Undefined 1 (2009) 1–5
IOS Press

Linked data for Potential Algal Biomass Production

Editor(s): Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited review(s): Anne Thessen, Arizona State University, USA; Boris Villazon-Terrazas, iSOCO, Madrid, Spain; Femke Reitsma, University of Canterbury, New Zealand

Monika Solanki [a,*], Johannes Skarka [b], Craig Chapman [c]
[a] *Knowledge Based Engineering Lab, Birmingham City University, Birmingham, United Kingdom*
E-mail: monika.solanki@bcu.ac.uk
[b] *Karlsruhe Institute of Technology, ITAS, Karlsruhe, Germany*
E-mail: johannes.skarka@kit.edu
[c] *Knowledge Based Engineering Lab, Birmingham City University, Birmingham, United Kingdom*
E-mail: craig.chapman@bcu.ac.uk

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, 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 *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 raw datasets which served as the basis for *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 *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 [5,6], government bodies [8] and dedicated non-profit organisations such as ABO (Algal Biomass Organisation) [3] and EABA (European Algal Biomass Associ-

*Principal and Corresponding author. Email: monika.solanki@bcu.ac.uk

[1] http://www.enalgae.eu/

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

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

ation)[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 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 [4]. Figure 1[5] 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.

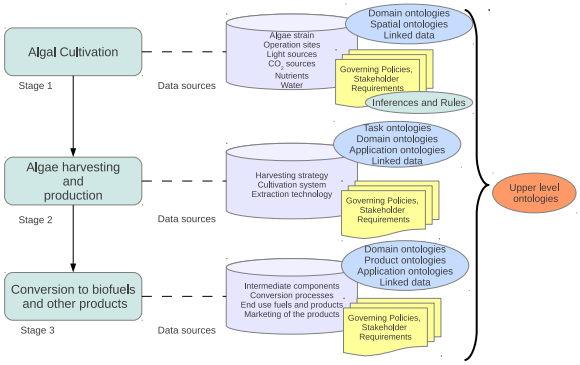

Fig. 1. The Algal biomass supply chain

Stage 1 encompasses the cultivation of algae. It involves setting up an algal cultivation site and incorporates datasets about location related geographical 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. 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[6] 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)[7] regions in NWE.

3. Transformation of Raw datasets to linked data

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 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 [8] 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

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

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

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

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

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

XPath [9] to selectively query the XML datasets[10] 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 [7] was developed that enabled the transformation for each of the datasets.

An architecture underlying the transformation process is depicted in 2. The main components of the application are

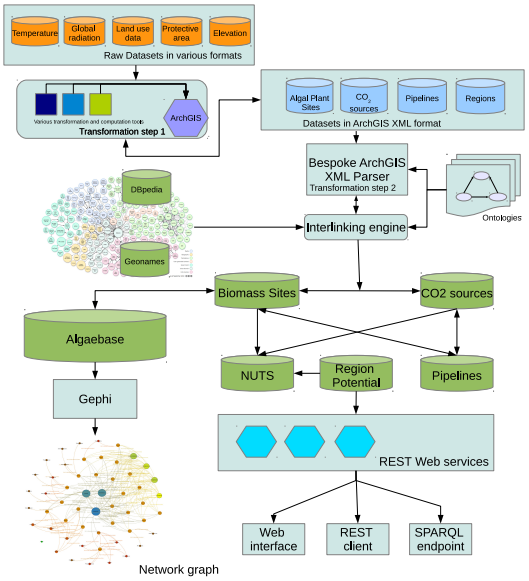

Fig. 2. Architecture of *LEAPS*

- **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. From the LOD cloud, we currently provide outgoing links to DBpedia[11] and Geonames[12].
- **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 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. 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[14] of Gephi[15] has been exploited to visualise the biological taxonomy of algae. This visualisation is also made available via the *LEAPS* interface.

