EFITA WCCA CIGR
2013 Conference
June 23-27, 2013
TORINO, ITALY
Sustainable Agriculture through ICT innovation

From Algal Biomass to Bioenergy via Semantic Web and Linked data

Monika Solanki*[1] and Johannes Skarka[2]

[1]Aston Business School
Aston University, Birmingham, UK
m.solanki@aston.ac.uk
[2] Karlsruhe Institute of Technology, ITAS
Karlsruhe, Germany
johannes.skarka@kit.edu

ABSTRACT

In this paper we present an account of the publication of a suite of datasets, *LEAPS*, that collectively enable the evaluation of potential algal biomass production sites in North West Europe (NWE). *LEAPS* forms the basis of a prototype Web application that enables stakeholders in the algal biomass domain to interactively explore via various facets, potential algal production sites and sources of their consumables across NWE.

Keywords: Algae, Biomass, Bioenergy, Triplification, Linked data, Ontologies, United Kingdom

1. INTRODUCTION

The last few decades have seen a consistent rise in energy and oil prices along with a significant depletion of fossil fuel resources. Recently algal biomass has been identified as a potential source of large scale production of biofuels. In order to derive fuels from biomass, algal operation plant sites are setup, that facilitate biomass cultivation and conversion of the biomass into end use products, some of which are biofuels. In their report the IEA Bioenergy Task 39 (Darzins, A. and Pienkos, P. and Edye. L. 2010.) point out that there is currently no comprehensive analysis on the resource potential of algal biomass available and emphasize the need for such a work.

In this paper we present *LEAPS* - **L**inked **E**ntities for **A**lgal **P**lant **S**ites, a suite of linked datasets that collectively enable the evaluation of the potential of algal biomass production sites in North West Europe (NWE). The framework underlying *LEAPS* has been developed within the context of the EnAlgae[1] project. In Section 2 we present the motivation behind curating the *LEAPS* dataset suite. Section 3 provides an account of the transformation of raw datasets to their linked data (Bizer, C. and Heath, T. and Berners-Lee, T. 2009.) counterparts. Section 4 describes our prototype Web application[2] and finally Section 5 presents conclusions and discusses future work.

* Corresponding author
[1] http://www.enalgae.eu/
[2] http://www.semanticwebservices.org/enalgae/

<M Solanki, J Skarka>. <"From Algal Biomass to Bioenergy via Semantic Web and Linked data">. EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy, 24-27 June 2013. The authors are solely responsible for the content of this technical presentation. The technical presentation does not necessarily reflect the official position of the Internation Commission of Agricultural and Biosystems Engineering (CIGR) and of the EFITA association, and its printing and distribution does not constitute an endorsement of views which may be expressed. Technical presentations are not subject to the formal peer review process by CIGR editorial committees; therefore, they are not to be presented as refereed publications.

2. MOTIVATION

The idea that algae biomass based biofuels could serve as an alternative to fossil fuels has been embraced by councils across the globe. Major companies, government bodies and dedicated non-profit organisations such as ABO (Algal Biomass Organisation) and EABA (European Algal Biomass Association) have been pushing the case for research into clean energy sources including algae biomass based biofuels.

Within the context of algae production, a major objective of the EnAlgae project is to create a network of pilot scale algal facilities across NWE in order to address the current lack of verifiable information on algal productivity. The integrated network incorporates an up to date inventory in which pilots collect and share data in a standardised manner and provide demonstrations to diverse project partners, observers and stakeholders.

One of the key gaps identified within the algal biomass domain is the lack of a semantically enriched infrastructure for sharing and reusing knowledge. An introspection of the algae-to-biofuels lifecycle reveals several layers where Semantic Web standards and linked data technologies could be very successfully applied and immensely benefit the community. Algal biomass data manifests itself across several facets. At a very high level, the value chain for algal biomass ranges from cultivation of algae to production of biofuels and other products from the cultivated biomass. Figure 1 depicts a schematic representation of the algal biofuel value chain stages and the contributions that Semantic Web and linked data could bring to each of the stages.

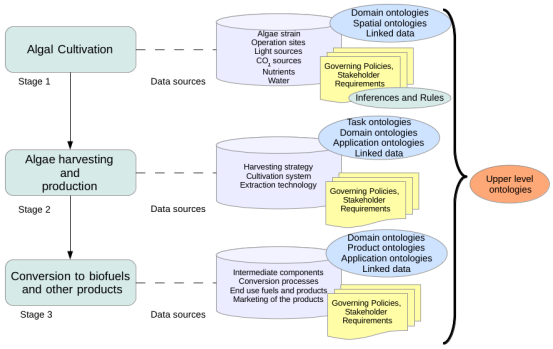

Figure 1. The Algal Biomass Supply Chain

Stage 1 encompasses the cultivation of algae. It involves setting up an algal cultivation site and incorporates datasets about location related geographical in- formation about the sites, locations of sources of light, CO_2, nutrients, water and labour. The linked data for datasets is described using domain specific and spatial ontologies. Stage 2 is concerned with the harvesting of algal biomass. Datasets and vocabularies related to harvesting strategies and extraction techniques are the key semantic outputs of this stage. Stage 3 involves the conversion of biomass to end use products such as biofuels and other constituents. Application ontologies and product ontologies such as GoodRelations will be crucial in describing the datasets for this stage.

