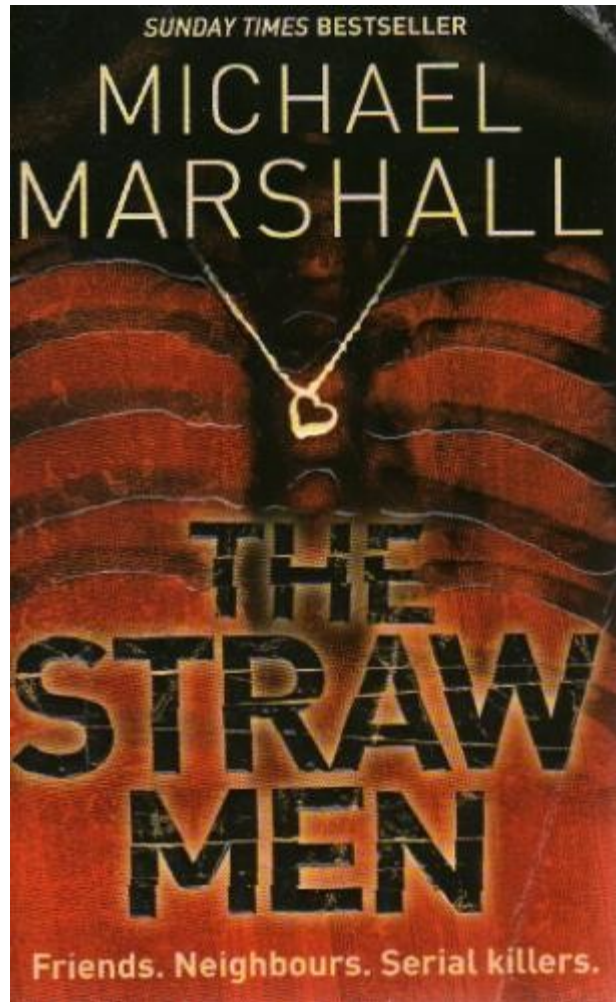


# Ontologies and Semantic Annotation

## Text Search isn't Enough



"I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota... I'm the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd."

*Michael Marshall, The Straw Men. HarperCollins Publishers, 2002.*

# ANNIE Annotations

German foreign minister Westerwelle visits Ghana.

William Hague and Angelina Jolie visit Eastern DRC.

Blackstone Group LP (BX) agreed to buy 23 industrial properties in southern Virginia and the Washington and Baltimore metropolitan areas from First Potomac Realty Trust (FPO) for \$241.5 million.

<input type="checkbox"/>	FirstPerson
<input type="checkbox"/>	JobTitle
<input checked="" type="checkbox"/>	Location
<input type="checkbox"/>	Lookup
<input type="checkbox"/>	Money
<input checked="" type="checkbox"/>	Organization
<input checked="" type="checkbox"/>	Person

- We know the type of named entity but nothing more
  - What kind of organization is Blackstone Group LP?
  - What is the job of William Hague?
  - Where is Eastern DRC, what does DRC stand for?
- => only semantics: choice of annotation type name
- => some knowledge hidden deep in JAPE & Code

# Need More Semantics:

- To co-reference DRC with “Democratic Republic of Congo”
- To avoid scattered knowledge in JAPE/Java?  
Cities are locations, cities have zip codes, ...
- To disambiguate: which “Washington” (state / city)?
- To use extracted information to allow for queries like:
  - European politicians who visited an African country?
  - Politicians and actors travelling together?
- To use extracted information to add information to our own Database/Knowledge base:
  - Add information about the buying-agreement to our data about Blackstone Group and First Potomac Realty Trust
  - Connect with trading information or other data we have

# Semantic Queries in Google

## [Paris convention and visitors office - Official website - Paris tourism](#)

[en.parisinfo.com/](http://en.parisinfo.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](http://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/](http://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](http://www.guardian.co.uk/travel/paris)

Latest news and comment on **Paris** from guardian.co.uk.

[co.uk/search?hl=en&tbo=d&biw=1081&bih=623&q=paris+weather](http://co.uk/search?hl=en&tbo=d&biw=1081&bih=623&q=paris+weather)



## 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 km<sup>2</sup>

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

**Local time:** Monday 23:12

## Points of interest



Eiffel Tower



Louvre




Disneyland  
Resort Paris




# Searching for Things, Not Strings

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

See results about



[University of Cambridge](#)  
The University of Cambridge is a public research university ...



[Cambridge](#)  
The city of Cambridge is a university town and the administrative ...

- Summaries of information about the entity being searched

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



## Anthony Blair

Anthony Charles Lynton Blair is a British Labour Party politician who served as the Prime Minister of the United Kingdom from 1997 to 2007. [Wikipedia](#)

**Born:** May 6, 1953 (age 59), [Edinburgh](#)

**Full name:** Anthony Charles Lynton Blair

**Parents:** [Hazel Corscadden](#), [Leo Blair](#)

**Siblings:** [William J. L. Blair](#)

**Children:** [Euan Blair](#), [Kathryn Blair](#), [Nicky Blair](#), [Leo Blair](#)

**Education:** [St John's College, Oxford](#) (1976), [Fettes College](#), [Chorister School](#), [University of Oxford](#)

### People also search for



[Gordon Brown](#)



[David Cameron](#)



[Margaret Thatcher](#)



[John Major](#)

Current **Tesco** employees who like **Horses**

Customer Service Assistant at Tesco

Likes **Horses** and **Dogs**

Studied **at** **at**

Lives in **Liverpool**

Listens to

Add Friend

Message

Search

Works at TESCO

Likes **Horses**

Studied **at** **at Uni. Wolverhampton**

Lives in

Listens to

Add Friend

Message

Search

Works at TESCO

Likes **Horses**

Studied **at**

Lives in

Listens to

Add Friend

Message

Search

Works at Tesco

Likes **Horses**

Studied **at**

Lives in **London, United Kingdom**

4 followers

Add Friend

Follow

Message

Search

General Assisant at Tesco

Likes **Horses**

Studies **Leeds Metropolitan University '13**

Lives in

In a Relationship • Female

Add Friend

Message

Search

More Than 100 People

View Grid

REFINE THIS SEARCH

Gender

Add...

Relationship

Add...

Current Employer

Tesco

Add

Position...

Employer Location...

Time Period...

Current City

Add...

Hometown

Add...

School

Add...

Friendship

Add...

Likes

Horses

Add

SEE MORE

EXTEND THIS SEARCH

More pages they like

Photos of these people

These people's friends

SEE MORE

Discover Something New

# Semantic Enrichment

- Textual mentions aren't actually that useful in isolation
  - knowing that something is a “Person” isn't very helpful
  - knowing which Person the mention refers to can be very useful
- Disambiguating mentions against an ontology provides extra context
- This is where **semantic enrichment** comes in
- The end product is a set of textual mentions linked to an ontology, otherwise known as **semantic annotations**
- Annotations on their own can be useful but they can also
  - be used to generate corpus level statistics
  - be used for further ontology population
  - form the basis of summaries
  - be indexed to provide semantic search



# Automatic Semantic Enrichment

- Use Text Mining, e.g.
  - Information Extraction – recognise names of people, organisations, locations, dates, references, etc.
  - Term recognition – identify domain-specific terms
- Automatically extend article metadata to improve search quality
- Example: using a customised GATE text mining pipeline to enrich metadata in the Envia environmental science repository

<http://www.bl.uk/reshelp/expertshelp/science/eventsandprojects/enviatbl/index.html>



☐ Whole words only

Show Advanced Filters

## Preliminary flood risk assessment : prepared to meet the Vale of Glamorgan Council's duties to manage local flood risk under the Flood Risk Regulations (2009)



**View/Open**  
<http://a0768b4a8a31e106d8b0-50dc802554eb38a24458b98ff72d550b.r19.cf3.rackcdn.com/flho1111bvet-e-e.pdf>

**Date**  
2011

**Author(s)**  
Vale of Glamorgan (Wales). Council

**Publisher**  
Barry : Vale of Glamorgan

**Citation**  
"2011. Preliminary flood risk assessment : prepared to meet the Vale of Glamorgan Council's duties to manage local flood risk under the Flood Risk Regulations (2009). Barry : Vale of Glamorgan."

**Description**  
Title from PDF cover (viewed on June 27, 2012).  
  
Includes bibliographical references (p. 29-30).

