



Semantic Enrichment and Search with Linked Data: A Case Study on Environmental Science Literature

Kalina Bontcheva



The
University
Of
Sheffield.



HR Wallingford
Working with water

Aims of the EnviLOD Project

Evaluate the potential of linked open data vocabularies to information discovery in environmental science

- Understand how environmental science vocabularies can be enriched and interlinked
- Develop intuitive semantic search methods
- Case study: the British Library information discovery tool for environmental science, Envia.
- Engage with domain experts and other stakeholders.

6 month research project, limited content scope, subject restricted to flooding.



Context : British Library Envia project

- Connect users to environmental information (reports, journal articles, data etc), wherever it is, or they are
- Flooding focus
- Beta version launched April 2013 (envia@bl.uk)
- Improve *discovery* and access of environmental information
 - Surface more *relevant* information
 - Less ‘filtering’ of results necessary



Environmental Information Discovery and Access

Search bar with a dropdown menu set to 'All', a 'Search' button, and a link to 'Advanced Search'.

Groundwater quality review. Furness and West Cumbria : analysis of the Furness and West Cumbria groundwater chemistry

DOWNLOAD DOCUMENT



LINK



<http://publications.environment-agency.gov.uk/pdf/GEHO1210BTHK-e-e.pdf>

CONTENT TYPE

Reports

PUBLISHER

Environment Agency

DATE

2011

SUBJECT

- England
- Groundwater
- Cumbria

[View Detailed Record](#)

Sparse Metadata

http://www.envia.bl.uk:8080/xmlui/handle/123456789/22

File Edit View Favorites Tools Help

BBC Weather Conference Aston Intern... Home - STM ES Staging Native Suggested Sites Web Slice Gallery Pin It


envia Environmental Information Discovery and Access

Home About Partners Contact us Help


All Search Advanced Search

Groundwater quality review. Furness and West Cumbria : analysis of the Furness and West Cumbria groundwater chemistry

DOWNLOAD DOCUMENT



LINK

 <http://publications.environment-agency.gov.uk/pdf/GEHO1210BTHK-e-e.pdf>

ABSTRACT

This report provides information about the baseline groundwater quality in the Water Framework Directive (WFD) groundwater monitoring units of Furness and West Cumbria. The data used in the report was taken from the Environment Agency's groundwater monitoring network, the purpose of this network is to comply with the agency statutory requirement to support other environmental protection duties and initiatives. Groundwater provides 13% of public water supplies in the North West Region. In addition, river and stream flows and wetland habitats are often heavily dependent on groundwater seepage and springs, especially during the drier months. The quantity and water quality of groundwater is therefore extremely important in maintaining these resources.

CONTENT TYPE

Reports

PUBLISHER

Environment Agency

DATE

2011

SUBJECT

- Cumbria
- Groundwater
- England

Manually adding abstract..

100%

Semantic Technologies

The screenshot displays the GATE software interface, which is used for text processing and semantic analysis. The interface is divided into several panes:

- Left Pane:** Contains a tree view of the project structure. It includes 'Applications', 'Language Resources' (with files like '013738113.pdf.txt_00', 'Thesis corpus', 'NERCMeteorologyCli', 'NERCEcologyEnviror', 'Dataset corpus', 'AlephReports2', and 'AlephReports1'), and 'Processing Resources' (with 'Datastores' and 'datastore').
- Top Pane:** Shows the 'Messages' tab with a list of files: 'datastore', 'AlephReports2', and '013738113.pdf.t...'. Below this is a toolbar with icons for 'Annotation Sets', 'Annotations List', 'Annotations Stack', 'Co-reference Editor', 'OAT', 'RAT-C', 'RAT-I', and 'Text'.
- Central Pane:** Displays the text document '013738113.pdf.txt_00'. The text is annotated with various semantic tags. For example, 'flooded' is tagged as 'Location', 'London' as 'Location', 'United Kingdom' as 'Location', 'England' as 'Location', 'Wales' as 'Location', 'River' as 'Location', 'Redbridge' as 'Location', 'North East London' as 'Location', 'Edmonton' as 'Location', 'Enfield' as 'Location', 'Richmond' as 'Location', 'Autumn 2000' as 'Date', '2002' as 'Date', 'Europe' as 'Location', and 'floods' as 'Location'. The text is also tagged with 'flooded' as 'Location' and 'flooded' as 'Location'.
- Right Pane:** A list of semantic types with checkboxes. The types are: Address, Date, FirstPerson, Identifier, JobTitle, Location, Lookup, Measurement, Money, Number, OntoRes, Organization, Percent, Person, PublicationAuthor, PublicationDate, PublicationPages, Ratio, Reference, RemoveAuthorsFrom, Sem_Location, Sem_Organisation, and Sem_Person. The 'Sem_Location' type is checked.
- Bottom Pane:** A 'Document Editor' tab showing 'Initialisation Parameters'.

