JISC

EnviLOD

Exploring the potential of Linked Open Data to Environmental Science

Kalina Bontcheva (University of Sheffield)
Johanna Kieniewicz (British Library)
Niraj Aswani (University of Sheffield)
Mike Wallis (HR Wallingford)

The University Of Sheffield.

BRITISH LIBRARY

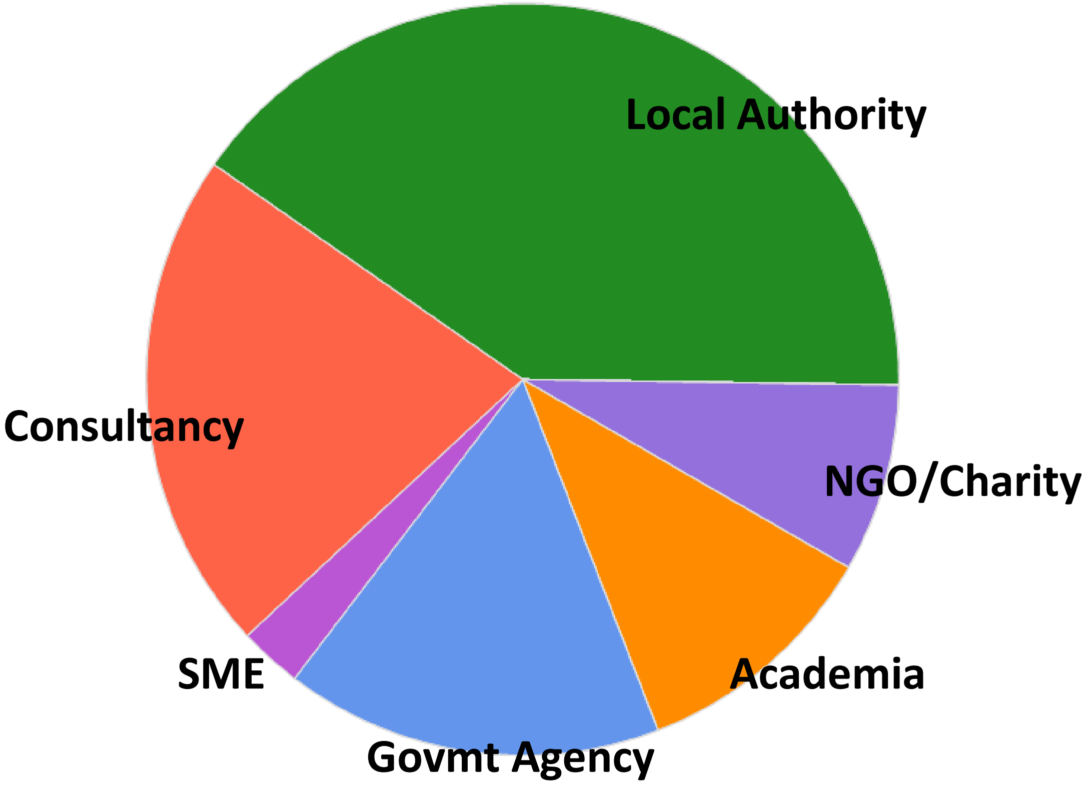
HR Wallingford
Working with water

Aims of the EnviLOD Project

Evaluate the potential of Linked Open Data (LOD) vocabularies to aid information discovery in environmental science

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

Survey of the flooding community



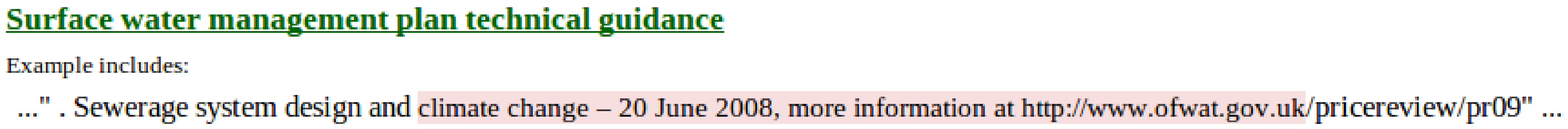
- Examples of search queries– what were they looking for, how did they search for it, what did they anticipate to be retrieved in return?
- Preferred searching methods, i.e. keyword vs. faceted navigation.

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

- The hits include references and URLs



Surface water management plan technical guidance

Example includes:

..." . Sewerage system design and climate change – 20 June 2008, more information at http://www.ofwat.gov.uk/pricereview/pr09" ...

- 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.

