Semantic Search

Kalina Bontcheva
University of Sheffield

PROMISE Winter School 2013
Bressanone, Italy

Outline of the Lecture

- What is Semantic Search? Why is it Useful?
- How does it work?
  - Semantic Annotation
  - Semantic Search
- How do we get the users to use it?
  - Faceted entity search
  - Form-based semantic constraints
  - Natural language queries
- Does it work? Peter Mika on Thursday 8:30am

Semantic Queries in Google

Paris convention and visitors office - Official website - **Paris** tourism
en.**paris**info.com/
Paris convention and visitors office diffuses all information to organise your stay or your trip in **Paris**: hotels and loadings, museums, monuments, going out, **...**
Our welcome centres - Paris Map - Transports and ... - Getting around - Book online

Paris - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Paris**
Coordinates: 48°51'24"N 2°21'03"E / 48.8567°N 2.3508°E / 48.8567; 2.3508. **Paris** is the capital and largest city of France. It is situated on the river **...**
List of tourist attractions in Paris - History of Paris - Demographics of Paris - Portal

Paris.com - **Paris** Travel Guide and hotel accommodation
www.**paris**.com/
Paris.com : **Paris**, France tourist services offering hotel accommodation, holiday apartments. We guide you to the best **Paris** city tours and things to do!

News for **paris**

Paris women finally allowed to wear trousers
BBC News - 21 minutes ago
The French government overturns a 200-year-old ban on women wearing trousers in the capital, **Paris**, dating from November 1800.

Skirts rule lifted: Centuries-old ban on women wearing trousers in **Paris** is finally axed
Mirror.co.uk - 3 hours ago
Women in **Paris** finally allowed to wear trousers
Telegraph.co.uk - 1 day ago

Paris | Travel | The Guardian
www.guardian.co.uk/travel/**paris**
Latest news and comment on **Paris** from guardian.co.uk.

Paris

Paris is the capital and largest city of France. It is situated on the river Seine, in northern France, at the heart of the Île-de-France region. The city of Paris, within its administrative limits, has a population of about 2,230,000. Wikipedia

Population: 2,234,105 (2009)

Area: 105.4 km²

Weather: 8°C, Wind SW at 10 mph (16 km/h), 71% Humidity

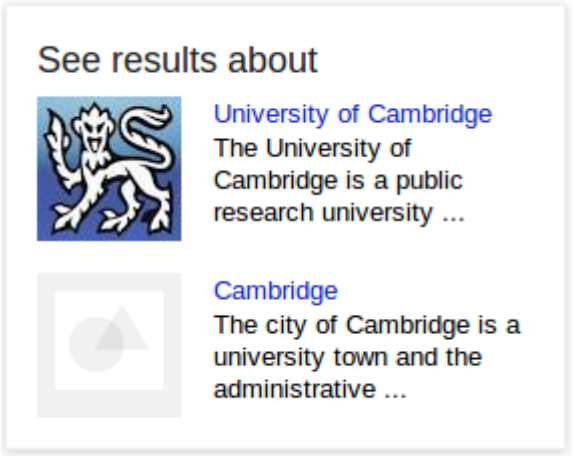
Local time: Monday 23:12

Points of interest



Searching for Things, Not Strings

- 500 million entities that Google "knows" about
- Used to provide more accurate search results



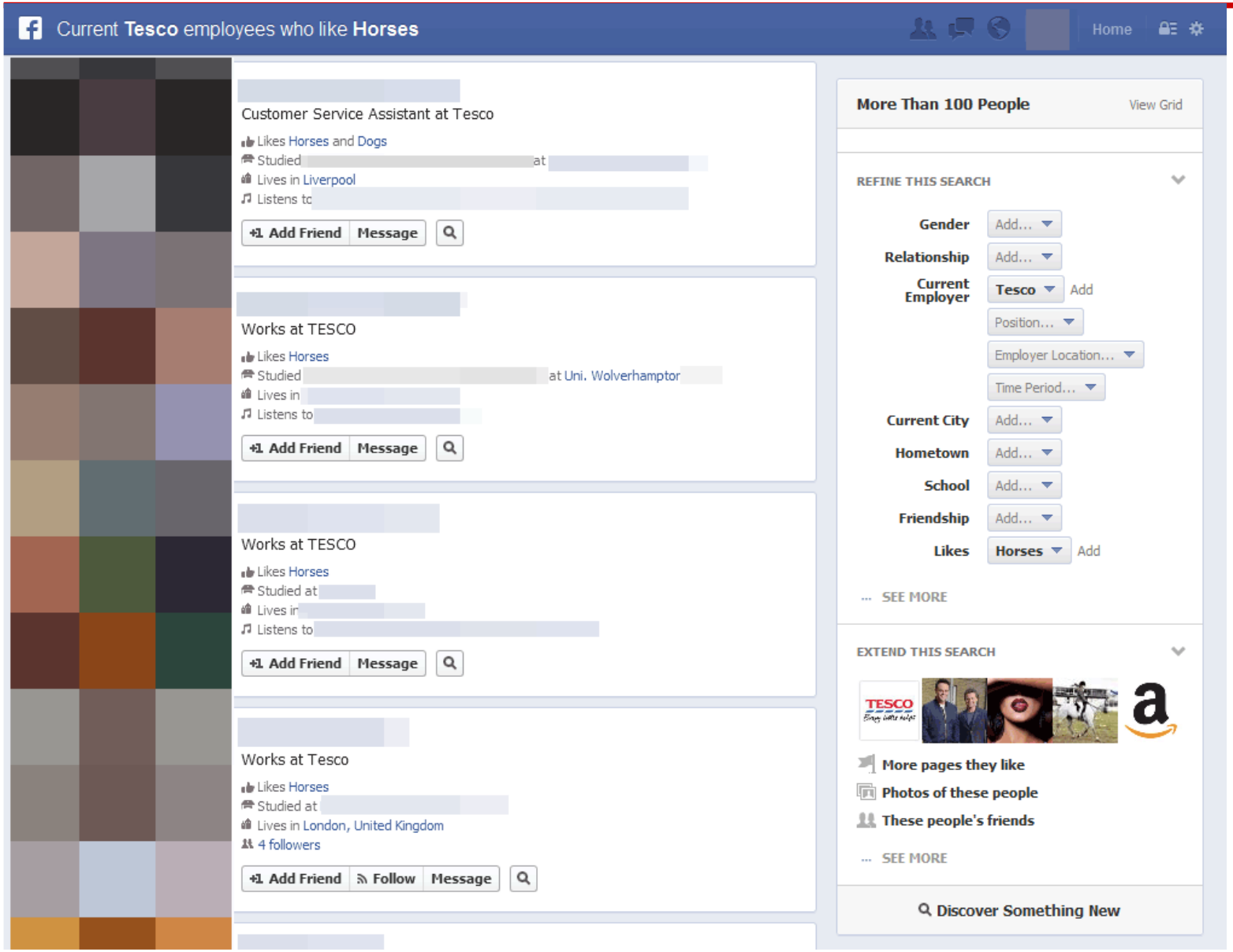
- Summaries of information about the entity being searched

http://googleblog.blogspot.it/2012/05/introducing-knc

University of Sheffield,
NLP

Facebook Graph Search: Now in Beta

http://actualfacebookgraphsearches.tumblr.com/



Embedded Semantics

- RDFa (or Resource Description Framework – in – attributes) – W3C standard
  - Adds a set of attribute-level extensions to HTML and XHTML, to describe rich metadata, embedded within Web documents

```
<p xmlns:dc="http://purl.org/dc/elements/1.1/"
   about="http://www.example.com/books/wikinomics">
  In his latest book
  <cite property="dc:title">Wikinomics</cite>,
  <span property="dc:creator">Don Tapscott</span>
  explains deep changes in technology,
  demographics and business.
  The book is due to be published in
  <span property="dc:date" content="2006-10-01">October 2006</span>.
</p>
```

- Schema.org – similar endevour
  - Bing, Google, Yahoo, and Yandex support it

Automatic Metadata Enrichment

- Use Text Mining, e.g.
  - Information Extraction – recognise names of people, organisations, locations, dates, references, etc.
  - Term recognition – identify domain-specific terms
- Extend automatically article metadata in digital libraries, to improve search quality
- Example from using a customised GATE text mining pipeline to enrich metadata in the Envia environmental science repository
  - http://www.bl.uk/reshelp/experthelp/science/eventsandprojects/enviatbl/index.html

FRM Act annual report to Parliament ... : Flood Risk Management (Scotland) Act 20

EDINBURGH : SCOTTISH GOVERNMENT

DOWNLOAD DOCUMENT

LINK

http://www.scotland.gov.uk/Publications/Recent

CONTENT TYPE

continuing
Electronic

DATE

2010

SUBJECT

- Flood damage prevention--Scotland--Periodicals
- Floods--Risk assessment--Scotland--Periodicals
- Flood control--Scotland--Planning--Periodicals
- Other social problems and services
- Flood
- Act
- Scotland
- risk
- plans
- Risk Management
- Scottish Government
- local authorities
- National Park
- flood protection

Why ontologies for 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**
  - So 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.
- **Knowledge source**
  - If I want to annotate *strikes* in baseball reports, the ontology will tell me that a *strike* involves a *batter* who is a *person*
  - In the text “BA went on strike”, using the knowledge that BA is a company and not a person, the IE system can conclude that this is not the kind of strike it is interested in

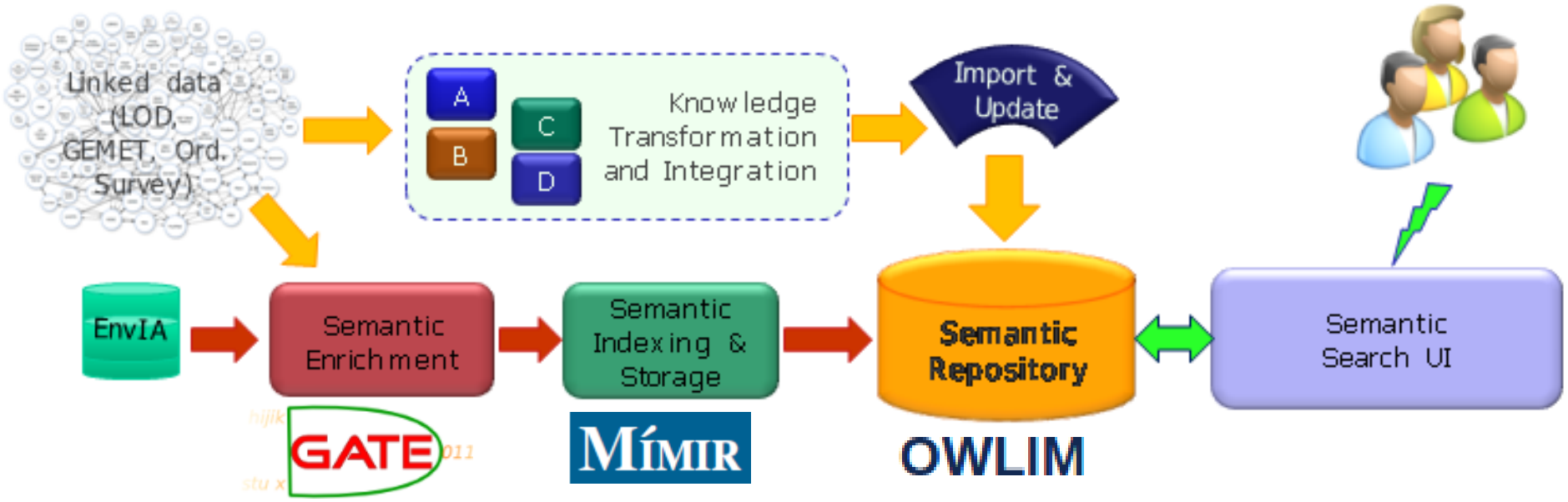
More semantic search examples


```
Q: {ScalarValue}{MeasurementUnit} ->
A: “12 cm”, “190 g”, “two hours”
---
Q: {Reference} ->
A: JP-A-60-180889
A: Kalderon et al. (1984) Cell 39:499-509
```

Outline of the Lecture

- What is Semantic Search? Why is it Useful?
- How does it Work?
  - Semantic Annotation
  - Semantic Search
- How do we get the users to use it?
  - Faceted entity search
  - Form-based semantic constraints
  - Natural language queries
- Does it work? Peter Mika on Thursday 8:30am

Example Semantic Search Architecture



What is Semantic Annotation

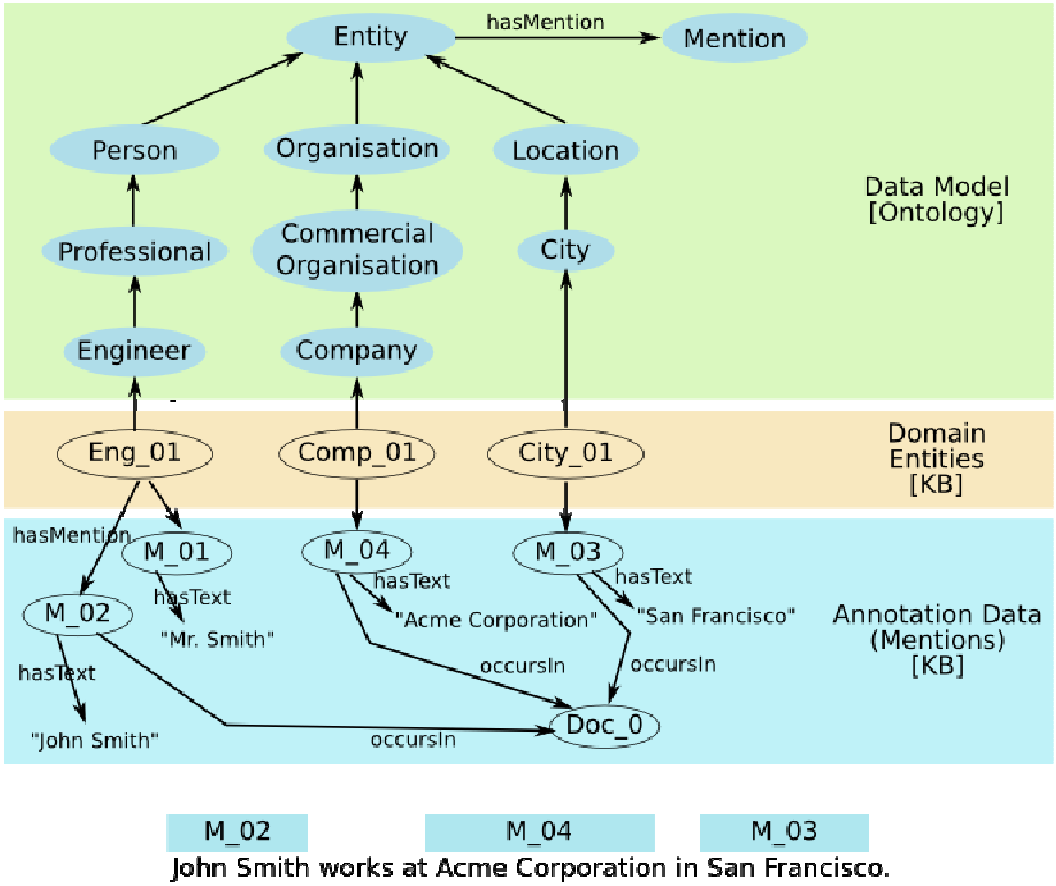
Annotation:

*The process of adding **metadata** to [parts of] a document.*

Semantic Annotation:

Annotation process where [parts of] the annotation schema (annotation types, annotation features) are ontological objects.

Semantic Annotation



Ontology – A Definition

- “An Ontology is a formal specification of a shared conceptualisation.” [Gruber]

What is an Ontology?

- Set of concepts (instances and classes)
- Relationships between them (is-a, part-of, located-in)
- Multiple inheritance
  - Classes can have more than one parent
  - Instances can have more than one class
- Ontologies are graphs, not trees

URIs, Labels, Comments

- The names of a classes or instance shown are URIs (Universal Resource Identifier)
  - http://dbpedia.org/page/Paris
- The linguistic lexicalisation is typically encoded in the **label** property, as a string
- The **comment** property is often used for documentation purposes, similarly a string
- Comments and labels are **annotation properties**

rdfs:label

- París
- Paříž
- Paris
- Paris
- París
- Pariisi
- Paris
- Párizs
- Parigi
- パリ
- 파리 (프랑스)
- Parijs
- Paris
- Paryż
- Paris
- Париж
- Paris
- Париж
- Paris
- 巴黎

Datatype Properties

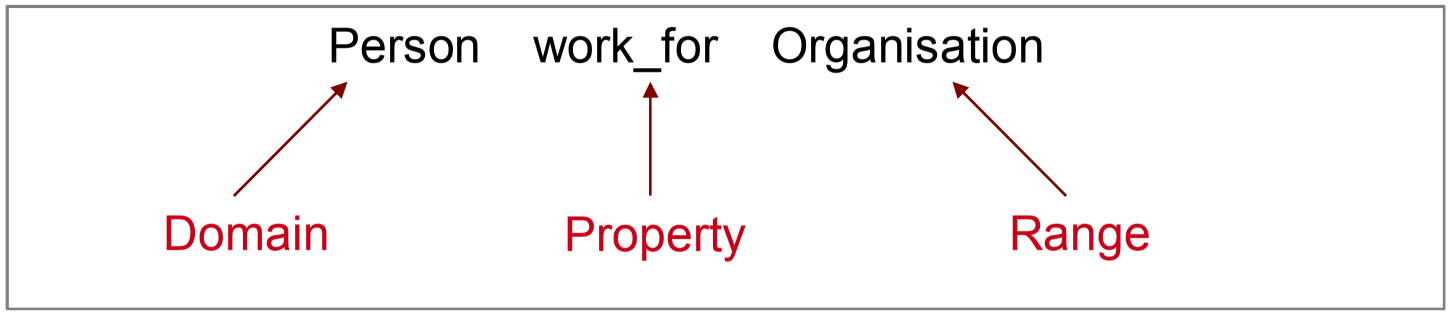
- Datatype properties link individuals to data values
- Datatype properties can be of type boolean, date, int,
  - e.g. a person can have an age property
- Available datatypes taken from XMLSchema

http://dbpedia.org/page/Paris

dbpprop:yearHighC	16 (xsd:integer)
dbpprop:yearLowC	9 (xsd:integer)
dbpprop:yearRecordHighC	40 (xsd:integer)
dbpprop:yearRecordLowC	-24 (xsd:integer)

Object Properties

- Object properties link instances together
- They describe relationships between instances, e.g. people work for organisations
- Domain is the subject of the relation (the thing it applies to)
- Range is the object of the relation (the possible"values")



Similar to domains, multiple alternative ranges can be specified by using a class description of the owl:unionOf, but this raises the complexity of the ontology and makes reasoning harder

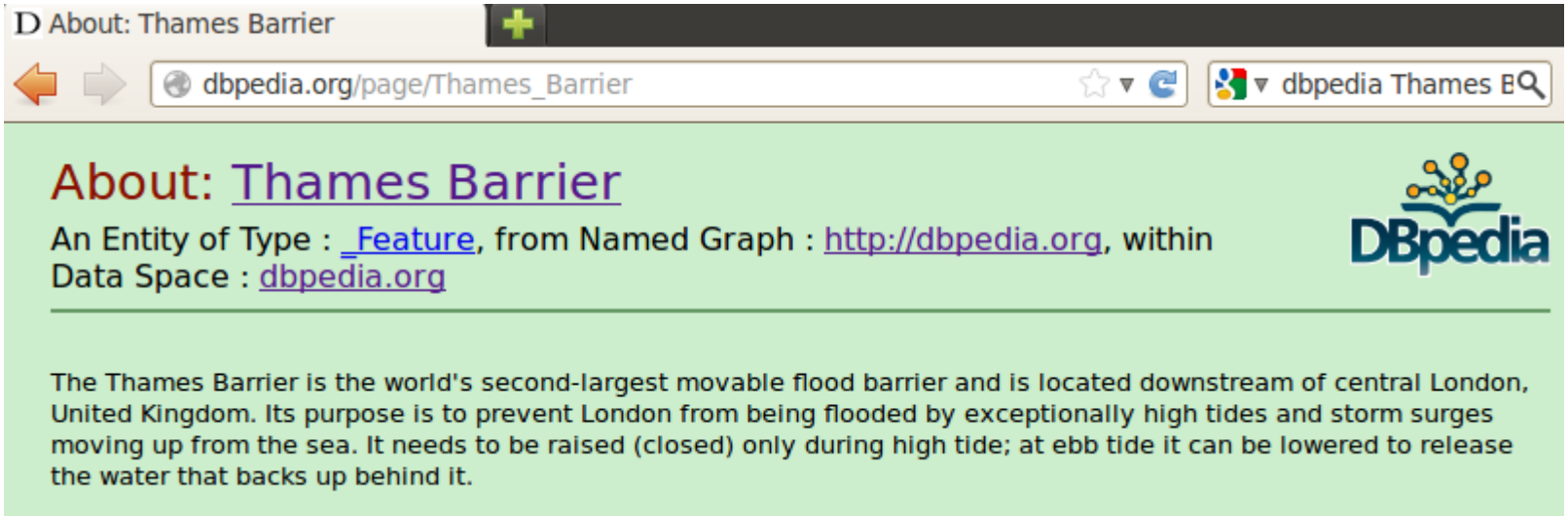
Linked Data: Design Principles

- Unambiguous identifiers for objects (resources)
  - Use URIs as names for things
- Use the structure of the web
  - Use HTTP URIs so that people can look up the names
- Make is easy to discover information about an object (resource)
  - When someone lookups a URI, provide useful information
- Link the object (resource) to related objects
  - Include links to other URIs

DBpedia

- Machine readable knowledge on various entities and topics, including:
  - 410,000 places/locations,
  - 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



. . .

