LEAPS: Realising the Potential of Algal Biomass Production through Semantic Web and Linked data

Monika Solanki*
KBE Lab
Birmingham City University
monika.solanki@bcu.ac.uk

Johannes Skarka
Karlsruhe Institute of Technology (ITAS)
johannes.skarka@kit.edu

Craig Chapman
KBE Lab
Birmingham City University
craig.chapman@bcu.ac.uk

ABSTRACT

Recently algal biomass has been identified as a potential source of large scale production of biofuels. Governments, environmental research councils and special interests groups are funding several efforts that investigate renewable energy production opportunities in this sector. However so far there has been no systematic study that analyses algal biomass potential especially in North-Western Europe. In this paper we present a spatial data integration and analysis framework whereby rich datasets from the algal biomass domain that include data about algal operation sites and CO_2 source sites amongst others are semantically enriched with ontologies specifically designed for the domain and made available as linked data. We then present a conceptual architecture and a prototype implementation of a GeoWeb service that provides querying and analysing capabilities over the linked datasets.

General Terms

Semantic Web, Linked data, Algal biomass, SPARQL, Ontologies

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. This has led to extensive research in the search and production of naturally viable and sustainable energy sources such as biofuels. 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 [4, 14], government bodies [19] and dedicated non-profit organisations such as ABO (Algal Biomass Organisation) [1] and EABA(European Algal Biomass Association)[2] have been pushing the case for research into clean energy sources including algae biomass based biofuels.

*Principal and corresponding author

[1] http://www.algalbiomass.org/

[2] http://www.eaba-association.eu/

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. Microalgal biomass production in Europe is seen as a promising option for biofuels production regarding energy security and sustainability. Since microalgae can be cultivated in photobioreactors on non-arable land this technology could significantly reduce the food vs. fuel dilemma. However, until now there has been no systematic analysis of the algae biomass potential for North-Western Europe. In [20], the authors assessed the resource potential for microalgal biomass but excluded all areas not between 37°N and 37°S, thus most of Europe. In their report the IEA Bioenergy Task 39 [5] 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.

The initiatives of producing biofuels from algae has immense research and commercial potential. It is also 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 grounding to the datasets for algal biomass so that they can be interlinked, shared and reused within the biomass community.

In this paper our objectives are two fold. Firstly, we motivate the use of Semantic Web [2] and linked data [3,9] technologies to integrate, analyse and visualise various facets of data from the algal biomass domain. We believe Semantic Web and linked data have immense potential to contribute towards making the systematic analysis of algal biomass potential available to stakeholders within a unified framework. Secondly, we propose to contribute algal biomass knowledge to the Semantic Web and the growing Web of data by

- 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 [6] Web service, *LEAPS*, (**L**inked **E**ntities for **A**lgal **P**lant **S**ites).

The paper is structured as follows: Section 2 motivates the idea of using Semantic Web and linked data for algal biomass. Section 3 discusses related work. Section 4 presents the requirements for ontological representation of algal biomass knowledge. Section 5 provides an account of the raw datasets and illustrates our methodology for producing linked data. Section 6 presents the framework and conceptual architecture for *LEAPS*. Section 7 illustrates an example of the queries used in *LEAPS* and finally Section 8 presents our conclusions.

2. SEMANTIC WEB AND LINKED DATA FOR ALGAL BIOMASS

One of the key gaps that have been 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 [12]. Each of the core tasks of cultivation, harvesting, processing and fuel production further involves several intermediate processes. Besides managing the process level domain knowledge from cultivation to marketing of the end products, every stage in the algal supply chain is governed by regulatory policies and strategies laid out by local government bodies and requirements defined by stakeholders. Each of the facets consumes and produces a large volume of unstructured data and information that opens up a huge potential for semantically grounded information extraction and knowledge representation technologies to be successfully applied. Figure 1[3] 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. As illustrated, at each stage of the value chain datasets can be described using well established as well as domain specific ontologies. Besides integrating stage specific and intra-stage datasets, contextually relevant datasets from the LOD cloud[4] can also be exploited for interlinking.