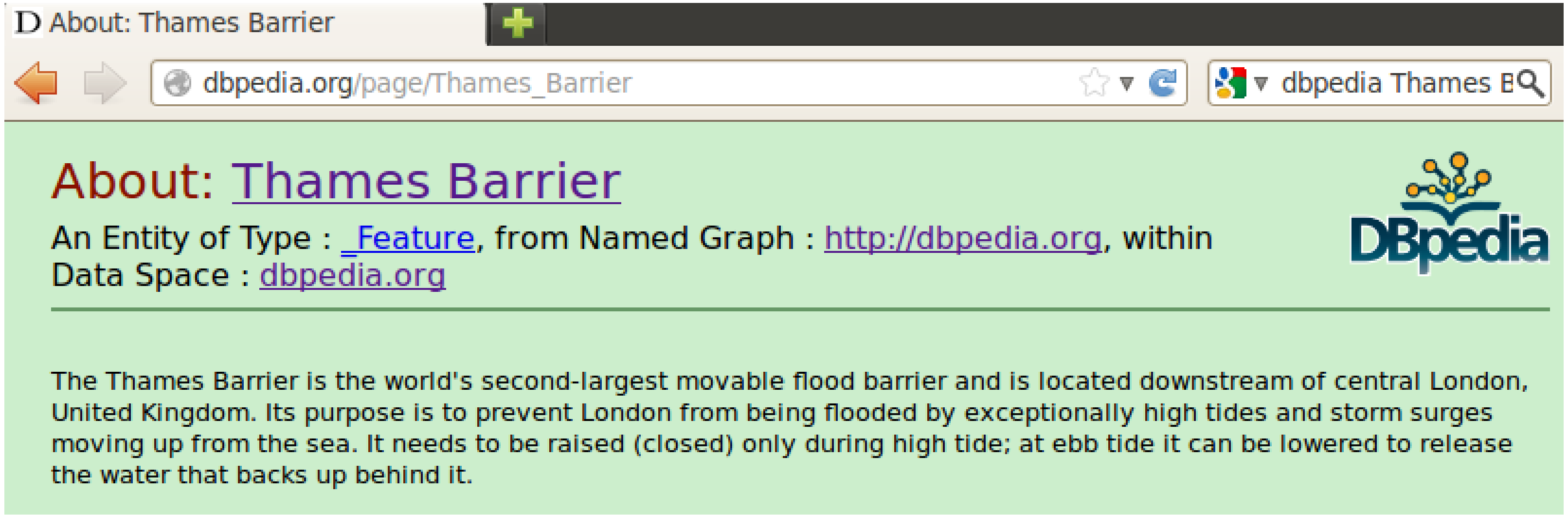
- GEMET (European Environment Agency Thesaurus)
- OS Hydrology Ontology
- GeoNAMES
- DBPedia

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

DBpedia

- Machine readable knowledge about 3.5 million entities, many relevant to EnviLOD:
  - 410,000 places,
  - 310,000 persons
  - 140,000 organisations
- For each entity we have:
  - Entity name variants (e.g. IBM, Int. Business Machines)
  - a textual abstract
  - reference(s) to corresponding Wikipedia page(s)
  - entity-specific properties (e.g. latitude and longitude for places)

Example from DBpedia

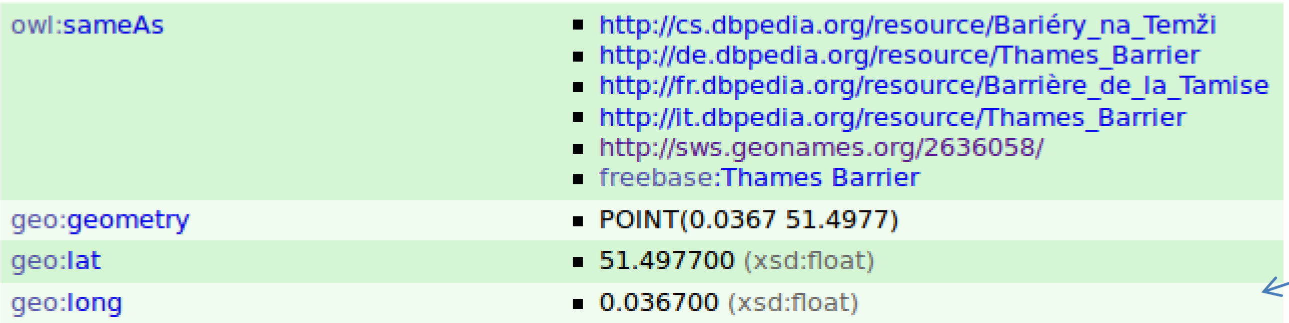


About: Thames Barrier

An Entity of Type : _Feature, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

The Thames Barrier is the world's second-largest movable flood barrier and is located downstream of central London, United Kingdom. Its purpose is to prevent London from being flooded by exceptionally high tides and storm surges moving up from the sea. It needs to be raised (closed) only during high tide; at ebb tide it can be lowered to release the water that backs up behind it.