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

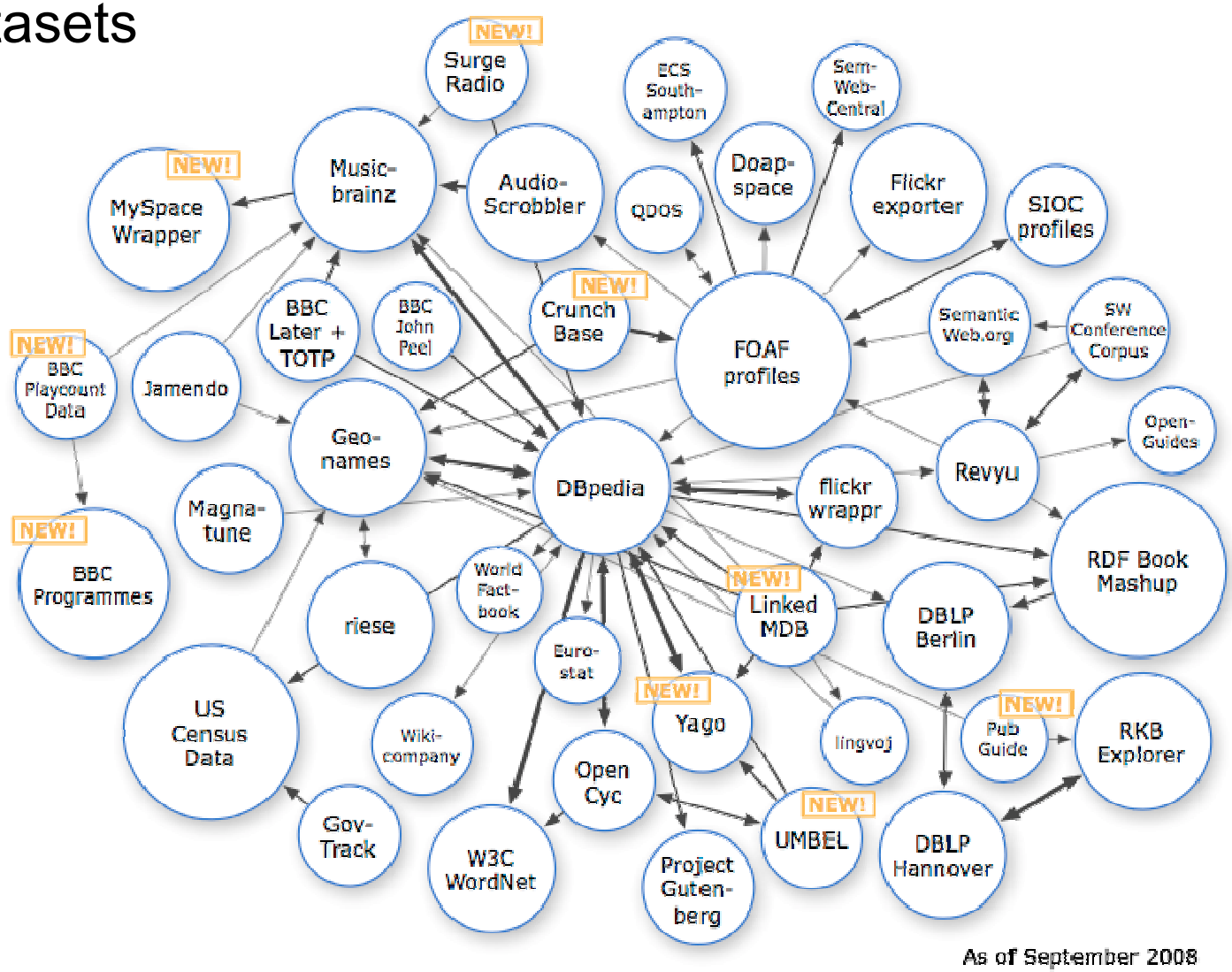
Latitude & Longitude

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 it as an additional knowledge source, but not as a primary source (DBpedia)

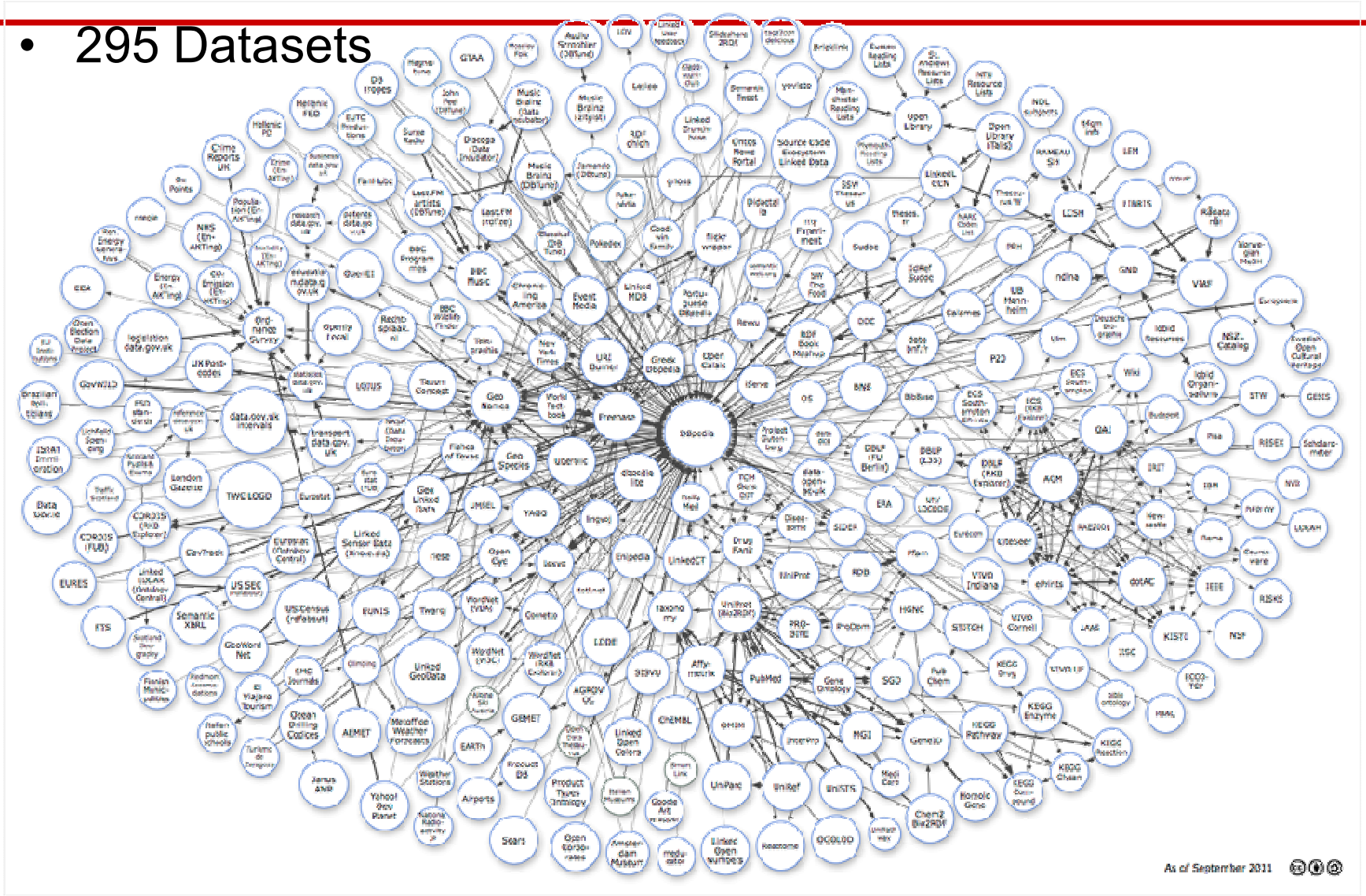
Linked Open Data: 2008

45 Datasets



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

Linked Open Data: Growth 2011



- 295 Datasets

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

Semantic Annotation In Brief

Information Extraction for the Semantic Web

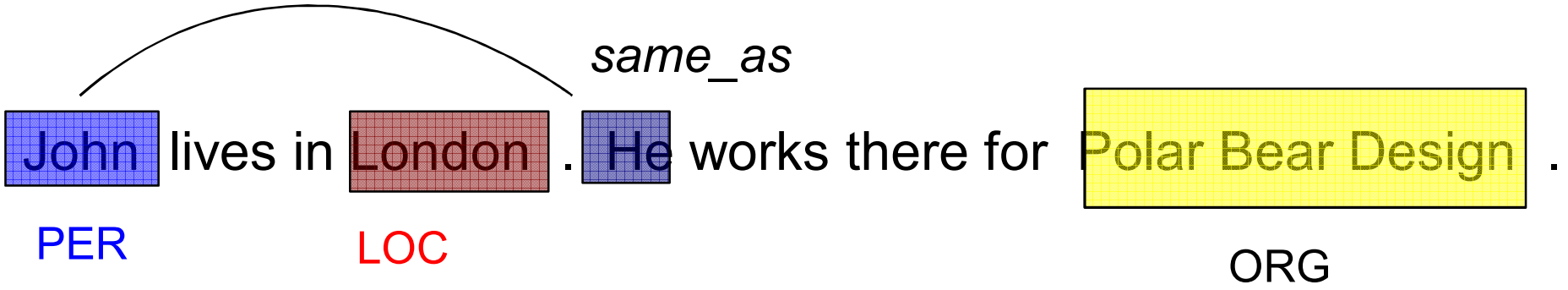
- Traditional IE is based on a flat structure, e.g. recognising Person, Location, Organisation, Date, Time etc.
- For the Semantic Web, we need information in a hierarchical structure
- Idea is that we attach semantic metadata to the documents, pointing to concepts in an ontology
- Information can be exported as an ontology annotated with instances, or as text annotated with links to the ontology

Traditional NE Recognition

John lives in London . He works there for Polar Bear Design .

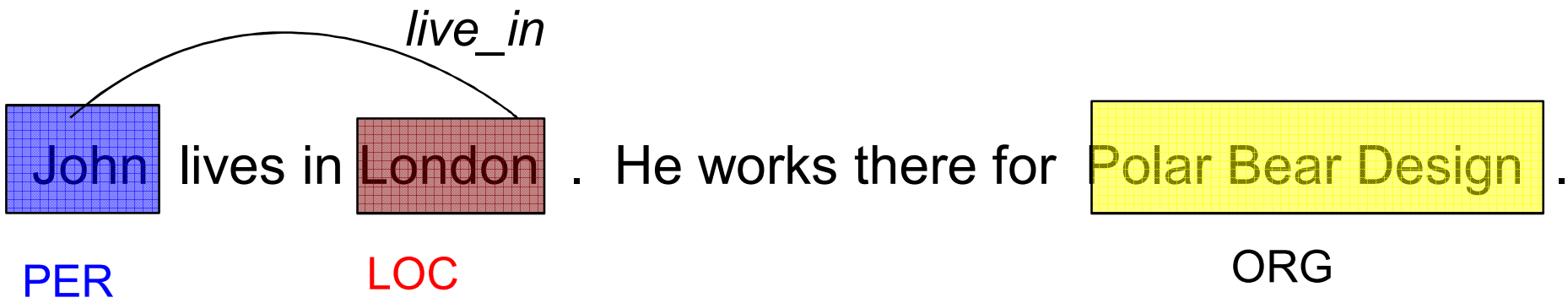
PERSON LOCATION ORGANISATION

Co-reference



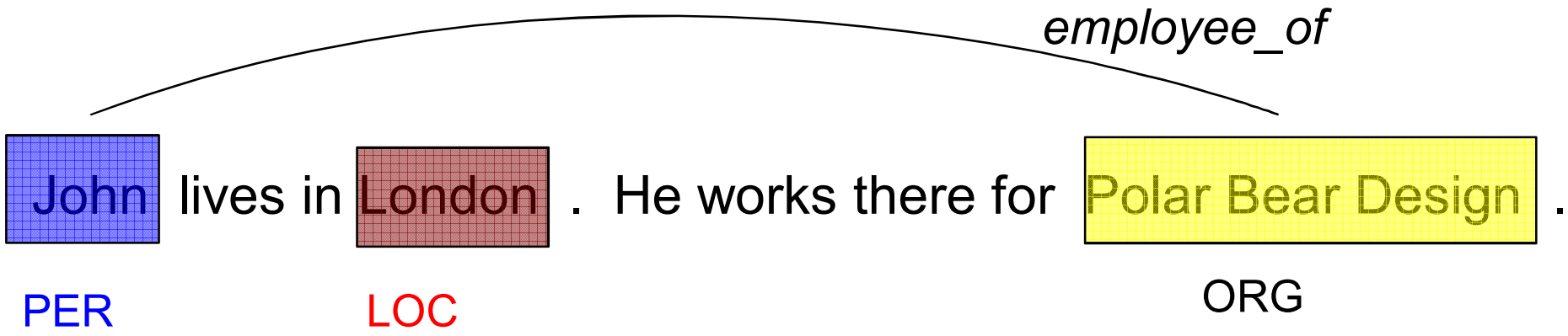

GATE

Relations


live_in
John
lives in
London
. He works there for
Polar Bear Design
.
PER
LOC
ORG

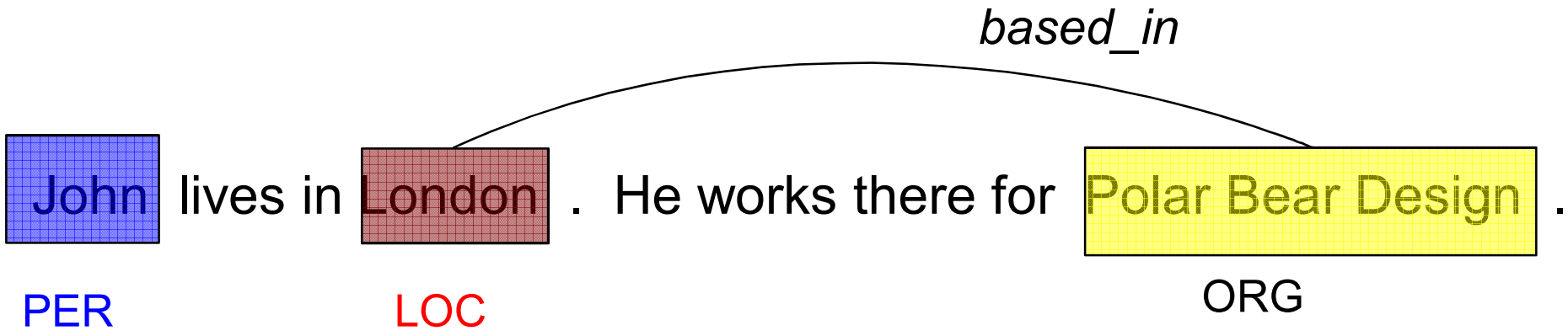

GATE

Relations (2)


employee_of
John
lives in
London
. He works there for
Polar Bear Design
.
PER
LOC
ORG

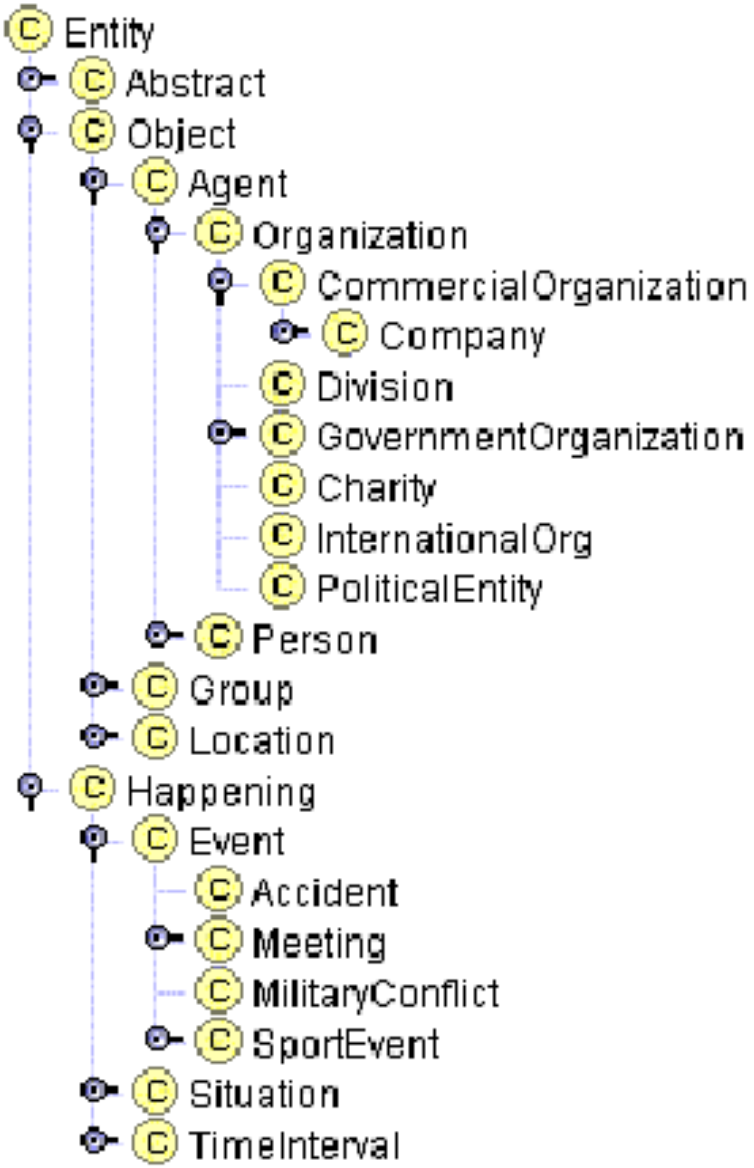

GATE

Relations (3)

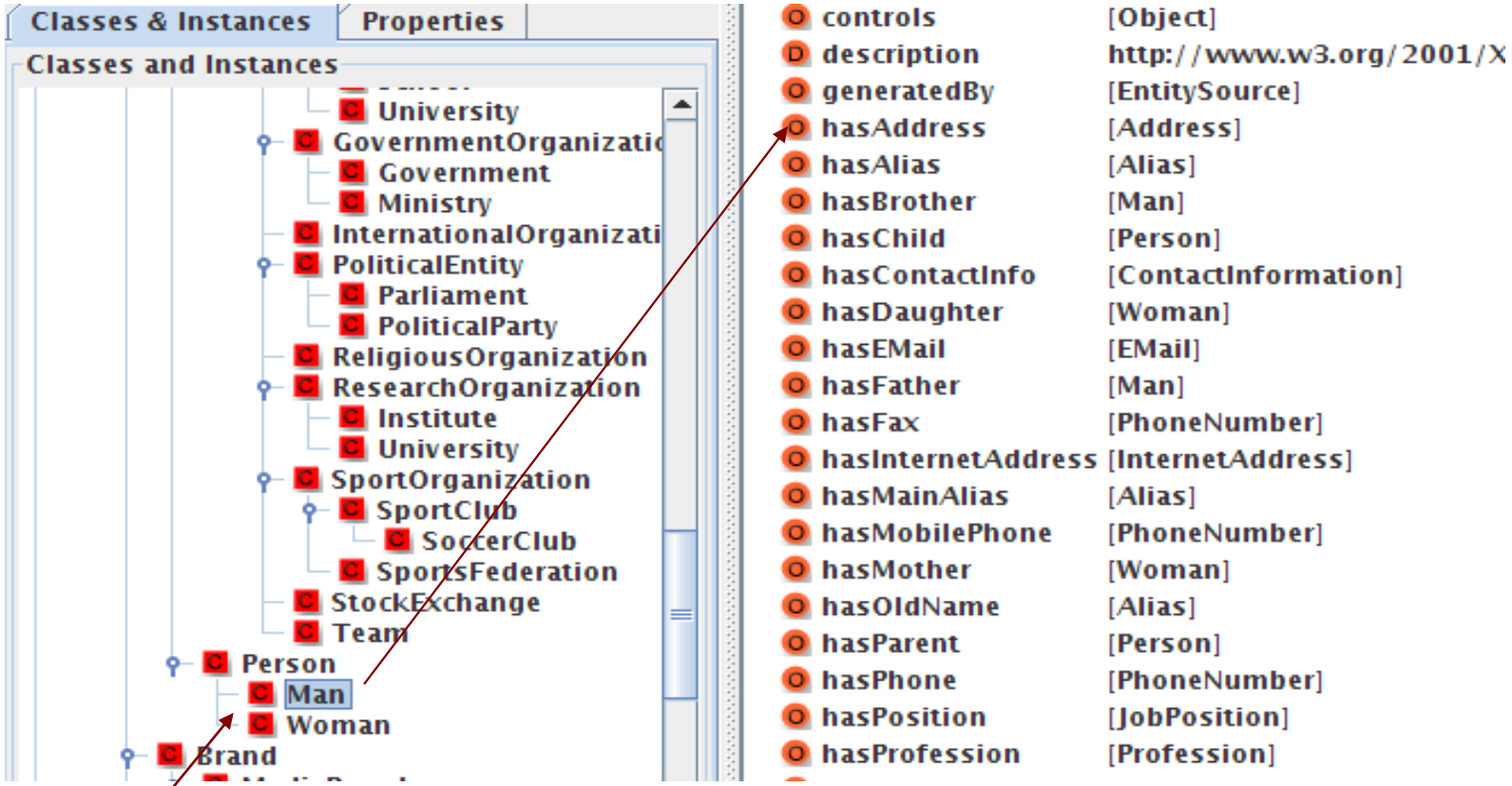

based_in
John
lives in
London
. He works there for
Polar Bear Design
.
PER
LOC
ORG