**Content Type**  
Report

**Pagination**  
1 online resource (31 p.)

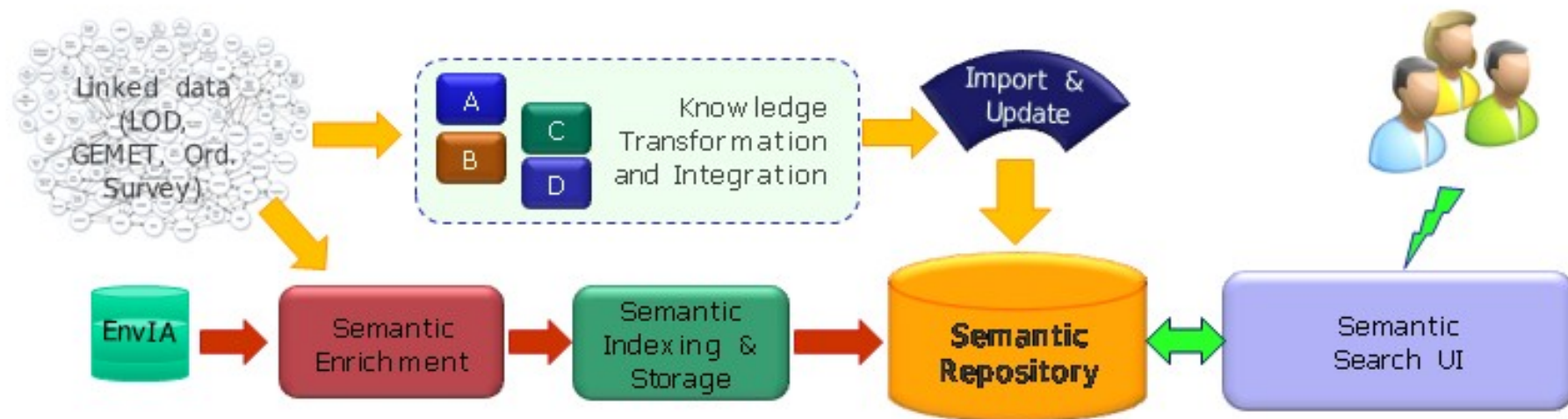
**Subject**  
Floods, Risk assessment, Wales, Vale of Glamorgan, Maps, Flood forecasting, Flood control, Planning

Subject
England (296)
Great Britain (276)
Flood control (258)
Floods (157)
Risk assessment (135)
... View More
Type
Report (1615)
Thesis (690)
Dataset (120)
atlas (13)
conference proceedings (3)
... View More
Publisher
Bristol : Environment Agency (355)
Luxembourg : Publications Office of the European Union (354)
London : Department for Communities and Local Government (51)

## 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

# Example Semantic Search Architecture



OWLIM

# 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: Basic Idea

- Link annotations to concepts in a knowledge base.
  - The annotated text is a “Mention” of a concept in the KB
  - We can use the knowledge associated with Mentions in our IE pipeline
    - e.g. Persons have JobTitles, Cities have zip codes
  - We can use the knowledge associated with Mentions for “Semantic Search”
  - We can use semantically annotated documents to add new facts to our knowledge base
- => We need some way to represent knowledge

# Knowledge Base

---

Would want to represent knowledge for this domain:

- Westerwelle:
  - has job Foreign minister of Germany → a politician
  - Germany → a country, in Europe
  - Member of the Free Democratic Party
  - Free Democratic Party → a political party
  - Political party → an organization
  - ...
- Blackstone Group L.P. → a private equity company
  - has NYSE symbol: BX
  - based in: New York City
  - New York City → a city
  - located in: New York State which is located in USA
  - ...

# Ontology

A formal way to represent knowledge as:

- Concepts of a domain or a set of domains  
“Agelina Jolie”, “Ghana”
- Relationships between concepts  
“New York City is located in New York State”
- Hierarchies of Concepts and Relationships  
“New York City is a City which is a Location”
- Associated Data  
“Blackstone Group has NYSE symbol BX”
- => most widely used formalism is RDF/OWL

# 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



# OWL Ontologies

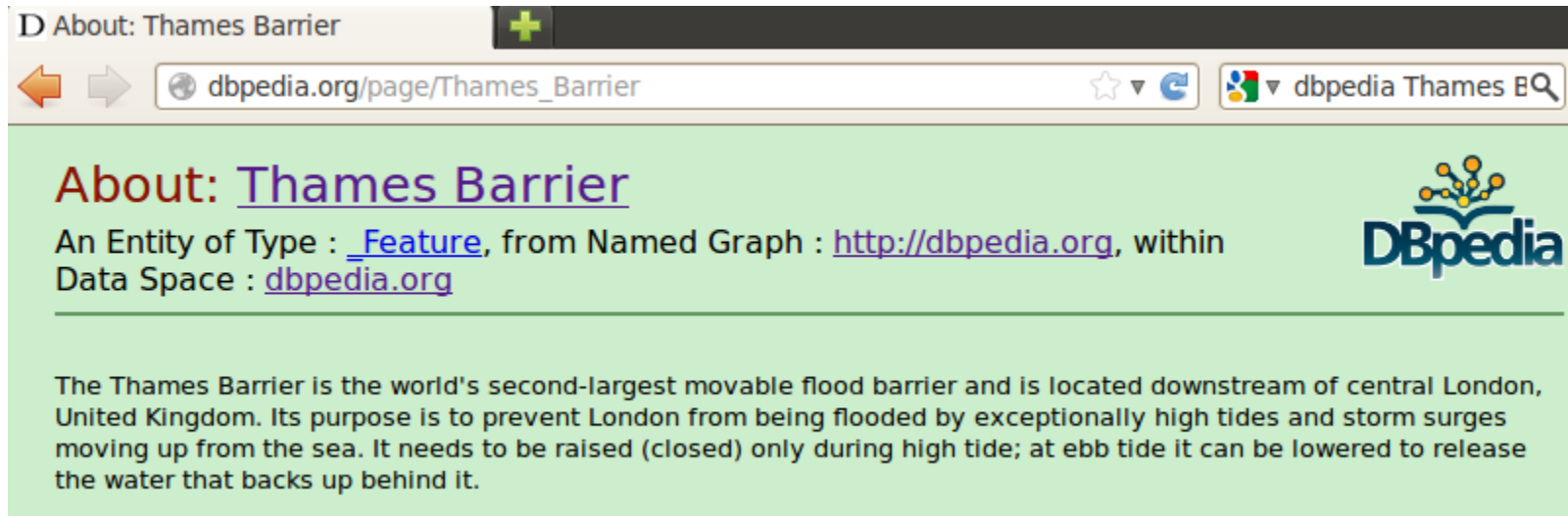
- OWL: Web Ontology Language
- Classes/Concepts and Individuals/Instances
- Properties:
  - DatatypeProperty: individual  $\rightarrow$  literal
  - ObjectProperty: individual  $\rightarrow$  individual
  - AnnotationProperty: resource  $\rightarrow$  literal, but no inference
- Inference/Reasoning:
  - Inheritance/Subsumption (classes and properties)
  - “Restrictions”: domain, range, allValuesFrom, hasValue ...infer class membership, property values
    - Open World Assumption: what isn’t asserted, we don’t know
    - Non Unique Name Assumption: different names may be used for same entity
- Classes can have more than one parent, Individuals can belong to more than one class  $\rightarrow$  OWL Ontologies are graphs, not trees



# 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



The screenshot shows a web browser window with the address bar displaying `dbpedia.org/page/Thames_Barrier`. The page title is "About: Thames Barrier". Below the title, it states: "An Entity of Type : [Feature](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)". The DBpedia logo is in the top right corner. The main content area has a light green background and contains a paragraph: "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://cs.dbpedia.org/resource/Bariéry_na_Temži)
- [http://de.dbpedia.org/resource/Thames\\_Barrier](http://de.dbpedia.org/resource/Thames_Barrier)
- [http://fr.dbpedia.org/resource/Barrière\\_de\\_la\\_Tamise](http://fr.dbpedia.org/resource/Barrière_de_la_Tamise)
- [http://it.dbpedia.org/resource/Thames\\_Barrier](http://it.dbpedia.org/resource/Thames_Barrier)
- <http://sws.geonames.org/2636058/>
- [freebase:Thames Barrier](#)