For each of the algae production site, information on biomass yield and site area are determined. Additionally, data on CO_2-providing industrial or power plants can be queried for each site and costs for CO_2-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 im-

[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
[11] http://dbpedia.org/
[12] http://sws.geonames.org/
[13] http://www.ontotext.com/owlim/editions
[14] http://wiki.gephi.org/index.php/SemanticWebImport
[15] https://gephi.org/

	Raw dataset (format)	Purpose	Source
1	Global radiation (GRIB [a])	Calculate the algal biomass yield	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 [3]).
2	Temperature (NetCDF [c])	Calculate the algal biomass yield	E-OBS dataset provided by the ECA&D project [d] [10], [1].
3	Data on land use (GeoTIFF)	Identify suitable land area	European Environment Agency (EEA), CORINE land cover raster dataset 2006 version 13 [e]. CORINE land cover raster dataset 2000 version 15 [f].
4	Protected areas (GIS vector data)	Identify suitable land area	World Database on Protected Areas (WDPA) provided on [g] by UNEP-WCMC [2].
5	Elevation (GIS raster data)	Identify suitable land area	Shuttle Radar Topography Mission (SRTM) data [9] by Global Land Cover Facility (GLCF) [h].
6	CO_2 sources (MS Access)	Identify sources of CO_2	European Pollution Release and Transfer Register (E-PRTR) provided by the EEA [i] in the version of 09 Nov 2009.

[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
[e] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster
[f] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2000-raster
[g] http://www.protectedplanet.net
[h] http://www.landcover.org
[i] http://bit.ly/v3dS2I

Table 1
Sources of raw datasets

mensely 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 [16], spatial relations ontology, [17] the Geonames ontology [18] and the NeoGeo ontology [19].

Geometries for algal plant sites and pipelines have been modelled using an extension of the NeoGeo ge-

[16] http://www.w3.org/2003/01/geo/wgs84_pos
[17] http://www.ordnancesurvey.co.uk/oswebsite/ontology/spatialrelations.owl
[18] http://www.geonames.org/ontology/ontology_v2.2.1.rdf
[19] http://geovocab.org/geometry

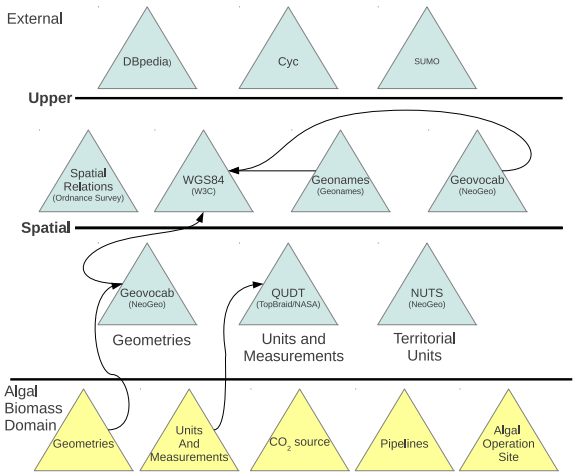

Fig. 3. Ontologies for algal biomass. Arrows indicate reuse

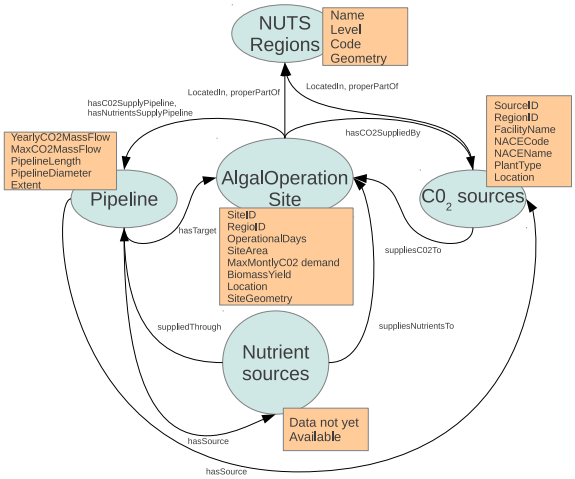

Fig. 4. A partial account of core concepts, their attributes and relationships

ometry ontology [20]. For the CO_2 sources, the geometry is modelled as a `Point` from the WGS84 ontology [21].

Modelling units and measurements for various attributes of the algal biomass datasets was non trivial. The QUDT ontology [22] for dimensions and units was extended to include bespoke units of measurements.

We developed conceptual OWL ontology schemas [23] 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[24] vocabulary.

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 [25], 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[26]. 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 the datasets.