Richer NE Tagging

- Attachment of instances in the text to concepts in the domain ontology
- Disambiguation of instances, e.g. Cambridge, MA vs Cambridge, UK

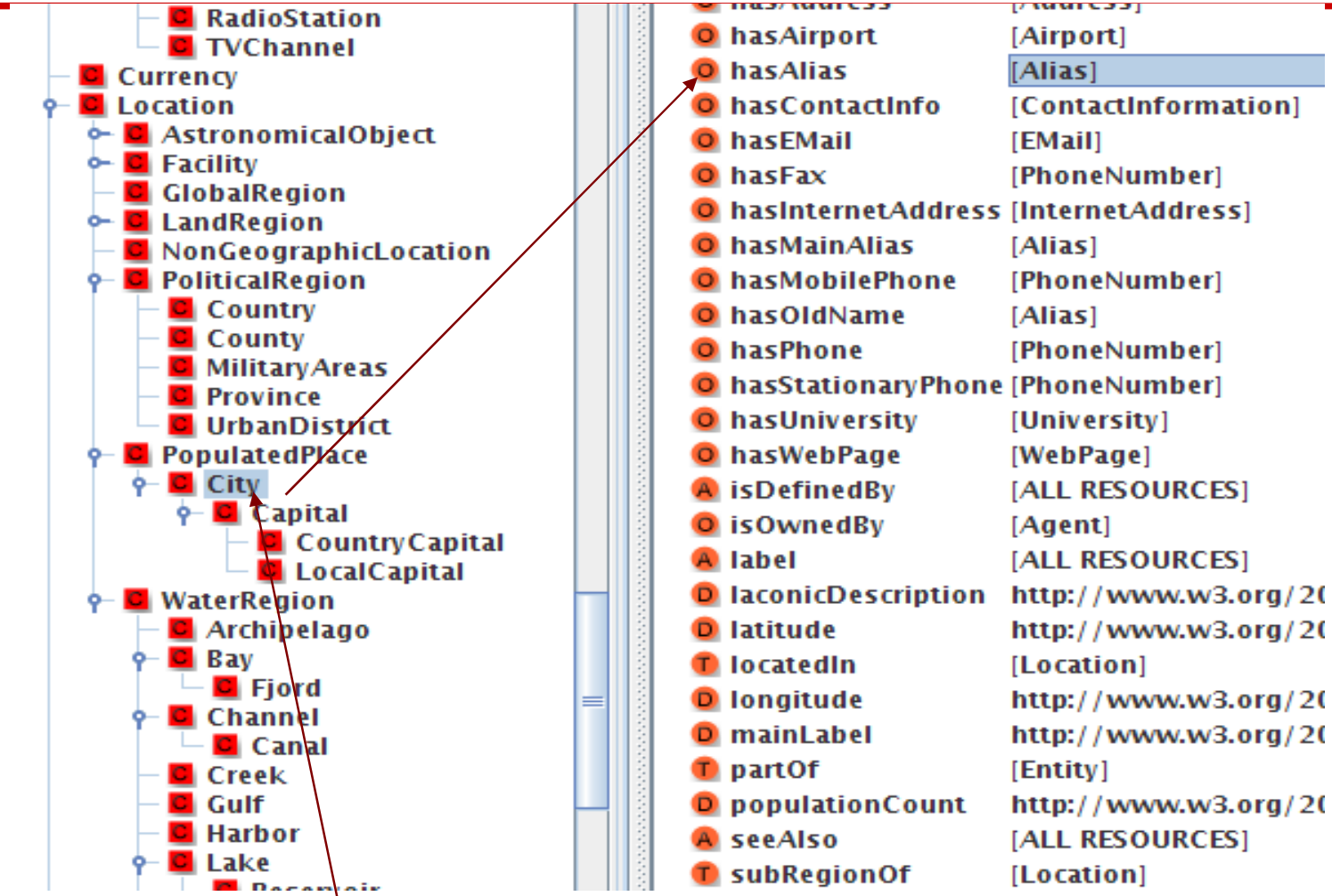


Ontology-based IE



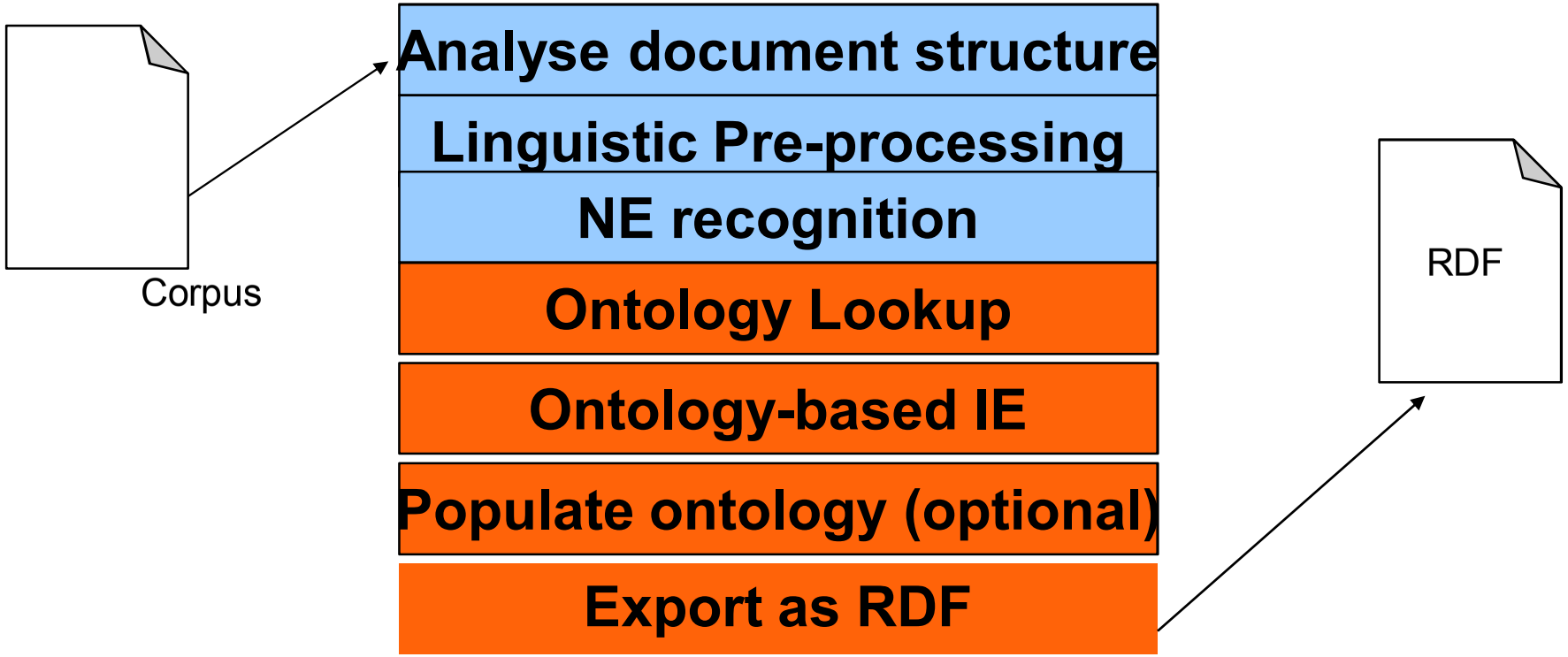
John lives in London. He works there for Polar Bear Design.

Ontology-based IE (2)



John lives in London. He works there for Polar Bear Design.

Typical Semantic Annotation pipeline



University of Sheffield,
NLP

OpenCalais

http://viewer.opencalais.com/

Paste text of http://www.membranes.com/

Since its founding in 1975, **Hydranautics** has been committed to the highest standards of **technology research**, produc
utics entered the reverse osmosis (RO) water treatment field in 1970, and is now one of the most respected and experie
y. **Hydranautics** became part of the **Nitto Denko Corporation** when it was acquired in 1987. **Hydranautics** corporate h
California in a 160,000 ft2 (14,684 m2) **manufacturing facility residing** on 14 acres, all owned by **Hydranautics**.

Hydranautics' continuing commitment to research and **technology results** in the ongoing development of a **range of s**
s' products are currently in use on seven continents throughout the world for **such diverse applications** as potable wat
astewater treatment, surface water treatment, seawater desalination, electronic rinse water, agricultural irrigation and

Comprehensive customer service and support are available virtually around the clock and around the world. **Hydranauti**
rk of worldwide sales offices throughout the **United States**, **Latin America**, **Europe** and **Asia**.

Entities:

- **City**
  - Oceanside,California,United
- **Company**
  - Hydranautics Inc
  - NITTO DENKO CORPORATION
- **Continent**
  - Asia
  - Europe
- **Country**
  - United States
- **Industry Term**
- **Province Or State**
  - California,United States
- **Region**
- **Technology**
  - wastewater treatment

Events & Facts:

- **Acquisition**
  - NITTO DENKO CORPORATION, 1987-00-00, in
- **Company Founded**
  - Hydranautics Inc, 1975
- **Generic Relations**
  - Hydranautics Inc, be
  - Hydranautics Inc, part of the Nitto Denko
  - Hydranautics Inc, commit
  - a network of worldwide sales offices,
  - Hydranautics Inc, the reverse osmosis, enter

Not easily customised/extended

Domain-specific coverage varies

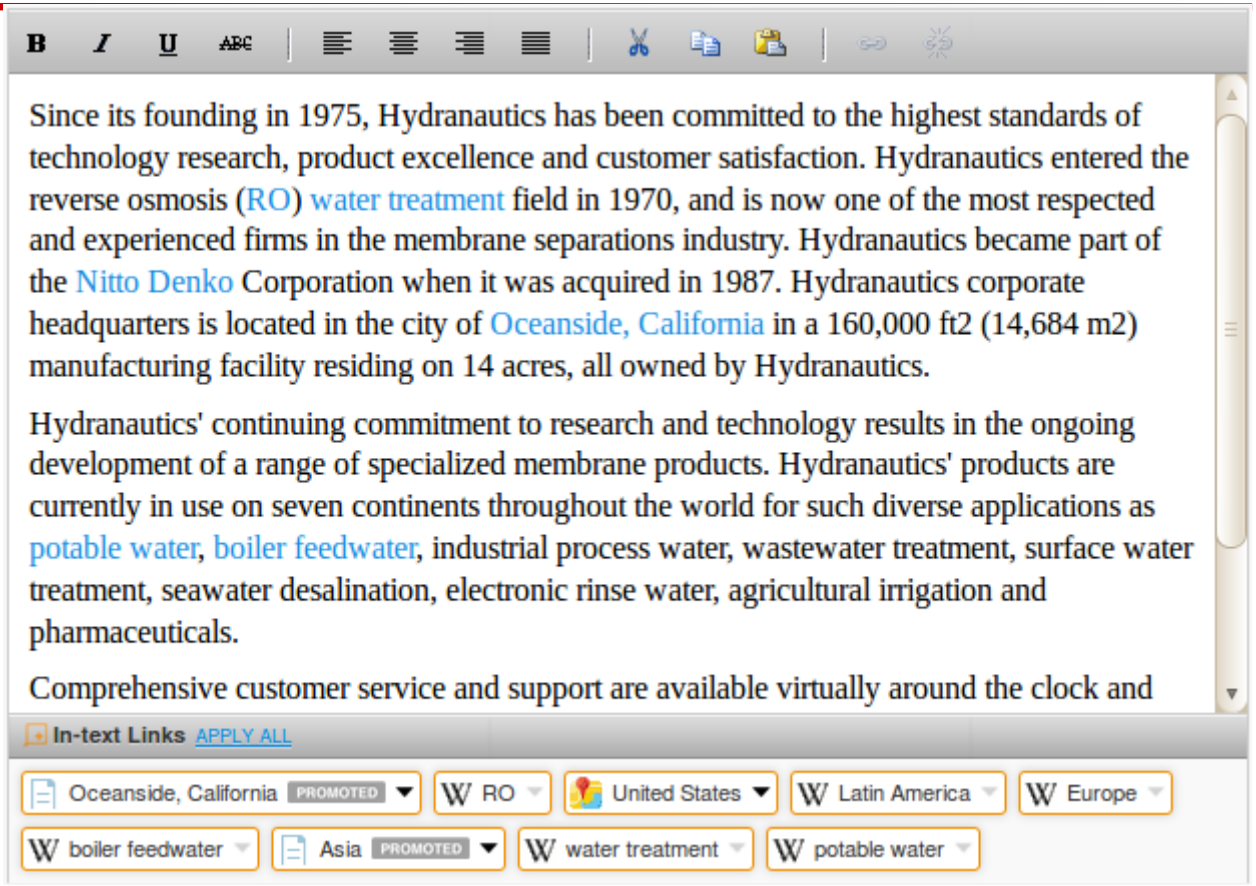
University of Sheffield, NLP

Zemanta

http://www.zemanta.com/demo/

Paste text from www.membranes.com

The main entity of interest: Hydranautics is missed



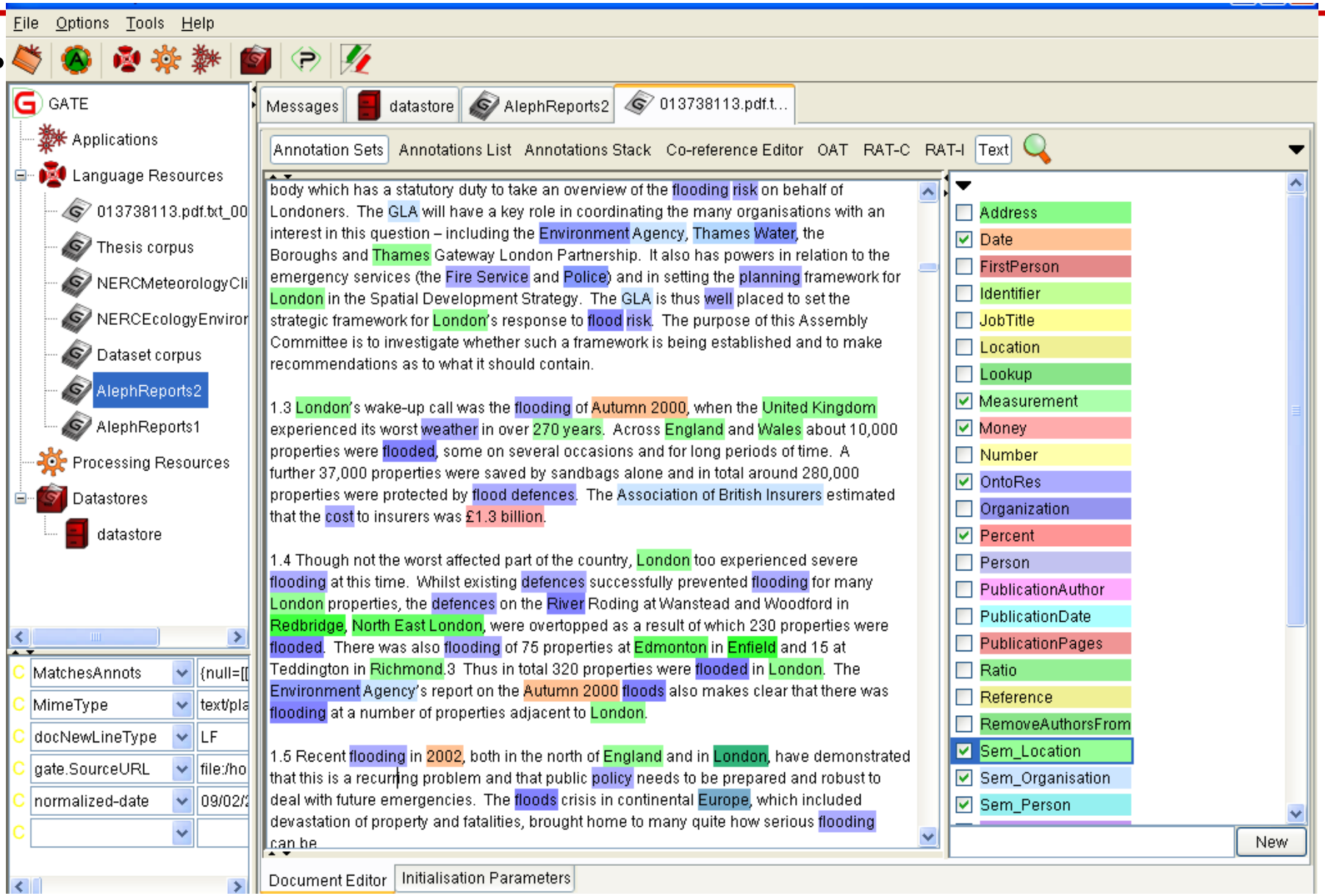
Common problem with general purpose, open-domain semantic annotation tools

Best results require bespoke customisation

Semantic annotation 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

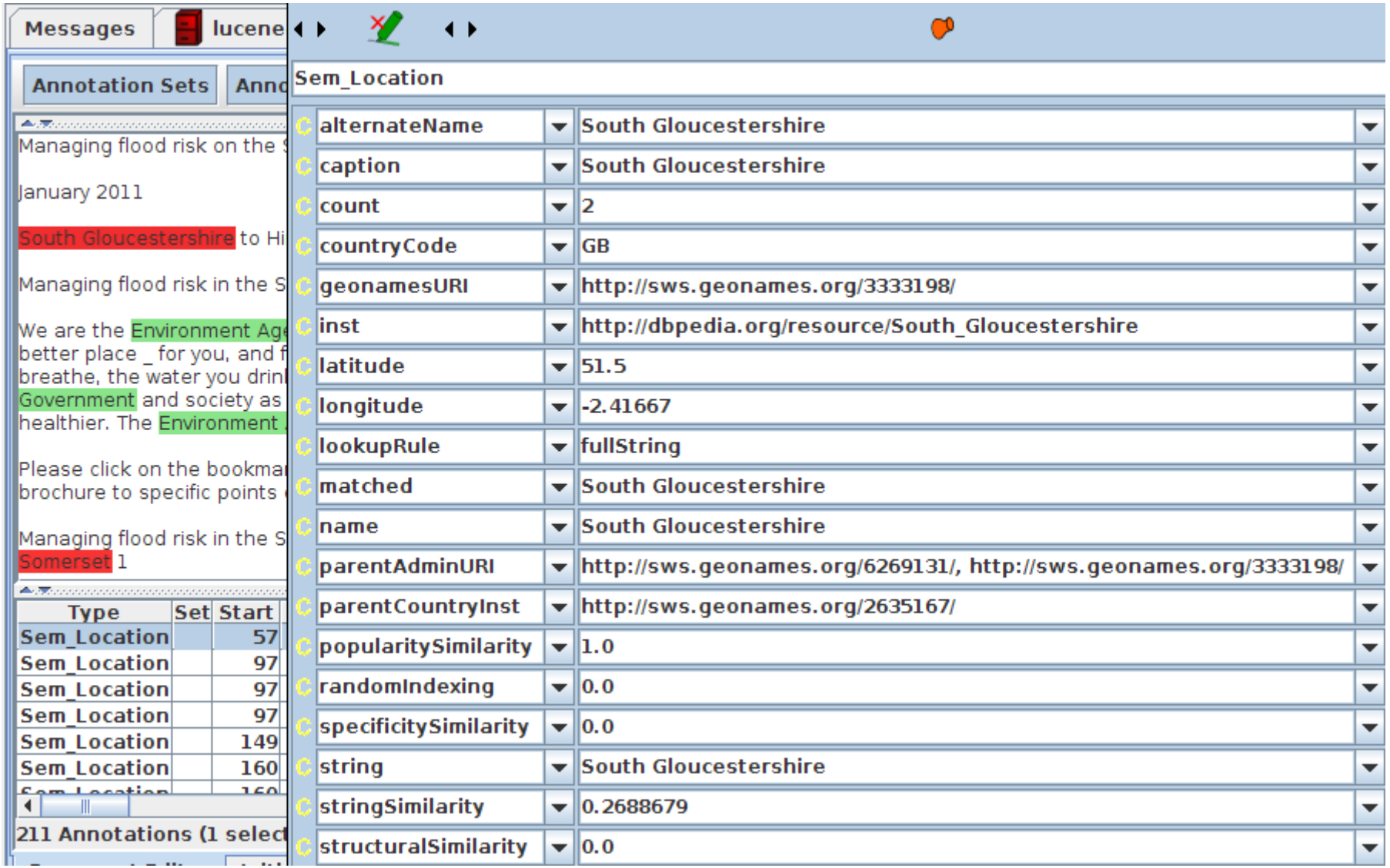
Semantic Annotations



Automatic Semantic Annotation

- 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 use knowledge to retrieve nearby places

“South Gloucestershire” Example



Semantic Annotation (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%

University of Sheffield,
NLP

The LODIE Service from GATE

http://demos.gate.c.uk/trendminer/obie/

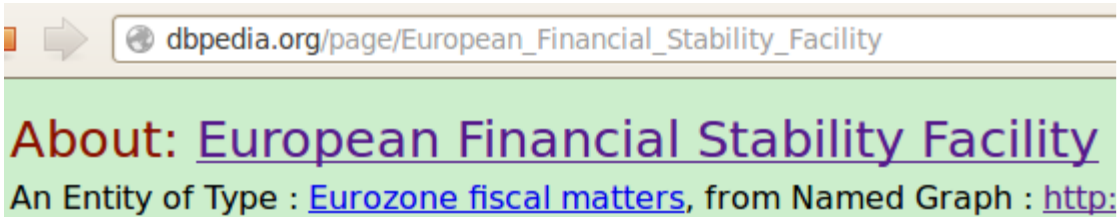
Ontology Based Information Extraction

A simple demo showing how the disambiguation service can be used to annotate documents against DBpedia.