…



owl:sameAs	▪ http://cs.dbpedia.org/resource/Bariéry_na_Temži ▪ http://de.dbpedia.org/resource/Thames_Barrier ▪ http://fr.dbpedia.org/resource/Barrière_de_la_Tamise ▪ http://it.dbpedia.org/resource/Thames_Barrier ▪ http://sws.geonames.org/2636058/ ▪ freebase:Thames Barrier
geo:geometry	▪ POINT(0.0367 51.4977)
geo:lat	▪ 51.497700 (xsd:float)
geo:long	▪ 0.036700 (xsd:float)

Links to GeoNames And Freebase

Latitude & Longitude

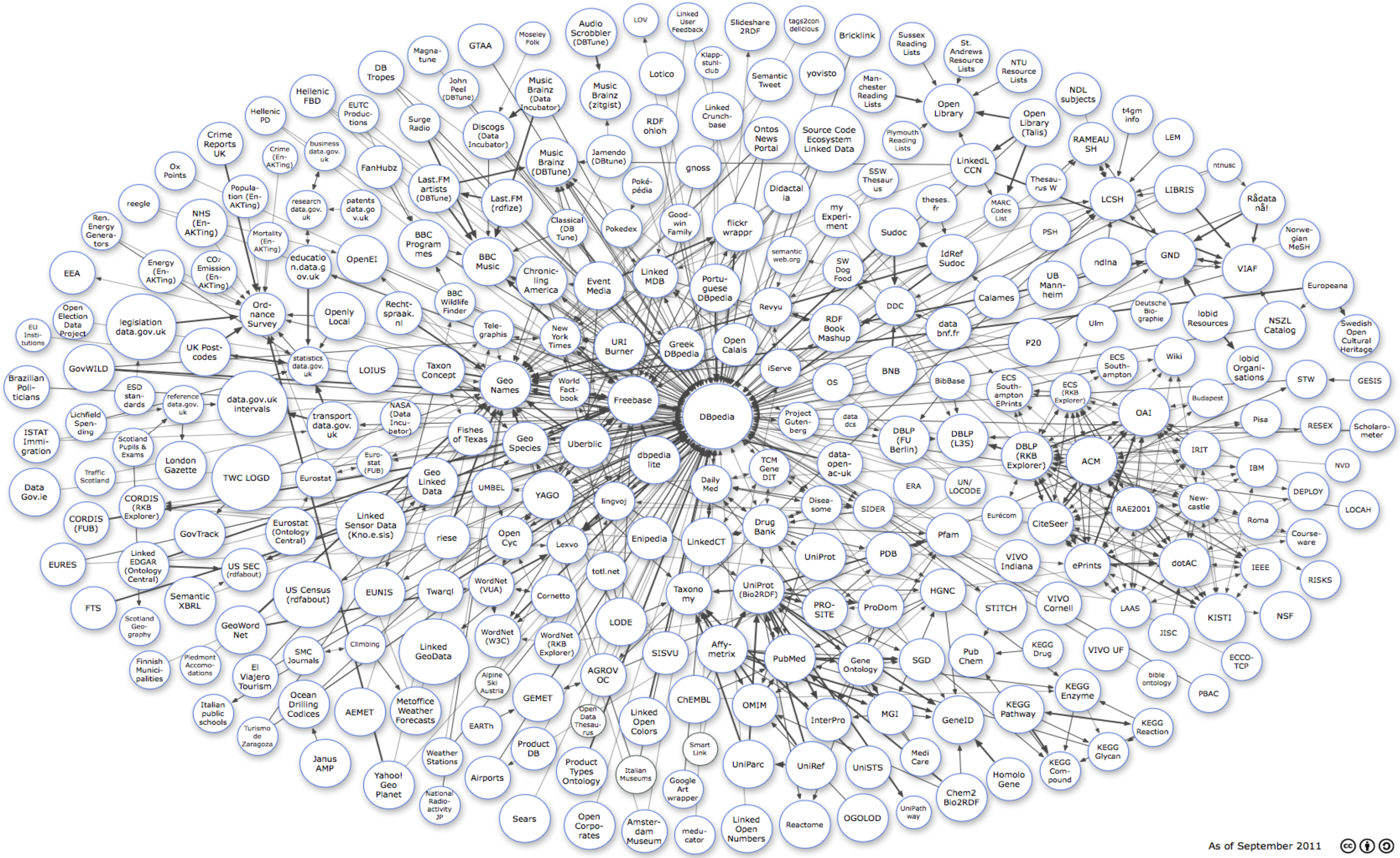
GeoNames

- 2.8 million populated places
  - 5.5 million alternate names
- Knowledge about NUTS country sub-divisions
  - use for enrichment of recognised locations with the implied higher-level country sub-divisions
- However, the sheer size of GeoNames creates a lot of ambiguity during semantic enrichment
- We use as an additional knowledge source, but not as a primary source (DBpedia)

Linking Knowledge from GeoNames and DBpedia

- For each DBpedia URI, find the corresponding Geonames URI (only for locations)
- Use GeoNames to obtain knowledge of:
  - Population
  - Latitude and longitude
  - Country code
  - Administrative regions

