LEAPS: A Semantic Web and Linked data framework for the Algal Biomass Domain

Monika Solanki[1] and Johannes Skarka[2]

[1] Birmingham City University, UK
monika.solanki@bcu.ac.uk
[2] Karlsruhe Institute of Technology, ITAS, Germany
johannes.skarka@kit.edu

Abstract. In this paper we present, *LEAPS*, a Semantic Web and Linked data framework for searching and visualising datasets from the domain of Algal biomass. *LEAPS* provides tailored interfaces to explore algal biomass datasets via REST services and a SPARQL endpoint for stakeholders in the domain of algal biomass. The rich suite of datasets include data about potential algal biomass cultivation sites, sources of CO_2, the pipelines connecting the cultivation sites to the CO_2 sources and a subset of the biological taxonomy of algae derived from the world's largest online information source on algae.

1 Motivation

Algal biomass holds huge promises. The use of microalgae as a food source for humans has been considered for overpopulated countries and for space travel since as early as 1961 [3]. If algae is grown under proper environmental conditions, the protein yield from it may be quite high. Algae have been collected for more than 4000 years in China and Japan for use as human food [3],

Recently 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 [1, 2], government bodies [4] and dedicated non-profit organisations such as ABO (Algal Biomass Organisation) [4] and EABA(European Algal Biomass Association)[5] have been pushing the case for research into clean energy sources including algae biomass based biofuels.

It is quickly evident that because of extensive research being carried out, the domain itself is a very rich source of information. Most of the knowledge is however largely buried in various formats of images, spreadsheets, proprietary data sources and grey literature that are not readily machine accessible/interpretable. A critical limitation that has been identified is the lack of a knowledge level infrastructure that is equipped with the capabilities to provide semantic

[3] http://www.botgard.ucla.edu/html/botanytextbooks/economicbotany/Algae/index.html
[4] http://www.algalbiomass.org/
[5] http://www.eaba-association.eu/

grounding to the datasets for algal biomass so that they can be interlinked, shared and reused within the biomass community.

Integrating algal biomass datasets to enable knowledge representation and reasoning requires a technology infrastructure based on formalised and shared vocabularies. Stakeholders in the domain who would benefit from such a structured, unambiguous and machine interpretable representation of data include researchers, algae producers and users, biofuels producers, oil companies, airline, cars and aerospace industry, national public authorities, international organisation and NGOs amongst others.

In this paper, we present *LEAPS*[6], a Semantic Web/Linked data framework for the representation and visualisation of knowledge in the domain of algal biomass. One of the main goals of *LEAPS* is to enable the stakeholders of the algal biomass domain to interactively explore, via linked data, potential algal sites and sources of their consumables across NUTS (Nomenclature of Units for Territorial Statistics)[7] regions in North-Western Europe.

Some of the objectives of *LEAPS* are,

- motivate the use of Semantic Web technologies and LOD for the algal biomass domain.
- laying out a set of ontological requirements for knowledge representation that support the publication of algal biomass data.
- elaborating on how algal biomass datasets are transformed to their corresponding RDF model representation.
- interlinking the generated RDF datasets along spatial dimensions with other datasets on the Web of data.
- visualising the linked datasets via an end user LOD REST Web service.
- visualising the scientific classification of the algae species as large network graphs.

The paper is structured as follows: Section 2 presents a brief overview of the dataset transformation process. Section 3 presents a description of the system architecture. Section 4 presents an overview of the querying mechanism underlying the *LEAPS* interface.

2 *LEAPS* Datasets

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 [8] 9.3.1.

The second step of lifting the data from XML to RDF is carried out using a bespoke parser that exploits XPath [9] to selectively query the XML datasets and generate linked data using the ontologies. While in most cases, transforming

[6] `http://www.semanticwebservices.org/enalgae`

[7] `http://bit.ly/I7y5st`

[8] http://www.esri.com/software/arcgis/index.html

[9] http://www.w3.org/TR/xpath/

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. In order to produce a linked data representation of the datasets, that directly interlinked the resources of sites, sources, pipelines and region potential to each other and their NUTS regions of location, a bespoke parser that utilised a complex underlying data structure to facilitate the transformation was implemented.