Summary of #Greece bailout plan: 109bn aid; maturity of future EFSF loans 15-30 years, lower rates expecte... (cont) http://deck.ly/~A0muj

↓ Disambiguate ↓

Summary of #Greece bailout plan: 109bn aid; maturity of future EFSF loans 15-30 years, lower rates expecte... (cont) http://deck.ly/~A0muj



Evaluation: LODIE on TAC-KBP'2010

	PER	LOC	ORG	TOTAL
DB Spotlight	0.97 / 0.40	0.82 / 0.46	0.86 / 0.31	0.85 / 0.39
Zemanta	0.96 / 0.84	0.89 / 0.62	0.82 / 0.57	**0.90** / 0.68
LODIE	0.81 / 0.82	0.73 / 0.76	0.56 / 0.59	0.71 / **0.74**
Zemanta ∩ LODIE	1.00 / 0.74	0.95 / 0.45	0.97 / 0.42	**0.97** / 0.54
Zemanta U LODIE	0.94 / 0.93	0.77 / 0.76	0.72 / 0.71	0.82 / **0.81**

Precision and Recall figures shown above

Similar results on a set of metadata records and scientific papers in environmental science (EnviLOD)

Candidate ambiguity is high = tough task

	TAC-KBP				
	PER	LOC	ORG	UKN	TOTAL
Entities	89	361	141	274	865
Avg. number of tokens	1.91	1.20	2.12	1.87	1.78
Candidate URIs	9,427	9,553	9,502	14,649	43,131
Avg. number cand. URIs	105.02	26.46	67.39	53.46	49.86
Unambig. candidates	3	10	3	43	59

Outline of the Lecture

- What is Semantic Search? Why is it Useful?
- How does it Work?
  - Semantic Annotation
  - Semantic Search
- How do we get the users to use it?
  - Faceted entity search
  - Form-based semantic constraints
  - Natural language queries
- Does it work? Peter Mika on Thursday 8:30am

Semantic Search: An Overview

MIMIR: Searching Text Mining Results

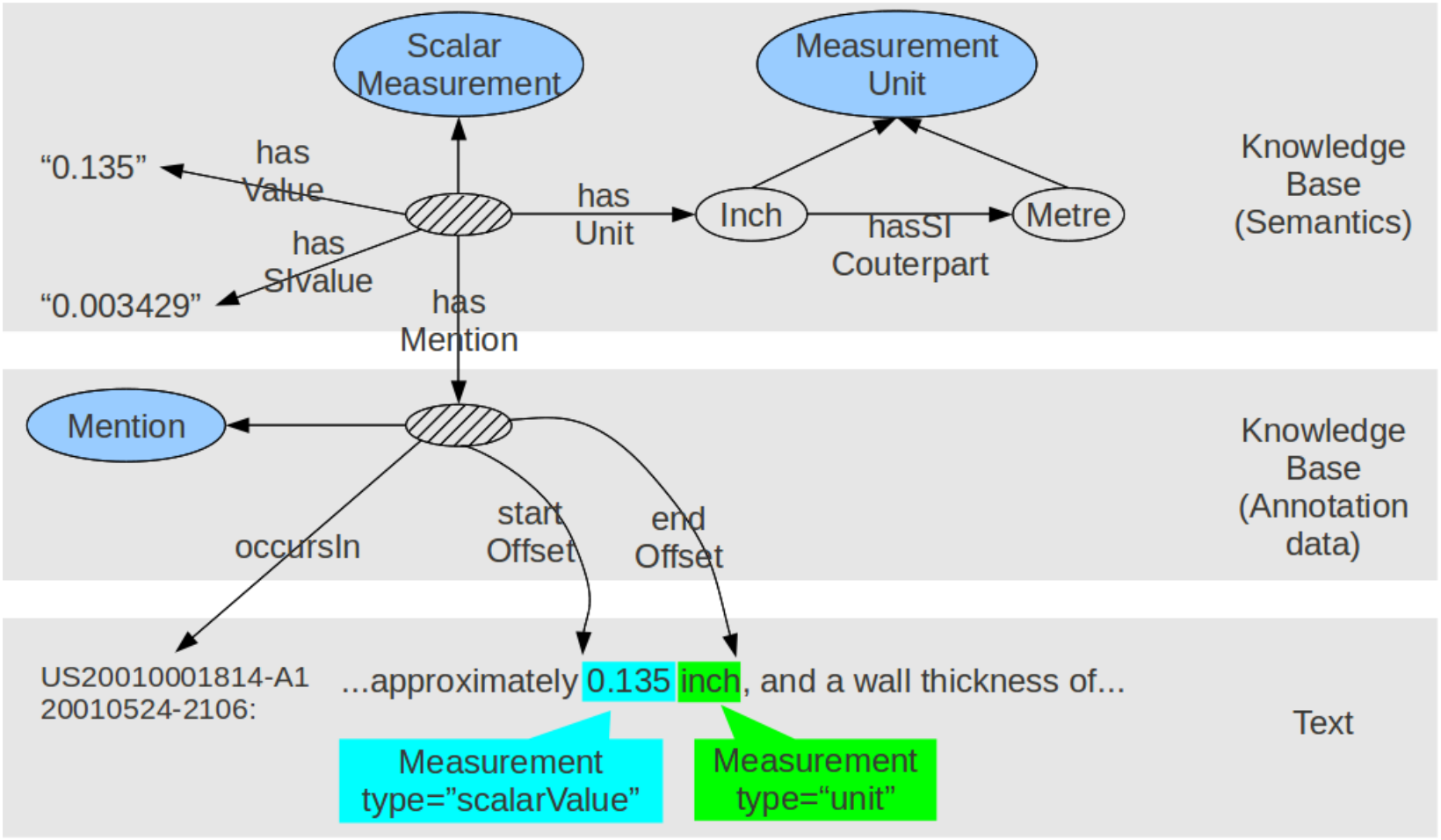
Searching and managing text annotations, semantic information, and full text documents in one search engine

Queries over annotation graphs

Regular expressions, Kleene operators

Designed to be integrated as a web service in custom end-user systems with bespoke interfaces

Modelling Text Annotations and Additional Knowledge Sources



MIMIR: Under the hood

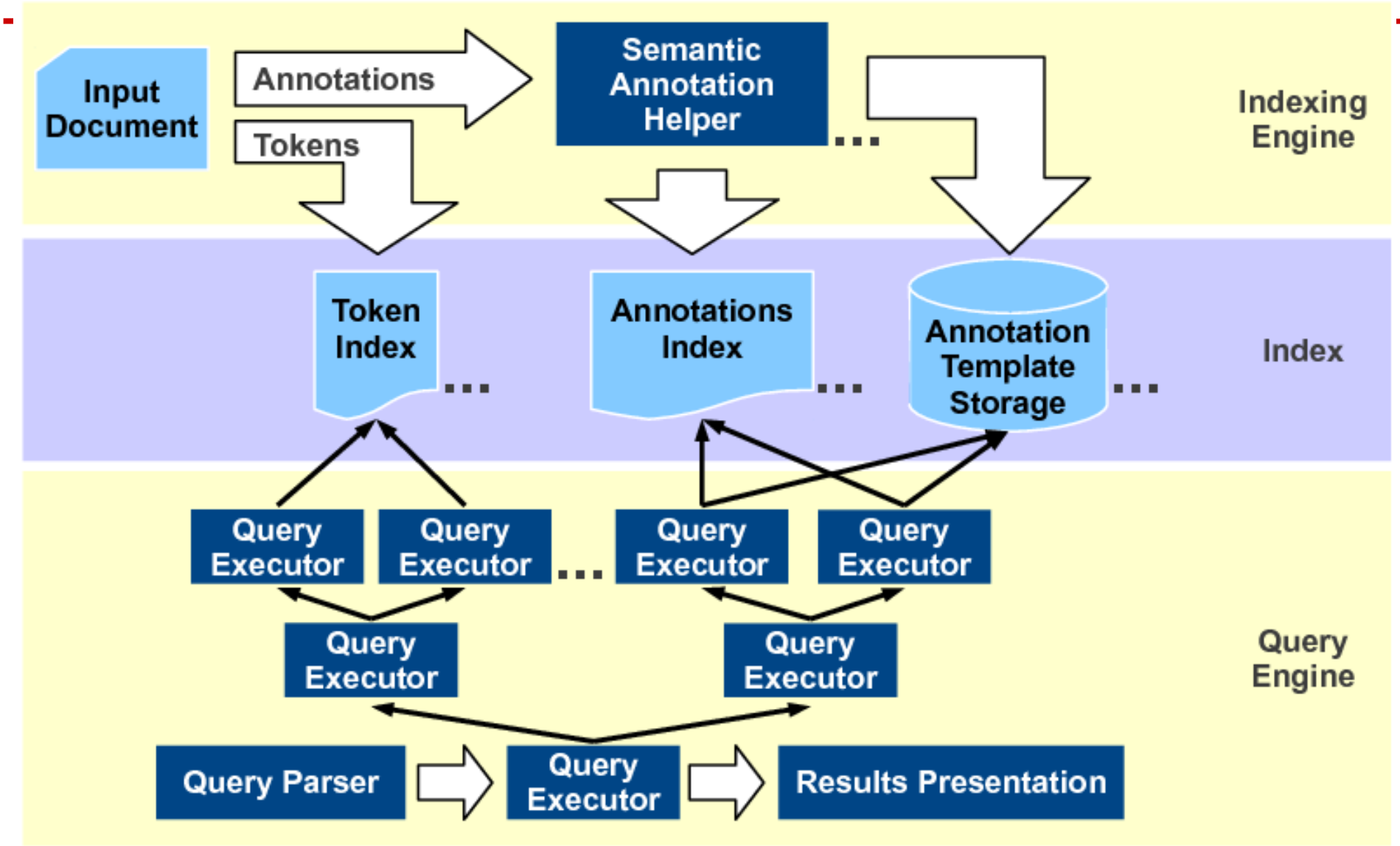
Uses a combination of an FTS engine (MG4J), a database for indexing the semantic annotations (H2), and a knowledge repository (OWLIM/Sesame)

Use database to indexing annotation kinds, then the additional linguistic data for words/annotations are stored in MG4J (POS, morphological root, etc)

Semantic query: against the database/ knowledge base for the additional knowledge

A query: decompose into the respective parts, query the appropriate component, then merge the results

Mimir


Input Document
Annotations
Tokens
Semantic Annotation Helper
Indexing Engine
Token Index
Annotations Index
Annotation Template Storage
Index
Query Executor
Query Executor
Query Executor
Query Executor
Query Executor
Query Executor
Query Engine
Query Parser
Query Executor
Results Presentation

Clustering and Scalability

All searches are local to a document, so we can use clusters of semantic repositories:

- Faster indexing (less data in each repository)
- Faster searching (search space is broken into slices that are searched in parallel – à la Google).
- Joining results is trivial: union of result sets.
- Simple scalability: just add more nodes.
- Search speed stays almost constant while the data increases (each individual repository has the same amount of data; there are just more repositories)

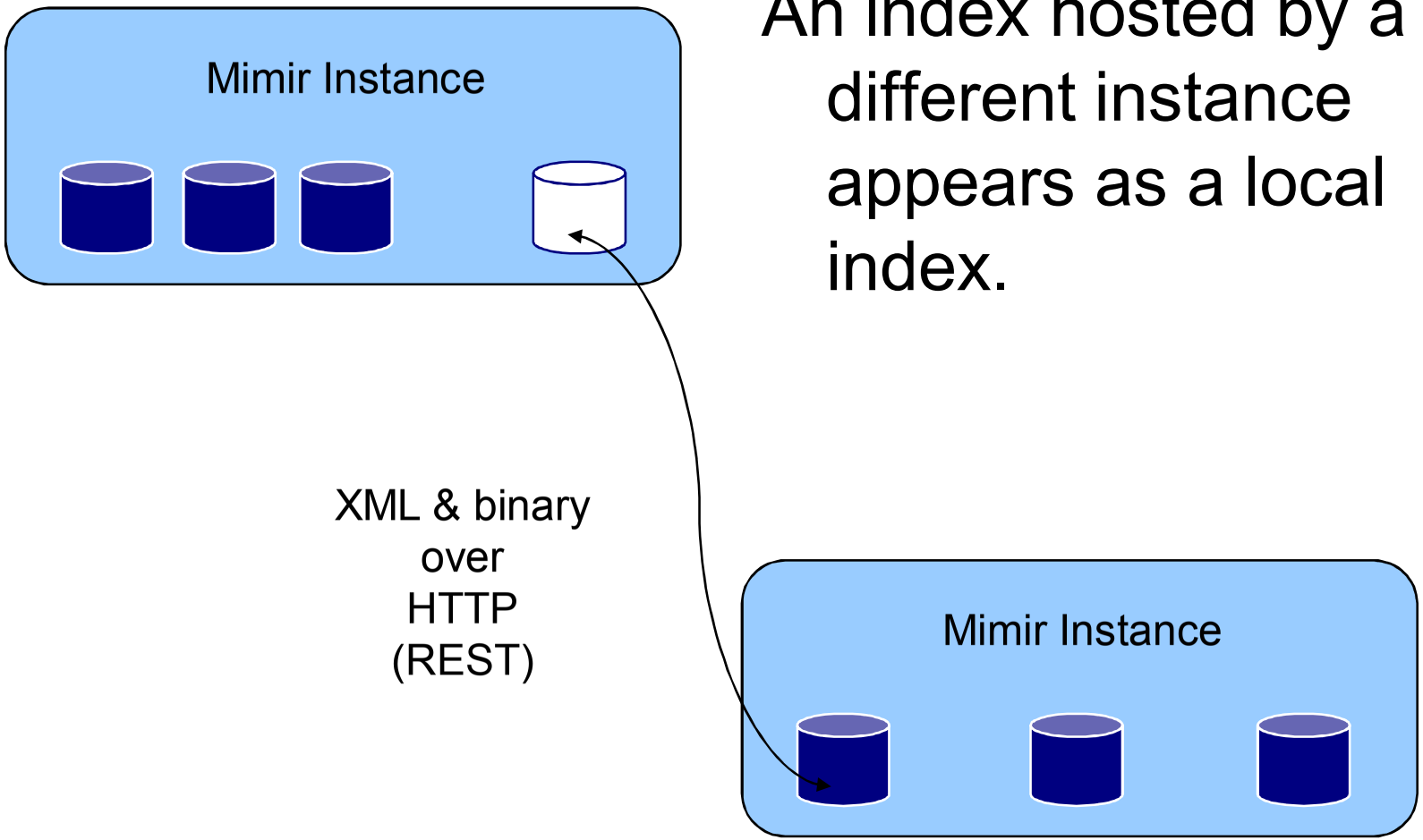
Ranking of Results

- By default documents are returned in index order
- The following ranking algorithms are also available (others can be implemented via plugins if required)
  - –Count
  - –Hit length
  - - TF.IDF
  - –BM25

Ranking (2)

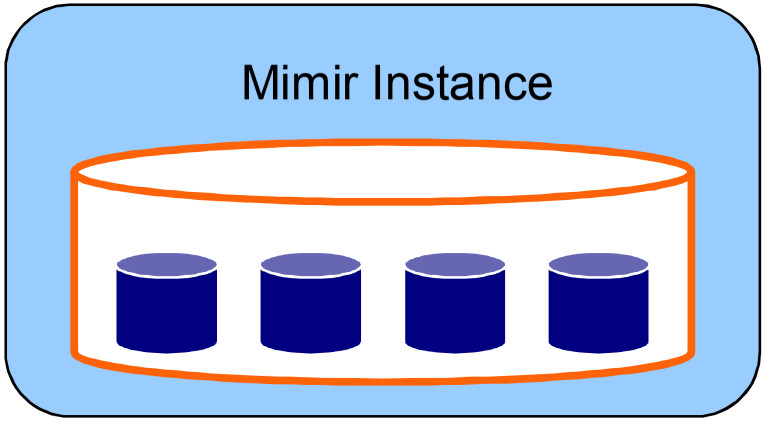
- These algorithms treat annotations within the queries in the same way as words, using index level frequencies etc.
- The ranking algorithm can be changed without re-indexing allowing for easy experimentation

Remote Indexes



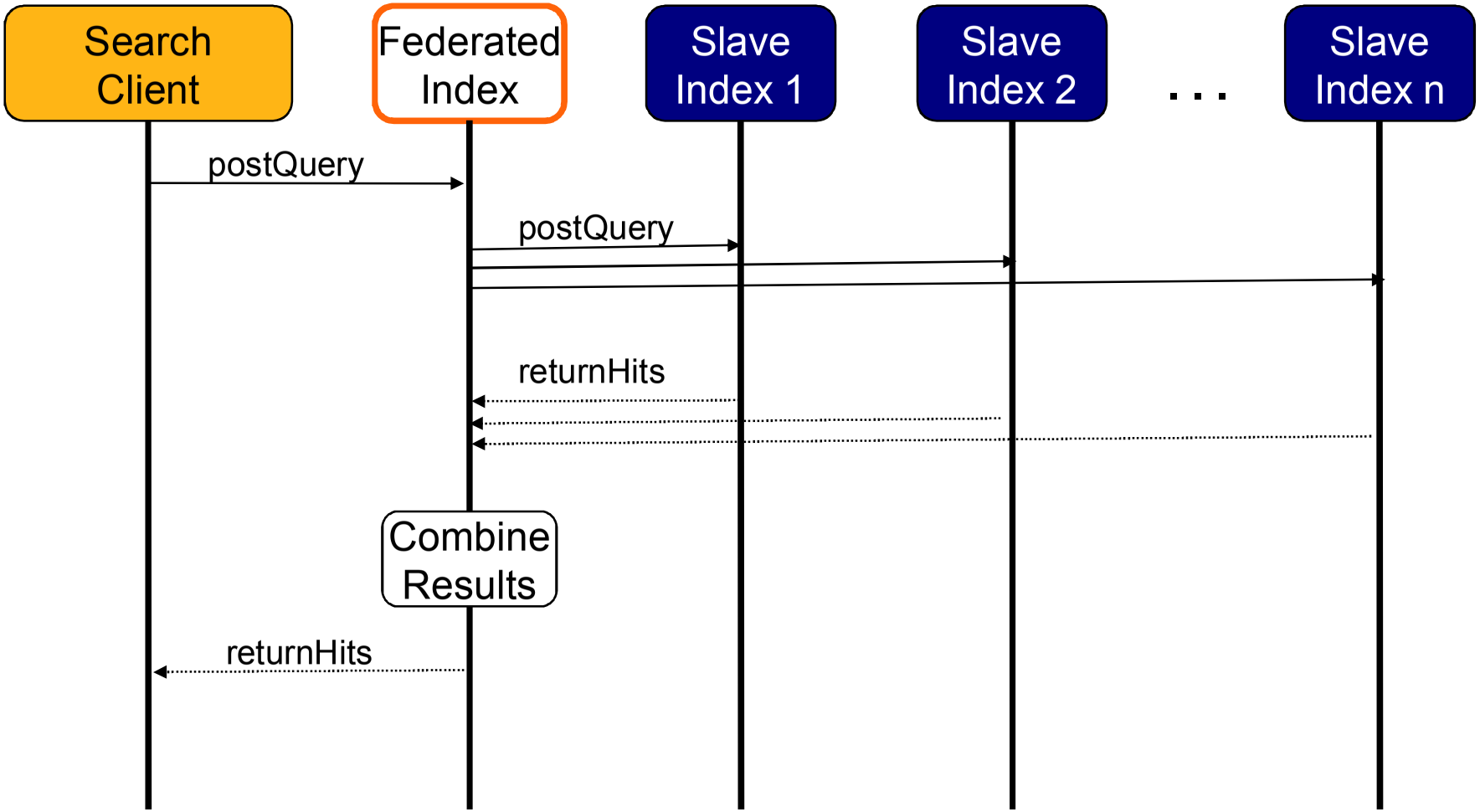
An index hosted by a different instance appears as a local index.

Federated Indexes



A set of indexes (of any kind) is grouped together into a federated index, exposing the normal functionality of a simple index.