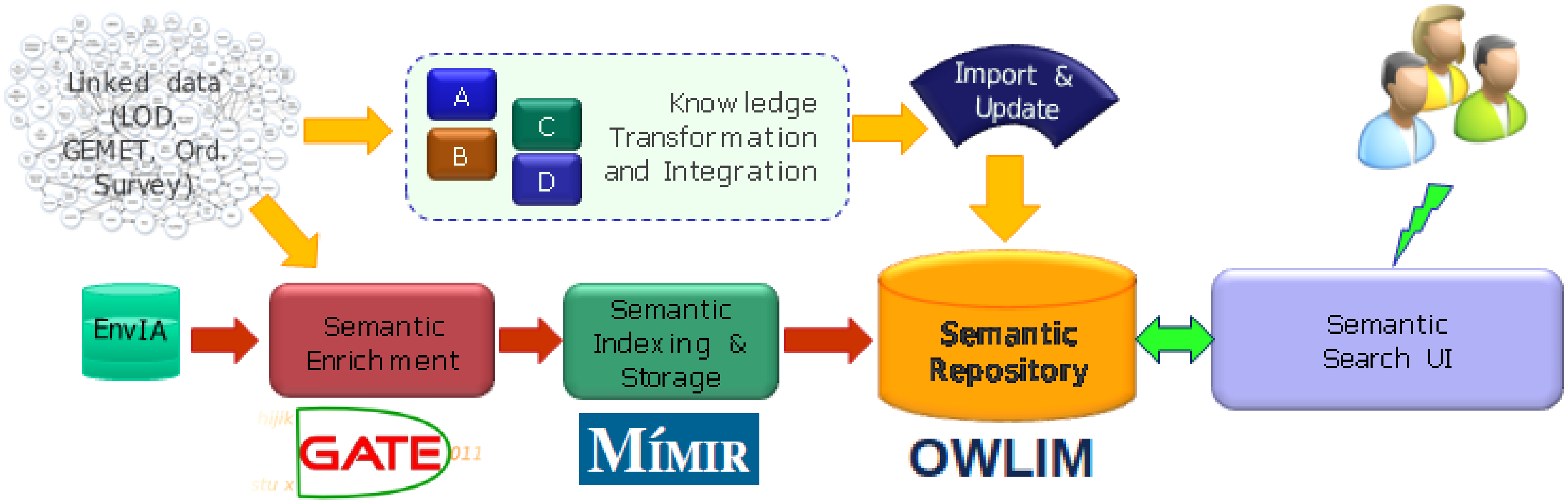
Linked Data Cloud



As of September 2011

Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

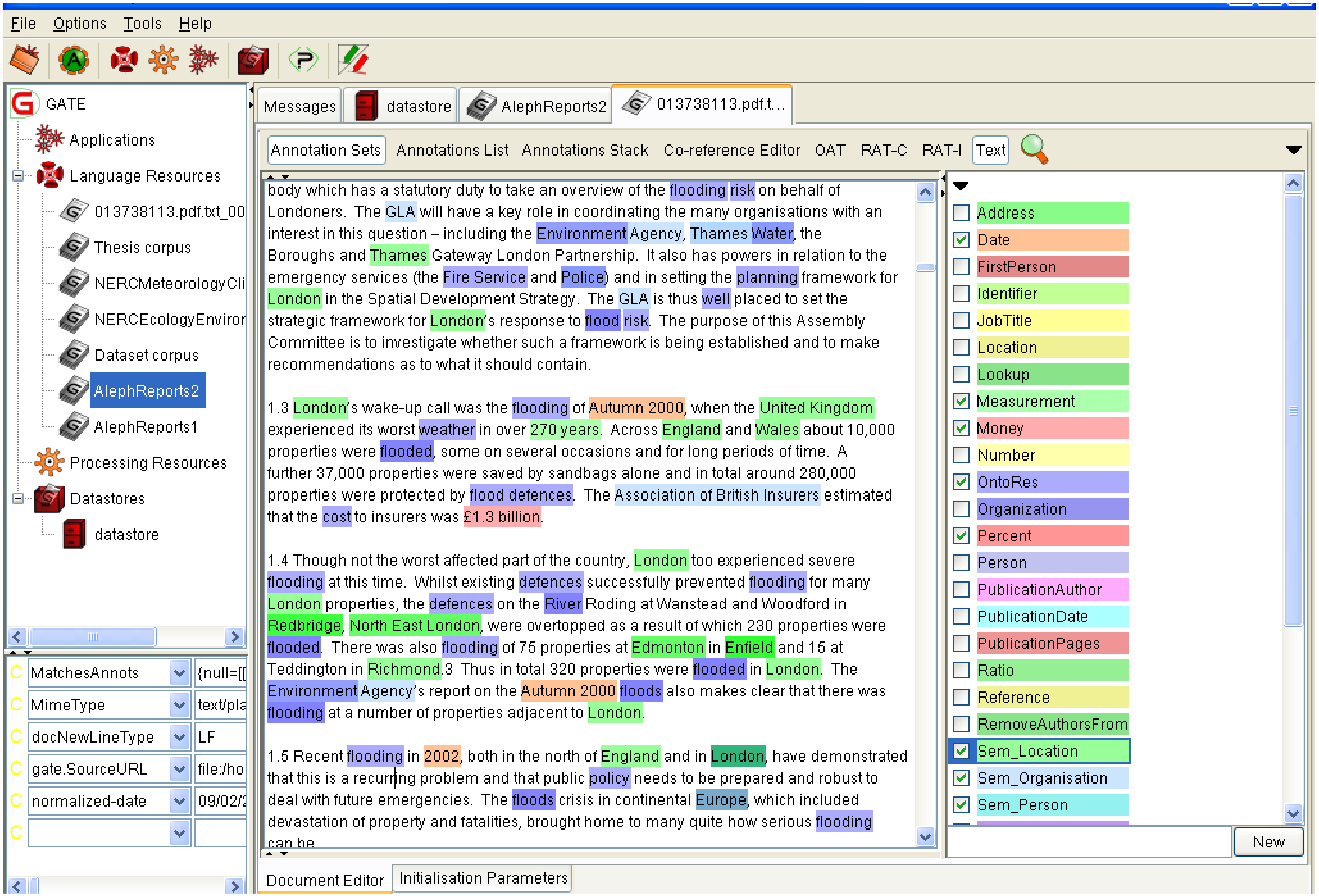
EnviLOD Architecture


Linked data (LOD, GEMET, Ord. Survey)
A
B
C
D
Knowledge Transformation and Integration
Import & Update
EnvIA
Semantic Enrichment
GATE
Semantic Indexing & Storage
MÍMIR
Semantic Repository
OWLIM
Semantic Search UI

Semantic enrichment with GATE

- GATE – General Architecture for Text Engineering
  - http://gate.ac.uk
  - Started in 1996, established; large developer community, incl. industrial committers (Ontotext, Intellius, SAIC)
- Tool for developing and deployment of Text Mining technology
- Used worldwide by many organisations to build bespoke solutions, e.g., TNA and Press Association