The transformation process yielded four datasets which were stored in distributed triple store repositories: Biomass production sites, CO_2 sources, pipelines 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 [10], was already available there was no SPARQL endpoint or service to query the dataset for region names. We retrieved the dataset dump and curated it 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. The transformed datasets interlinked resources defining sites, CO_2 sources, pipelines, regions and NUTS data using link predicates defined in the ontology network.

Datasets about algae cultivation can become more meaningful and useful to the biomass community, if they are integrated with datasets about algal strains. This can help the plant operators in taking judicious decisions about which strain to cultivate at a specific geospatial location. Algaebase[11] provides the largest online database of algae information. While Algaebase does not make RDF versions of the datasets directly available through its website, they can be programmatically retrieved via their LSIDs (Life Science Identifiers) from the LSID Web resolver [12] made available by Biodiversity Information Standards (TDWG)[13] working group.

We retrieved RDF metadata for 113061 species of algae[14] and curated in our triple store. We then used the Semantic import plugin with Gephi to visualise the biological taxonomy of the algae species.

3 System Description

LEAPS provides an integrated view over multiple heterogeneous datasets of potential algal sites and sources of their consumables across NUTS regions in North-Western Europe. Figure 1 illustrates the conceptual architecture of *LEAPS*. The main components of the application are

[10] `http://nuts.geovocab.org/`

[11] http://www.algaebase.org/about/

[12] http://lsid.tdwg.org/

[13] http://www.tdwg.org/

[14] The retrieval algorithm ran on an Ubuntu server for three days

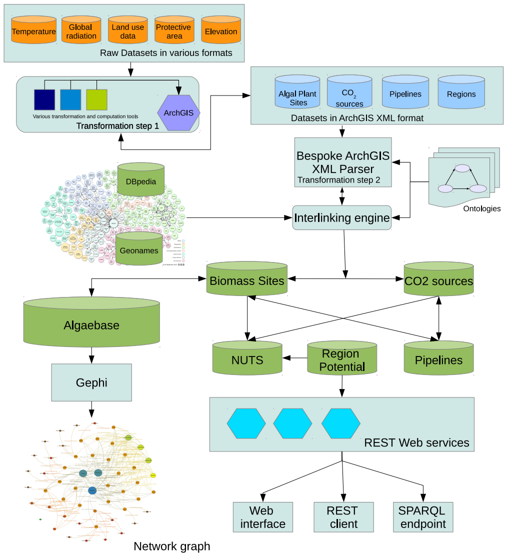

Fig. 1. Architecture of *LEAPS*

- **Parsing modules**: As shown in Figure 1, 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. From the LOD cloud, we currently provide outgoing links to DBpedia[15] and Geonames[16].
- **Triple store**: The linked datasets are stored in a triple store. We use OWLIM SE 5.0 [17].
- **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 repositories are available. Snorql has been customised as the front end for the endpoint. An endpoint for federated queries is planned to be implemented as part of future work.
- **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, ontolgoies and SPARQL endpoints.

[15] http://dbpedia.org/About

[16] http://sws.geonames.org/

[17] http://www.ontotext.com/owlim/editions

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[18] of Gephi[19] has been exploited to visualise the biological taxonomy of algae. This visualisation is also made available via the *LEAPS* interface.

4 Queries powering *LEAPS*

Since our objective was to assess the potential of the production of algal biomass in NUTS regions of North Western Europe, most of the queries over the datasets are based on retrieving knowledge centered around location information. The queries are federated across the various repositories holding the linked data. As an example consider the informal query,

Which are the algal operation sites with CO_2 sources that have CO_2 emissions less than 130000 kgs, where total costs of supplying CO_2 is lower then 5000 GBP per ton of CO_2, areal yield is greater than 30 tons per hectare and which are located within the NUTS region "UKM61"? Supplement the data with supporting information about the region.

