



Module 6: Ontologies and Semantic Annotation





About this tutorial

- This tutorial will be a mixture of explanation, demos and hands-on work
- Things for you to try yourself are in red
- It assumes basic familiarity with the GATE GUI and with ANNIE and JAPE; no Java expertise



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.

- FirstPerson**
- JobTitle**
- Location**
- Lookup**
- Money**
- Organization**
- 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/

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



[Eiffel Tower](#)



[Louvre](#)




[Disneyland](#)




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



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](#)

University of Sheffield, NLP Facebook Graph Search



Facebook interface showing search results for "Current Tesco employees who like Horses".

Search Results:

- Profile 1:** Customer Service Assistant at Tesco. Likes Horses and Dogs. Studied at [redacted] at [redacted]. Lives in Liverpool. Listens to [redacted].
- Profile 2:** Works at TESCO. Likes Horses. Studied at [redacted] at Uni. Wolverhampton. Lives in [redacted]. Listens to [redacted].
- Profile 3:** Works at TESCO. Likes Horses. Studied at [redacted]. Lives in [redacted]. Listens to [redacted].
- Profile 4:** Works at Tesco. Likes Horses. Studied at [redacted]. Lives in London, United Kingdom. 4 followers.

Refinement and Extension Options:

- REFINE THIS SEARCH:** Gender, Relationship, Current Employer (Tesco), Position, Employer Location, Time Period, Current City, Hometown, School, Friendship, Likes (Horses).
- EXTEND THIS SEARCH:** More pages they like, Photos of these people, These people's friends.

Navigation: Home, Add Friend, Message, Search, Discover Something New.



Semantic Annotation: Basic Idea/Vision

- 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

Use an 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



OWL Ontologies - RDF(S)

- Based on RDF(S) - Resource Description Framework (Schema):
 - Everything is identified by an URI: <http://dbpedia.org/page/Paris>
 - Everything can be expressed as triples of the form *Subject Predicate Object*:
 - :NewYork rdf:type :City .
 - :City rdfs:subClassOf :Location .
 - :Location a rdfs:Class .
 - :BlackstoneGroup :hasNyseSymbol "BX" .
 - Simple vocabulary to express things:
 - rdf:type = "belongs to a class"
 - rdf:Class = "the class of all classes"
 - "BX" = the literal string "BX"



OWL Ontologies - RDF(S)

- All resources identified by URIs
Different URIs may refer to the same resource
- Resources that are “Individuals” can be grouped into “Classes” and relate to other things and to values by “Properties”.
- Values represented through “Literals”:
 - “BX” - a string (untyped literal)
 - “New York State”@en – string with language tag (untyped)
 - “Guido Westerwelle”^^xsd:string – typed literal
 - “24”^^xsd:integer
- :A rdf:type :B – :A is contained in class :B
- :B rdf:type rdfs:Class – :B is an RDFS Class
- :B rdfs:subClassOf :C – all members of :B are in :C



OWL Ontologies

- OWL: Web Ontology Language
- Classes/Concepts and Individuals/Instances
- Properties:
 - DatatypeProperty: individual → literal
 - ObjectProperty: individual → individual
 - AnnotationProperty: resource → 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 → OWL Ontologies are graphs, not trees



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

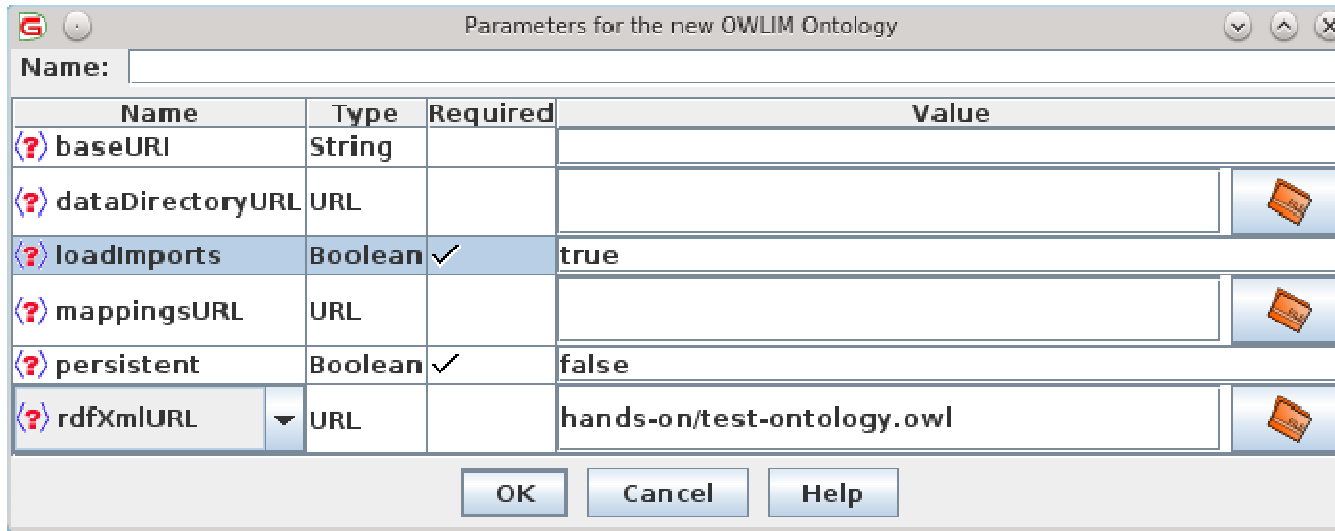











GATE Ontology Implementation



- 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:  Language Resources
 test-ontology.owl_00016