- A free open source framework (LGPL) and graphical development environment
- Includes Information Extraction in many languages
- Component based, easy mix between OS and proprietary plugins

Semantic Annotations



MatchesAnnots	{null=[[
MimeType	text/pla
docNewLineType	LF
gate.SourceURL	file:/ho
normalized-date	09/02/2

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.3 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.

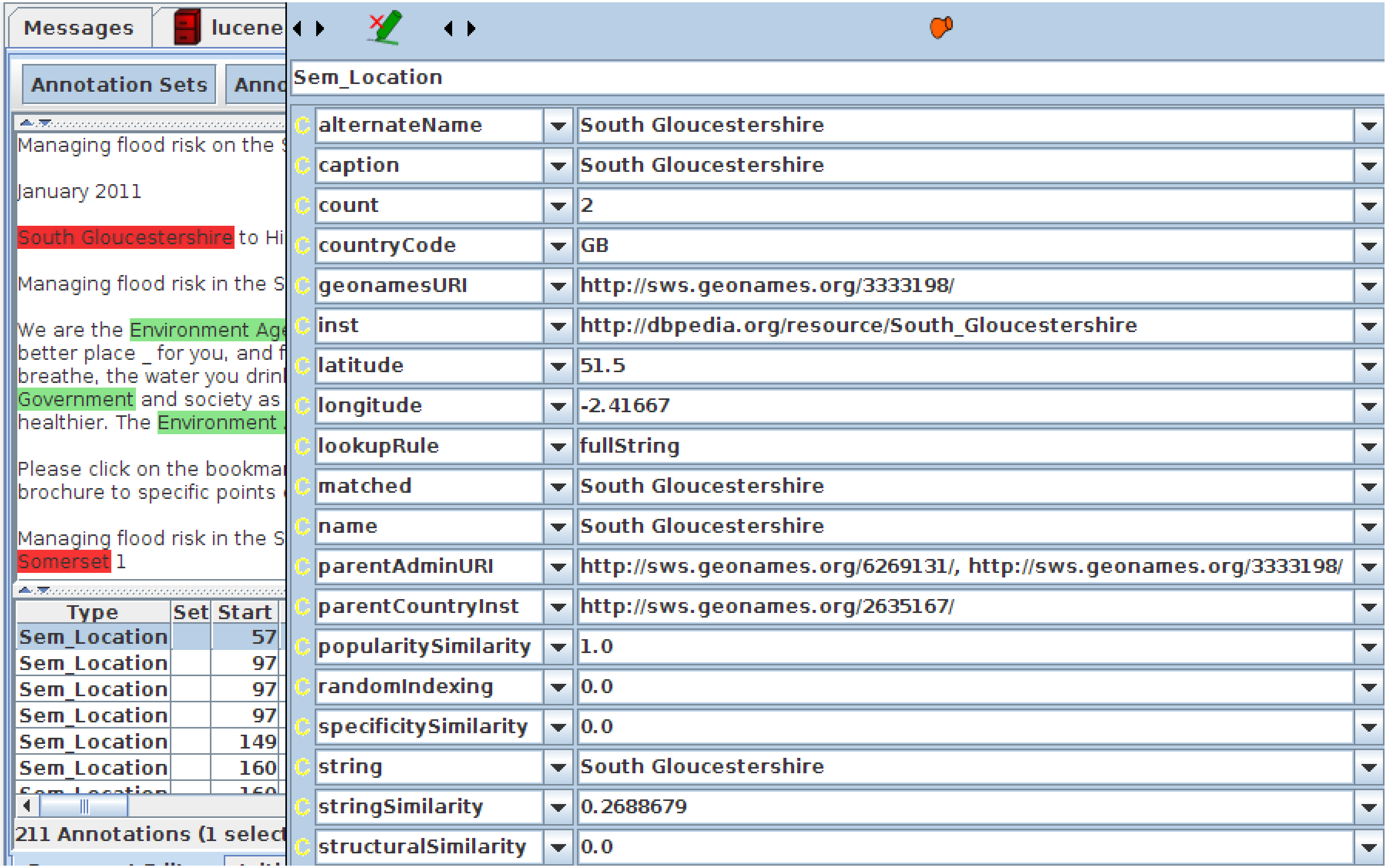
1.5 Recent flooding in 2002, both in the north of England and in London, have demonstrated that this is a recurring problem and that public policy needs to be prepared and robust to deal with future emergencies. The floods crisis in continental Europe, which included devastation of property and fatalities,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