The above query is federated between various datasets: the sites dataset provides location data (lat., lng. for the sites) and data about areal yield, the CO_2 sources dataset provides CO_2 emission data for the sources and the pipelines dataset provides information about the total cost of supplying CO_2 to the sites. The NUTS regions dataset includes coreferences to the DBpedia and Geonames dataset, which provides the supporting information required to supplement the results retrieved from the query. A SPARQL representation of the query is listed below.

```
PREFIX site:<http://biomass.org/algae/ontologies/biomass#>
PREFIX co2:<http://biomass.org/algae/ontologies/co2source#>
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX qudt:<http://qudt.org/1.1/vocab/dimensionalunit#>
PREFIX omgeo: <http://www.ontotext.com/owlim/geo#>
PREFIX pipe:<http://biomass.org/algae/ontologies/pipeline#>
SELECT DISTINCT ?siteID ?sourceID ?emissionValue
       ?arealYield ?totalCO2CostValue ?related
WHERE {
   SERVICE <http://localhost/repositories/biomass>
   { ?site a site:OperationSite;
     site:inNUTSRegion ?region;
```

[18] http://wiki.gephi.org/index.php/SemanticWebImport

[19] https://gephi.org/

```
    geo:location ?loc. ?loc
    geo:lat ?lat.
    ?loc geo:long ?long.
    ?site site:hasSiteID ?siteID;
    site:hasArealYield ?z.
    ?z qudt:quantityValue ?y.
    ?y qudt:numericValue ?arealYield.
    ?y qudt:unit ?unit.
 }
 SERVICE <http://localhost/repositories/co2source>
 { ?source a co2:CO2Source;
   co2:hasSourceID ?sourceID;
   co2:hasCO2Emission ?emission.
   ?emission qudt:quantityValue ?emissionQty.
   ?emissionQty qudt:numericValue ?emissionValue.
 }
 SERVICE <http://localhost/repositories/pipeline>
 { ?pipe a pipe:Pipeline;
   pipe:hasSiteID ?siteID;
   pipe:hasSourceID ?sourceID;
   pipe:hasTotalCO2Cost ?cost.
   ?cost qudt:quantityValue ?qty.
   ?qty qudt:numericValue ?totalCO2CostValue.
   ?qty qudt:unit ?totalCO2CostUnit.
 }
 SERVICE <http://localhost/repositories/region>
 { regionID a ramon:NUTSRegion;
   owl:sameAs ?related
 }
 FILTER((?emissionValue < 130000)
       && (contains(str(?region), "UKM61"))
       && (?arealYield > 30)
       && (?totalCO2CostValue < 5000) )
}
```

The Web interface of the application highlights several applications of precompiled federated queries. A SPARQL endpoint that allows executing bespoke federated queries is planned as an extension of the application.

5 Application access

LEAPS[20] is available on the Web. The interface currently provides visualisation and navigation of the algae cultivation datasets in a way most intuitive for the phycologists. 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

[20] http://www.semanticwebservices.org/enalgae

queries because of their federated nature, however optimisation work on the retrieval mechanism is in progress to enable faster retrieval of information.

Acknowledgments

The research described in this paper is partly supported by the Energetic Algae project (EnAlgae), a 4 year Strategic Initiative of the INTERREG IVB North West Europe Programme.

References

1. A. H. Claire Smith. Research needs in ecosystem services to support algal biofuels, bioenergy and commodity chemicals production in the uk. Technical report, NNFCC, 2011.
2. Oilgae. Oilgae comprehensive report, energy from algae: Products, market, processes and strategies. Technical report, Oilgae, 2011.
3. R. C. Powell and E. M. Nevels. Algae feeding in humans. *Journal of Nutrition*, 1961.
4. U.S. Department of Energy. National Algal Biofuels Technology Roadmap. Technical report, accessed June 2012.