In this paper we showcase the publication of linked data for some of the datasets from stage 1. The objective of *LEAPS* is to enable the stakeholders of the algal biomass domain to interactively explore, via linked data, potential algal production sites and sources of their consumables across NUTS (Nomenclature of Units for Territorial Statistics) regions in NWE.

3. TRANSFORMATION OF RAW DATASETS TO LINKED DATA

All the datasets were openly available in non-RDF formats with various origins (Solanki, M. and Skarka J. and Chapman, C. 2013). By performing potential analysis on different NUTS levels, regions with high potential were identified. The calculations were based on high resolution (300 m) data on possible algae production sites and data on CO_2 sources.

The transformation of the raw datasets to linked data takes place in two steps. The first part of the data processing and the potential calculation are performed in a GIS-based model which was developed for this purpose using ArcGIS 9.3.1. Raw datasets with various origins and formats were first transformed using bespoke computational algorithms to an ArcGIS specific XML format. This step is very crucial for two main reasons:

- It brings uniformity in the format of representation of the datasets.
- In the process of transformation, important computations that are part of the final datasets are performed.

The second step of lifting the data from XML to RDF is carried out using a bespoke parser that exploits XPath to selectively query the XML dataset and generate linked data using the ontologies illustrated in Figure 3 and a linking engine. While in most cases, transforming XML datasets to their linked data counterparts is done assuming a simplistic one-to-one mapping between the XML elements and RDF entities, in our scenario, the original data sources had several limitations and a one-to-one transformation was not possible. A bespoke engine (Solanki, M. and Skarka, J. and Chapman, C. 2012) was developed that enabled the transformation for each of the datasets.

3.1 Architecture

An architecture underlying the transformation process is depicted in Figure 2.

The main components of the application are

- **Parsing modules**: As shown in Figure 2, the parsing modules are responsible for lifting the data from their original formats to RDF. The lifting process takes place in two stages to ensure uniformity in transformation.

- **Linking engine**: The linking engine along with the bespoke XML parser is responsible for producing the linked data representation of the datasets. The linking engine uses ontologies, dataset specific rules and heuristics to generate interlinking between the five datasets.

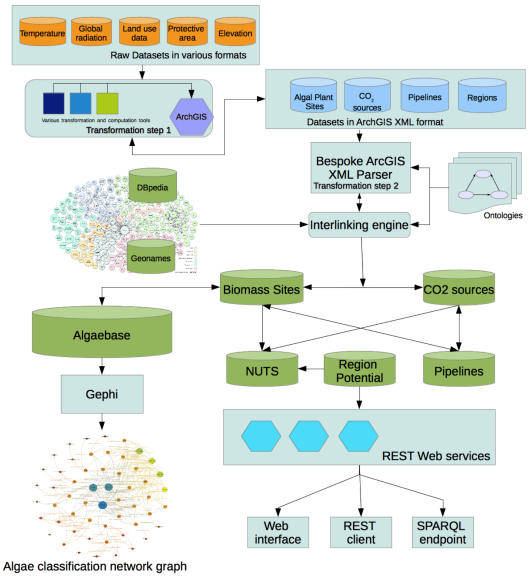

Figure 2. The *LEAPS* Architecture

- **Triple store**: The linked datasets are stored in a triple store. We use OWLIM SE 5.0 13.
- **Web services**: Several REST Web services have been implemented to provide access to the linked datasets.
- **SPARQL endpoints**: SPARQL endpoints that provide access to individual dataset repositorie are available. Snorql has been customised as the front end for the endpoint.
- **Ontologies**: A suite of OWL ontologies for the algal biomass domain have been designed and made available.
- **Interfaces**: The Web interface provides an interactive way to explore various facets of sites, sources, pipelines, regions, ontologies and SPARQL endpoints. The map visualisation has been rendered using Google maps. Besides the SPARQL endpoint and the interactive Web interface, a REST client has been implemented for access to the datasets. Query results are available in RDF/ XML, JSON, Turtle and XML formats.
- **Biological taxonomy visualisation**: A subset of the Algaebase database which is the largest information source of algae on the Web, has been retrieved and curated in our triple store. This dataset when integrated with the dataset for algal cultivation site, can inform stakeholders about the strains of algae that can be harvested on that site. Further, the Semantic Import plugin of Gephi has been

exploited to visualise the biological taxonomy of algae. This visualisation is also made available via the LEAPS interface