Stage 1 encompasses the cultivation of algae. It involves setting up an algal cultivation site and incorporates datasets about location related geographical information 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. In this paper we showcase the publication of linked data for datasets from stage 1.

[3]All figures in the paper are available at `http://purl.org/biomass/LEAPSFigures`

[4]`http://www4.wiwiss.fu-berlin.de/lodcloud/state/`

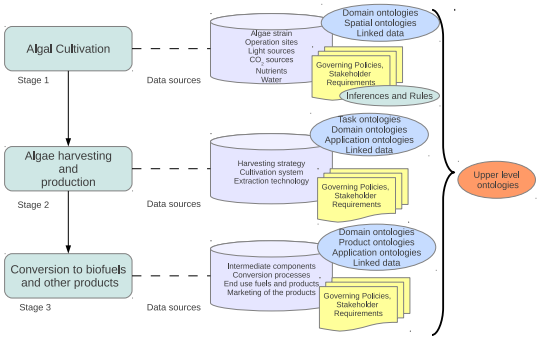

Figure 1: The Algal biomass supply chain

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. Besides domain specific ontologies, task and application ontologies for the processes involved in this stage will play a crucial role in modelling knowledge.

Stage 3 involves the conversion of biomass to end use products such as biofuels and other constituents. Marketing of the products is also an integral part of this stage. Application ontologies and product ontologies such as GoodRelations[5] will be crucial in describing the datasets for this stage.

Upper level ontologies such as SUMO [6] and BFO [7] can be employed as metadata specification at all the three stages. Regulatory policies or domain specific rules can be specified either as part of the vocabulary description using standards such as OWL 2 RL[8] or as rules using RuleML[9], SWRL[10] or RIF[11]. We propose to address this as part of our future work as an extension to *LEAPS*.

Linked data applications built to serve the stakeholders rely on an unambiguous and contextual representation of the domain knowledge in order to provide reliable and unambiguous answers to data driven questions raised by stakeholders. Below we enumerate a few examples of informal queries that could be potentially raised against the knowledge in stage 1.

- 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.
- Which are the top ten algal operation sites with the lowest impact on global warming potential?
- For a given algal operation site which are the first five

[5]http://purl.org/goodrelations/v1
[6]http://www.ontologyportal.org/
[7]http://www.ifomis.org/bfo
[8]`http://www.w3.org/TR/owl-profiles/#OWL_2_RL`
[9]`http://ruleml.org/`
[10]`http://www.w3.org/Submission/SWRL/`
[11]`http://www.w3.org/2005/rules/`

most cost effective combinations of light, water, nutrients and CO_2 sources?

In the following sections we illustrate how we describe and expose some of the datasets from stage 1 of the algal value chain as linked data, which facilitates the answering of queries such as above as well as the visualisation of results.

3. RELATED WORK

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. The closest attempt to build a taxonomy of algal strains is the AquaFuels[12] project. The taxonomy is however made available as a PDF file rather than a SKOS representation. This paper is a first attempt in formalising biomass knowledge as OWL ontologies and publishing the datasets described using the ontologies, as linked data to be shared and reused by the algal biomass community.

Outside the Semantic Web and linked data technological infrastructure and within the broader context of bioenergy and biomass, various efforts [13], [14] have been made to expose biofuels and biomass datasets to the stakeholders. The BioEnerGIS Project [15] has developed a GIS based Decision support tool, *BIOPOLE* [11] to locate the most suitable sites for biomass plants feeding district heating systems. The project focuses on biomass potential in the target regions of Lombardy (Italy), Northern Ireland, Slovenia and Wallonia (Belgium). The data is stored in Microsoft access while the Web interface of BIOPOLE has been developed using Google maps and open layers. Clearly, the above applications are very limited in their capabilities to provide useful support to the community in terms of sharing and reusing knowledge.