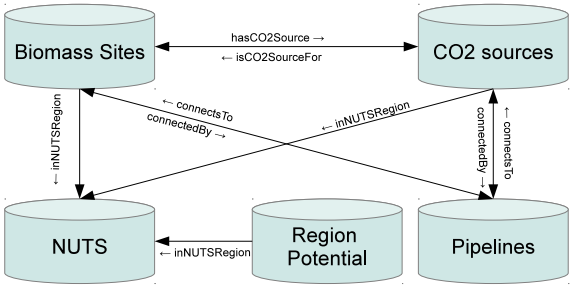

Fig. 5. Linked datasets for algal biomass

[20] `http://geovocab.org/geometry`
[21] `http://www.w3.org/2003/01/geo/wgs84\_pos`
[22] `http://qudt.org/1.1/vocab/dimensionalunit`
[23] Ontologies are available at `http:/purl.org/biomass/ontologies`
[24]
[25] `http://nuts.geovocab.org/`

[26] 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.

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
- CO_2 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.

	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

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.

Dataset	Number of Statements
Algae sites dataset	84703
CO_2 sources	14238
Pipelines	261680
NUTS	14469
Global radiation	24738
Biomass potential	4557

Table 2
Dataset statement statistics

Dataset availability

The *LEAPS*[27] application 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 queries because of their federated nature, however optimisation work on the retrieval mechanism is in progress to enable faster retrieval of information.

Besides the application Web interface, datasets are available for querying via the dedicated triple store Web interface[28].

VOID description of the datasets

The VoID[29] descriptions of some of the datasets in the *LEAPS* suite have been made available [30]. Once the datasets are made public the VoID descriptions will be updated.

Linking to other datasets

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 CO_2 sources, pipeline datasets and the nuts region dataset. Back links to the datasets have been provided from the CO_2 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 [31] with a Web interface built over REST-

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

[28] http://www.semanticwebservices.org/openrdf-workbench/repositories/NONE/repositories

[29] http://www.w3.org/TR/void/

[30] http://purl.org/biomass/void

[31] http://purl.org/biomass/LEAPSDemo

ful 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, 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. For the stakeholders

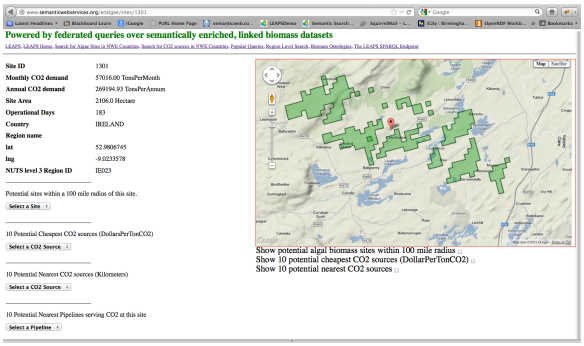

Fig. 7. A typical site as visualised with the *LEAPS* Web interface

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.

7. Discussion

While *LEAPS* currently provides integrated information about algal plant sites, CO_2 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 [32] 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.

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 NWE 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.
- 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.

[32] http://www.algaebase.org

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[33] 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[34] can be seen via the Web application. The reasoning infrastructure in *LEAPS* is currently based on implicit OWL 2[35] DL inferences. Work is also in progress on exploiting rule based reasoning to model domain specific constraints.

[33] http://www.aquafuels.eu/

[34] Visualisations are powered by Sgvizler, http://code.google.com/p/sgvizler/

[35] http://www.w3.org/TR/owl2-direct-semantics/

References

[1] M.R. Haylock, N. Hofstra, A.M.G. Klein Tank, E.J. Klok, P.D. Jones, and M. New. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, 2008.

[2] IUCN and UNEP. *The World Database on Protected Areas (WDPA)*. UNEP-WCMC, Cambridge, UK, 2010.

[3] E. Kalnay et al. The NCEP / NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, 77:437–471, 1996.

[4] Teresa M. Mata, António A. Martins, and Nidia. S. Caetano. Microalgae for biodiesel production and other applications: A review. *Renewable and Sustainable Energy Reviews*, 2010.

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

[6] Claire Smith and Adrian Higson. Research Needs in Ecosystem Services to Support Algal Biofuels, Bioenergy and Commodity Chemicals Production in the UK. Technical report, NNFCC, 2011.

[7] Monika Solanki, Johannes Skarka, and Craig Chapman. 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.

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

[9] USGS. *Shuttle Radar Topography Mission, 3 Arc Second scene SRTM_ff03_nYYYeXXX (covering Europe), filled finished 2.0*. Global Land Cover Facility, University of Maryland, College Park, Maryland, 2006.

[10] E.J.M. van den Besselaar, M.R. Haylock, G. van der Schrier, and A.M.G. Klein Tank. A European Daily High-resolution Observational Gridded Data set of Sea Level Pressure. *J. Geophys. Res.*, 2011.