3.2 Vocabularies

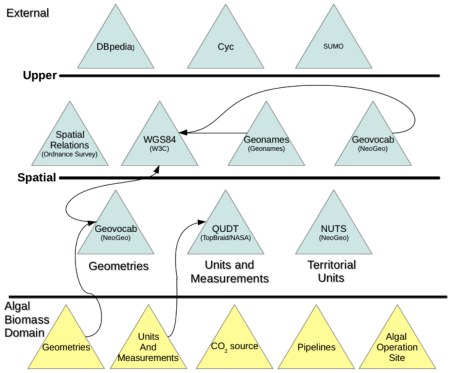

Figure 3. Vocabularies for Algal Biomass. Arrows indicate reuse

LEAPS utilises a set of several well established and domain specific vocabularies as illustrated in Figure 3. Spatial data has been modelled using a combination of several ontologies namely, WGS84 ontology, spatial relations ontology, the Geonames ontology and the NeoGeo ontology.

Geometries for algal plant sites and pipelines have been modelled using an extension of the NeoGeo geometry ontology. For the CO2 sources, the geometry is modelled as a Point from the WGS84 ontology. Modelling units and measurements for various attributes of the algal biomass datasets was non trivial. The QUDT ontology for dimensions and units was extended to include bespoke units of measurements.

To the best of our knowledge, there has been no effort so far within the algal biomass community that exploits the potential of linked data and Semantic Web technologies for the structured representation and sharing of knowledge. Therefore there are no controlled vocabularies and ontologies available to be readily reused or extended. We developed conceptual OWL ontology schemas for algal plant site, CO_2 sources, regions and pipelines. Figure 4 illustrates some of the core concepts, their relationships and attributes. The figure also shows the relationship with the NUTS vocabulary.

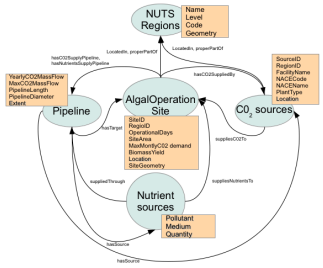

Figure 4. A partial account of core concepts, their attributes and relationships

---

<Insert here the number of your paper in format: **C0XXX**
<M Solanki, J Skarka>. <"From Algal Biomass to Bioenergy via Semantic Web and Linked data">. EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy, 24-27 June 2013.

3.3 The *LEAPS* Dataset

The transformation process yielded four datasets which were stored in distributed triple store repositories: Biomass production sites, CO_2 sources, pipelines for the CO_2 sources and region potential. We stored the datasets in separate repositories to simulate the realistic scenario of these datasets being made available by distinct and dedicated dataset providers in the future. While a linked data representation of the NUTS regions data, was already available there was no SPARQL endpoint or service to query the dataset for region names. We retrieved the dataset dump.

In order to improve the query retrieval performance, we pruned the dataset to include only the regions in NWE. We then curated the pruned dataset in our local triple store as a separate repository. The NUTS dataset was required to link the biomass production sites and the CO_2 sources to regions where they would be located and to the dataset about the region potential of biomass yields. We further enhanced and augmented the NUTS dataset, with data on global radiation. The transformed datasets, interlinked resources defining sites, CO_2 sources, pipelines, regions and NUTS data using link predicates defined in the ontology network depicted in Figure 3. Figure 5 illustrates the linkages between some of the datasets.

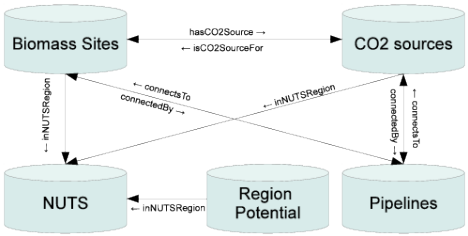

Figure 5. Some of the interlinked datasets

4. PROTOTYPE APPLICATION

The *LEAPS* integrated datasets enables a screening for promising individual sites, provides base data for more detailed planning purposes and would be immensely useful to stakeholders in research, national councils and industry.
We have developed a prototype application with a Web interface built over RESTful Web services that exposes the LEAPS datasets via various facets. The Web interface provides an interactive way to explore various facets of sites, sources, pipelines, regions, ontologies and SPARQL endpoints. Figure 6 provides exemplar screen shots of the application. The map visualization has been rendered using Google maps. Besides the SPARQL endpoint and the interactive Web interface, a REST client has been implemented for access to the datasets. Query results are available in RDF/XML, JSON, Turtle and XML formats.