A closely related effort on a much larger scale is the Bioenergy knowledge discovery framework (BioKDF) [16] from the U.S. department of Energy. The portal provides a large suite of tools, apps and spatial visualisation "data" maps, which provide extensive querying options over the datasets powering the portal. Although the application presents an impressive array of features, the data made available through the interfaces cannot be linked to external resources and there is no evidence of ontologies being used to back the data models.

On the other hand Semantic Web technologies have been successfully applied and several successful end user linked data applications [17] can be found in domains such as government, news, education, travel, music, archaeology [18] and health care and life sciences amongst others. A project somewhat closely related to our work is the Reegle energy portal [18]. The project provides various energy related datasets as linked open data and a SPARQL endpoint to access the datasets. It also provides a thesaurus for more than 3000 energy terms. In [21] the authors present a linked data infrastructure for data integration and resource discovery on the Smart energy grid.

Some of the early works that discuss the exploitation of linked data for building Web applications are [7,8]. The paper and report respectively highlights tasks that need to be undertaken in order to make the data compliant with linked data principles and make it amenable to discovery and usage. More recently [9] provides a comprehensive overview on various facets of linked data which includes its publication, consumption and applications.

Besides being a framework that showcases the benefits of structured knowledge representation within the algal biomass domain, *LEAPS* is also a GeoWeb application. Extensive work has been done in the Geospatial community towards the development of vocabularies, publication and visualisation of linked data for geographical information. In [10] the authors present an approach whereby geospatial Web resources such as location maps are treated as data in Semantic Web applications. The key idea is that these resources, termed as geospatial *proxies*, could complement the semantic description of entities in some scenarios, i.e. the linked data representation of the entities would include pointers to these proxies. In another work [1] OpenStreetMap datasets are transformed to the RDF data model.

4. MODELLING ALGAL BIOMASS KNOWLEDGE

In this section we first outline a set of ontological requirements for the representation of knowledge in stage 1 of the algal value chain. We focus on the requirements that are crucial to the analysis of biomass potential. Guided by the requirements, we then propose a set of ontologies. The set comprises of well established ontologies that we have reused and domain specific ontologies built ground-up. Together these ontologies support the publication of algal biomass linked data. We then propose some guidelines for URI design patterns for algal biomass community.

4.1 Ontological requirements

The domain and scope of ontologies for algal biomass is driven by several key requirements as outlined below.

- **Spatiality**: The analysis of the biomass potential is highly dependent on the locality of the possible algae cultivation site as well as the location of the sources of consumables (CO_2, nutrients and water). It is therefore important to choose ontologies that can model spatial concepts and relationships.
- **Geometries**: Another important factor in the calculation of the biomass potential is the area of the cultivation site. In order to represent various areal configurations, an ontology that can represent extents, polygons, linear and ring arrays is required.

[12]http://www.aquafuels.eu/
[13]http://maps.nrel.gov/biomass
[14]http://www.cifor.org/bioenergy/maps/
[15]http://www.bioenergis.eu/
[16]http://bit.ly/voHmzx
[17]http://bit.ly/voHmzx
[18]http://data.reegle.info

- **Units and Measurements**: Besides conventional measurement units such as Kgs for quantities and hectares for area, the cost analysis includes quantities of consumables specified in bespoke units of measurements, i.e., Kgs/hectare or Kgs/annum. The ontology should be expressive enough to be able to model such units.
- **Territorial units for statistics**: In order to assess the overall economic potential of a region for algal biomass cultivation, each of the potential sites in that region have to be linked to their NUTS identifier. By performing potential analysis on different NUTS-levels, regions with high potential can be identified. An ontology that covers the core concepts of the NUTS system is therefore an integral requirement.
- **Domain specific knowledge**: For each algae production site, information on biomass yield and site area are determined. Additionally, data on CO_2-providing industrial or power plants are available for each site and costs for CO_2-supply can be calculated. Total pipelines costs are determined based on the cost of the pipes, the compression and the capture rates. Thus domain specific ontologies with levels of expressivity that allow the modelling of attributes and relationships for various biomass potential calculations are needed.