- □ Address
- ☑ Date
- □ FirstPerson
- □ Identifier
- □ JobTitle
- □ Location
- □ Lookup
- ☑ Measurement
- ☑ Money
- □ Number
- ☑ OntoRes
- □ Organization
- ☑ Percent
- □ Person
- □ PublicationAuthor
- □ PublicationDate
- □ PublicationPages
- □ Ratio
- □ Reference
- □ RemoveAuthorsFrom
- ☑ Sem_Location
- ☑ Sem_Organisation
- ☑ Sem_Person

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



Managing flood risk on the

January 2011

South Gloucestershire to Hi

Managing flood risk in the S

We are the Environment Age
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

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%

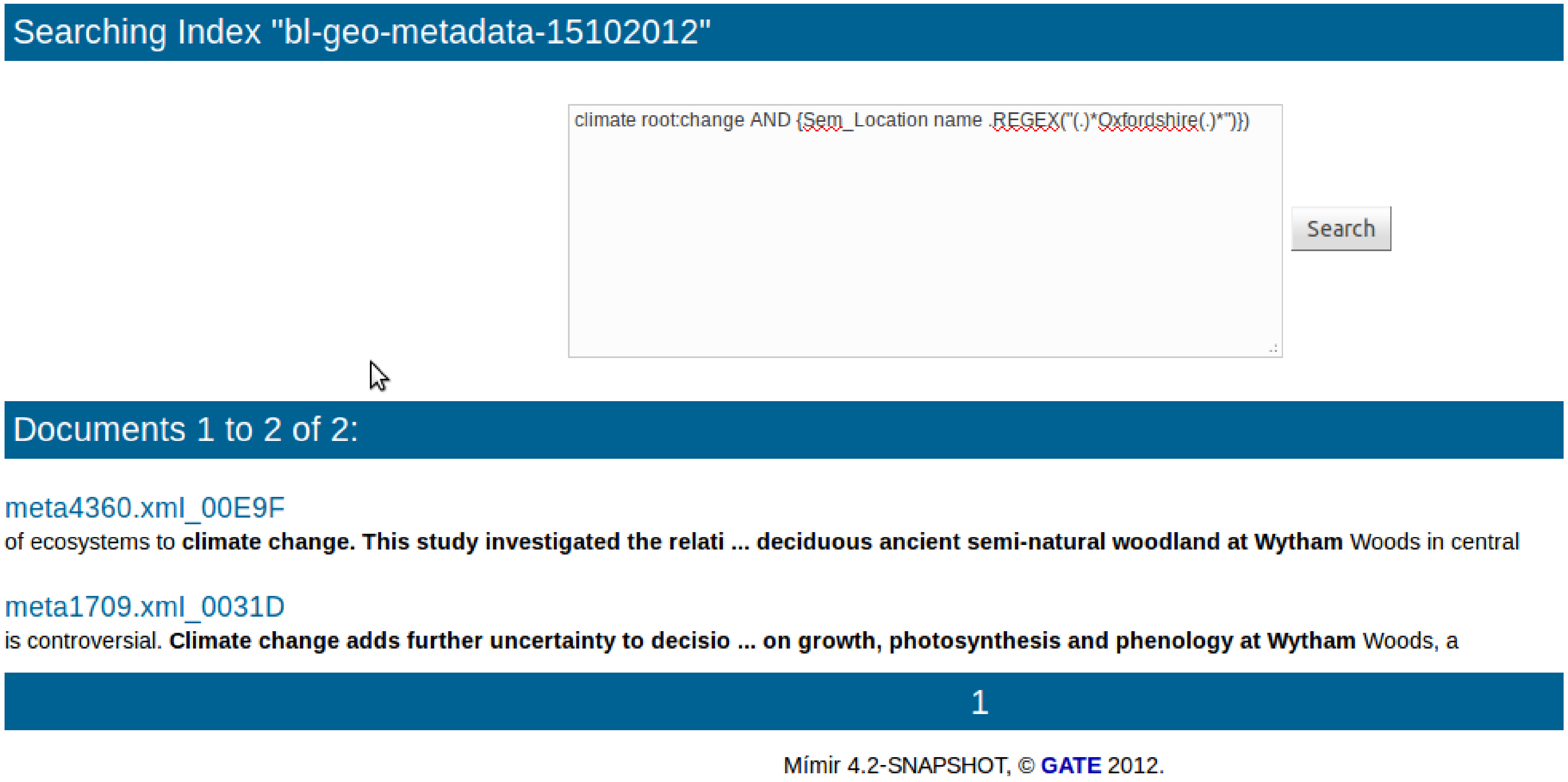
Semantic Search

- **Semantic annotation**: rather than just annotating the word "Cambridge" as a location, link it to an ontology instance
  - Differentiate between *Cambridge, UK* and *Cambridge, Mass*.
- **Semantic search via reasoning**
  - Now we can infer that this document mentions a city in Europe.
  - Ontologies tell us that this particular Cambridge is part of the country called the UK, which is part of the continent Europe.
- Semantic search matches not against the strings, but against their meaning
- Additional knowledge from DBpedia and other linked LOD resources can be brought in, to improve search results