Federated Searching



Building a Mimir Index

Mímir server instance:

- On the GATECloud.net
- Build your own from sources

Index template

A source of annotated GATE documents

A method for pushing documents to the index:

- Mímir Client PR, running inside Developer
- Mímir output handler in GATE Cloud Paralleliser
- Mímir output handler on GATECloud.net

Mímir Index Template

Specifies which:

- Token features should be indexed
- (semantic) annotation types should be indexed
- features for each annotation type
- features of the document

… should be indexed.

Example Index Template


```
tokenASName = ""

tokenAnnotationType = ANNIEConstants.TOKEN_ANNOTATION_TYPE

tokenFeatures = {

  string()

  category()

  root()

}

semanticASName = "“

semanticAnnotations = {

  index {

    annotation helper:new DefaultHelper(annType:'Sentence')
```

Example Index Template (2)


```
    annotation helper:new DefaultHelper(annType:'Person', nominalFeatures:['gender'])

    annotation helper:new DefaultHelper(annType:'Location', nominalFeatures:['locType'])

    annotation helper:new DefaultHelper(annType:'Organization')

    annotation helper:new DefaultHelper(annType:'Percent')

    annotation helper:new DefaultHelper(annType:'Money')

    annotation helper:new DefaultHelper(annType:'Date', nominalFeatures:['kind'])

    annotation helper:new DefaultHelper(annType:'Address', nominalFeatures:['kind'])

  }

}
```

Scaling Up

- We annotated 1.08 million web pages using a GATE language analysis pipeline.
  - Documents crawled using Heritrix10 , with total content size of 57 GiB or 6.6 billions plain text characters.
  - The indexing server has 2 Intel Xeon 2.8GHz CPUs 11 GB of RAM, and runs 64 bit Ubuntu Linux. Indexing process was 94 hours.
- We also indexed 150 million similar web pages, using two hundred Amazon EC2 Large Instances running for a week to produce a federated index
- Mimir runs on GateCloud.net, so easy to scale up

University of Sheffield,
NLP

MIMIR Matches Sequences: 81 hits

Harriet Harman

Search

Documents 1 to 20 of 81:

UK childcare needs to be more affordable - CentreForum (cached)
quality' MP **Harriet Harman** was the architect

Birth weight among social mobility checks - Nick Clegg (cached)
's deputy leader **Harriet Harman** said Mr Clegg

Ed Miliband's shadow cabinet and ministerial teams (cached)
Miliband Opposition leader **Harriet Harman** Deputy Leader & in 2011. **HARRIET HARMAN** - DEPUTY LEADER

PM's response to Skinner Commons question 'shameful' (cached)
. Deputy leader **Harriet Harman** wrote on Twitter

Daily Politics and Sunday Politics highlights of 2012 (cached)
Sunday April 29 **Harriet Harman** on Hunt, Cameron and Clegg **Harriet Harman** struggles with bank on health by **Harriet Harman** PMQs: Harriet ...

Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance' (cached)
Shadow Culture Secretary **Harriet Harman** says Jeremy Hunt culture secretary, **Harriet Harman**, told the

University of Sheffield,
NLP

Harriet Harman says: 29 hits

Harriet Harman says

Search

Documents 1 to 20 of 29:

Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance' (cached)
Shadow Culture Secretary **Harriet Harman says** Jeremy Hunt was

Ed Miliband defends Iraq war condemnation (cached)
, says Harman **Harriet Harman says** Labour will be

Ed Miliband tells Labour: We're the optimists now (cached)
, says Harman **Harriet Harman says** Labour will be

Labour must have credible deficit plan, says Darling (cached)
, says Harman **Harriet Harman says** Labour will be

David Miliband says he won't join brother Ed's team (cached)
, says Harman **Harriet Harman says** Labour will be

Balls: Labour must fight cuts 'every inch of the way' (cached)
, says Harman **Harriet Harman says** Labour will be

University of Sheffield,
NLP

Harriet Harman root:say : 38 hits

Harriet Harman root:say

Search

Documents 21 to 18 of 38:

David Cameron criticised for 'calm down dear' jibe (cached)
former equality minister **Harriet Harman said** Mr Cameron's

Queen's Speech: Biggest change to voter registration (cached)
, Labour's **Harriet Harman said** the government was

Harriet Harman struggles with bank bonus and job figures (cached)
in Coventry, **Harriet Harman said**: "I

PMQs: Harriet Harman and Nick Clegg on unemployment (cached)
Labour, but **Harriet Harman said** unemployment was falling

Leveson Inquiry: Jeremy Hunt fair on BSkyB, says top civil servant (cached)
Shadow culture secretary **Harriet Harman said**: "David

Jeremy Hunt: I followed due process over BSkyB (cached)
But Labour's **Harriet Harman said** Mr Hunt had

Ed Miliband 'will marry' but politics 'got in the way' (cached)
, says Harman **Harriet Harman says** Labour will be

University of Sheffield,
NLP

{Person} root:say – 3980 hits

{Person} root:say

Search

Documents 1 to 20 of 3980:

Apple's Sir Jonathan Ive reaffirms desire to stay at company (cached)
Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

Diamond Jubilee Tube train was faulty (cached)
be happening' **Ms Siggs said**: "It

Warning over deep-ocean stowaways (cached)
embarrassment. But **Dr Voight says** the experience is it," **Dr Voight said**. "We

School building system not fit for purpose, review says (cached)
shadow education secretary **Andy Burnham said** Mr Gove had . General secretary **Chris Keates said** the capital budget BCSE) director **Ty Goddard said** there was "

EU wants Greece to stay in eurozone, says Van Rompuy (cached)
's Europe editor **Gavin Hewitt says** the crisis gives UK Prime Minister **David Cameron said** "there was . German Chancellor **Angela Merkel said** the bonds,

Huhne partner loses privacy case (cached)
000 costs. **Ms Trimingham said** this could become Daily Mail's **Andrew Pierce said** it was a Speaking outside court **Ms Trimingham said** she was disappointed

University of Sheffield,
NLP

{Person} AND root:say – 11803 hits

Stone Roses reunion gig hailed by fans (cached)
by fans By **Ian Youngs Entertainment reporter, BBC News The gi ... re. "They've never played so well together," said** 43-year-old together," **said 43-year-old Andrew Rudder**, from Ashton 43-year-old **Andrew Rudder, from Ashton under Lyne. But opinion ... s voice. "He can't sing but he never could," said** Tom Six, **...**

000011_http://www.bbc.co.uk/news/northern_ireland/ (cached)
02:07 **Michaela McAreavey trial - the first day Watch 02:51 Minister says** she has ' :51 Minister **says she has 'heard enough' Watch 01:12 Flat bombs find - man arrested Watch 01:39 Michaela**'s husband braves

Warning over deep-ocean stowaways (cached)
using the famous **Alvin sub say** the vehicle picked famous Alvin sub **say the vehicle picked up limpets from a depth of ... s had to cope with huge pressure changes as Alvin** conducted its dives pressure changes as **Alvin conducted its dives. The researchers report ... matter of some embarrassment. But Dr Voight says** the experience is **...**

In pictures: Royal arts gathering (cached)
Queen. Sir **Paul McCartney, who was among the musicians the event, said** he was " the event, **said he was "a big fan" of the monarch. Artist David Hockney** shared a few

School building system not fit for purpose, review says (cached)
purpose, review **says Accessibility links Skip to content Skip to l ... ucation & Family Home World UK England N. Ireland** Scotland Wales Business World UK England **N. Ireland Scotland Wales Business Politics Health ... building system not fit for purpose, review says** Some schools awaiting purpose, review **says Some schools awaiting rebuilds rely on tempor ... government-commissioned review by Sebastian James** of Dixons Group **...**

Hewlett-Packard to cut 27,000 jobs by the end of 2014 (cached)
World UK England **N. Ireland Scotland Wales Business Politics Health ... cut 27,000 jobs by end of 2014. The company said** the cuts - . The company **said the cuts - about 8% of its workforce - will r ... riod fell 3% on a year ago to $30.7bn. Meg Whitman**, HP's year. HP **said in a statement that the money would be reinve ... riod fell 3% on a year ago to $30.7bn. Meg Whitman**, HP's **...**

{Person} [0..5] root: say – 5495 hits

Documents 1 to 20 of 5495:

Apple's Sir Jonathan Ive reaffirms desire to stay at company (cached)
Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

Diamond Jubilee Tube train was faulty (cached)
be happening' **Ms Siggs said**: "It

Warning over deep-ocean stowaways (cached)
using the famous **Alvin sub say** the vehicle picked embarrassment. But **Dr Voight says** the experience is it," **Dr Voight said**. "We

School building system not fit for purpose, review says (cached)
government-commissioned review by **Sebastian James of Dixons Group said** value for money by Education Secretary **Michael Gove, Mr James said**. Schools with shadow education secretary **Andy Burnham said** Mr Gove had **...**

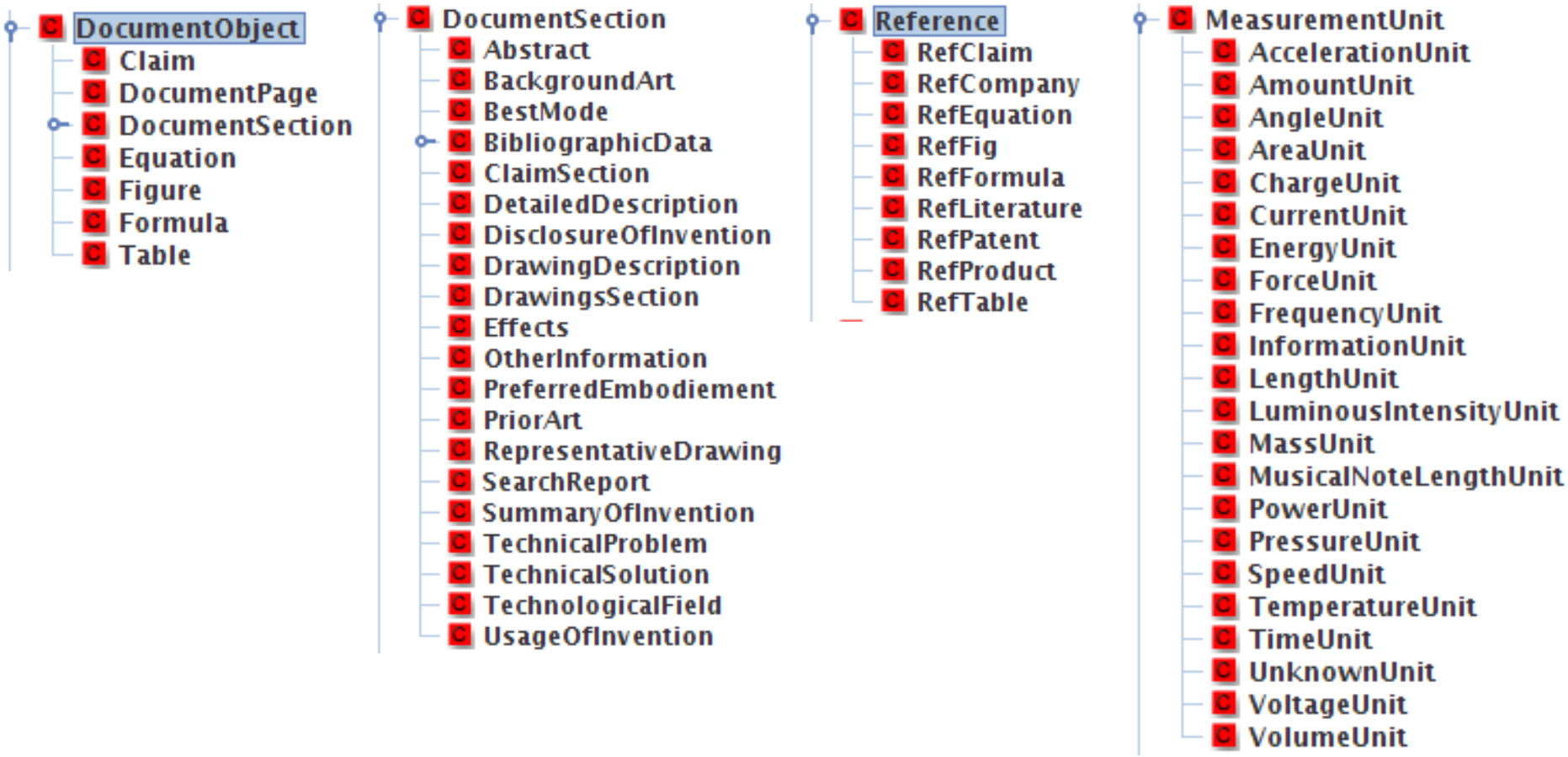
EU wants Greece to stay in eurozone, says Van Rompuy (cached)
, European Council **President Herman Van Rompuy has said**. He was European Council President **Herman Van Rompuy has said**. He was 's Europe editor **Gavin Hewitt says** the crisis gives **...**

Huhne partner loses privacy case (cached)
000 costs. **Ms Trimingham said** this could become Daily Mail's **Andrew Pierce said** it was a Speaking outside court **Ms Trimingham said** she was disappointed **...**

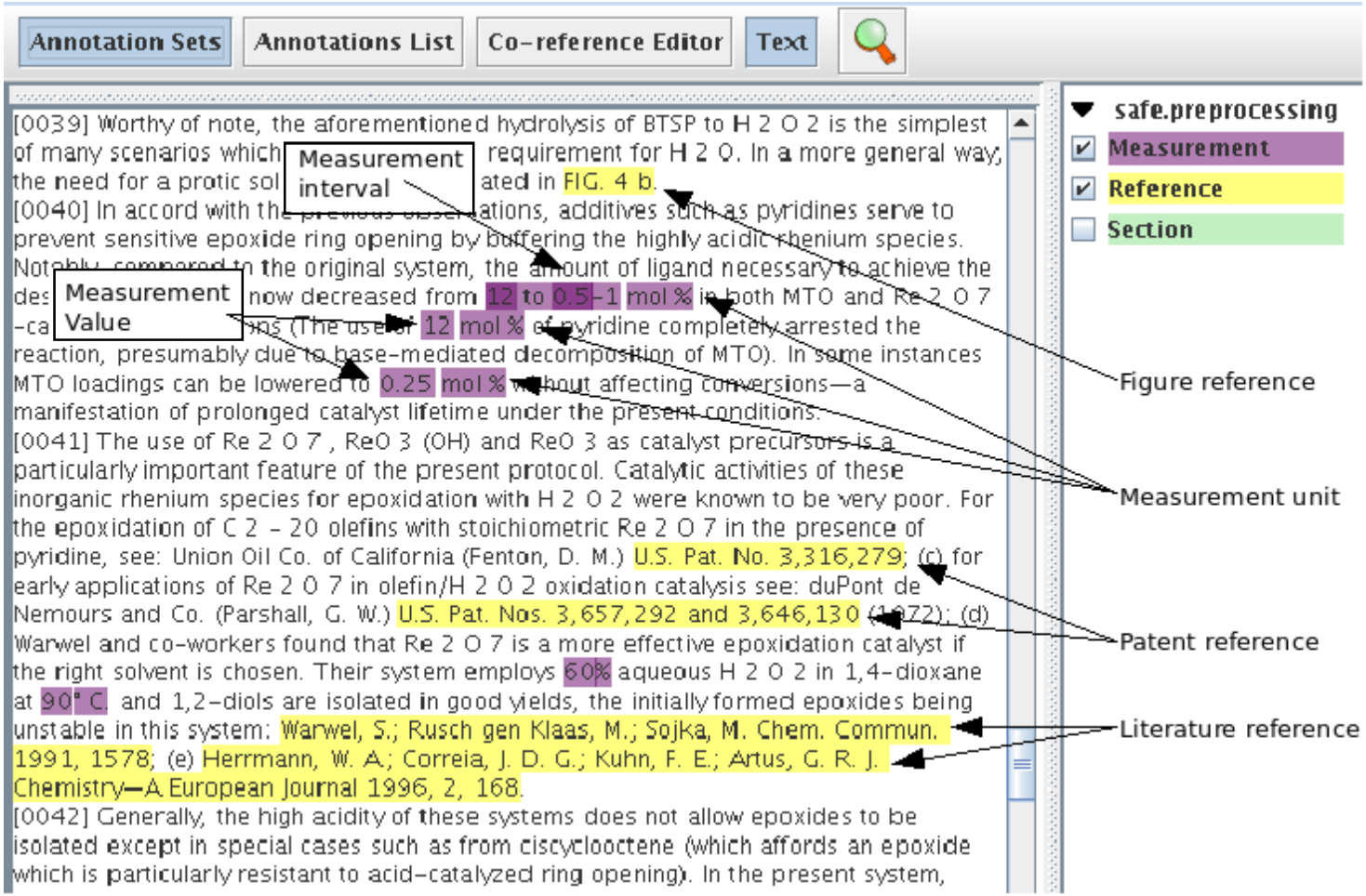
Migration to UK more than double government target (cached)
. Immigration Minister **Damian Green said**: "Our figures. Chairman **Sir Andrew Green said**: "You

No 'inappropriate' government contact, News Corp lobbyist tells Leveson (cached)
adviser. Fred **Michel said** he did not 's team. **Adam Smith stood down after saying** his e-mails with inquiry, Mr **Michel says** he did not **...**

Patent Annotation Data Model



An Example Text



Hands-On MIMIR

http://demos.gate.ac.uk/mimir/patents/search/index

Text. Matches plain text.

Example: nanomaterial

Linguistic variations of text

Example: (root:nanomaterial | root:nanoparticle)

Annotation. Matches semantic annotations.

Syntax: {Type feature1=value1 feature2=value2...}

Example: {Abstract lang="DE"}

Sequence Query. Sequence of other queries.

Syntax: Query1 [n..m] Query2...

Example: from **{Measurement} [1..5] {Measurement}**

Inclusion Queries

IN Query. Hits of one query only if in hits of another.

Syntax: Query1 IN Query2

Example: (root:nanomaterial | root:nanoparticle) IN {Abstract}

Number of times these words are mentioned in patent abstracts (as well as links to the actual documents)

OVER Query. Hits of a query, only if overlapping hits of another.

Syntax: Query1 OVER Query2

Example: {Abstract} OVER (root:nanomatrial | root:nanoparticle)

Finds all abstracts that contain nanomaterial(s) or nanoparticle(s)

Date restrictions


```
(
   {Abstract lang="EN"} OVER
      (root:nanomaterial | root:nanoparticle )
)
IN
{PatentDocument date > 20050000}
```

YYYYMMDD

Find references to literature or patents in the prior art or background sections, which contain nanomaterial/nanoparticle


```
({Reference type="Literature"}
   |
  {Reference type="Patent"}
) IN
  ({Section type="PriorArt"}
     |
    {Section type="BackgroundArt"}
  )
    OVER
    (root:nanomaterial | root:nanoparticle)
```

Queries Using External Knowledge

{Measurement spec="1 to 100 volts“}

Uses GNU Units (http://www.gnu.org/software/units/) to convert measurements and normalise them to ISI units

{Measurement spec="1 to 100 kg m^2 / A s^3"}

Example hits: 10 volts, 2V, +20 and -20 volts; ±10V; +/- 100V; +3.3 volts

Queries Using External Knowledge (2)

{Measurement spec="1 to 100 m / s"}

Example hits: 40 km/hr, 60m/min, 100cm/sec, 60 fps;
10 to 2000 cm/sec

Searching LOD with SPARQL

- SQL-like query language for RDF data
- Simple protocol for querying remote databases over HTTP
- Query types
  - *select* –projections of variables and expressions
  - *construct*–create triples (or graphs) based on query results
  - *ask*–whether a query returns results (result is true/false)
  - *describe*– describes resources in the graph

SPARQL Example


```
# Software companies founded in the US
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbp-ont: <http://dbpedia.org/ontology/>
PREFIX geo-ont: <http://www.geonames.org/ontology#>
PREFIX umbel-sc: <http://umbel.org/umbel/sc/>

SELECT DISTINCT ?Company ?Location
WHERE {
   ?Company rdf:type dbp-ont:Company ;
            dbp-ont:industry dbpedia:Computer_software ;
            dbp-ont:foundationPlace ?Location .
   ?Location geo-ont:parentFeature dbpedia:United_States .
}
```

SPARQL Results

FactForge RDF Search and E

SPARQL Query

Results for your query (81) - Edit query

View as Exhibit

Company	Location
dbpedia:Redxpress	dbpedia:Glen_Allen,_Missouri
dbpedia:Borland	dbpedia:California
dbpedia:Lawson_Software	dbpedia:Minneapolis
dbpedia:Exterro,_Inc.	dbpedia:Oregon
dbpedia:Tableau_Software	dbpedia:Seattle
dbpedia:NeuroDimension	dbpedia:Florida
dbpedia:Computer_Usage_Company	dbpedia:New_York_City
dbpedia:Macromedia	dbpedia:San_Francisco
dbpedia:Core_International,_Inc	dbpedia:Florida
dbpedia:Cerulean_Studios	dbpedia:Connecticut
dbpedia:Cornerstone_OnDemand	dbpedia:California

- Try at: http://factforge.net/sparql

Documents mentioning Persons born in Sheffield

{Person sparql = "SELECT ?inst WHERE { ?inst :birthPlace <http://dbpedia.org/resource/Sheffield>}"}

Ed Miliband's shadow cabinet and ministerial teams (cached)
and employment minister **David Blunkett** before becoming an

The BBC web document does not mention Sheffield at all:
http://www.bbc.co.uk/news/uk-politics-11494915

The relevant text snippet is:

HILARY BENN - SHADOW COMMUNITIES AND LOCAL GOVERNMENT SECRETARY

As the son of former Labour cabinet minister Tony Benn, the MP for Leeds Central is part of a political dynasty. Regarded as more pragmatic than his father, he was a union official and special adviser to then education and employment minister David Blunkett before becoming an MP in 1999. Well-regarded as international development and environment secretary under Gordon Brown despite having a generally low profile. At the **age of 53** stood for the deputy leadership in 2007, coming fourth. One of Ed Miliband's primary supporters in the leadership contest.

