated Greenhouse Systems), a datahub was developed with a Common Greenhouse Ontology (CGO) as a solution to this challenge together with main greenhouse construction companies and equipment suppliers. First, some related work on horticultural ontologies is described. The remainder of the paper describes the Datahub, CGO and some application results.

1.1 Related Work on Horticultural Ontologies

In agriculture, several efforts have been made to develop ontologies for the domain. A few webportals provide search engines in a larger set of agricultural ontologies, such as AgroPortal (http://agroportal.lirmm.fr) and GODAN (https://vest.agrisemantics.org). However, they do not contain an ontology that describes the concepts in and around a greenhouse. In addition, (Rehman, 2015) gives a brief overview. AGROVOC and the Advanced Ontology Service (AOS) project was proposed by the Food and Agriculture Organization of the United Nations (FAO) for the development of agricultural ontologies based on their multilingual thesaurus as described by (Soergel et al., 2004). AGROVOC is a multilingual agricultural thesaurus and contains over 32,000 concepts in 27 languages. It comes close to an ontology and is the largest available agricultural thesaurus that is still being maintained. Smaller and older examples of ontologies are the PLANTS ontology (Goumopoulos et al., 2004), OntoCrop (Maliappis, 2009), Crop-Pest Ontology (Beck et al., 2005), Irrigation Ontology (Cornejo et al., 2005), AgriOnto (Xie et al., 2008), ONTAgri (Rehman and Shaikh, 2011) and the Crop Research Ontology (http://www.cropontology.org/). In (Roussey et al., 2013) and (Amarger et al., 2014) a small crop production ontology is described as well as an approach to use ontology design patterns for combining different ontologies into one. Most of these ontologies are either out-of-date, not maintained anymore, only partly available or simply not covering the specific greenhouse domain that is in scope. Therefore, we developed our own Common Greenhouse Ontology (CGO) to support the integration of data in and around the greenhouse. Where possible, the CGO reuses these existing ontologies that describe part of the domain.

2. The DDINGS Datahub

The DDINGS datahub connects and combines various data sources and enables data analysis for better decision support of the grower. In order to show the feasibility of our approach and the datahub, data from a trial experiment with different fertilizer recipes for a small plant crop was used. Data on water and leaf chemical analysis and crop growth parameters were measured. These data sources were mapped to the CGO and made available via our datahub. In addition, a linear regression algorithm was used to find the relation between a specific nutrient in the water and leaf chemical analysis and the crop growth.

2.1 The Datahub Concept

The type of customers of greenhouse construction company is shifting from growers to investors. Investors demand guarantees on production in a greenhouse build with certain construction concepts. Therefore, greenhouse construction companies need to shift from providing a turnkey greenhouse construction company to a service provider that supports the grower in its activities. As a result, growers, investors and construction companies want remote monitoring of the greenhouse and its connected devices and equipment. They want insight in the performance of the greenhouse in terms of yield related to growing strategies and greenhouse configuration. They need to tackle anomalies in the climate or crop in the greenhouse by combining data about growing strategies and greenhouse configuration. Moreover, they want to optimize greenhouse construction concepts using intelligence from available data with new ICT analysis techniques, like deep learning.

A lot of data is already available in separate data sources that can help to deal with these challenges. However, a Datahub is needed to connect and combine these data sources (see Figure 1). Our Datahub:

1. Saves development time/money by connecting a data source ***once*** and reuse it for ***multiple*** applications.
2. Aligns ***semantics*** of different data sources to get unambiguous meaning of information.
3. Provides a ***standardized*** interface to speed up application development on top of it.
4. Offers ***storage*** capabilities for long-term historic data capturing and comparison.

5. Enables ***data analysis*** over historic and real-time data of multiple combined data sources.
6. Secures data access via ***authentication*** and ***mandates*** for data users.