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. On the left, a tree view shows the ontology structure with categories like Applications, Language Resources, and Processing Resources. The main workspace is divided into three panes: 'Classes & Instances', 'Properties', and 'Resource Information'. The 'Classes & Instances' pane shows a hierarchy where 'Country' is selected, and a context menu is open over it. The 'Properties' pane shows a list of properties such as 'hasSon', 'hasSpouse', and 'hasCapital'. The 'Resource Information' pane shows details for the selected class, including its URI, type, and direct types.

Classes & Instances

- Classes and Instances
 - Canyon
 - WaterBank
 - Waterfalls
 - South
 - NonGeographicLocation
 - PoliticalRegion
 - Country
 - Algeria
 - Amer
 - Argel
 - Aust
 - Belgium
 - Brazil
 - Britain
 - Canada
 - Denmark
 - England
 - France
 - Germany
 - India
 - Iran
 - Ireland
 - Italy
 - Japan
 - Netherlands
 - Nigeria
 - Northern Ireland
 - Qatar
 - Russia
 - South Africa
 - South Korea
 - Spain
 - Sweden

Properties

- hasSon
- hasSpouse
- label
- hasEMail
- hasAddress
- informationResourceIdentifier
- hasCapital
- hasContactInfo
- isDefinedBy
- partOf
- hasCurrency
- More >

Resource Information

Algeria
URI
TYPE
Direct Types
Country
All Types
Location
Country
PoliticalRegion

[Man]
[ALL CLASSES]
[PhoneNumber]
[ALL RESOURCES]
[ALL CLASSES]
[EMail]
[Address]
[InternetAddress]
[Man]
[PhoneNumber]
[ALL RESOURCES]
[ALL CLASSES]
http://www.w3.org/2001/XMLSchema
http://www.w3.org/2001/XMLSchema
[Capital]
[ALL RESOURCES]
[Woman]
[ALL CLASSES]
[ALL RESOURCES]



Modelling social media with ontologies

- SIOC and SIOC Types Ontologies
- SIOC (Semantically-Interlinked Online Communities) Core Ontology provides concepts and properties, describing information from online communities (e.g. wikis, weblogs)
 - Documentation: <http://rdfs.org/sioc/spec/#sec-modules>
 - Ontology namespace: <http://rdfs.org/sioc/ns#>
- SIOC Types adds extensions for Twitter modelling
 - Ontology namespace: <http://rdfs.org/sioc/types#>
- Open the SIOC Types ontology in GATE (in hands-on), by giving the URL as an RDF/XML parameter to the OWLIM Ontology LR
- Double click to view the ontology