Outline of the Lecture

- What is Semantic Search? Why is it Useful?
- How does it Work?
  - Semantic Annotation
  - Semantic Search
- How do we get the users to use it?
  - Faceted entity search
  - Form-based semantic constraints
  - Natural language queries
- Does it work? Peter Mika on Thursday 8:30am

User Interfaces for SPARQL-based Semantic Search

- SPARQL-based semantic searches, tapping into LOD resources are extremely powerful
- However, impossible to write by the vast majority of users
- User interfaces for SPARQL-based semantic search:
  - Faceted searches (see ExoPatent next)
  - Form-based searches (see EnviLOD)
  - Text-based searches (natural language interfaces for querying ontologies), e.g. FREyA

Faceted Search: ExoPatent Example

http://exopatent.ontotext.com

Use semantic information to expose linkages between douments based on the intersecting relationships between various sets of data from

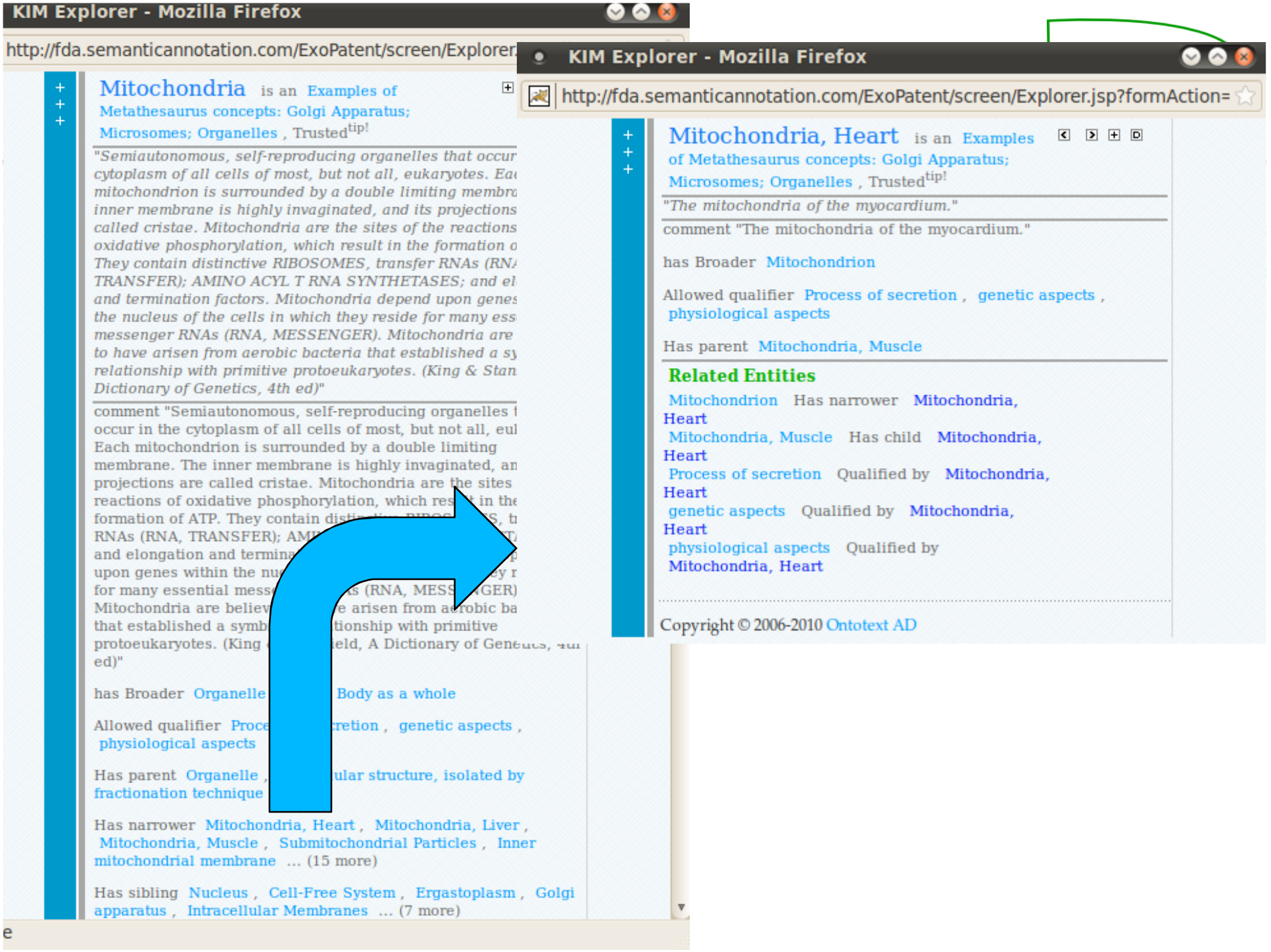
- FDA Orange Book (23,000 patented drugs)
- Unified Medical Language System (UMLS) – database of medical terms (370,000)
- Patent bibliographic information

Search for diseases, drug names, body parts, references to literature and other patents, numeric values, ranges

Demo uses a small set of patents (40,000)

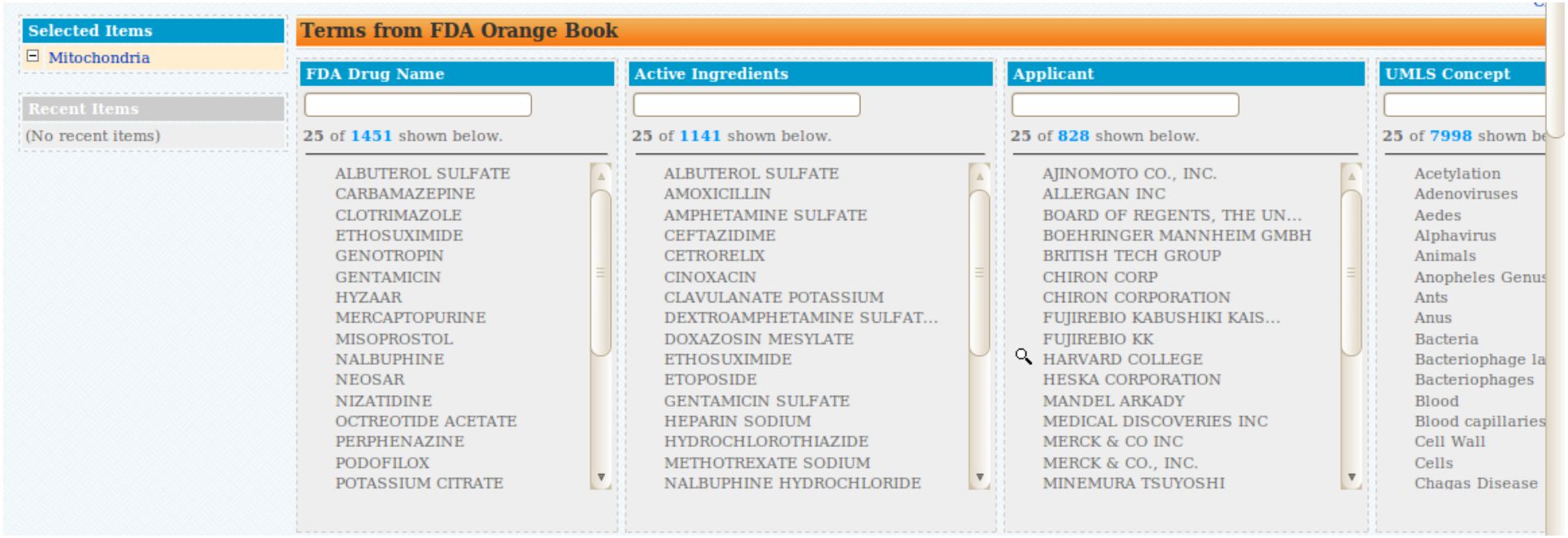
University of Sheffield,
NLP

ExoPatent: Faceted Search


KIM Explorer - Mozilla Firefox
http://fda.semanticannotation.com/ExoPatent/screen/Explorer
Mitochondria is an Examples of Metathesaurus concepts: Golgi Apparatus; Microsomes; Organelles , Trusted tip!
"Semiautonomous, self-reproducing organelles that occur
cytoplasm of all cells of most, but not all, eukaryotes. Ea
mitochondrion is surrounded by a double limiting membra
inner membrane is highly invaginated, and its projections
called cristae. Mitochondria are the sites of the reactions
oxidative phosphorylation, which result in the formation o
They contain distinctive RIBOSOMES, transfer RNAs (RNA
TRANSFER); AMINO ACYL T RNA SYNTHETASES; and el
and termination factors. Mitochondria depend upon genes
the nucleus of the cells in which they reside for many ess
messenger RNAs (RNA, MESSENGER). Mitochondria are
to have arisen from aerobic bacteria that established a sy
relationship with primitive protoeukaryotes. (King & Stan
Dictionary of Genetics, 4th ed)"
comment "Semiautonomous, self-reproducing organelles
occur in the cytoplasm of all cells of most, but not all, eul
Each mitochondrion is surrounded by a double limiting
membrane. The inner membrane is highly invaginated, an
projections are called cristae. Mitochondria are the sites
reactions of oxidative phosphorylation, which res in the
formation of ATP. They contain distinctive RIBOS, t
RNAs (RNA, TRANSFER); AMI TA
and elongation and termina
upon genes within the nu ey r
for many essential mess s (RNA, MESS GER)
Mitochondria are believ e arisen from aerobic ba
that established a symb tionship with primitive
protoeukaryotes. (King ield, A Dictionary of Genetics, 4th
ed)"
has Broader Organelle Body as a whole
Allowed qualifier Proce cretion , genetic aspects , physiological aspects
Has parent Organelle , ular structure, isolated by fractionation technique
Has narrower Mitochondria, Heart , Mitochondria, Liver , Mitochondria, Muscle , Submitochondrial Particles , Inner mitochondrial membrane ... (15 more)
Has sibling Nucleus , Cell-Free System , Ergastoplasm , Golgi apparatus , Intracellular Membranes ... (7 more)
KIM Explorer - Mozilla Firefox
http://fda.semanticannotation.com/ExoPatent/screen/Explorer.jsp?formAction=
Mitochondria, Heart is an Examples of Metathesaurus concepts: Golgi Apparatus; Microsomes; Organelles , Trusted tip!
"The mitochondria of the myocardium."
comment "The mitochondria of the myocardium."
has Broader Mitochondrion
Allowed qualifier Process of secretion , genetic aspects , physiological aspects
Has parent Mitochondria, Muscle
Related Entities
Mitochondrion Has narrower Mitochondria, Heart
Mitochondria, Muscle Has child Mitochondria, Heart
Process of secretion Qualified by Mitochondria, Heart
genetic aspects Qualified by Mitochondria, Heart
physiological aspects Qualified by Mitochondria, Heart
Copyright © 2006-2010 Ontotext AD

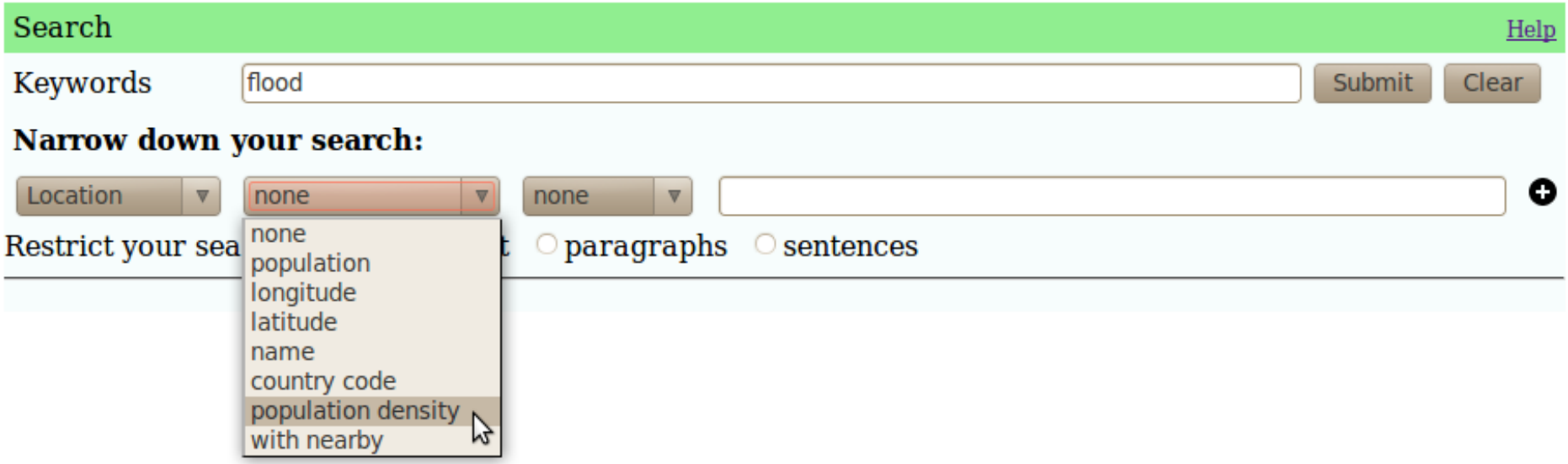
GATE

Find all applicants who filed patents related to mitochondria, as well as drug names and active ingredients



Semantic Search over Content and Annotations: Example

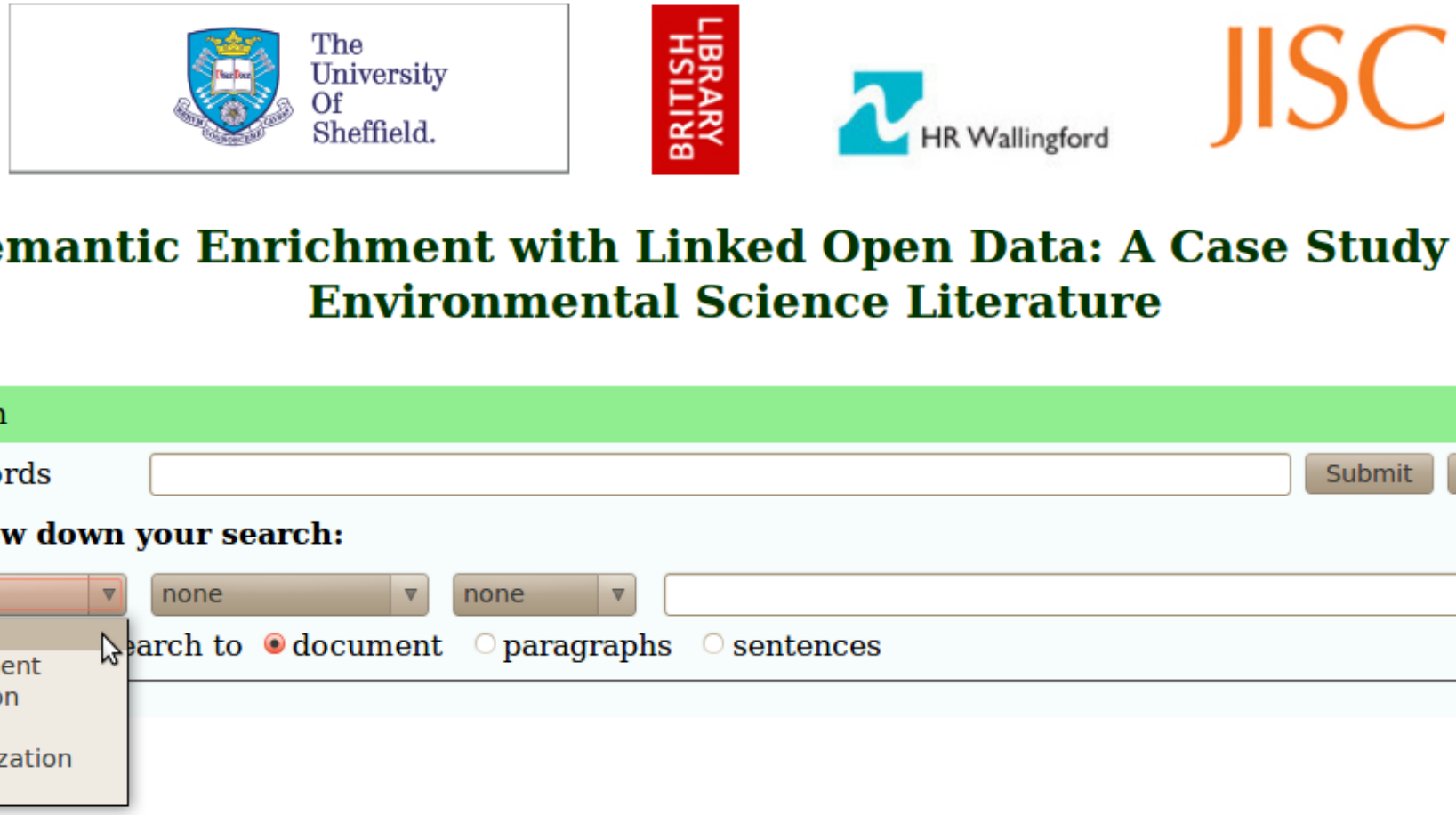
Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature



http://demos.gate.ac.uk/trendminer/envilod

University of Sheffield,
NLP

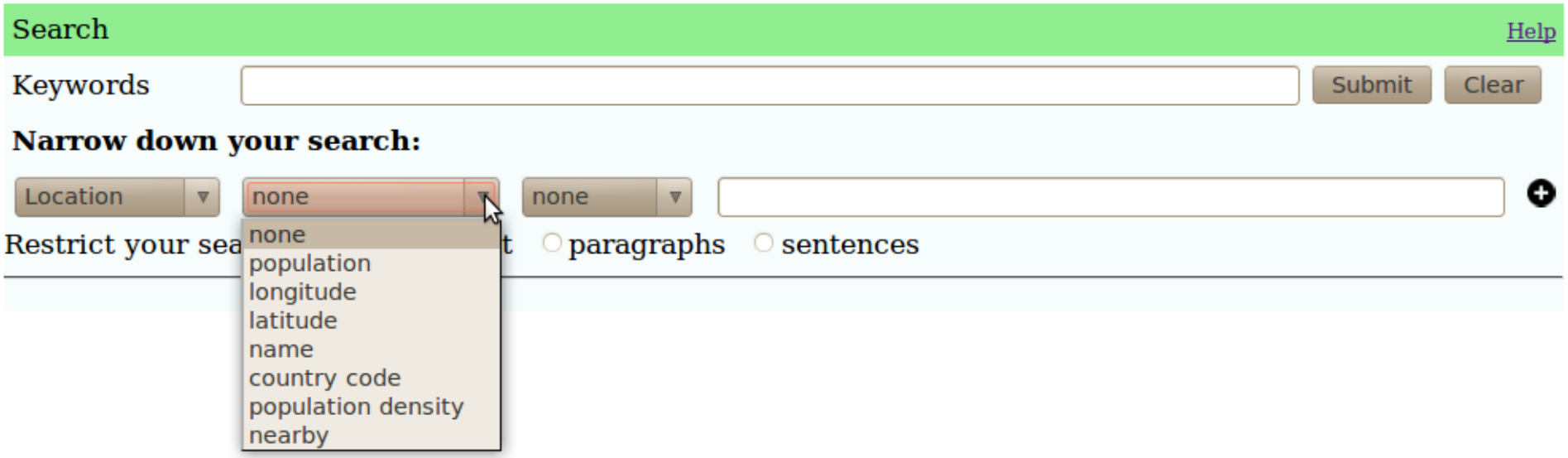
EnviLOD Semantic Search UI



University of Sheffield,
NLP

EnviLOD Semantic Search UI

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



University of Sheffield,
NLP

EnviLOD Semantic Search UI

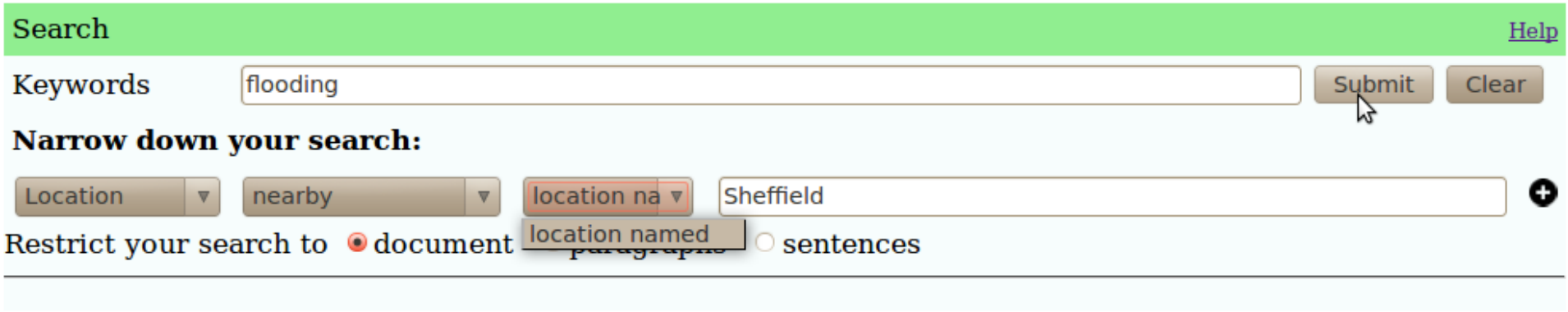
The University Of Sheffield.