4.2 Ontologies for Algal Biomass

After establishing the requirements, domain and scope, an extensive search of existing ontology repositories was undertaken in order to reuse or extend any previously defined and closely related vocabularies. *LEAPS* utilises a set of several well established and domain specific vocabularies as illustrated in Figure 2.

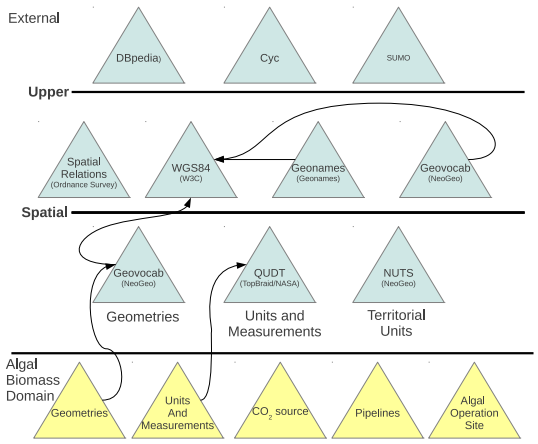

Figure 2: Ontologies for algal biomass. Arrows indicate reuse

Spatial data has been modelled using a combination of several ontologies namely, WGS84 ontology [19], spatial relations ontology, [20] the Geonames ontology [21] and the NeoGeo ontology [22].

Geometries for algal plant sites and pipelines have been modelled using an extension of the NeoGeo geometry ontology [23]. For the CO_2 sources, the geometry is modelled as a `Point` from the WGS84 ontology [24].

Modelling units and measurements for various attributes of the algal biomass datasets was non trivial. The QUDT ontology [25] for dimensions and units was extended to include bespoke units of measurements.

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. For this purpose, the ontology space was also explored using Semantic Web search engines such as Sindice[26], SIG.MA[27], Watson[28] and Bioportal [29]. As noted in Section 1, due to the almost non existent uptake of structured knowledge representation in this domain, the search did not reveal any promising domain vocabularies that could be utilised off-the-shelf or that could be extended. Most of the results yielded terms derived using automated concept extraction techniques over Wikipedia and WordNet. Many of the search results yielded URIs which did not resolve to any concrete ontologies. Given this limitation, barring a few principal terms which were extended from upper level ontologies such as SUMO and BFO, the concepts and relationships for algal biomass had to be defined from ground-up in accordance to the principles of ontology development proposed in [13]. 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.

We developed conceptual OWL ontology schemas [30] for algal plant site, CO_2 sources, regions and pipelines. While the raw datasets provided some guidelines about entities and their attributes, the relationships between the entities within and across datasets had to be established based on the informal queries as exemplified in Section 2. Figure 3 illustrates some of the core concepts, their relationships and attributes. The figure also shows the relationship with the NUTS vocabulary.

4.3 Designing URIs for Algal Biomass Data

The first two principles of linked data signify the importance of URIs for publishing datasets on the Web of data.

- Use URIs as names for things.

[19] `http://www.w3.org/2003/01/geo/wgs84_pos`
[20] `http://www.ordnancesurvey.co.uk/oswebsite/ontology/spatialrelations.owl`
[21] `http://www.geonames.org/ontology/ontology_v2.2.1.rdf`
[22] `http://geovocab.org/geometry`
[23] http://geovocab.org/geometry
[24] http://www.w3.org/2003/01/geo/wgs84_pos
[25] http://qudt.org/1.1/vocab/dimensionalunit
[26] http://sindice.com/
[27] http://sig.ma/
[28] http://watson.kmi.open.ac.uk/WatsonWUI/
[29] http://bioportal.bioontology.org/
[30] Ontologies are available at `http:/purl.org/biomass/ontologies`