MIMIR: Semantic Search Platform

- Searching and managing text annotations, semantic information, and full text documents in one search engine
- Queries over annotation graphs
- Regular expressions, Kleene operators
- Provides a Google-like search UI, currently experimental
- Designed to be integrated as a web service in custom end-user systems with bespoke interfaces
- Open source (see http://gate.ac.uk)

Climate change in Oxfordshire



Searching Index "bl-geo-metadata-15102012"

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

Documents 1 to 2 of 2:

meta4360.xml_00E9F

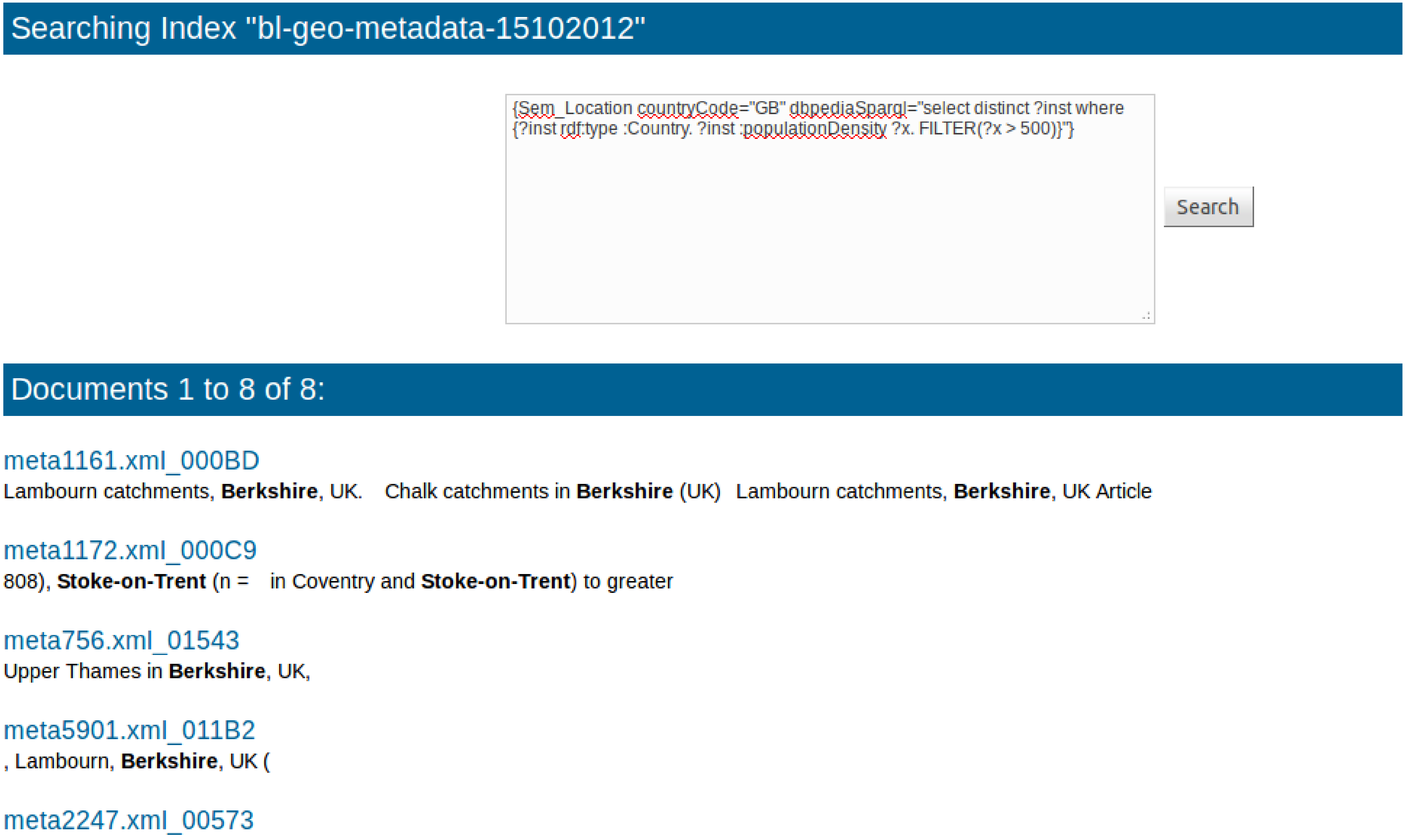
of ecosystems to **climate change. This study investigated the relati ... deciduous ancient semi-natural woodland at Wytham** Woods in central

meta1709.xml_0031D

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

Mímir 4.2-SNAPSHOT, © **GATE** 2012.