The text in the central pane is as follows:

body which has a statutory duty to take an overview of the flooding risk on behalf of Londoners. The GLA will have a key role in coordinating the many organisations with an interest in this question – including the Environment Agency, Thames Water, the Boroughs and Thames Gateway London Partnership. It also has powers in relation to the emergency services (the Fire Service and Police) and in setting the planning framework for London in the Spatial Development Strategy. The GLA is thus well placed to set the strategic framework for London's response to flood risk. The purpose of this Assembly Committee is to investigate whether such a framework is being established and to make recommendations as to what it should contain.

1.3 London's wake-up call was the flooding of Autumn 2000, when the United Kingdom experienced its worst weather in over 270 years. Across England and Wales about 10,000 properties were flooded, some on several occasions and for long periods of time. A further 37,000 properties were saved by sandbags alone and in total around 280,000 properties were protected by flood defences. The Association of British Insurers estimated that the cost to insurers was £1.3 billion.

1.4 Though not the worst affected part of the country, London too experienced severe flooding at this time. Whilst existing defences successfully prevented flooding for many London properties, the defences on the River Roding at Wanstead and Woodford in Redbridge, North East London, were overtopped as a result of which 230 properties were flooded. There was also flooding of 75 properties at Edmonton in Enfield and 15 at Teddington in Richmond. Thus in total 320 properties were flooded in London. The Environment Agency's report on the Autumn 2000 floods also makes clear that there was flooding at a number of properties adjacent to London.

1.5 Recent flooding in 2002, both in the north of England and in London, have demonstrated that this is a recurring problem and that public policy needs to be prepared and robust to deal with future emergencies. The floods crisis in continental Europe, which included devastation of property and fatalities, brought home to many quite how serious flooding can be.

Groundwater quality review. Furness and West Cumbria : analysis of the Furness and West Cumbria groundwater chemistry - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Groundwater quality review. Furness and W... +

vlb-esstaging.bl.uk:8090/xmlui/handle/123456789/25

Most Visited Getting Started Latest Headlines Customize Links Ovid: Search Results Import to Mendeley ScienceServer Form Templates - ESP...

BRITISH LIBRARY **envia** Environmental Information Discovery and Access


Home About Partners Contact us Help

All Search Advanced Search


Groundwater quality review. Furness and West Cumbria : analysis of the Furness and West Cumbria groundwater chemistry

BRISTOL : ENVIRONMENT AGENCY

DOWNLOAD DOCUMENT



LINK

 <http://publications.environment-agency.gov.uk/pdf/GEHO1210BTHK-e-e.pdf>

CONTENT TYPE

Monographic

DATE

2011

SUBJECT

- Groundwater--England--Cumbria
- Economic geology
- monitoring
- Environment
- groundwater
- pesticides
- statistics
- Environment Agency
- Groundwater Quality
- Aromatic compounds
- groundwater quality
- Land use

View Detailed Record

<http://vlb-esstaging.bl.uk:8090/xmlui/bitstream/handle/123456789/25/015740974.pdf?sequence=1>

Automatically generating keywords

User Survey: Key Findings

- Types of search queries related to flooding
 - Regulation & policy; Research; Risk; Spending, etc.
 - Location-based search needed
 - Keyword-based searches preferred amongst users
- Example queries:
 - How is flood defence spending prioritised in non UK countries?
 - The top ten flood risk areas in Oxfordshire?
 - Where in the UK has surface water flooding taken place since 2007?
 - Where has there been flooding near Sheffield?