Links to GeoNames  
And Freebase

geo:geometry

- POINT(0.0367 51.4977)

Latitude & Longitude

geo:lat

- 51.497700 (xsd:float)

geo:long

- 0.036700 (xsd:float)

# 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)

# Ontologies in GATE

- Can use OWL-Lite ontologies as language resources  
(→ Plugin Ontology)
- Ontology Editor, Ontology Annotation Tool, Relation Annotation Tool (→ Plugin Ontology\_Tools)
- Ontology-enabled JAPE, JAPE Plus
- LKB Gazetteer (→ Plugin Gazetteer\_LKB)  
OntoRoot Gazetteer (→ Plugin Gazetteer\_Ontology\_Based)
- Ontology-based evaluation  
(→ Plugin Ontology\_BDM\_Computation)
- Java API for ontology manipulation, triple manipulation, SPARQL queries

# GATE Ontology Implementation

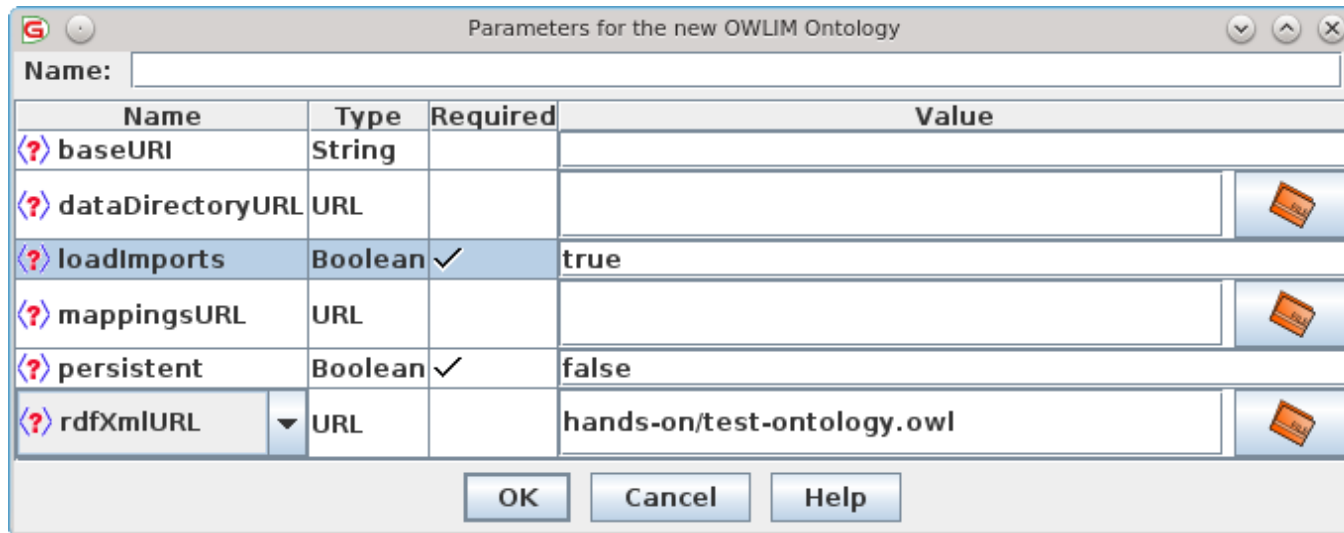
---







- Based on Sesame and the OWLIM-Lite SAIL (Storage and Inference Layer) implementation from Ontotext
- Fast in memory repository, scales to millions of statements (depending on RAM)
- In addition to local file ontology, can connect to server:
  - OWLIM Lite
  - OWLIM SE/Enterprise: commercial product, persistent and scalable implementation for huge (billion triples) ontologies
- Java API represents OWL concepts (ontology, property, literal) as Java objects
- Also provides support for SPARQL and manipulating Triples directly



# Load Ontology

- Need plugin Ontology
- For Editor, also need plugin Ontology\_Tools
- Language Resource → New → OWLIM Ontology



Name	Type	Required	Value
 baseURI	String		
 dataDirectoryURL	URL		
 loadImports	Boolean	✓	true
 mappingsURL	URL		
 persistent	Boolean	✓	false
 rdfXmlURL	URL		hands-on/test-ontology.owl

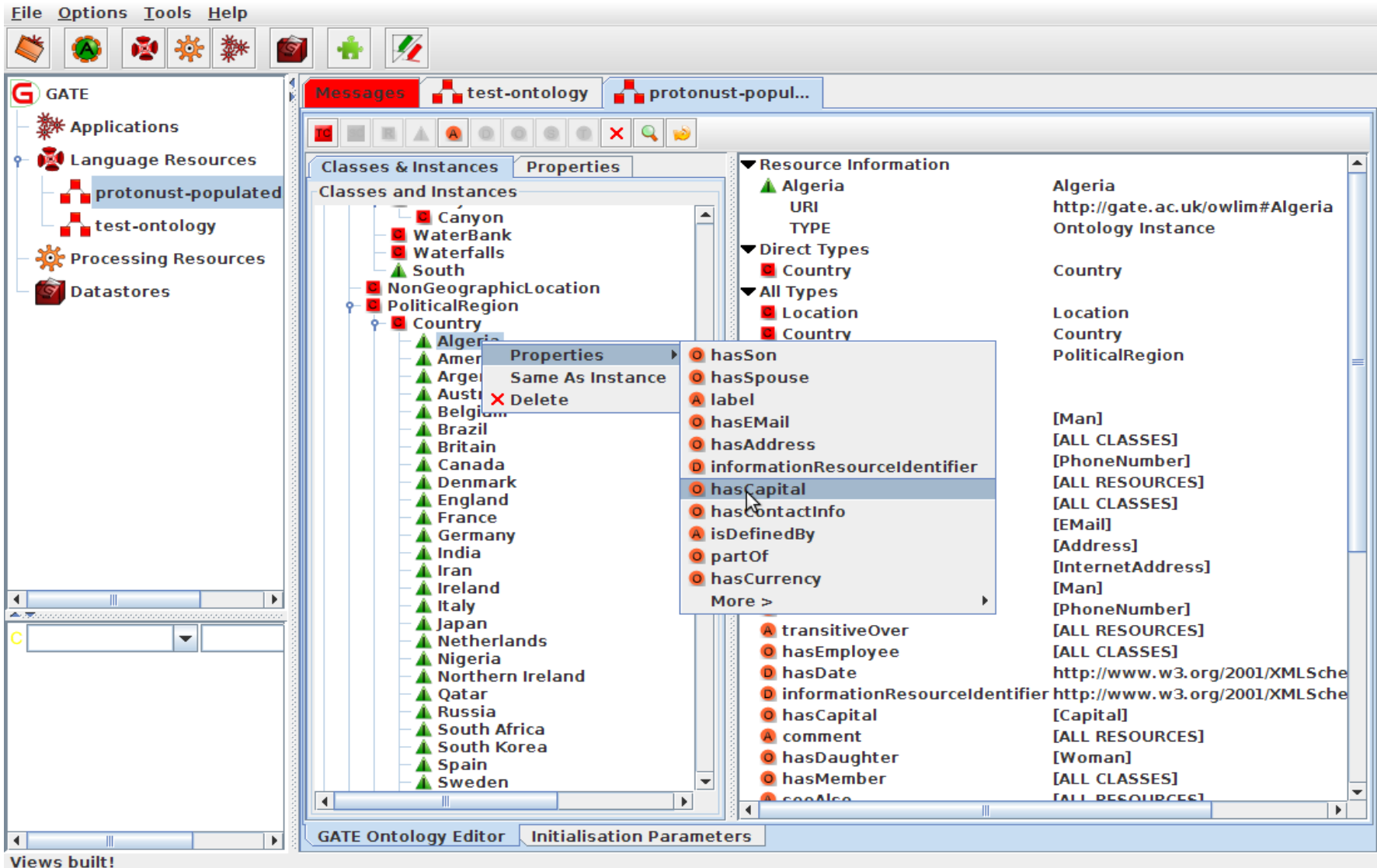
- Loaded:

# Ontology Viewer/Editor

---

- Basic viewing of ontologies
- Some edit functionalities:
  - create new concepts and instances
  - define new properties and property values
  - deletion
- Some limitations of what's supported, basically chosen from practical needs for semantic annotation
- Not a Protégé replacement