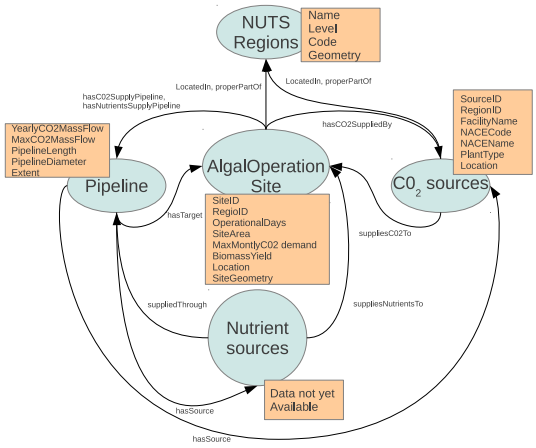

Figure 3: A partial account of core concepts, their attributes and relationships

- Use HTTP URIs, so that people can look up those names.

URIs provide a common mechanism to identify the same "Thing" across the Web. Guidelines for URI design for public sector information [15] and location data [16] have been published by the Chief Technology Officer Council, UK. Inspired by these guidelines and in order to expose algal biomass datasets as linked data, 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
- CO_2 sources
- Pipelines.

Figure 4 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.

	URI type	Description	Example
	Identifier	An identifier for an algal plant site with site ID 546	http://data.biomass.org/algae/sites/id/546
		An identifier for a CO_2 source with a source ID 6122	http://data.biomass.org/CO2sources/id/6122
		An identifier for sites in a region with NUTS ID UKM66	http://data.biomass.org/algae/sites/nuts/id/UKM66
		An identifier for a pipeline connecting algal plant site with site ID 546 to a CO_2 source with source ID 6122	http://data.biomass.org/pipeline/id/pipe_546_6122
	Document	A document about an algal plant site with site ID 546	http://data.biomass.org/algae/sites/doc/546
		A document about sites in a region with NUTS ID UKM66	http://data.biomass.org/algae/sites/nuts/doc/UKM66
	Representation	Representation returned when RDF is requested	http://data.biomass.org/algae/sites/doc/546/site546.rdf
		Representation returned when JSON is requested	http://data.biomass.org/algae/sites/doc/546/site546.json
	Ontology	An identifier for the concept AlgalOperationSite	http://vocab.biomass.org/algae/def/AlgalOperation/AlgalOperationSite
		An identifier for the property hasCO2Source	http://vocab.biomass.org/algae/def/AlgalOperation/hasCO2Source

Figure 4: Representative URIs for Algal Biomass Plant Site

5. LIFTING XML DATASETS TO RDF BASED LINKED DATA

As highlighted in Section 2, in this paper we focus on some of the datasets that contribute to stage 1 of the algal value chain. The datasets for stage 1 are utilised to make potential calculations for the production of algal biomass. By performing potential analysis on different NUTS (Nomenclature of Units for Territorial Statistics)[31] 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 CO_2 sources.

An account of the raw sources of the datasets along with their purpose is available at [17]. 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. 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 [32] 9.3.1. Raw datasets with various origins and formats are first transformed using bespoke computational algorithms to an ArchGIS specific XML format. This step is very crucial for two main reasons: It brings uniformity in the format of representation of the datasets and 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 [33] to selectively query the XML datasets and generate linked data using the ontologies illustrated in Figure 2 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. The XML data sources related the biomass production sites and the CO_2 sources via the pipeline dataset, i.e., the pipeline dataset included for each pipeline, the integer IDs of the production site and source it connected. There was no direct relationship between the production sites and the CO_2 sources. This meant that in order to query for all sources that supplied CO_2 to a specific site, the query had to be made via the pipeline dataset. Further the sites and the source datasets included only the String literal identifier of the NUTS region where they were located. 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

[31] `http://bit.ly/I7y5st`
[32] http://www.esri.com/software/arcgis/index.html
[33] http://www.w3.org/TR/xpath/

by distinct and dedicated dataset providers in the future. While a linked data representation of the NUTS regions data [34], 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. Figure 5 illustrates the linkages between the datasets.