Problems with keyword only search

- Cannot find answers for location-based queries
 - Flooding in places near Sheffield/Oxford/etc.
 - Flooding in places with population < 15,000
- Misses out relevant documents
 - “Climate change AND Oxfordshire” returns no hits even though docs mention Wytham Woods and Banbury

Environmental Vocabularies

*A **controlled vocabulary** is a list of established terms used in the indexing and retrieving of information. They enable better querying of information, and for hierarchical relationships to be developed*

*Ex: The **Thames Barrier** is a **flood defence** along the **Thames**, which is a river in **England**.*

- GEMET (European Environment Agency Thesaurus)
- OS Hydrology Ontology
- GeoNAMES
- DBPedia

*In the EnviLOD project, we use a **Linked Open Data** approach to improve information discovery*

Automatic Semantic Enrichment

- Locations (linked to DBpedia and GeoNames)
 - Markup the place name itself (e.g. Norwich) with the corresponding DBpedia and GeoNames URIs
 - Also use knowledge of the implied reference to the levels 1, 2, and 3 sub-divisions from the Nomenclature of Territorial Units for Statistics (NUTS). For Norwich, these are East of England (UKH – level 1), East Anglia (UKH1 – level 2), and Norfolk (UKH13 – level 3).
 - Similarly knowledge of nearby places
 - Use ontology classes to categorise rivers

“South Gloucestershire” Example

Messages

Lucene

Annotation Sets

Annotation

Managing flood risk on the S

January 2011

South Gloucestershire to Hi

Managing flood risk in the S

We are the Environment Ag

better place _ for you, and f

breathe, the water you drin

Government and society as

healthier. The Environment

Please click on the bookma

brochure to specific points

Managing flood risk in the S

Somerset 1

Type	Set	Start
Sem_Location		57
Sem_Location		97
Sem_Location		97
Sem_Location		97
Sem_Location		149
Sem_Location		160
Sem_Location		160

211 Annotations (1 select

Sem_Location

alternateName	South Gloucestershire
caption	South Gloucestershire
count	2
countryCode	GB
geonamesURI	http://sws.geonames.org/3333198/
inst	http://dbpedia.org/resource/South_Gloucestershire
latitude	51.5
longitude	-2.41667
lookupRule	fullString
matched	South Gloucestershire
name	South Gloucestershire
parentAdminURI	http://sws.geonames.org/6269131/, http://sws.geonames.org/3333198/
parentCountryInst	http://sws.geonames.org/2635167/
popularitySimilarity	1.0
randomIndexing	0.0
specificitySimilarity	0.0
string	South Gloucestershire
stringSimilarity	0.2688679
structuralSimilarity	0.0

Semantic Enrichment (2)

- Organisations (linked to DBpedia)
 - Names of companies, government organisations, committees, agencies, universities, and other organisations
- Dates
 - Absolute (e.g. 31/03/2012) and relative (yesterday)
- Measurements and Percentages
 - e.g. 8,596 km² , 1 km, one fifth, 10%

Climate change in Oxfordshire

Searching Index "bl-geo-metadata-15102012"

```
climate root:change AND {Sem_Location name .REGEX("(.*Oxfordshire(.)*")}}
```

Search

Documents 1 to 2 of 2:

[meta4360.xml_00E9F](#)

of ecosystems to **climate change**. This study investigated the relati ... **deciduous ancient semi-natural woodland at Wytham** Woods in central

[meta1709.xml_0031D](#)

is controversial. **Climate change** adds further uncertainty to decisio ... on growth, photosynthesis and phenology at **Wytham** Woods, a