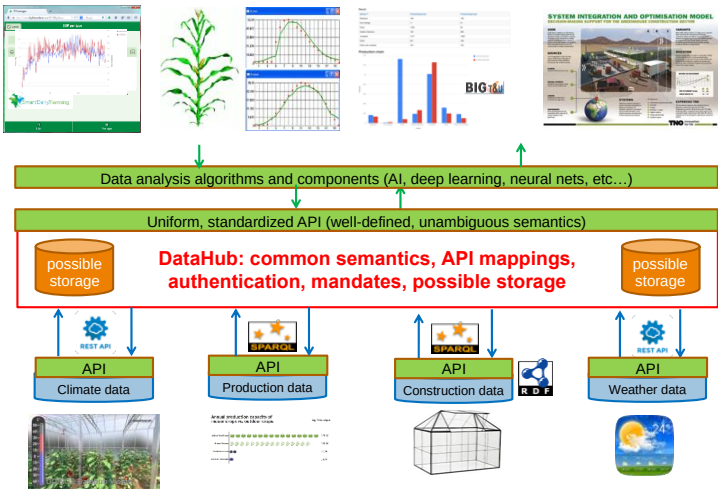

Figure 1. Sketch of the DDINGS Datahub and its context and functionality

The datahub connects relevant data sources and uses a CGO expressed in the Ontology Web Language (OWL) for semantic alignment. The CGO is used to make data accessible in generic, common terms that describe the concepts in and around the greenhouse, such as the construction and topology, the systems present, the crop, the growing system, the fertigation and nutrients strategies, possible diseases, labour planning etc. The CGO is represented using linked data technology based on RDFand is stored in an Apache Jena Fuseki triplestore that forms the main component of our Datahub. The Datahub provides a SPARQL interface that can be used to query the CGO and its concepts and relations. It contains a mapping from the CGO to the specific terms in the data sources made accessible. Finally, the datahub provides data analysis components on top of this interface, to support the grower in making better decisions using machine learning algorithms, such as linear regression, clustering and decision trees.

2.3 The Common Greenhouse Ontology

A greenhouse is a concept which in itself contains many other concepts: crops, technical systems, insects, and more. For each of these concepts one could think up an ontology, therefore existing ontologies were reused. The CGO serves two purposes: the first is to give a truthful representation of the different elements in a greenhouse, the second to support the integration of data such that our datahub can easily access the data.

Different kinds of data need to be able to integrate with the CGO: for example, the thickness of the stem of a flower, the color of a leaf, or the nutrition level of a water sample. One thing that these data types have in common is that they are somehow generated through an observation by some sensor. The semantic sensor network (SSN) ontology was used to describe sensors and their observations (Compton et al., 2012). An observation of some feature of interest, which has an observable property, is made by a sensor that implements a certain procedure to get a result. For example, when the color of a leaf is observed, the leaf is a *Feature of interest*, the color is the *Observable property*, the observer can be a human or a technical *Sensor*, who uses a color scale as a *Procedure* to determine the color. The *Result* of the observation can in this case be expressed in a numerical value and some color code. To express the results of observations, the ontology of units of measure (OM) by (Rijgersberg et al., 2013) was used. This ontology contains a class *Measure*, which is a combination of a numerical value and a measuring unit and can be seen as a subclass of the result class of the SSN ontology.

On top of these ontologies, several classes were added to describe characteristics of a greenhouse, such as the Greenhouse class itself, a *Plant* class which denotes plants and its subclass *Crop*, and

Flower. Another class that was added is the *Part* class, which requires some explanation. A lot of the data that needs to be expressed in the ontology is about a part of a whole, for instance the leaf or the stem of a flower. Also other parts of wholes in the greenhouse domain such as construction parts or system parts need to be modelled. Therefore, it is useful to gather these parts under one class instead of making many different classes. The work of (Rector et al., 2005) was used as an inspiration for the part/whole relations that were defined. In relation to the SSN ontology, the parts are possible features of interest for observations.

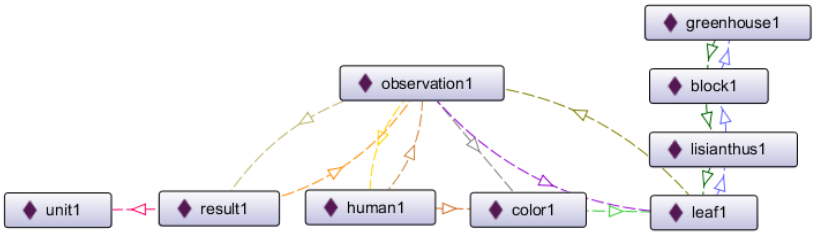

Figure 2. A High-Level Snapshot of an Instantiation of the Common Greenhouse Ontology

Data can be expressed in our CGO concepts and it is shown in Figure 2 how the color of a leaf fits in the ontology through individuals. The area on the right-hand side shows individuals in the CGO that have a part-whole relation with each other. This relation is shown with blue and green arrows. Greenhouse1 has a part (has_part) that is named block1, block1 has a part lisianthus1, which is a flower, and this flower has a part leaf1: the leaf of the flower. This leaf can then be used as a feature of interest for a certain observation: observation1. The observation is made of some property, color1, by a sensor, human1, and gives a result, result1. This result is expressed by a numerical value and a unit as defined by the OM ontology.