BRITISH LIBRARY HSIRB

HR Wallingford

JISC

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



University of Sheffield,
NLP

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 wetlan 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 servi

(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

…and the underlying SPARQL Query


```
{Sem_Location dbpediaSparql="select distinct ?inst
where {
  {{ ?inst <http://dbpedia.org/property/north> ?loc} UNION
    { ?inst <http://dbpedia.org/property/east> ?loc } UNION
    { ?inst <http://dbpedia.org/property/west> ?loc } UNION
    { ?inst <http://dbpedia.org/property/south> ?loc } UNION
    { ?inst <http://dbpedia.org/property/northeast> ?loc } UNION
    { ?inst <http://dbpedia.org/property/northwest> ?loc } UNION
    { ?inst <http://dbpedia.org/property/southeast> ?loc } UNION
    { ?inst <http://dbpedia.org/property/southwest> ?loc }
FILTER(REGEX(STR(?loc), \"Sheffield\", \"i\"))} }"} AND (root:"flood")
```

Ongoing work: Use GeoSparql instead and be able to specify distances and reason with the richer information in GeoNames

University of Sheffield,
NLP

Flooding in Oxford

Keywords flood Submit Clear

Narrow down your search:

Location name contains Oxford

Restrict your search to document paragraphs sentences

Mimir Query:

Showing 1 to 3 of 3 hits. Pages: 1

The governments response to Sir Michael Pitt?s review of the summer 2007 Floods : progress report

Example hits:

..." fund early action to tackle flood risk. Applications were due by 30 November. Successful applic...olk 16. Northumberland 17. North Yorkshire 18. Nottinghamshire 19. Oxfordshire4 20. Somerset 21" ....." 4 Includes all authorities in Oxfordshire 5 Includes responses from all authorities in Suffolk...mation about this publication and copies are available from: Flood Management Division 2D Ergon" ...

Keywords: (other metadata)

Flooding in London : a London Assembly scrutiny report

Example hits:

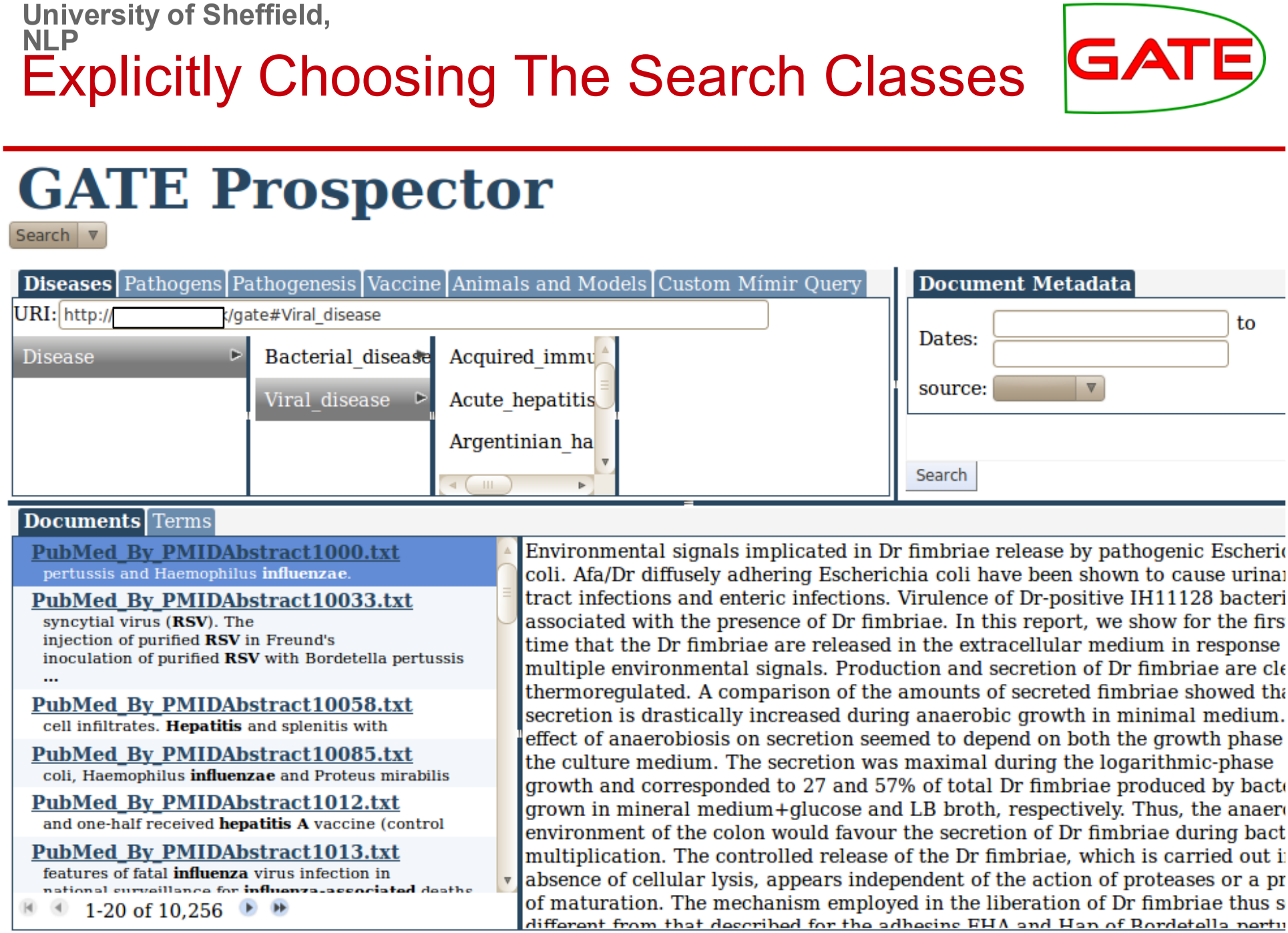
..." study in 2000 by the Flood Hazard Research Centre at Middlesex University of the longer-term health effects of the 1998 flooding in Banbury and Kidlington in the Thames" ....." 1998 flooding in Banbury and Kidlington in the Thames region. Studies in flood-affected areas r...ced by the engineering profession on the human distress caused by flooding – its social impact." ...

Keywords: Flood damage prevention--England--London, Floods--England--London, Flood control--England--London, Emergency management--England--

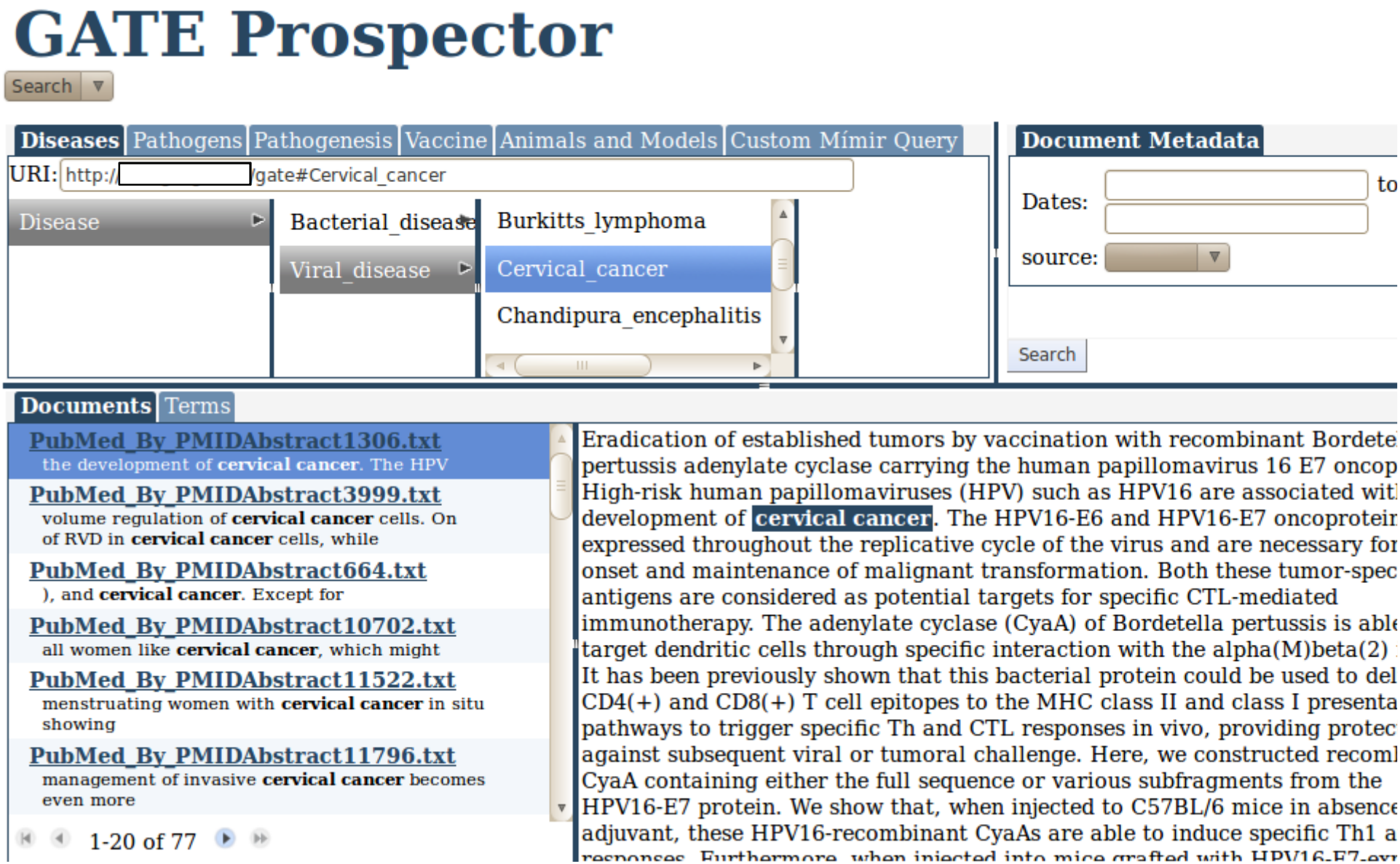
Some Caveats….

- EnviLOD demonstrator built in a 6 month project
- VERY limited environmental science content indexed
  - Some Defra, Environment Agency, Scottish Government report
- PDFs of the articles are not connected to

Explicitly Choosing The Search Classes

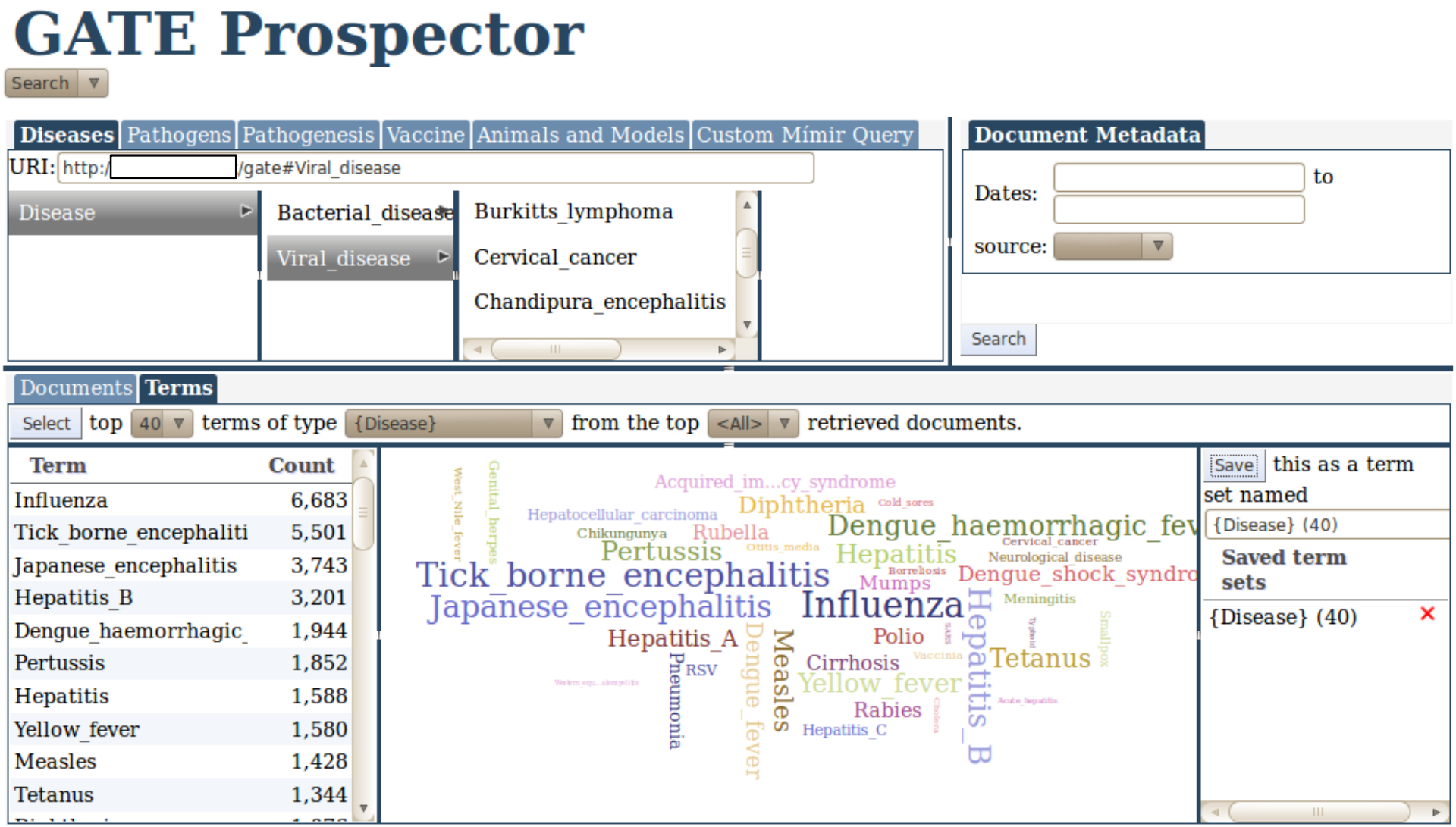


Choosing A Specific Instance



University of Sheffield,
NLP

What diseases are in these documents?



University of Sheffield,
NLP

What Pathogens?

GATE

GATE Prospector

Search

Diseases | **Pathogens** | Pathogenesis | Vaccine | Animals and Models | Custom Mímir Query

URI: http://fera.gsi.gov.uk/gate#Pathogen

Pathogen ▸ | Bacteria ▸ | Virus ▸

Document Metadata

Dates: to

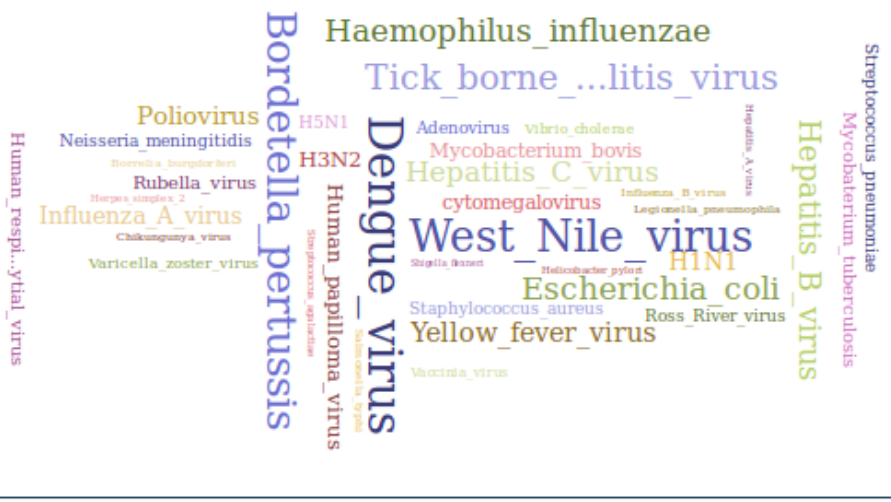
source:

Search

Documents | **Terms**

Select top 40 terms of type {Pathogen} from the top <All> retrieved documents.

Term	Count
Dengue_virus	17,867
West_Nile_virus	11,726
Bordetella_pertussis	9,909
Japanese_encephalitis_vi	5,093
Human_immunodeficien	4,939
Tick_borne_encephalitis	3,899
Haemophilus_influenzae	2,709
Escherichia_coli	2,342
Hepatitis_B_virus	2,043
Hepatitis_C_virus	1,567



Save this as a term set named

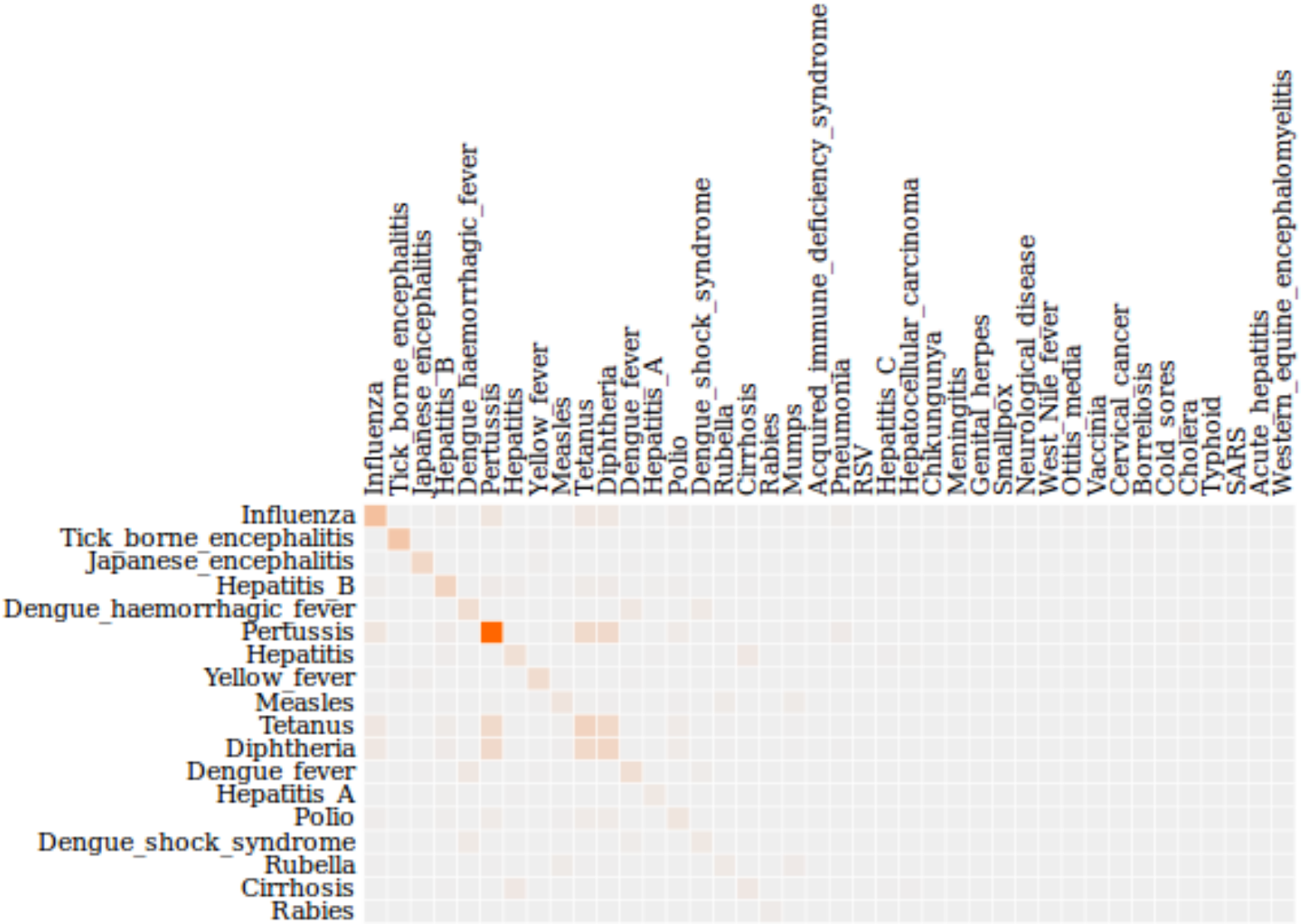
{Pathogen} (40)