# Ontology Editor

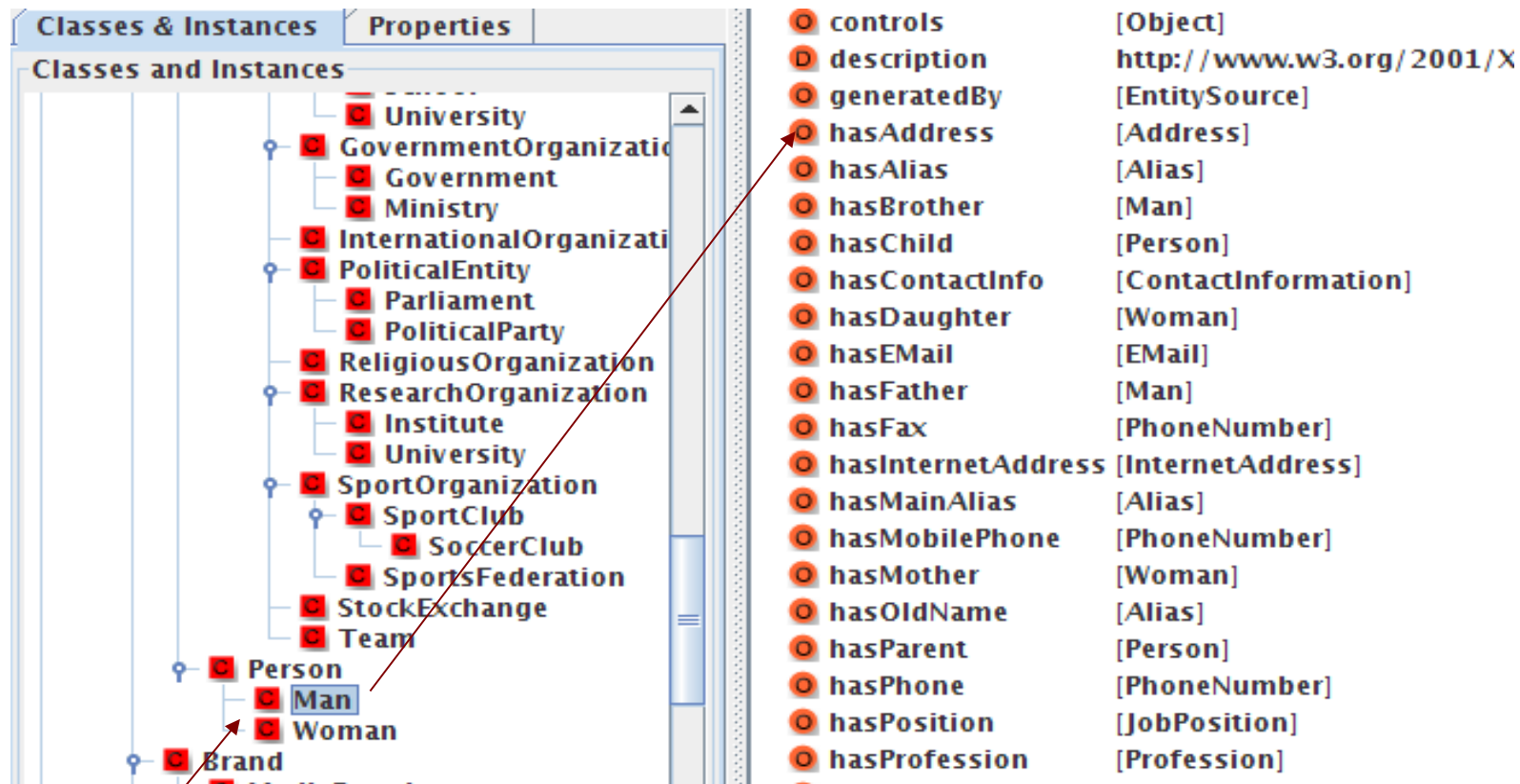


The screenshot displays the GATE Ontology Editor interface. The left sidebar shows the project structure with 'Language Resources' containing 'protonust-populated' and 'test-ontology'. The main workspace is divided into three panes:

- Classes & Instances:** A tree view showing a hierarchy of classes. 'Country' is selected, and a context menu is open over it. The menu options include:
  - Properties
  - Same As Instance
  - Delete
  - hasSon
  - hasSpouse
  - label
  - hasEmail
  - hasAddress
  - informationResourceIdentifier
  - hasCapital** (highlighted)
  - hasContactInfo
  - isDefinedBy
  - partOf
  - hasCurrency
  - More >
  - transitiveOver
  - hasEmployee
  - hasDate
  - informationResourceIdentifier
  - hasCapital
  - comment
  - hasDaughter
  - hasMember
  - seeAlso
- Properties:** A pane showing the properties of the selected class. It lists 'Country' and 'Location' as direct types, and 'Country' and 'PoliticalRegion' as all types.
- Resource Information:** A pane showing details for the selected resource 'Algeria'. It includes the URI 'http://gate.ac.uk/owlim#Algeria', the type 'Country', and a list of instances: 'Algeria', 'Amer', 'Arge', 'Aust', 'Belg', 'Brazil', 'Britain', 'Canada', 'Denmark', 'England', 'France', 'Germany', 'India', 'Iran', 'Ireland', 'Italy', 'Japan', 'Netherlands', 'Nigeria', 'Northern Ireland', 'Qatar', 'Russia', 'South Africa', 'South Korea', 'Spain', and 'Sweden'.

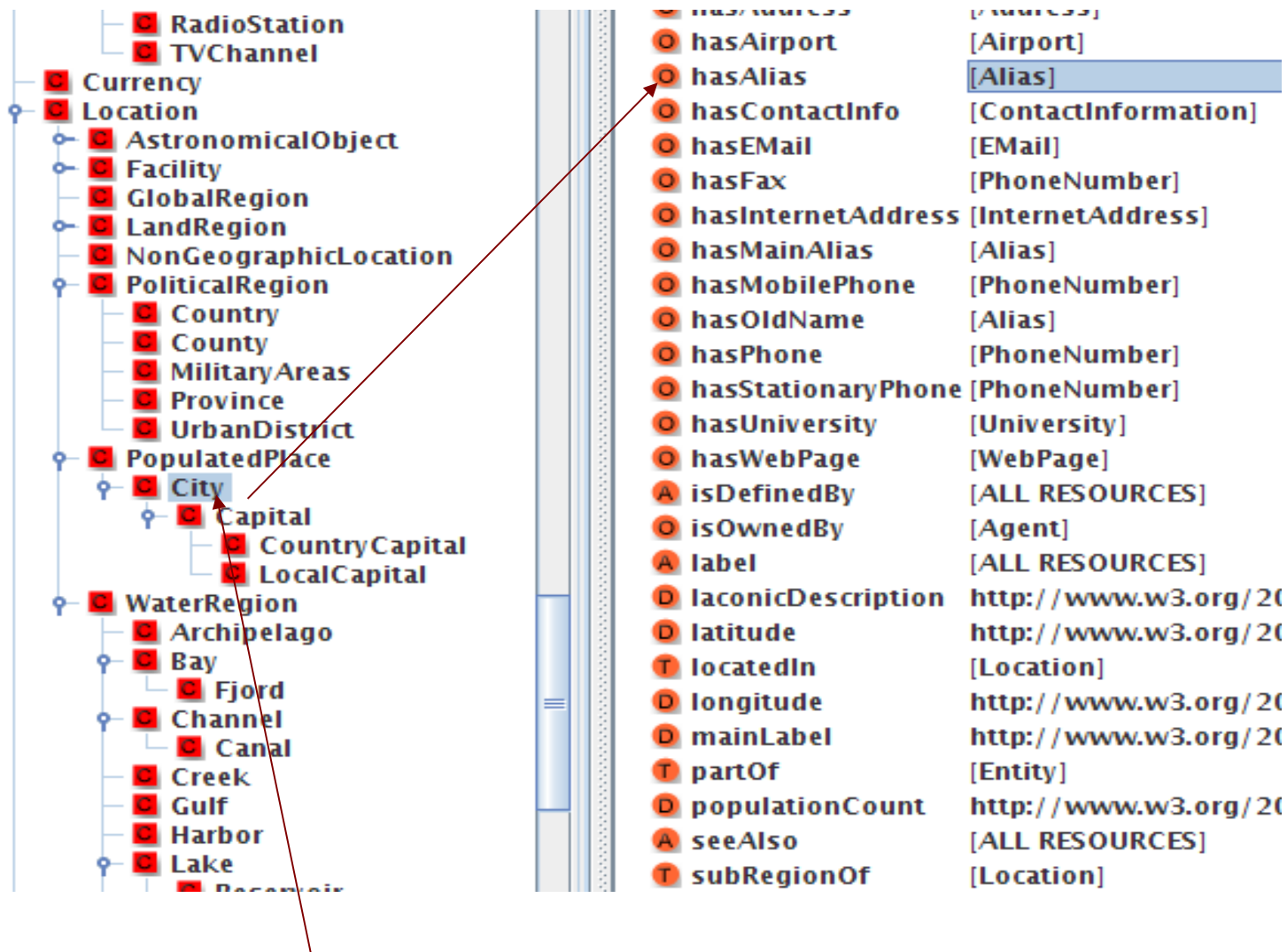
The bottom status bar indicates 'Views built!' and 'GATE Ontology Editor Initialisation Parameters'.