3. APPLICATION

3.1 Use Case Analysis

In order to show the feasibility of our Datahub, data was used from a half-year trial experiment. The experiment focused on 6 different fertilizer recipes for Lisianthus, a small plant crop. During three consecutive crop rounds of 6-8 weeks, data related to (1) water chemical analysis, (2) leaf chemical analysis and (3) crop growth parameters were measured in CSV format. This data was transformed to RDF using the LODRefine tool (https://sourceforge.net/projects/lodrefine/) and an RDF skeleton mapped to the CGO and stored it as 3 different data sources into an Apache Jena Fuseki triplestore. As shown in Figure 1, the Datahub mediates between the data sources at the bottom and the applications at the top that give earned value to the users. Data analysis of the datasets collected was done during the three consecutive crop rounds. The resulting application uses the Datahub to retrieve measurements of plant characteristics and analyse and visualize these to the user. A linear regression analysis component was used to find the relation between a specific nutrient in the water and leaf chemical analysis versus the crop growth.

Data analysis was started with consulting the growers about what plant characteristics are most important for them and what is considered to be an optimal plant with respect to, for example, stem thickness, number and colour of leaves and plant length. The result of this consultation was that in general growers prefer fast growing larger plants. Also, higher branch weight, greener leaves and faster blooming is important. The latter because it reduces the heating costs considerably. Based on this consultation, 1) analysis was done on which of the nutrient recipes produced the largest plants and 2) find the influence of the level of natrium on the growth rate. The following Python packages were used: *Numpy*, a fundamental package for scientific computing, *Pandas*, a library providing high-performance, easy-to-use data structures and data analysis tools, *Matplotlib*, a Python 2D plotting library which produces publication quality figures and *Sklearn*, a simple and efficient tool for data mining and data analysis.

During the pre-processing phase, the measurements of the crop were collected and normalized in such a way that it supported the two analysis questions mentioned above. Each row of the resulting table captures per measurement on a particular plant the crop round, the nutritional recipe, the supposed level of natrium in this recipe, the length on that particular date, the age of the plant and its average growth per day since the previous measurement. For the first phase of our analysis, only the measurements from the last week of the crop round were taken to determine the final length of the plant and correlate these with the different recipes. For the second phase, the age of the plant and the average growth per day were taken and correlated with the level of natrium in their nutritional recipe.

3.2 Results

From the first analysis that tried to determine which of the 6 different nutritional recipes produced the largest plants, correlations were found between the height of the plants and the recipe they received. The second phase of the analysis produced Figure 3.

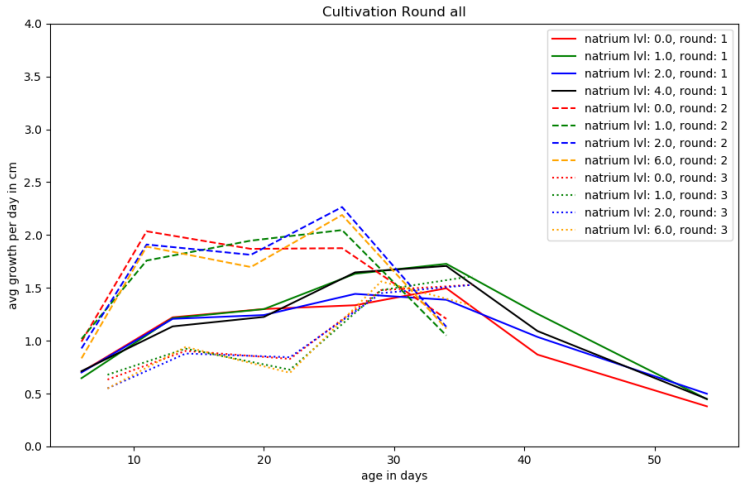

Figure 3. Data Analysis Results for the Nutrient Use case

The age, growth rate and the level of natrium the plants received are correlated. The horizontal axis represents the age in days of the plants, the vertical axis represents the average growth per day of the plant, the color of the lines represents the level of natrium in their recipe and the dash-ness of the lines represent the crop round in which the plants took part. Although higher levels of natrium did not inhibit the growth rate of the plants, as was expected, the figure very clearly shows the influence of the seasons to the growth pattern of the plants. These results have been visualised via the front-end of the datahub.