The interface currently provides visualisation and navigation of the algae cultivation datasets in a way most intuitive for the phycologists. For the stakeholders in the biomass domain, the application provides an integrated view over multiple heterogeneous datasets of potential algal sites and sources of their consumables across NUTS regions

in NWE. The application has been demonstrated to several stakeholders of the community at various algae-related workshops and congresses. They have found the navigation very useful and made suggestions for future dataset aggregation. At the time of this writing, data retrieval is relatively slow for some queries because of their federated nature, however optimisation work on the retrieval mechanism is in progress to enable faster retrieval of information.

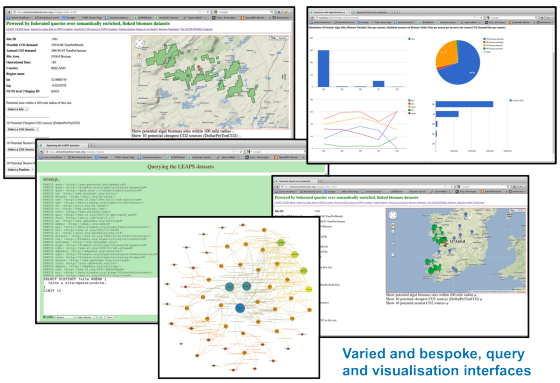

Figure 6. Exemplar Screenshots from the Prototype Application

Besides the application Web interface, datasets are available for querying via the dedicated triple store Web interface. The VoID descriptions of some of the datasets in the *LEAPS* suite have been made available. Once the datasets are made public the VoID descriptions will be updated.

6. CONCLUSIONS

In this paper we presented a framework, *LEAPS,* that exploits Semantic Web and linked data for making the analysis of biomass potential in NWE available to the stakeholders. While *LEAPS* currently provides integrated information about algal plant sites, CO2 sources and the pipelines connecting them, there are several other datasets such as nutrients, water supply and their associated sources which need to be integrated once they become available.

One of the core datasets which should be made available as linked data is that of algal strains that can be cultivated on the plant sites. We have recently curated the Algaebase dataset as linked data. In the near future, experiments would be carried out on the potential sites to establish the algal strains that can be cultivated there. The algal strains from the Algaebase dataset will then be integrated within *LEAPS* to link the potential biomass production sites with the algal strains they could produce. We believe this will go a long way in providing the stakeholders, information about the kind of algae that can be cultivated on potential sites, thereby helping in a more accurate analysis of the economic potential of producing biofuels from Algae.

A limitation of *LEAPS* is the low number of out- going links it provides with other datasets. Currently *LEAPS* links to DBpedia, Geonames and the NUTS datasets. Three main reasons can be identified for the shortcoming in linkages:

– The lack of motivation within the algal biomass community to open up and share data.
– The lack of shared vocabularies and uptake of Se- mantic Web and linked data technologies within the community. *LEAPS* is the first dataset suite to be exposed as linked data using RDF.
– *LEAPS* is a newly curated dataset. Its availability as a data source to which other datasets can pro- vide outgoing links needs to be widely advertised both within and across the domain.

In order to increase the uptake and showcase the potential of *LEAPS* we have been presenting the application at various algae congresses and workshops. This also informs us about any related datasets that can be integrated within *LEAPS*. We are working with biologists in the domain to facilitate the process of making the taxonomy from the AquaFuels33 project avail- able as SKOS models. Multifaceted visualisation of the integrated datasets is another area that we are currently focusing on to motivate the idea of interlinking datasets. A few examples of these visualisations34 can be seen via the Web application. The reasoning infrastructure in *LEAPS* is currently based on implicit OWL DL inferences. Work is also in progress on exploiting rule based reasoning to model domain specific constraints.

x. REFERENCES[1]

Bizer, C. and Heath, T. and Berners-Lee, T. 2009. Linked data - the story so far. International Journal on Semantic Web and Information Systems, 2009.

Darzins, A. and Pienkos, P. and Edye. L. 2010. Current Status and Potential for Algal Biofuels Production. IEA Bioenergy Task 39, 2010.

Solanki, M. and Skarka J. and Chapman, C. 2012. LEAPS: Realising the Potential of Algal Biomass Production through Semantic Web and Linked data. In *I-Semantics 2012: Proceedings of the 8th International Conference on Semantic Systems*. ACM ICP Series, 2012

Solanki, M. and Skarka J. and Chapman, C. 2013. Linked data for Potential Algal Biomass Production. In Semantic Web Journal, IOS Press, 2013.

[1] Due to space restrictions we do not include references to ontologies, vocabularies and third party software used in the development of LEAPS. Detailed references to these can be found in the last two papers in the References section.