Appendix

A Justification: Minimum requirements

In this section we provide a justification of *LEAPS* wrt. the minimum requirements set out by the Semantic Web challenge.

- **The application**: *LEAPS* is a Web application[21] for stakeholders in the domain of algal biomass. Stakeholders include investors interesting in funding algae cultivation systems, plant operators responsible for running the algae cultivation plant, phycologists interested in visualising and querying information on algae strains, local councils and government bodies responsible for laying out policies and regulations for algae cultivation.
- **Information sources**: An account of the raw sources of the datasets along with their purpose is available[22]. All the datasets were openly available in non-RDF formats with various origins. The transformation of the raw datasets to linked data takes place in two steps as illustrated in Section 2.
- **Ontologies**: *LEAPS* utilises a set of several well established and domain specific vocabularies. Spatial data has been modelled using a combination of several ontologies namely, WGS84 ontology [23], spatial relations ontology, [24] the Geonames ontology [25] and the NeoGeo ontology [26].

[21] http://www.semanticwebservices.org
[22] http://purl.org/biomass/leaps/rawDatasets
[23] `http://www.w3.org/2003/01/geo/wgs84_pos`
[24] `http://www.ordnancesurvey.co.uk/oswebsite/ontology/spatialrelations.owl`
[25] `http://www.geonames.org/ontology/ontology_v2.2.1.rdf`
[26] `http://geovocab.org/geometry`

Geometries for algal plant sites and pipelines have been modelled using an extension of the NeoGeo geometry ontology [27]. For the CO_2 sources, the geometry is modelled as a `Point` from the WGS84 ontology [28].

Modelling units and measurements for various attributes of the algal biomass datasets was non trivial. The QUDT ontology [29] for dimensions and units was extended to include bespoke units of measurements.

Data retrieved from the Algaebase[30] database has been modelled using a combination of DCterms [31] and Darwin core [32].

While it was relatively easy to discover ontologies for modelling spatial knowledge, units and measurements, discovering vocabularies conceptualising the domain knowledge for algal biomass was non trivial. We developed conceptual OWL ontology schemas [33] for algal plant site, CO_2 sources, regions and pipelines. The design of the ontologies was very strongly guided by feedback from questionnaires made available to the stakeholders, interviews with domain experts, providers of raw datasets and grey literature from the algal biomass and biofuels domain.

B Justification: Desirable features

- *LEAPS* is the **first** Semantic Web/Linked data application in the domain of algal biomass. The ontologies developed within the framework for the domain, have been built from ground up as ontological modelling of algal biomass knowledge has not been undertaken so far.
- *LEAPS* provides a set of tailored, well defined and functional interfaces for the visualisation of algal biomass information. The design of these interfaces have been undertaken in association with stakeholders in the domain, to provide an environment suited to their requirements.
- *LEAPS* has been designed with scalability and modularity as its primary underlying features. As data about sources of water and sources of nutrients become available, these would be added to the *LEAPS* triple store and served through additional *LEAPS* interfaces. While, right now, only a subset of algaebase information is being visualised using Gephi, the integration of the entire dataset is easily feasible because of the scalable design.
- Ranking of results is an implicit part of the *LEAPS* interface, i.e., it is possible to search for the top ten sites for algal biomass production, top ten sites for CO_2 consumption, cheapest and nearest CO_2 sources.
- Unlike many SW applications, *LEAPS* has been built around openly available and real world data on potential sites in NWE where algae can be cultivated. This provides a huge potential for commercial uptake of the applications by industries where algae is utilised as a source of raw material, e.g., the bioenergy and the pharmaceutical industries.

[27] http://geovocab.org/geometry

[28] http://www.w3.org/2003/01/geo/wgs84_pos

[29] http://qudt.org/1.1/vocab/dimensionalunit

[30] http://www.algaebase.org/about/

[31] http://purl.org/dc/terms/

[32] http://rs.tdwg.org/dwc/terms/

[33] Ontologies are available at `http:/purl.org/biomass/ontologies`