Linked data for Potential Algal Biomass Production

Monika Solanki \(^a, \ast\),
\(^a\) Knowledge Based Engineering Lab, Birmingham City University, Birmingham United Kingdom
E-mail: monika.solanki@bcu.ac.uk

Abstract. In this paper we present an account of the publication of a suite of datasets, \textit{LEAPS}, that collectively enable the evaluation of potential algal biomass production sites in North Western Europe. \textit{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 North-Western Europe.

Keywords: Algae, Biomass, Energy, triplification, linked data, Ontologies

1. Introduction

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 this paper we present \textit{LEAPS} - Linked Entities for Algal Plant Sites, a suite of linked datasets that collectively enable the evaluation of the potential of algal biomass production sites in North Western Europe (NWE). The framework underlying \textit{LEAPS} has been developed within the context of the EnAlgae\(^1\) project. In Section 2 we present the motivation behind curating the \textit{LEAPS} dataset suite. Section 3 provides an account of the raw datasets which served as the basis for \textit{LEAPS}. Section 4 illustrates the vocabularies used in the datasets. Section 5 discusses characteristics of the dataset. Section 6 describes the prototype Web application built over \textit{LEAPS}\(^2\). Section 7 outlines limitations and finally Section 8 presents conclusions and discusses future work.

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 [6,1], government bodies [8] and dedicated non-profit organisations such as ABO (Algal Biomass Organisation)\(^3\) and EABA (European Algal Biomass Association)\(^4\) 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

---

\(^*\)Corresponding author. E-mail: monika.solanki@gmail.com.
\(^1\)http://www.enalgae.eu/
\(^2\)http://purl.org/biomass/LEAPSDemo
\(^3\)http://www.algalbiomass.org/
\(^4\)http://www.eaba-association.eu/
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 [5]. Figure 1 depicts a schematic representation of the algal biomass supply chain.

Fig. 1. The Algal biomass supply chain

Table 1 highlights the sources of the datasets along with their purpose. All the datasets were openly available in non-RDF formats with various origins. By performing potential analysis on different NUTS levels, regions with high potential can be identified. The calculations are based on high resolution (300 m) data on possible algae production sites and data on CO₂ 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 are 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 datasets 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

---

5 All figures in the paper are available at http://purl.org/biomass/LEAPSFigures

6 http://purl.org/goodrelations/v1


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

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

10 A snippet of one of the XML dataset is available at http://www.purl.org/biomass/XMLSnippet
our scenario, the original data sources had several limitations and a one-to-one transformation was not possible. A bespoke engine [7] was developed that enabled the transformation for each of the datasets.

An architecture underlying the transformation process is depicted in 2.

For each of the algae production site, information on biomass yield and site area are determined. Additionally, data on CO₂-providing industrial or power plants can be queried for each site and costs for CO₂-supply can be calculated. Thus, the data 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.

It is worth noting that the process of aggregating data for nutrients and water sources for each of the algae production site is in progress. Once these become available they would be integrated with the other datasets as outlined further in Section 5.

4. Vocabularies

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 [11], spatial relations ontology, [12] the Geonames ontology [13] and the NeoGeo ontology [14].

Geometries for algal plant sites and pipelines have been modelled using an extension of the NeoGeo geometry ontology [15]. For the CO₂ sources, the geometry is modelled as a Point from the WGS84 ontology [16].

Modelling units and measurements for various attributes of the algal biomass datasets was non trivial. The QUDT ontology [17] for dimensions and units was extended to include bespoke units of measurements.

We developed conceptual OWL ontology schemas [18] for algal plant site, CO₂ 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 [19] vocabulary.

---

**Fig. 2. Architecture of LEAPS**

**Fig. 3. Ontologies for algal biomass. Arrows indicate reuse**

[16] http://www.w3.org/2003/01/geo/wgs84\_pos
[18] Ontologies are available at http://purl.org/biomass/ontologies
[19]
<table>
<thead>
<tr>
<th>Raw dataset (format)</th>
<th>Purpose</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Global radiation (GRIB (^a))</td>
<td>Calculate the algal biomass yield</td>
<td>NCEP-NCAR 50-year reanalysis provided by the CISL Research Data Archive (^b). DSWRF (downward shortwave radiation flux at the surface) in dataset number ds090.2 (^4).</td>
</tr>
<tr>
<td>2 Temperature (NetCDF (^c))</td>
<td>Calculate the algal biomass yield</td>
<td>E-OBS dataset provided by the ECA&amp;D project (^d) ([10],[2]).</td>
</tr>
<tr>
<td>3 Data on land use (GeoTIFF)</td>
<td>Identify suitable land area</td>
<td>European Environment Agency (EEA), CORINE land cover raster dataset 2006 version 13 (^e). CORINE land cover raster dataset 2000 version 15 (^f).</td>
</tr>
<tr>
<td>4 Protected areas (GIS vector data)</td>
<td>Identify suitable land area</td>
<td>World Database on Protected Areas (WDPA) provided on (^g) by UNEP-WCMC ([3]).</td>
</tr>
<tr>
<td>5 Elevation (GIS raster data)</td>
<td>Identify suitable land area</td>
<td>Shuttle Radar Topography Mission (SRTM) data ([9]) by Global Land Cover Facility (GLCF) (^h).</td>
</tr>
<tr>
<td>6 CO(_2) sources (MS Access)</td>
<td>Identify sources of CO(_2)</td>
<td>European Pollution Release and Transfer Register (E-PRTR) provided by the EEA (^i) in the version of 09 Nov 2009.</td>
</tr>
</tbody>
</table>