# 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.



## Hands-on: loading an ontology

---

- Load plugin Ontology (for basic ontology support)
- Load plugin Ontology\_Tools (for a simple ontology editor)
- Language Resource → New → OWLIM Ontology
- Fill in the parameter “rdfXmlURL” by navigating through the hands on material to [pubmed-hands-on/FastVacOntology\\_Nov2013v3.owl](#)
- Double-click the ontology to view it
- Scroll through the ontology
- Search for something (e.g. Yersinia Pestis)
- Look at its ancestors and descendants
- Look at its properties, including “label”



# Semantic Annotation

Print

Greece v Argentina: Who wins on penalties?  
By Robert Plummer Business reporter, BBC News  
Anyone examining the precedents for the Greek financial crisis might well be amused by the draw for next month's football World Cup matches.  
Greece's players celebrated after qualifying for the 2010 World Cup

For, as fate would have it, Greece's foes in Group B include the country that last suffered a comparable economic fiasco: Argentina.

In the worst-case scenario, Argentina's recent past is Greece's future.

The peso collapse, massive default and subsequent social and political unrest that rocked Argentina in 2001-2002 are being seen by many economists as an awful warning for the politicians in Athens and Brussels.

As far as football is concerned, t  
and final group match.

But the day of decision for the G  
stave off default by honouring bo

The EU and the IMF have agreed

Location

class	http://dbpedia.org/ontology/Place	X
inst	http://dbpedia.org/resource/Brussels	X
locType	other	X
matches	[6413, 6412]	X
rule	LKB_Location	X
		X

Open Search & Annotate tool

Type	Set	Start	End
Location		1222	1228
Location		1222	1228
Location		1222	1228
Location		1222	1228
Location		1222	1228
Location		1222	1228
Location		1222	1228
Location		1222	1228
Location		1233	1241
Organization		1556	1558

Location

class	http://dbpedia.org/ontology/Place	inst	http://dbpedia.org/resource/Brussels	locType	other	matches	[6413, 6412]	rule	LKB_Location
-------	-----------------------------------	------	--------------------------------------	---------	-------	---------	--------------	------	--------------