4. CONCLUSIONS

Lessons learned show that making data sources linkable and combining them in a Datahub with ontologies is feasible. A common ontology enables semantic alignment of the data in different sources towards its users. As an important result, analysis components can use this semantically aligned data unambiguously. Nevertheless, pre-processing of data towards an analysis algorithm remains an important and cumbersome task. Since it is often unclear upfront which combination of features will yield the best results, the preprocessing phase should provide a way to choose these features flexibly and this puts requirements on the technology used during preprocessing. Our conclusion is that the combination of RDF/OWL and SPARQL gives the flexibility that is necessary for a proper data analysis, while still providing normalized and cleaned data. Another conclusion can be drawn on the effects of keeping data at its source on data analysis. This yields a different approach as opposed to centralizing data for analysis. For instance, a reasoning component that orchestrates the data analysis process can

be used to retrieve distributed data on request during the analysis without storing the data first in a central data source. Our Datahub is currently being extended with a few use cases that shows this behavior. Finally, a difference in modelling approaches was encountered between the domain expert and the ontology designer, as they tend to look differently at the concepts to be modelled in the ontology. For instance, a gap existed between model technical terms like 'Part' or 'Sensor' (to also include human observators) and domain terminology who rejects some of the model-technical terms. Our solution was to introduce a 'model-technical term' flag that was put on concepts in the ontology that were necessary from a modelling perspective but not well recognized by domain experts. Currently, our Datahub is also being extended with other external data sources with different API's.

REFERENCES

Rehman A.U. (2015) 'Smart Agriculture: An Approach Towards Better Agriculture Management', OMICS group eBooks, https://www.esciencecentral.org/ebooks/smart-agriculture-an-approach-towards-better-agriculture-management/pdf/agricultural-ontologies.pdf, Foster City, USA.

Soergel D, Lauser B, Liang A, Fisseha F, Keizer J, et al. (2004) 'Reengineering Thesauri for New Applications: the AGROVOC Example', Journal of Digital Information 4: 1-23.

Goumopoulos C, Christopoulou E, Drossos N, Kameas A (2004) 'The Plants system: enabling mixed societies of communicating plants and artefacts', Ambient Intelligence p: 184-195.

Maliappis M.T. (2009) 'Using Agricultural Ontologies', Metadata and Semantics p: 493-498.

Beck H.W., Kim S., Hagan D. (2005) 'A Crop-Pest Ontology for Extension Publications', in 2005 EFITA/WCCA Joint Congress on IT in Agriculture, Vila Real, Portugal p: 1169-1176.

Cornejo C, Beck H.W., Haman D.Z., Zazueta F.S. (2005) 'Development and Application of an Irrigation Ontology', in Joint conference, 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment, 3rd World Congress of Computers in Agriculture and Natural Resources, Vila Real, Portugal.

Xie N., Wang W, Yang Y (2008) 'Ontology-based Agricultural Knowledge Acquisition and Application', Computer and Computing Technologies in Agriculture 1: 349-357.

Rehman A.U., Shaikh Z.A. (2011) 'ONTAgri: Scalable Service Oriented Agriculture Ontology for Precision Farming', in 2011 International Conference on Agricultural and Biosystems Engineering (ICABE 2011), Advances in Biomedical Engineering p: 411-413.

Roussey C., Chanet J-P., Cellier V., Amarger F. (2013) 'Agronomic Taxon', in Proceedings of the Second International Workshop on Open Data, BNF Paris, France.

Amarger F., Chanet J-P., Haemmerle O., Hernandez N., Roussey C. (2014) 'SKOS Sources Transformations for Ontology Engineering: Agronomical Taxonomy Use case', in 2014 Proceedings of Metadata and Semantics Research Conference (MTSR 2014), Karlsruhe, Germany, Springer, Communications in Computer and Information Science, Volume 478, pp 314-328.

Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., ... & Huang, V. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Web semantics: science, services and agents on the World Wide Web*, *17*, 25-32.

Rijgersberg, H., Van Assem, M., & Top, J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, *4*(1), 3-13.

Rector, A., Welty, C., Noy, N., & Wallace, E. (2005). Simple part-whole relations in OWL Ontologies.