Saved term sets

{Disease} (40) ×

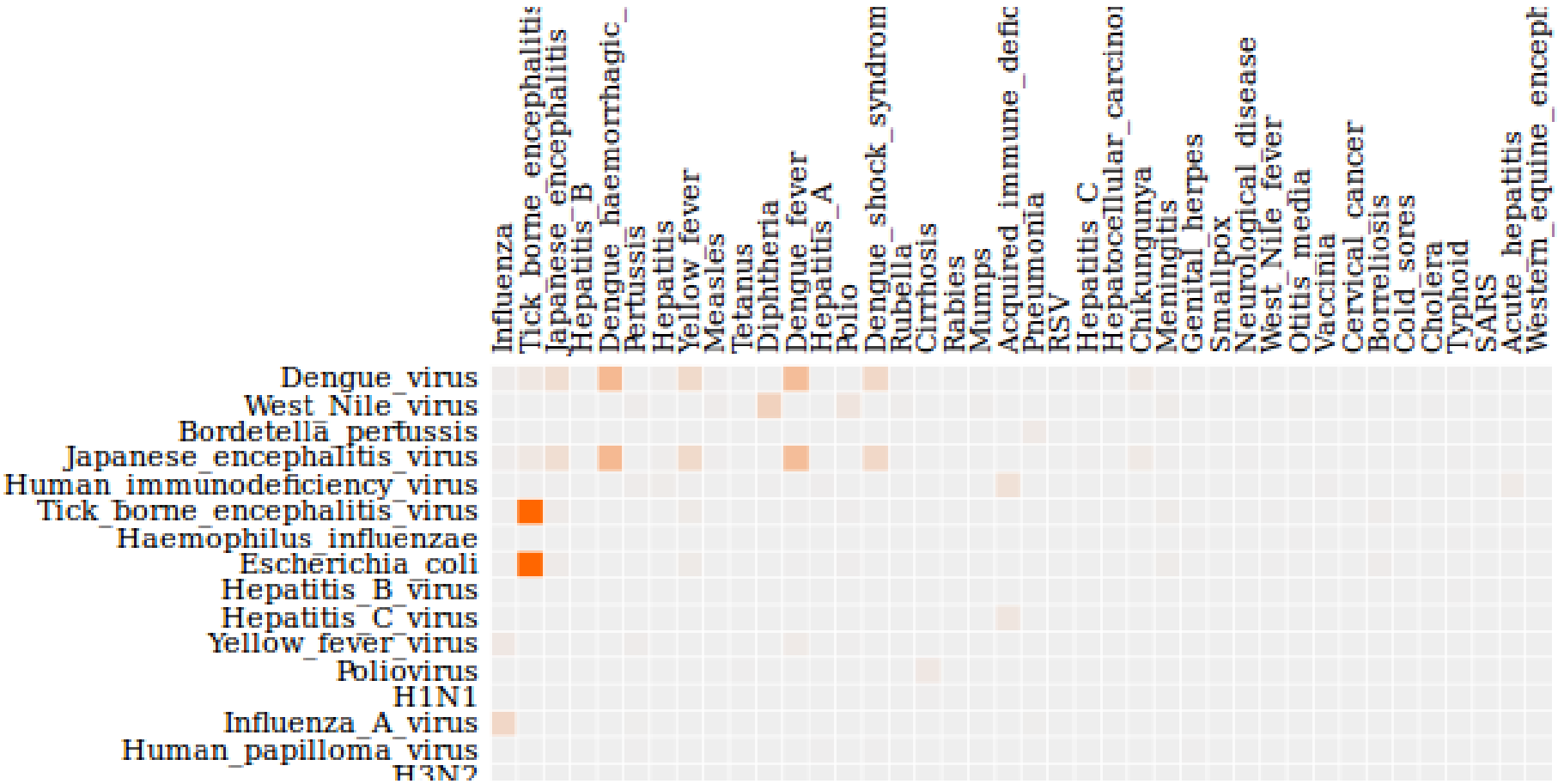
University of Sheffield,
NLP

Disease vs Disease Co-ocurrences



University of Sheffield,
NLP

Diseases vs Pathogens



Natural Language Query Interfaces

- SPARQL is complex to write, so instead let's use NL queries
  - *The Modigliani test: "tell me the locations of all the original paintings of Modigliani" (Richard MacManus, ReadWriteWeb)*
- For research on this topic see:

- Damljanovic, Agatonovic, Cunningham. Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction. Proc. Of the 7th Extended Semantic Web Conf. (ESWC). 2010.
- Wang, C., Xiong, M., Zhou, Q., Yu, Y.: Panto: A portable natural language interface to ontologies. In: The SemanticWeb: Research and Applications. pp. 473|487. Springer (2007)
- Tablan, V., Damljanovic, D., Bontcheva, K.: A natural language query interface to structured information. In: Proc. of the 5h European Semantic Web Conference (ESWC 2008). Lecture Notes in Computer Science, vol. 5021, pp. 361-375.
- Lopez, V., Uren, V., Motta, E., Pasin, M.: Aqualog: An ontology-driven question answering system for organizational semantic intranets. Web Semantics: 5(2), 72- 105. June 2007.
- Kaufmann, E., Bernstein, A., Fischer, L.: NLP-Reduce: A naive but domain independent natural language interface for querying ontologies. In: Proceedings of the European Semantic Web Conference ESWC 2007, Innsbruck, Austria.
- Kaufmann, E., Bernstein, A., Zumstein, R.: Querix: A natural language interface to query ontologies based on clarification dialogs. In: 5th International Semantic Web Conference (ISWC 2006). pp. 980 -981. 2006.

Passing the Modigliani test

Source: http://blog.larkc.eu/: "LDSR Passes the Modigliani Test for Semantic Web", more than 1h to generate a SPARQL query

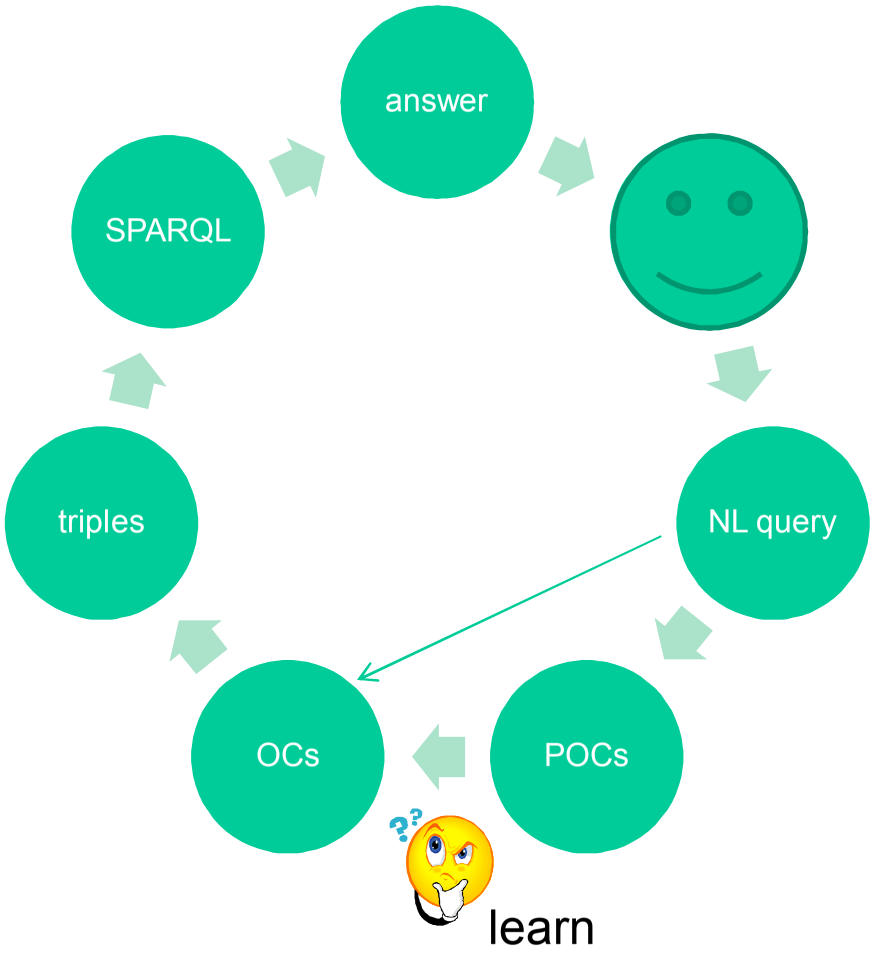
```
PREFIX fb: <http://rdf.freebase.com/ns/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbp-prop: <http://dbpedia.org/property/>
PREFIX dbp-ont: <http://dbpedia.org/ontology/>
PREFIX umbel-sc: <http://umbel.org/umbel/sc/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ot: <http://www.ontotext.com/>

SELECT DISTINCT ?painting_l ?owner_l ?city_fb_con ?city_db_loc
      ?city_db_cit
WHERE {
  ?p fb:visual_art.artwork.artist dbpedia:Amedeo_Modigliani ;
     fb:visual_art.artwork.owners [
      fb:visual_art.artwork_owner_relationship.owner ?ow ] ;
     ot:preferredLabel ?painting_l.
     ?ow ot:preferredLabel ?owner_l .
  OPTIONAL { ?ow fb:location.location.containedby [
      ot:preferredLabel ?city_fb_con ] } .
  OPTIONAL { ?ow dbp-prop:location ?loc. ?loc rdf:type umbel-
      sc:City ; ot:preferredLabel ?city_db_loc }
  OPTIONAL { ?ow dbp-ont:city [ ot:preferredLabel ?city_db_cit ] }
}
```

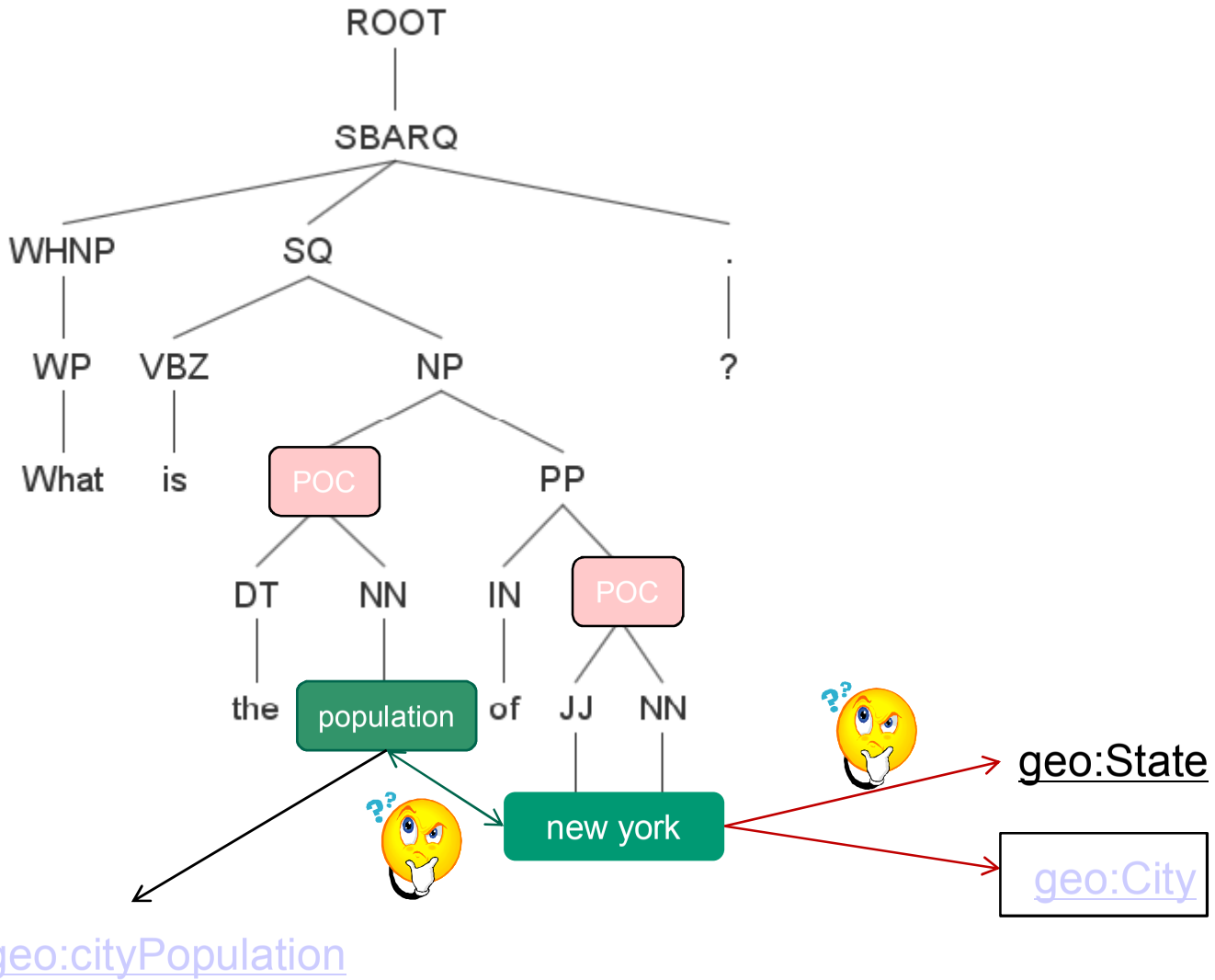
The FREyA System [Damljanovic 2010]

Potential Ontology Concept (POC)

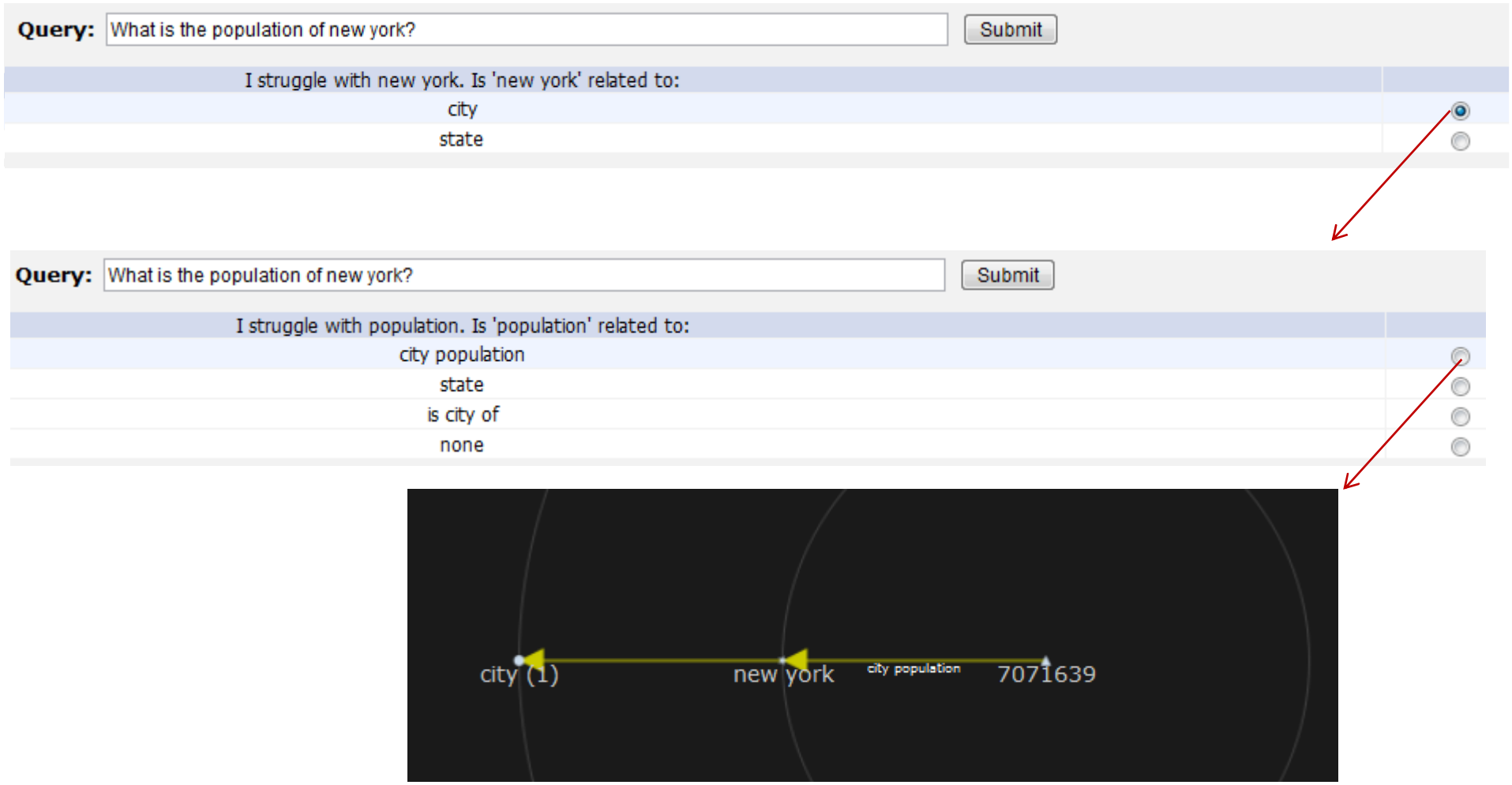
Ontology Concept (OC)



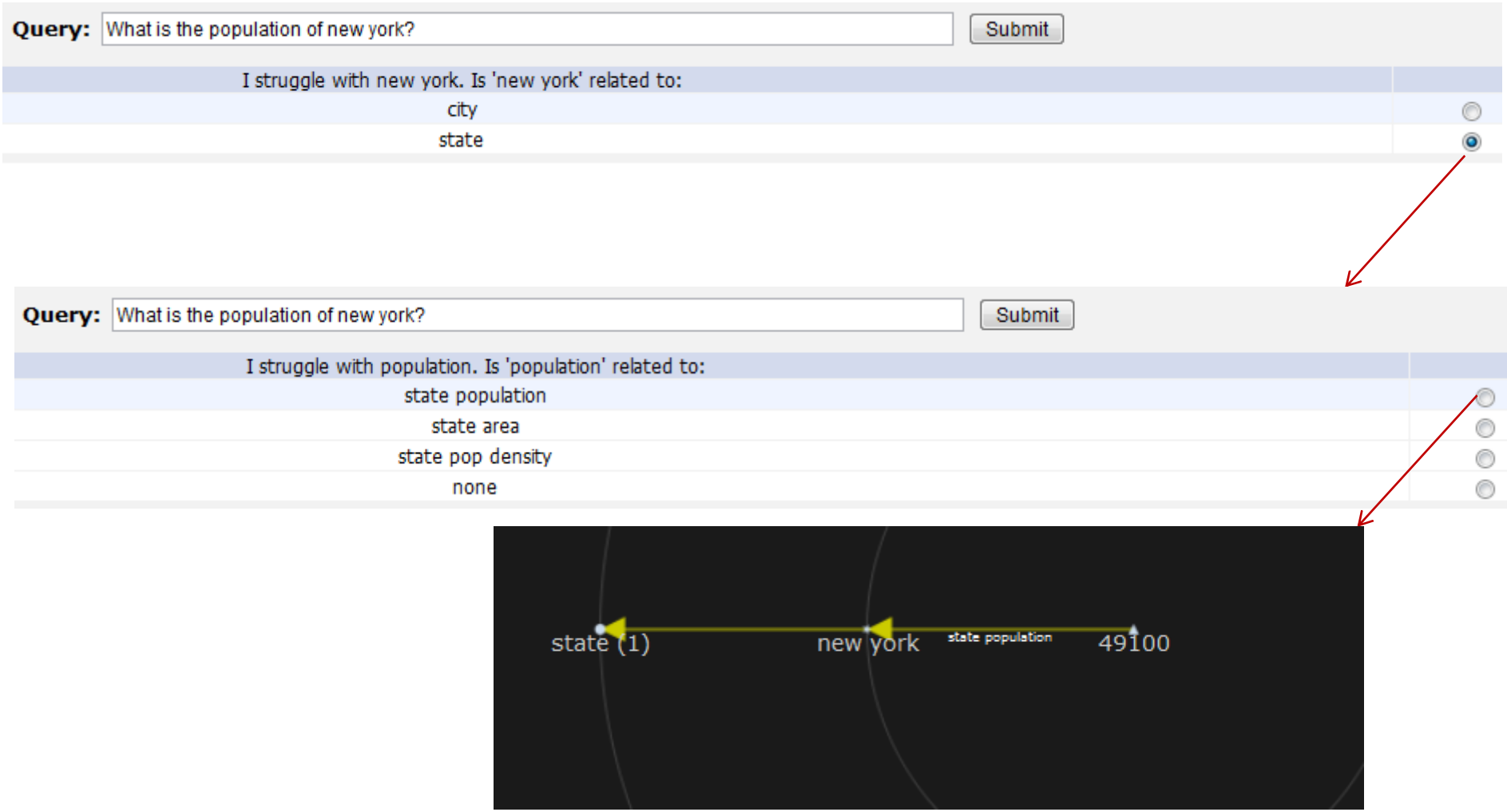
FREyA: Mapping POC to OCs



FREyA: New York is a city



New York is a state



Outline of the Lecture

- What is Semantic Search? Why is it Useful?
- How does it Work?
  - Semantic Annotation
  - Semantic Search
- How do we get the users to use it?
  - Faceted entity search
  - Form-based semantic constraints
  - Natural language queries
- Does it work? Peter Mika on Thursday 8:30am

Thank you!

Any Questions?

GATE Summer school on Text Mining
June 3rd – 7th, Sheffield, UK
http://gate.ac.uk

@GateAcUk