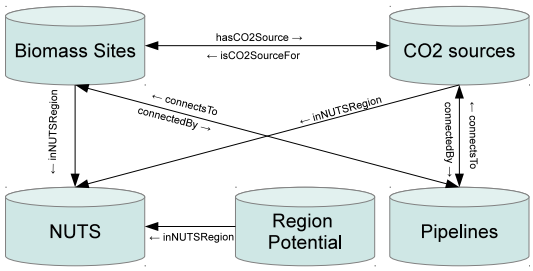

Figure 5: Linked datasets for algal biomass

The 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.

6. THE LEAPS FRAMEWORK

LEAPS[35] is an end user LOD application with a Web interface built over RESTful Web services. For the stake holders in the biomass domain, it 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 6 illustrates the conceptual architecture of *LEAPS*.

The main components of the application are

- Parsing modules: As shown in Figure 6 and discussed in Section 5, 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[36] and Geonames[37].

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

[35] The application will be made available online shortly on a dedicated Web server. A video demonstrating the Web interface of the application is available at `http://purl.org/biomass/LEAPSDemo`.

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

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

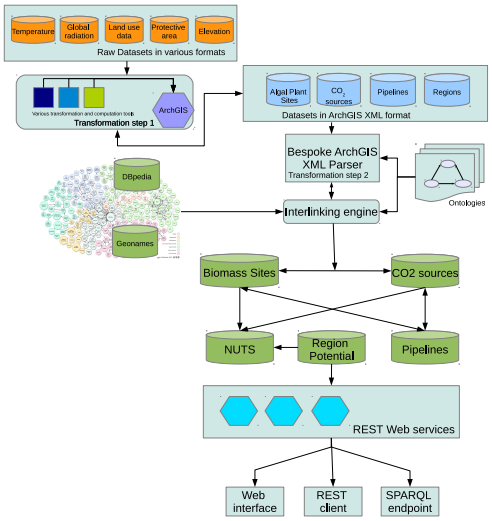

Figure 6: Architecture of *LEAPS*

- Triple store: The linked datasets are stored in a triple store. We use OWLIM SE 5.0 [38].
- 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. 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.

7. QUERYING LINKED ALGAL BIOMASS DATA

In this section we highlight the benefits of representing algal biomass datasets as linked data and using Semantic Web standards while querying the knowledge base. As outlined in Section 3, conventional frameworks for querying over biomass datasets do not explore knowledge available outside the datasets, because of their inherent limitation of lack of linkages in the dataset to external sources of information. Since we provide links to major data hubs such as DBpedia and Geonames we are able to provide useful background knowledge along with the core data requested by the stakeholders.

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

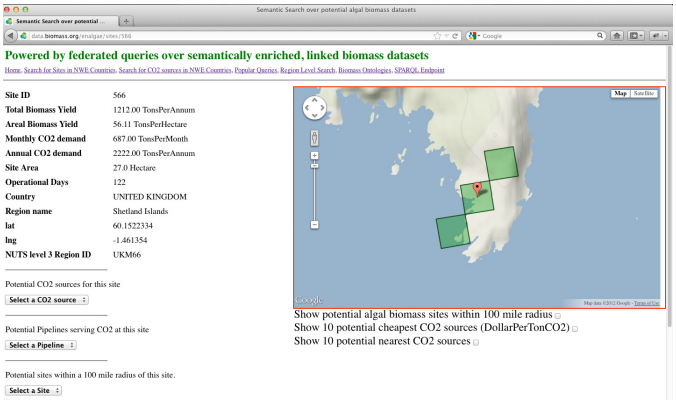

Figure 7: A typical site as visualised with the *LEAPS* Web interface

With our set of ontologies and URI patterns in place, we generated linked datasets using the architecture outlined in Section 6 and the approach illustrated in 5. 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 highlighted in Section 2,

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;
     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.

8. CONCLUSIONS

Investigations into using algal biomass as an alternative source of fuel is gaining widespread momentum. As research in the sector progresses, a wealth of information will be available that could be exploited by domain specific applications. In order to facilitate further research and benefit commercial setups, the accumulated knowledge needs to be made accessible in a machine interpretable and integrated format such that it can be easily shared and reused by stakeholders of the domain.