Open Search & Annotate tool

☒ Content

☐ Date

☐ Document

☐ DocumentClassification

☐ DocumentDate

☐ DocumentTitle

☐ FirstPerson

☐ JobTitle

☒ Location

☐ Lookup

☐ Measurement

☐ Money

☐ Number

☒ Organization

☐ Person

☐ Ratio

☐ Sentence

☐ SpaceToken

☐ Split

☐ Temp

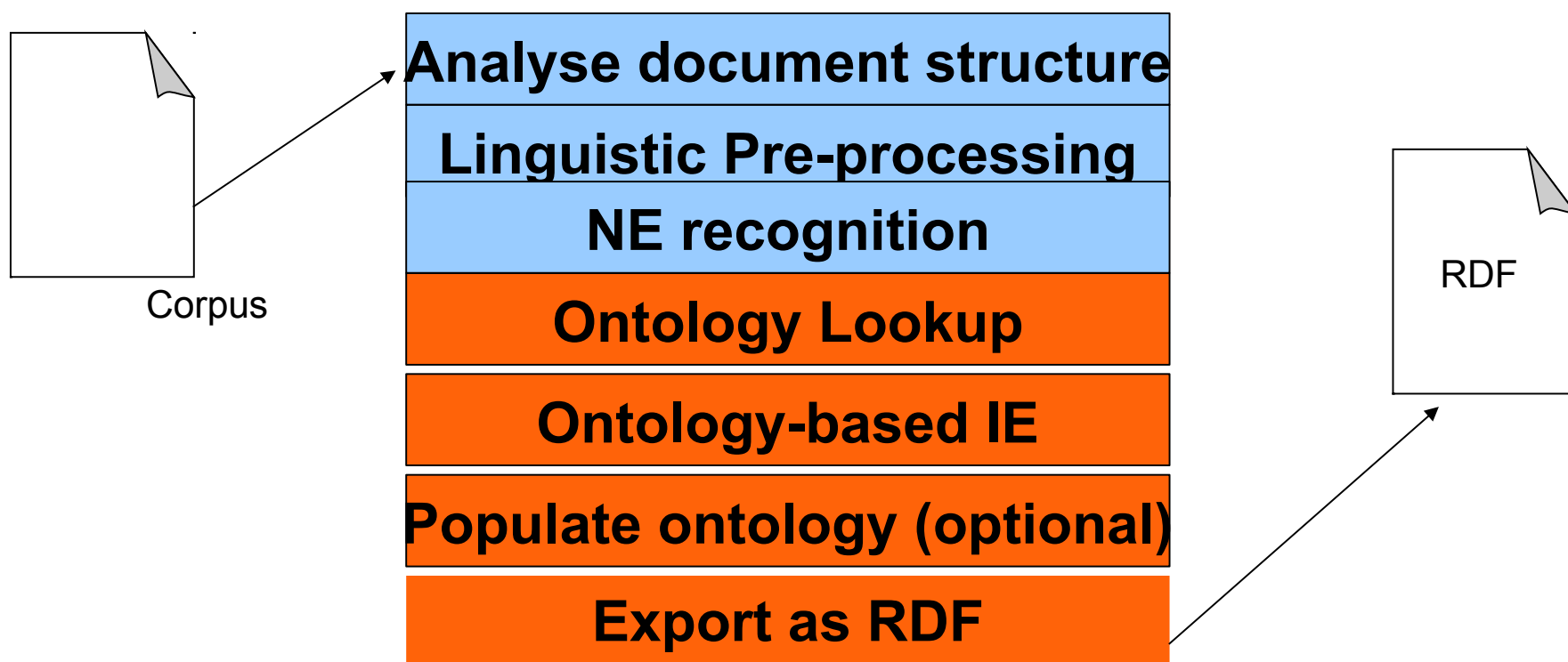
☐ Title

☐ Token

☐ Unknown

☒ Original markups

# Typical Semantic Annotation pipeline

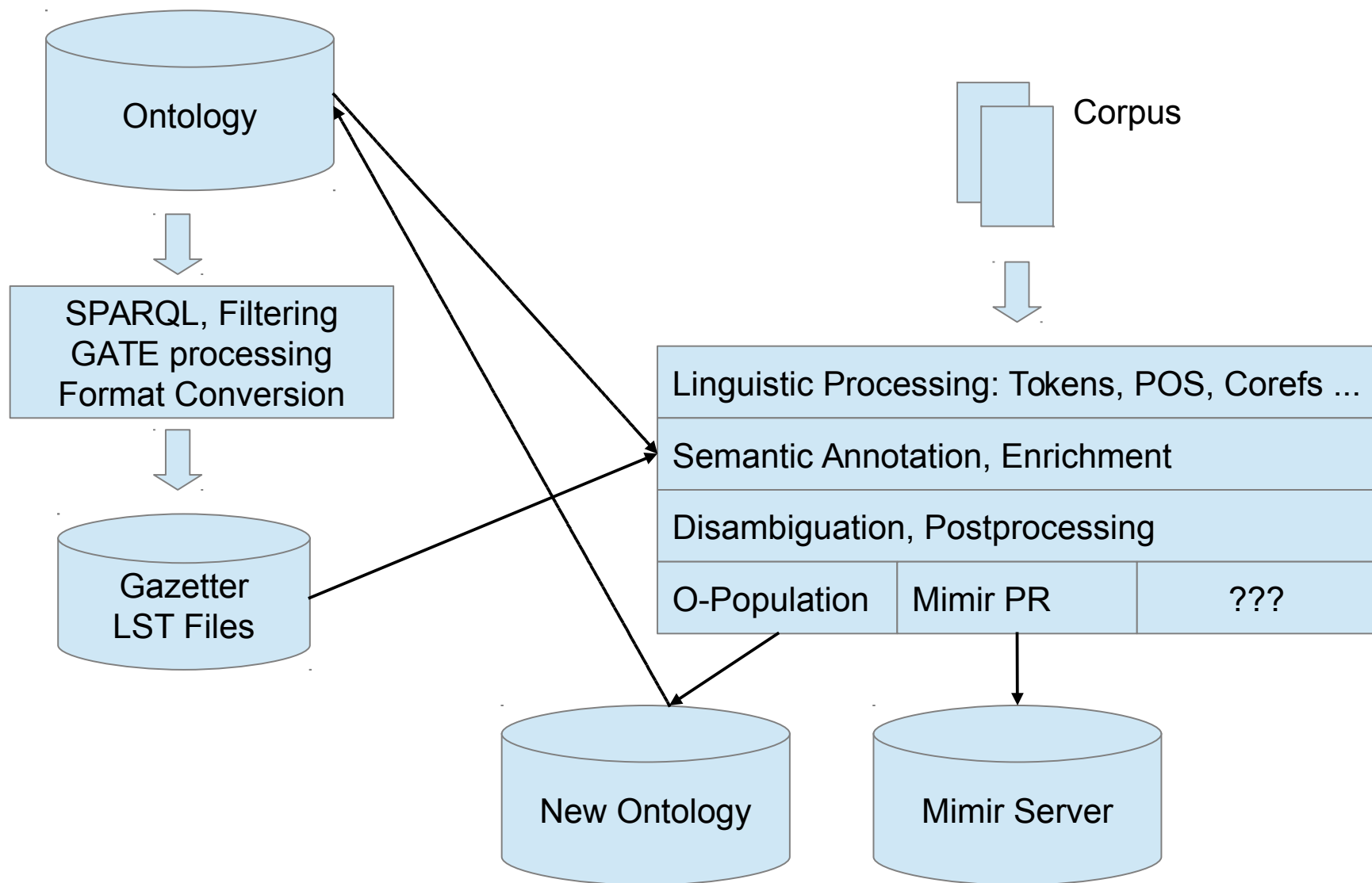


## Semantic Annotation: How

---

- Manually  
GATE: ontology based annotation using OAT/RAT or through crowdsourcing
- Automatically
  - Gazetteer/rule/pattern based
  - Classifier (ML) based
  - Combination of the two

# Semantic Annotation: The Big Picture



# GATE: Automatic Semantic Annotation

---

- Ontology aware Gazetteers:
  - OntoRoot Gazetteer
  - LKB Gazetteer
  - Other gazetteers, using inst/class features
- Ontology aware JAPE
- Semantic Enrichment: LKB Gazetteer, JAPE

## Ontology Lookup: OntoRoot Gazetteer

---

- Finds mentions in the text matching classes, instances, data property values and labels in the ontology
- Matching can be done between any morphological or typographical variant (e.g. upper/lower case, CamelCase)
- Converts CamelCase names, hyphens, underscores
- Morphological analysis is performed on both text and ontology, then matching is done between the two at the root level.
- Text is annotated with features containing the root and original string(s)
- Creates a gazetteer PR that can be used with the FlexibleGazetteerPR

## OntoRoot Gazetteer

---













- Lives in the Gazetteer\_Ontology\_Based plugin
- Generates candidate gazetteer list from ontology
- Runs the Tokeniser, POS tagger, Morphological Analyser to create lemmas from the labels and the fragment identifiers of all classes and instances and then creates lists to match against the text
  - Gordon\_Brown, GordonBrown → Gordon Brown
- Note that the gazetteer produced is stored in memory only and cannot be edited  
→ limited size!
- Must use tokeniser, sentence splitter, POS tagger and morphological analyser first: so we get “root” (lemma) feature!



# Init-time OntoRoot params

Parameters for the new Onto Root Gazetteer

Name:

Name	Type	Required	Value
 caseSensitive	Boolean	✓	false
 considerHeuristicRules	Boolean	✓	false
 considerProperties	Boolean	✓	true
 morpher	Morph	✓	<none>
 ontology	Ontology	✓	<none>
 posTagger	POSTagger	✓	<none>
 propertiesToExclude	String		
 propertiesToInclude	String		
 separateCamelCasedWords	Boolean	✓	true
 tokeniser	DefaultTokeniser	✓	<none>
 typesToConsider	Set		<input type="text"/>
 useResourceUri	Boolean	✓	true

OK Cancel Help

One or more of "class", "instance", "property"

Ontology LR

POS Tagger

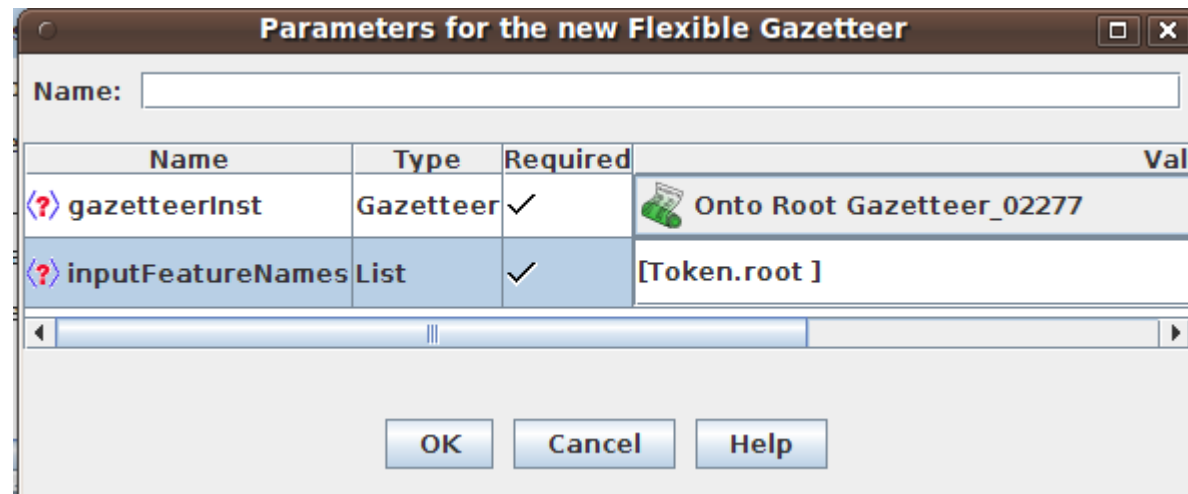
Tokeniser

## Running the OntoRoot gazetter

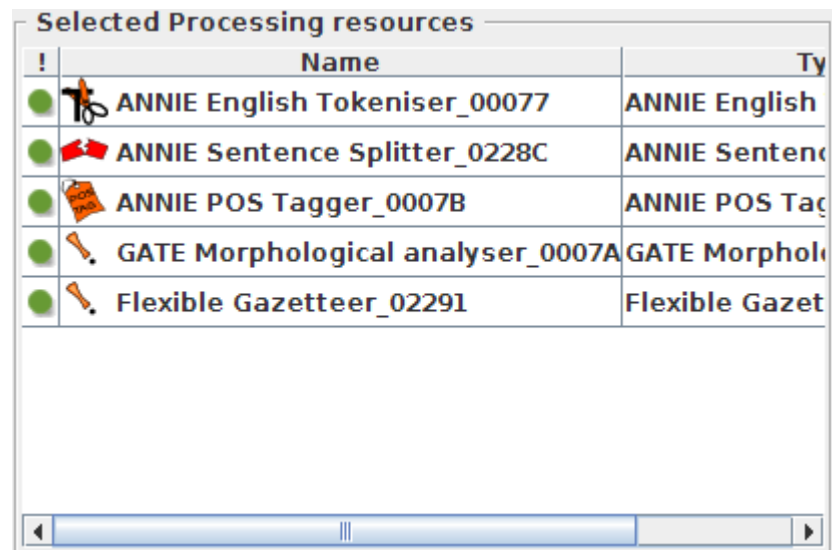
- If mostly matching proper names, then add to your application and run like the ANNIE gazetter
- It will match against the document text as it is, which is not ideal if matching against terms (“leaders” should match “leader”: need lemma/root)
- To find root we need: Tokeniser, Sentence Splitter, POS tagger, and Morphological Analyser
- To match the root and not the text, use Flexible Gazetteer PR with OntoRoot as the embedded gazetter
- Flexible Gazetteer delegates to OntoRoot Gazetteer: Flexible Gazetteer is the one that needs to be added to the application!  
→ If Flexible Gazetteer is used, no need to add OntoRoot Gazetteer to application.