\(^a\)http://badc.nerc.ac.uk/help/formats/grib/
\(^b\)http://dss.ucar.edu
\(^c\)http://www.leokrut.com/leonetcdf/netcdfformat.html
\(^d\)http://eca.knmi.nl
\(^g\)www.protectedplanet.net
\(^h\)www.landcover.org
\(^i\)http://bit.ly/v3dS2I

| 5. The LEAPS datasets |

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 \(20\), 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 inform 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 \(21\). The transformed datasets, interlinked resources defining sites, CO\(_2\) sources, pipelines, regions and NUTS data us-

\(^20\)http://nuts.geovocab.org/

\(^21\)global radiation is another term for solar radiation. It is the sum of short wavelength incoming radiation to the earth’s surface and consists of the direct and the diffuse sunlight.
Fig. 4. A partial account of core concepts, their attributes and relationships

ing link predicates defined in the ontology network depicted in Figure 3. Figure 5 illustrates the linkages between the datasets.

![Diagram](https://via.placeholder.com/150)

**Fig. 5. Linked datasets for algal biomass**

**URIs**

We propose URI patterns for the datasets used in this paper to be reused across the sector. Note that while we would like the URIs to be persistent, they may evolve as the uptake of linked data within the algal biomass community gains momentum.

In particular we propose URI sets for

- Algal plant sites
- CO₂ sources
- Pipelines.

Figure 6 exemplifies some of the URIs minted for real world algal biomass entities which have unique identifiers and which are uniquely located in a certain NUTS region. It also illustrates the definition of a conceptual entity and a relationship within the algal plant site ontology.

![Diagram](https://via.placeholder.com/150)

**Fig. 6. Representative URIs for Algal Biomass Plant Site**

**Statement level statistics**

Statement level statistics for the various datasets in LEAPS are indicated in Table 2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algae sites dataset</td>
<td>84703</td>
</tr>
<tr>
<td>CO₂ sources</td>
<td>14238</td>
</tr>
<tr>
<td>Pipelines</td>
<td>261680</td>
</tr>
<tr>
<td>NUTS</td>
<td>14469</td>
</tr>
<tr>
<td>Global radiation</td>
<td>24738</td>
</tr>
<tr>
<td>Biomass potential</td>
<td>4557</td>
</tr>
</tbody>
</table>

**Table 2**

**Dataset statement statistics**

**Dataset availability**

Due to project level, consortium restrictions, we cannot expose the complete LEAPS dataset at the time of this writing, however we have made available, snippets 22 of the various datasets incorporated in LEAPS. Discussions are in the progress with the consortium to make the datasets available before the end of the project.

**VOID description of the datasets**

The VoID23 descriptions of some of the datasets in the LEAPS suite have been made available 24. Once the datasets are made public the VoID descriptions will be updated.

---

22 [http://purl.org/biomass/LEAPSDatasetSnippets](http://purl.org/biomass/LEAPSDatasetSnippets)
23 [http://www.w3.org/TR/void/](http://www.w3.org/TR/void/)
24 [http://purl.org/biomass/void](http://purl.org/biomass/void)
Datasets within the LEAPS suite have been linked to each other via the link predicates in the domain specific vocabularies created for the datasets. The sites datasets have been linked to the CO₂ sources, pipeline datasets and the nuts region dataset. Back links to the datasets have been provided from the CO₂ and the pipelines dataset.

Though we have achieved good interlinking between the datasets in the LEAPS suite, we have only been able to link externally to DBpedia and Geonames. Reasons for the low external linkages have been discussed in Section 7.

6. A 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 ²⁵ 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 7 illustrates a typical site. 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. For the stakeholders

7. Discussion

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 ²⁶ 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 outgoing 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 Semantic 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 provide outgoing links needs to be widely advertised both within and across the domain.

We strongly believe that the LEAPS dataset suite will prove a major milestone in generating the much needed awareness about linked data and its benefits within the algal biomass community. We are working with biologists in the domain to address the above limitations.

---

²⁵ http://purl.org/biomass/LEAPSDemo
²⁶ http://www.algaebase.org
8. Conclusions and Future Work

In this paper we presented a framework LEAPS that exploits Semantic Web and linked data for making the analysis of biomass potential in North-Western Europe available to the stakeholders. Specifically, the framework contributes by

- enabling the screening of data for promising individual plant sites and provides base data for more detailed planning purposes.
- proposing a set of domain specific ontologies for algal plant sites, CO₂ and pipelines to be shared and extended by the community.
- defining a linked data publishing architecture that transforms raw data in disparate formats to a uniform XML representation.
- using a set of well established and domain specific ontologies as metadata to transform it further into linked data.
- providing various data access options such as a SPARQL endpoint, an interactive Google map interface and a REST API for making the data accessible to stakeholders.

As discussed in Section 7, several limitations need to be overcome at various levels in order to fully realise the vision of have an open platform for publishing and consuming algal biomass datasets, both within and across community.

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 AquaFuels 27 project available 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 visualisations 28 can be seen via the Web application. The reasoning infrastructure in LEAPS is currently based on implicit OWL 2 29 DL inferences. Work is also in progress on exploiting rule based reasoning to model domain specific constraints.

References