Documents mentioning locations in UK where the population density is more than 500 people per square km

Searching Index "bl-geo-metadata-15102012"

```
{Sem_Location countryCode="GB" dbpediaSpargl="select distinct ?inst where  
{?inst rdf:type :Country. ?inst :populationDensity ?x. FILTER(?x > 500)}"}
```

Search

Documents 1 to 8 of 8:

[meta1161.xml_000BD](#)

Lambourn catchments, **Berkshire**, UK. Chalk catchments in **Berkshire** (UK) Lambourn catchments, **Berkshire**, UK Article

[meta1172.xml_000C9](#)

808), **Stoke-on-Trent** (n = in Coventry and **Stoke-on-Trent**) to greater

[meta756.xml_01543](#)

Upper Thames in **Berkshire**, UK,

[meta5901.xml_011B2](#)

, Lambourn, **Berkshire**, UK (

[meta2247.xml_00573](#)

EnviLOD Semantic Search UI



The
University
Of
Sheffield.



HR Wallingford

JISC

Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature

Search [Help](#)

Keywords

Narrow down your search:

Location ▼ none ▼

none ▼

+

Restrict your search to ☐ paragraphs ☐ sentences

none
population
longitude
latitude
name
country code
population density
with nearby

EnviLOD Semantic Search UI (2)



Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature

Search [Help](#)

Keywords

Narrow down your search:

Location ▼

nearby ▼

location na ▼

+

Restrict your search to ☒ document

location named

☐ sentences

Example Results

Development and flood risk : : practice guide

Example hits:

..." flood defences, or to flood alleviation schemes which provide benefit to the wider community. An example is provided below. Case study The Avenue Site, Chesterfield – example of organisations working" " working together to help reduce flood risk and create wetland habitats This ongoing project is involving the restoration and de-contamination of a former major coking works to the south of Chesterfield by the East Midlands Development" " of new wetland, a flood storage area and a restored section of the River Rother. The project will result in reductions in flood risk downstream in Chesterfield. A steering group comprising" ...

Keywords: Flood control--Great Britain, Flood damage prevention--Great Britain, Floodplain management--Great Britain, Other social problems and services (other metadata )

Lower Derwent flood risk management strategy : non-technical summary

Example hits:

..." the Environment Agency. Managing Flood Risk in Derby and the Lower Derwent The" " . We cannot prevent floods. There

Searching PubMed for viral diseases

FastVac QueriesFastVac Concepts and InstancesCustom Query

URI: #Viral_disease

Antigen

Disease

Disease_pathogenesis

Immune_response

Bacterial_disease

Viral_disease

Dengue

Tick_borne_encephalitis

Viral_haemorrhagic_fever

Viral_hepatitis

Document Metadata

Dates: 01/01/2009 to 31/12/2009

source: PubMed

type: article

SearchCustomize query

DocumentsTerms

HCV Genome and Life Cycle

patients with acute hepatitis C developing chronic HCV result in liver cirrhosis, hepatic failure hepatic failure or hepatocellular carcinoma, which are

Development of an Infectious HCV Cell Culture System

from a fulminant hepatitis C patient. The

Erratum.

in patients with AIDS" (20

Sexual Practices That May Favor the Transmission of HIV in a Rural Community in Nigeria.

HIV) and Acquired Immune Deficiency Syndrome (AIDS)