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

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

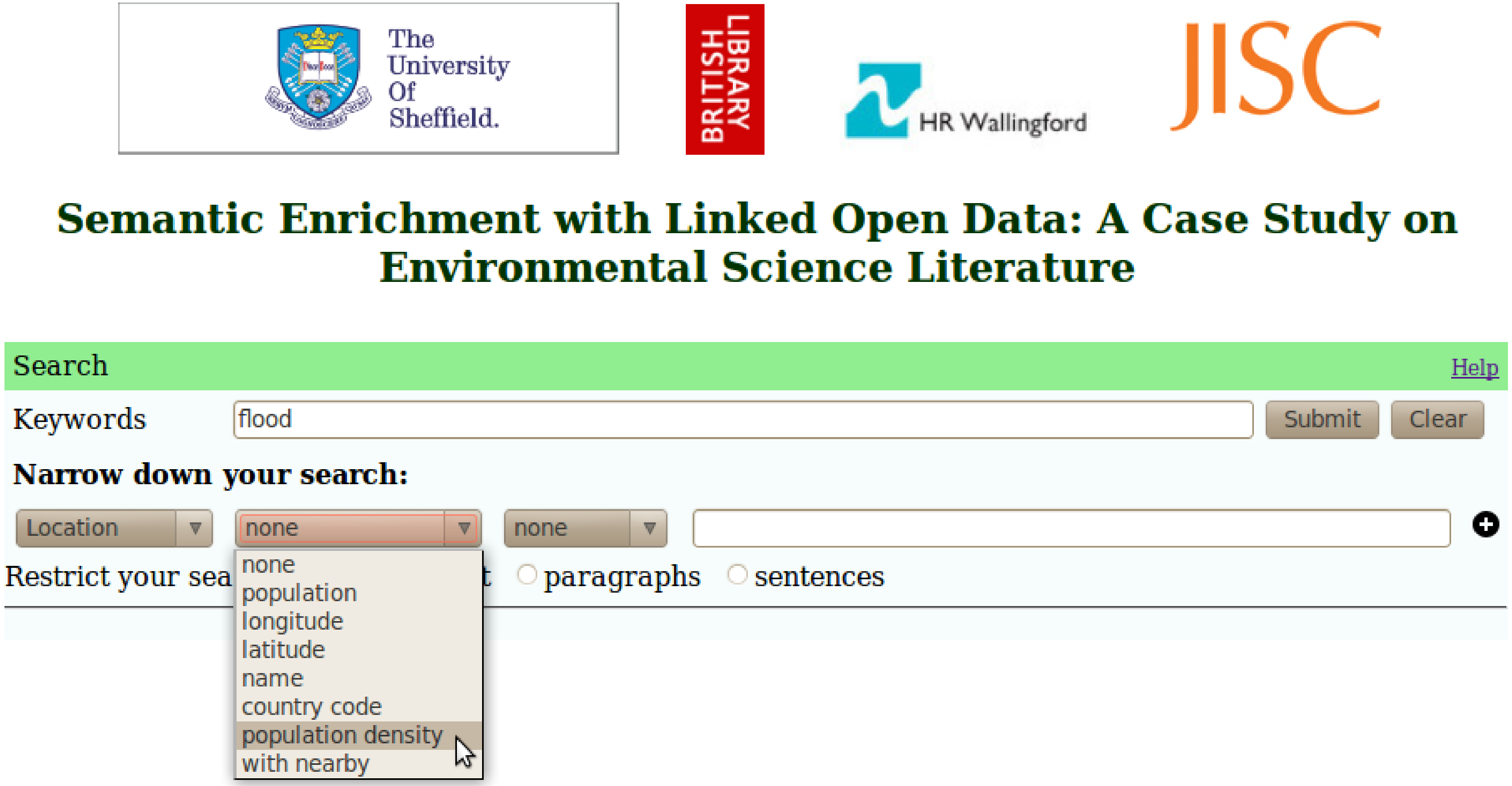
Upper Thames in **Berkshire**, UK,

meta5901.xml_011B2

, Lambourn, **Berkshire**, UK (

meta2247.xml_00573

EnviLOD Semantic Search UI



The University Of Sheffield. BRITISH LIBRARY HR Wallingford JISC

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

Search

Keywords: flood

Narrow down your search:

Location ▼ | none ▼ | none ▼

Restrict your sea… ○ paragraphs ○ sentences

Some Caveats….

- Demonstrator built in a 6 month project
- VERY limited content indexed
  - Some Defra, Environment Agency, Scottish Government reports
  - Some NERC metadata from the NORA repository
- PDFs of the articles are not connected to

Your thoughts…..

- Usability?
- How it discovers information?
- Is this something you could imagine being useful?
- What would need to change to make it useful

Thank you!

- More info on EnviLOD
  - johanna.kieniewicz@bl.uk
  - K.Bontcheva@dcs.shef.ac.uk
- On British Library Envia
  - Envia@bl.uk
  - johanna.kieniewicz@bl.uk