# OntoRoot Application in GATE

Create a Flexible Gazetteer with an OntoRoot inside it



Build a GATE application with the PRs shown



# Output Example

standing for election across the country.

David Cameron was the first of the main UK party leaders to cast their vote. The Tory leader went to a community hall in Witney, Oxfordshire, shortly after 1030 BST, accompanied by his wife Samantha.

**Lookup**

URI	http://gate.ac.uk/example#
classURI	http://gate.ac.uk/example#
classURIList	[http://gate.ac.uk/example#
heuristic_level	0
majorType	
type	instance

Open Search & Annotate tool

**Lookup**

URI	http://gate.ac.uk/example#Leader	X
heuristic_level	0	X
majorType		X
type	class	X
		X

Open Search & Annotate tool

Lookup	672	685	9704	{URI=http://gate.ac.uk/example#David_Cameron, classURI=http://g
Lookup	721	728	9705	{URI=http://gate.ac.uk/example#Leader, heuristic_level=0, majorTy
Lookup	758	764	9706	{URI=http://gate.ac.uk/example#Leader, heuristic_level=0, majorTy

- The URI feature contains the matched class or instance URI
- The type feature is either class or instance
- Instances have also features classURI and classURIList

# Hands-on: OntoRootGazetteer

- Load Gazetteer\_Ontology\_Based plugin, Format\_PubMed, ANNIE and Tools plugins
- Create Document Reset, Tokeniser, Sentence Splitter, POS Tagger, and Morphological Analyser (all with defaults)
- Create a new corpus pipeline called “text”, and add the above PRs to it in that order
- Create another new corpus pipeline called “ontology” and add the Tokeniser, POS Tagger, and Morphological Analyser in that order
- Create an OntoRootGazetteer, configuring these parameters:
  - ontology – the FastVac ontology
  - RootFinderApplication – the “ontology” pipeline
  - propertiesToInclude:
    - <http://www.w3.org/2000/01/rdf-schema#label>, [http://fera.gsi.gov.uk/gate#Latin\\_name](http://fera.gsi.gov.uk/gate#Latin_name)
  - separateCamelCaseWords – false
  - useResourceUri - false

## Hands-on: OntoRoot (contd.)

---

- Create a FlexibleGazetteer PR:
  - add Token.root to inputFeatureNames
  - choose the OntoRoot gazetteer as gazetteerInst
- Add Flexible Gazetteer to the “text” pipeline
- In a text editor, examine a document from the corpus in the hands on material, pubmed-hands-on/corpus
- They are in native PubMed format
- Create a corpus and populate it from this corpus, setting encoding to UTF-8
- Examine a document in GATE. Look at the original markup annotations, and the document features
- Run the “text” pipeline over the corpus and inspect the resulting Lookup annotations

## Conventions in GATE

- We use “Mention” annotations to reflect the fact that the text mentions a particular instance or a class
- The Mention annotations have two special features:
  - *class* = class URI from the ontology
  - *inst* = instance URI from the ontology (if available)  
e.g. Mention {class=Leader, inst=Gordon\_Brown}
- It's important **not** to use *class* and *inst* as features unless you're dealing with ontologies, as these are predefined names in several tools
- OntoRoot Gazetteer does not follow the conventions

## Compatibility with OntoRootGazetteer

---

- The OntoRootGazetteer always puts the matching resource (class or individual) URI in a feature called “URI” and the kind of match in a feature called “type”. For individuals it also creates the features “classURI” and “classURIList”
- But GATE/JAPE requires these features to be called **class** and **inst**
- So we need a JAPE grammar to first change the names of these features



# JAPE grammar to change feature names

Phase: LookupRename

Input: Lookup

Options: control = appelt

Rule: RenameLookup

```
(  
  {Lookup.type == instance}
```

finds all Lookups which OntoRoot gazetteer  
created from ontology instances

```
):match
```

```
-->
```

```
:match{
```

```
  for (Annotation lookup : matchAnnots) {  
    FeatureMap theFeatures = lookup.getFeatures();
```

add a new feature **class** with the  
value of the original classURI feature

```
    theFeatures.put(  
      "class", theFeatures.get("classURI"));
```

do the same  
for inst

```
    theFeatures.put("inst", theFeatures.get("URI"));
```

```
  }
```

```
}
```

## Ontology Aware JAPE

---

- JAPE transducers have a run-time parameter which is an ontology
- [Note that the ANNIE NE Transducer] does not have this parameter, so you cannot use it for ontology-aware JAPE]
- By default it is left blank, so not used during LHS matching
- When an ontology is provided, the **class** feature can be used on the LHS of a JAPE rule
- When matching the **class** value, the ontology is checked for subsumption: any subclass on the left side of “==” matches
- e.g. {Lookup.class == Person} will match a Lookup annotation with **class** feature, whose value is either Person or any subclass of it

## Ontology-aware JAPE example

Phase: OntoMatching

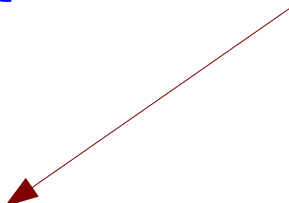
Input: Lookup

Options: control = appelt

Rule: PersonLookup

```
(  
  {Lookup.class == Person}
```

Matches the class Person  
or any of its subclasses

A thin red arrow originates from the text "Matches the class Person or any of its subclasses" and points diagonally down and to the left, ending at the condition "{Lookup.class == Person}" within the rule definition.

```
):person
```

```
-->
```

```
:person.Mention =  
  {class = :person.Lookup.class,  
    inst = :person.Lookup.inst}
```

Adds class and instance information  
as features on the Mention annotation

A thin red arrow originates from the text "Adds class and instance information as features on the Mention annotation" and points diagonally down and to the left, ending at the start of the rule body ":person.Mention =".

## Ontology-aware JAPE example

---

```
{Lookup.class == "http://example.com/stuff#Person"}
```

Matches this class or any subclass in the ontology

```
{Lookup.class == "Person"}
```

If the string is not a full URI, JAPE adds the default namespace from the ontology, looks up that class in the ontology, and matches it or any subclasses. Be very careful if your ontology uses more than one namespace!