See original document

HCV Genome and Life Cycle

21250393

Chevaliez S, Pawlotsky JM

Hepatitis C virus (HCV) infection afflicts more than 170 million people worldwide, with the great majority of patients with acute hepatitis C developing chronic HCV infection. It can ultimately result in liver cirrhosis, hepatic failure or hepatocellular carcinoma, which are responsible for hundreds of thousands of deaths each year. Despite the discovery of HCV over 15 years ago, our knowledge of the HCV lifecycle has been limited by our inability to grow the

source: PubMed

id: 21250393

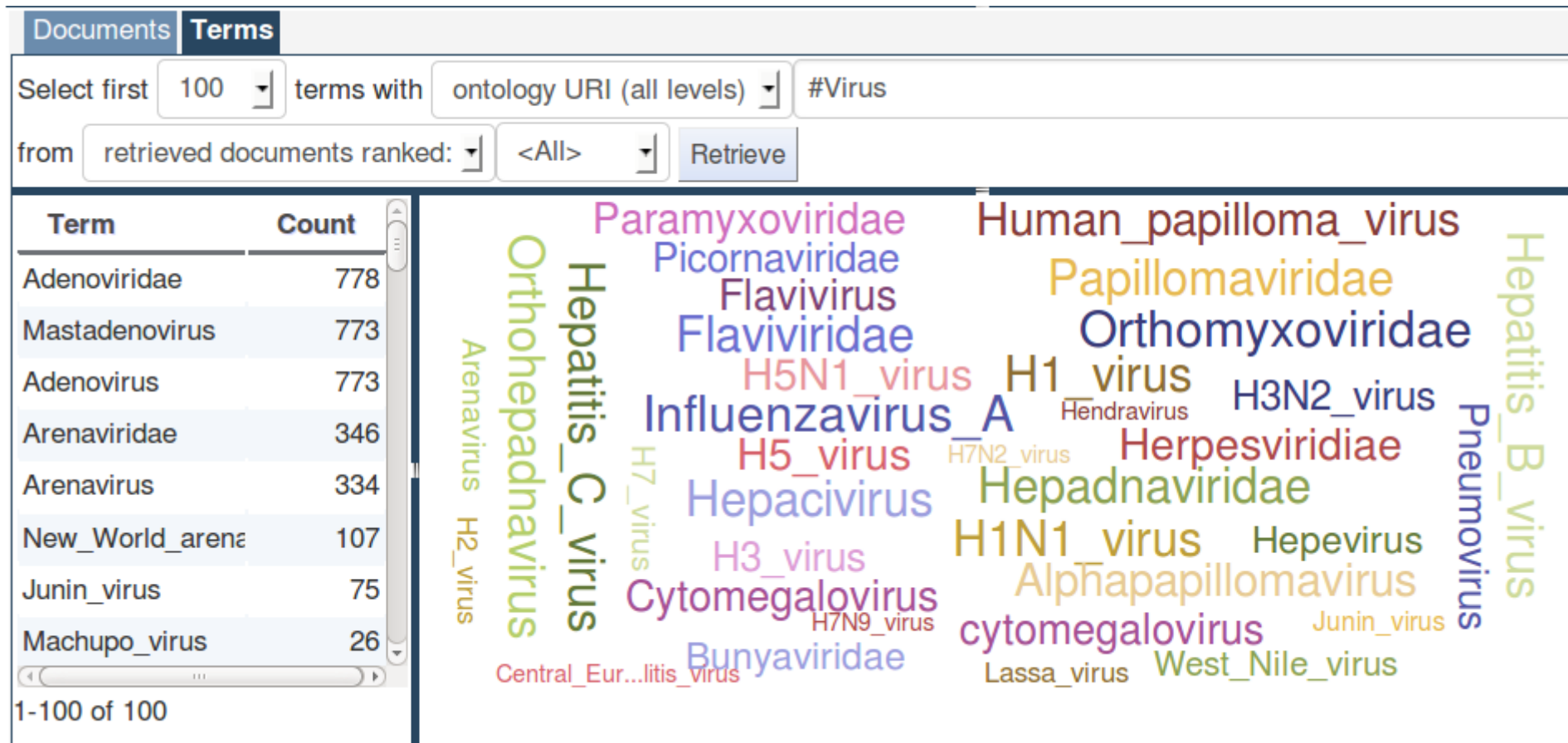
type: article

date: 20110121

1-20 of 35,853

Download

Top 100 Viruses mentioned in the selected PubMed abstract



Diseases vs Pathogens