The Algal biomass community currently does not employ any knowledge representation techniques to formalise and structure valuable knowledge harnessed through their operations. In this paper we present 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, CO_2 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.
- n 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.

In order to increase the uptake of Semantic web and linked data by the algal biomass community, lots more needs to be done. While *LEAPS* can currently provide integrated information about algal plant sites, CO_2 sources and the pipelines connecting them, there are several other datasets 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. We are working with biologists in the domain to facilitate the process of making the taxonomy from the AquaFuels project available as SKOS models. 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.

Multifaceted visualisation of the integrated datasets is another area that we are currently focusing on to motivate the idea of interlinking datasets. The reasoning infrastructure in *LEAPS* is currently based on implicit OWL 2 DL inferences. Work is also in progress on exploiting rule based reasoning to model domain specific constraints.

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.

9. REFERENCES

[1] S. Auer, J. Lehmann, and S. Hellmann. Linkedgeodata: Adding a spatial dimension to the web of data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, Berlin, Heidelberg, 2009. Springer-Verlag.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

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

[4] 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.

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

[6] R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.

[7] M. Hausenblas. Exploiting linked data to build web applications. *IEEE Internet Computing*, 13:68–73, July 2009.

[8] M. Hausenblas. Linked data applications - the genesis and the challenges of using linked data on the web. Technical report, Digital Enterprise Research Institute Galway (DERI), 2009.

[9] T. Heath and C. Bizer. *Linked Data - Evolving the Web into a Global Data Space*. Morgan & Claypool Publishers, 2011.

[10] F. J. Lopez-Pellicer, M. J. Silva, M. Chaves, F. J. Zarazaga-Soria, and P. R. Muro-Medrano. Geo linked data. In *Proceedings of the 21st international conference on Database and expert systems applications: Part I*, DEXA'10, pages 495–502, Berlin, Heidelberg, 2010. Springer-Verlag.

[11] G. Maffeis and A. Boccardi". *"A GIS-based Decision Support System (DSS) for BioEnerGIS"*, September 2011.

[12] T. M. Mata, A. A. Martins, and N. S. Caetano. Microalgae for biodiesel production and other applications: A review. *Renewable and Sustainable Energy Reviews*, 2010.

[13] N. F. Noy and D. L. Mcguinness. Ontology development 101: A guide to creating your first ontology. Technical report, Stanford Center for Biomedical Informatics Research (BMIR), 2001.

[14] Oilgae. Oilgae comprehensive report, energy from algae: Products, market, processes and strategies. Technical report, Oilgae, 2011.

[15] Public Sector Information Domain of the CTO Council's cross Government Enterprise Architecture. Designing URI Sets for the UK Public Sector. Technical report, Chief Technology Officer Council, 2009.

[16] Public Sector Information Domain of the CTO Council's cross Government Enterprise Architecture. Designing URI Sets for Location. Technical report, Chief Technology Officer Council, 2011.

[17] J. Skarka. Original data sources for leaps. `http://purl.org/biomass/leaps/rawDatasets`, 2012.

[18] M. Solanki, Y. Hong, and K. Rebay-Salisbury. SEA: A Framework for Interactive Querying, Visualisation and Statistical Analysis of Linked Archaeological Datasets. In *CAA: Proceedings of the 39th annual conference on Computer Applications and Quantitative Methods in Archaeology*, 2011.

[19] U.S. Department of Energy. National Algal Biofuels Technology Roadmap. Technical report, accessed June 2012.

[20] T. van Harmelen and H. Oonk. *Microalgae biofixation processes: Application and potential contributions to greenhouse gas mitigation options*. TNO Built Environment and Geosciences, Apeldoorn, 2006.

[21] A. Wagner, D. Anicic, R. Stühmer, N. Stojanovic, A. Harth, and R. Studer. Linked data and complex event processing for the smart energy grid. In *Proc. of Linked Data in the Future Internet at the Future Internet Assembly*, 2010.