## Templates to simplify namespaces

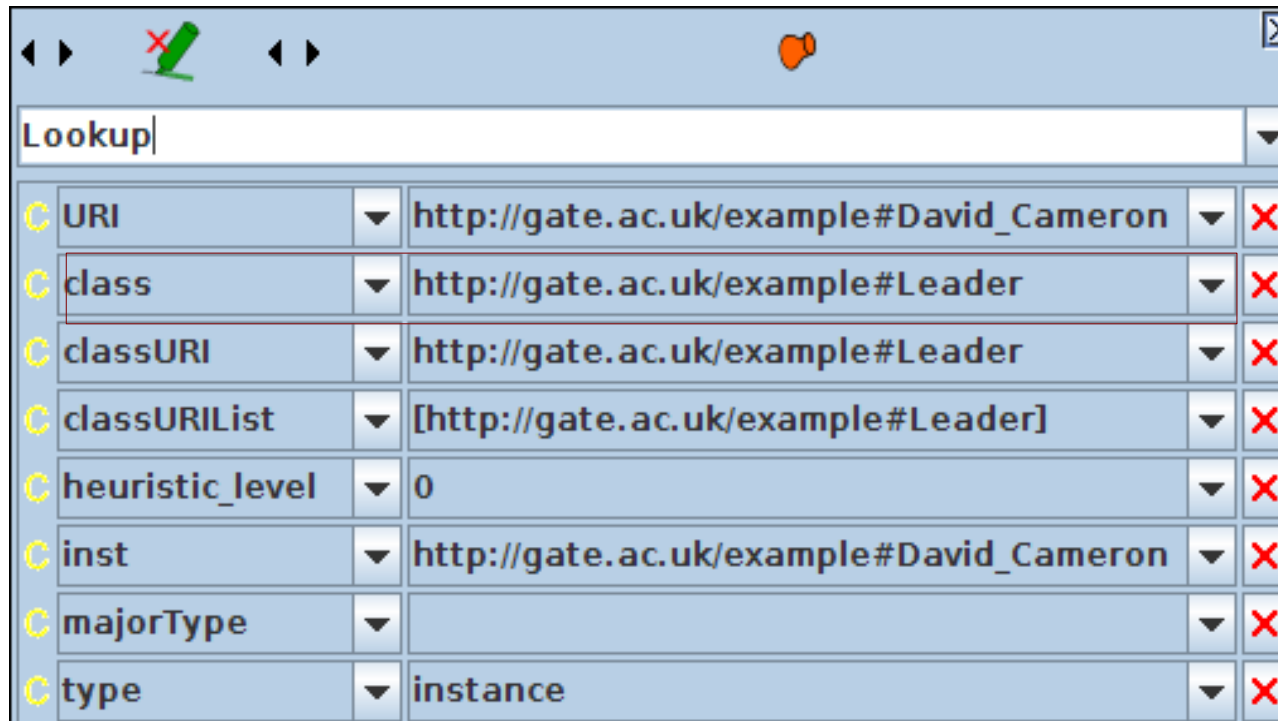
---

```
Template: protont =  
    "http://proton.semanticweb.org/2005/04/protont#${n}"  
...  
{Lookup.class == [protont n=Person]}  
...  
{Lookup.class == [protont n=Location]}
```

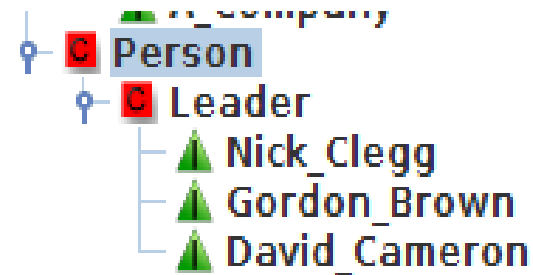
```
Template: protont =  
    "http://proton.semanticweb.org/2006/05/protont#${n}"  
...
```

# Matching subclasses

David Cameron was the first of the main UK party leaders...



Property	Value	Status
URI	http://gate.ac.uk/example#David_Cameron	✗
class	http://gate.ac.uk/example#Leader	✗
classURI	http://gate.ac.uk/example#Leader	✗
classURIList	[http://gate.ac.uk/example#Leader]	✗
heuristic_level	0	✗
inst	http://gate.ac.uk/example#David_Cameron	✗
majorType		✗
type	instance	✗



The rule matches because Leader is a subclass of Person

## Hands-on: ontology-aware JAPE

- Load the JAPE transducer *rename-lookup-features.jape* and add to the end of your “text” pipeline
- Run the modified pipeline to see how the Lookup annotations for Lookups now have class features
- Examine the JAPE file *pathogen-onto-matching.jape*
- Load the JAPE transducer *pathogen-onto-matching.jape* and add it to the end of the “text” pipeline as before.
- In the pipeline, select the FastVac ontology as the ontology runtime parameter of the transducer
- Run the modified pipeline to see how it creates new *Pathogen* annotations

# Semantic Annotation with other tools: 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, producing utics entered the reverse osmosis (RO) water treatment field in 1970, and is now one of the most respected and experienced. Hydranautics became part of the Nitto Denko Corporation when it was acquired in 1987. Hydranautics corporate l 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 water 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 States
- ☒ **Company**
  - ☒ Hydranautics Inc
  - ☒ NITTO DENKO CORPORATION
- ☒ **Continent**
  - ☒ Asia
  - ☒ Europe
- ☒ **Country**
  - ☒ United States
- ☒ **Industry Term**
  - ☒ wastewater treatment
- ☒ **Province Or State**
  - ☒ California, United States
- ☒ **Region**
- ☒ **Technology**

**Events & Facts:**

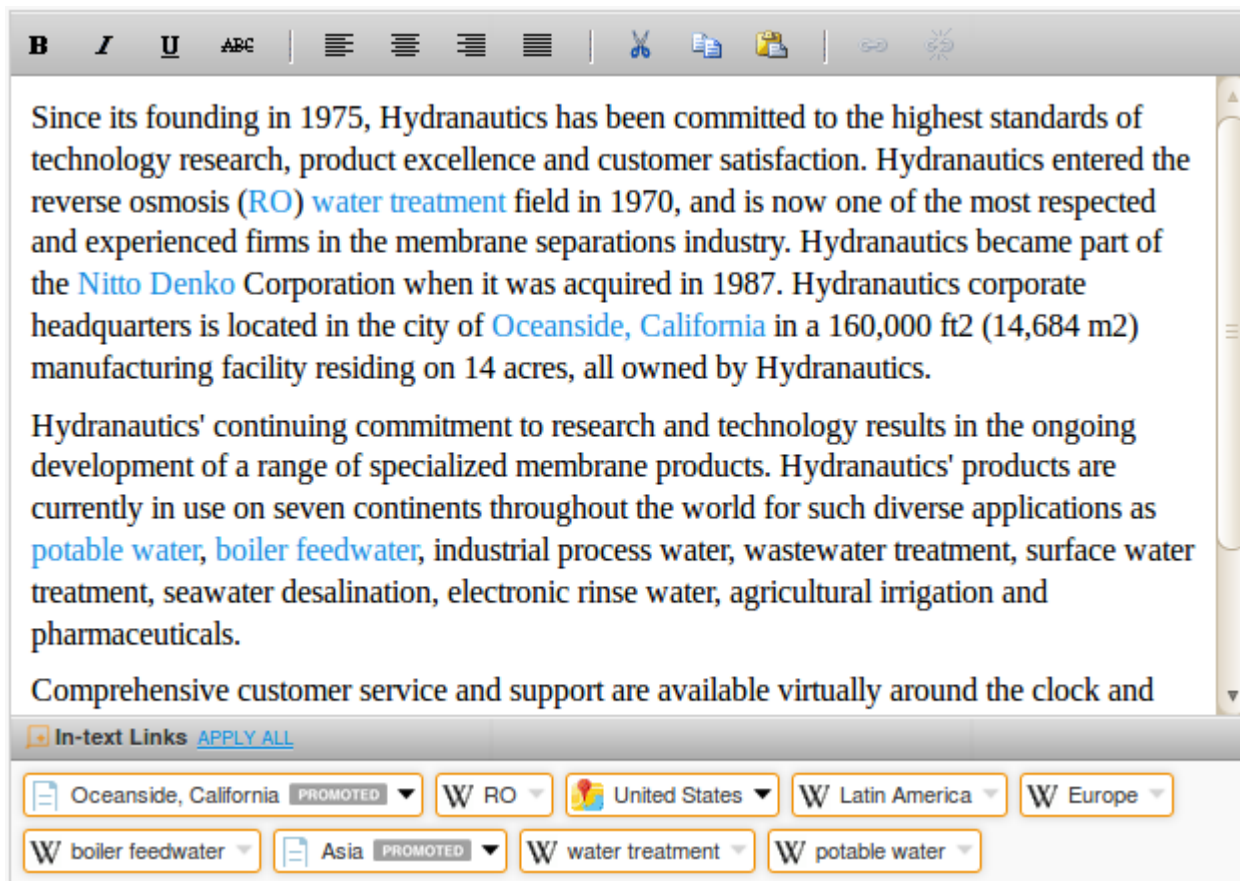
- ☒ **Acquisition**
  - ☒ NITTO DENKO CORPORATION, 1987-00-00, in
- ☒ **Company Founded**
  - ☒ Hydranautics Inc, 1975
- ☒ **Generic Relations**
  - ☒ Hydranautics Inc, be
  - ☒ Hydranautics Inc, part of the Nitto Denko
  - ☒ Hydranautics Inc, commit
  - ☒ a network of worldwide sales offices,
  - ☒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

Not easily customised/extended  
Domain-specific coverage varies



# Zemanta

- Paste text from [www.membranes.com](http://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



---

QUESTIONS?