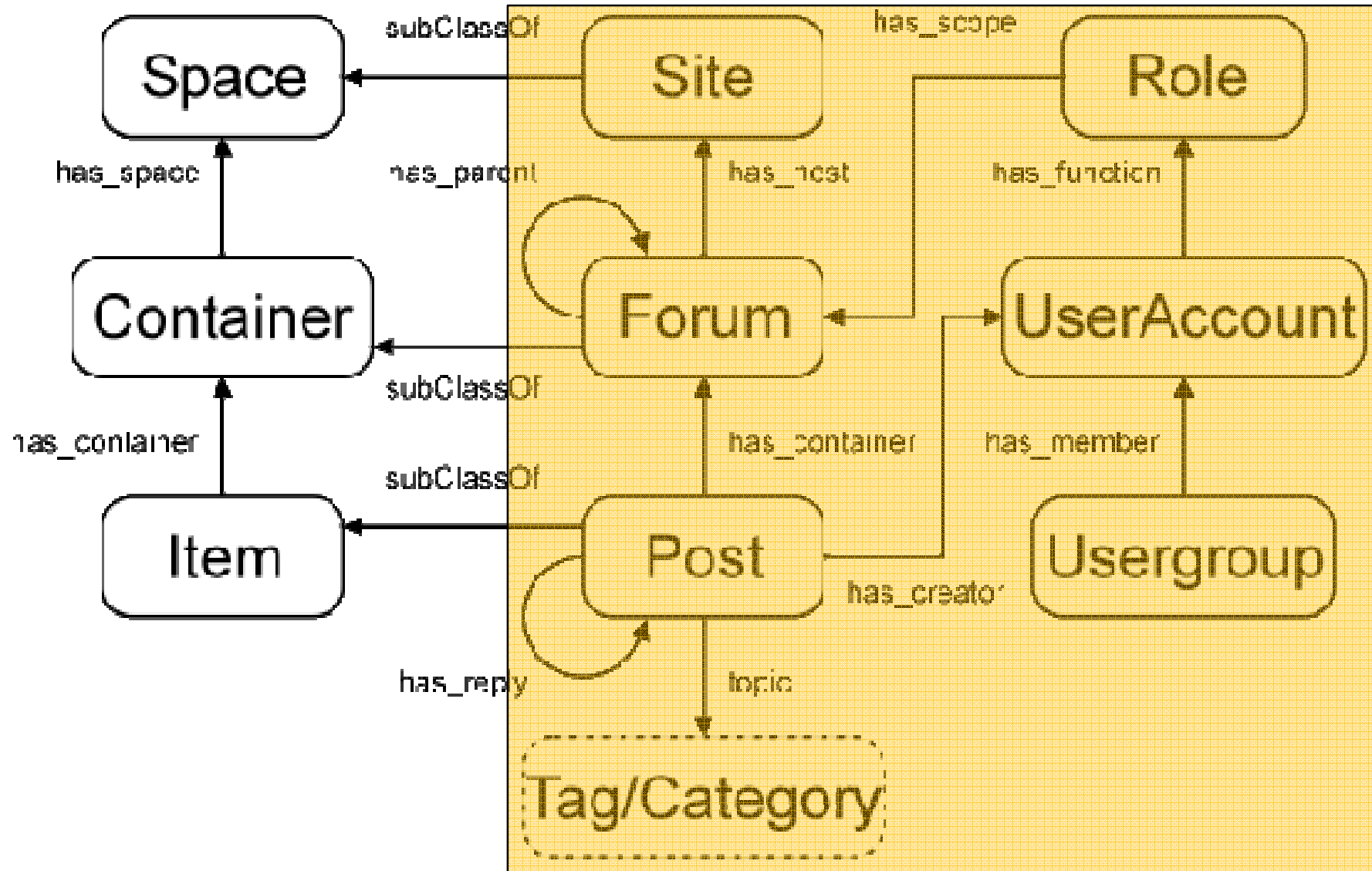


MicroblogPost and some properties

The screenshot shows the GATE software interface. The 'Classes & Instances' panel on the left displays a tree view of classes. The 'Properties' panel on the right lists various properties with their data types and URIs.

Property Name	Data Type	URI / Value
comment	[ALL RESOURCES]	
content	http://www.w3.org/2001/XMLSchema#string	
content_encoded	http://www.w3.org/2001/XMLSchema#string	
created_at	http://www.w3.org/2001/XMLSchema#string	
date	http://www.w3.org/2001/XMLSchema#string	
depiction	[ALL CLASSES]	
description	http://www.w3.org/2001/XMLSchema#string	
description	http://www.w3.org/2001/XMLSchema#string	
earlier_version	[Item]	
embeds_knowledge	[Graph]	
feed	[ALL CLASSES]	
group_of	[ALL CLASSES]	
has_container	[Container]	
has_creator	[UserAccount]	
has_discussion	[ALL CLASSES]	
has_function	[Role]	
has_group	[ALL CLASSES]	
has_modifier	[UserAccount]	
has_owner	[UserAccount]	
has_part	[ALL CLASSES]	
has_reply	[Item]	
has_space	[Space]	
id	http://www.w3.org/2001/XMLSchema#string	
ip_address	http://www.w3.org/2001/XMLSchema#string	
isDefinedBy	[ALL RESOURCES]	
label	[ALL RESOURCES]	
last_activity_date	http://www.w3.org/2001/XMLSchema#string	
last_reply_date	http://www.w3.org/2001/XMLSchema#string	
later_version	[Item]	
latest_version	[Item]	
link	[ALL CLASSES]	

SIOC: High Level Overview





Modelling Twitter with SIOCT

- Users modelled through the <http://rdfs.org/sioc/ns#UserAccount> class
- Useful properties for modelling tweet user info
 - sioc:description: corresponds to the description JSON entry
 - sioc:name, sioc:email, sioc:id
- Properties for relating users to users: follows
- Properties for relating users to tweets:
creator_of(UserAccount, Post/MicroblogPost)
- Modelling tweets: [http://rdfs.org/sioc/types# MicroblogPost](http://rdfs.org/sioc/types#MicroblogPost)
 - sioc:content, sioc:embeds_knowledge, sioc:has_creator, sioc:has_reply, sioc:links_to, sioc:topic



A word of warning:

- Watch out for the namespaces!
- Some are from SIOC, others – SIOCT, and yet others from other imported ontologies, like SKOS
- E.g. <http://rdfs.org/sioc/ns#UserAccount>
- Vs <http://rdfs.org/sioc/types#MicroblogPost>
- In JAPE rules, you need to:
 - Either specify the complete URIs, including the namespaces (unless it is the sioct, which is the default name space for this ontology)
 - Or use templates to shorten the NS URIs and make the JAPES more readable



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, the draw for the final group match.

But the day of decision for the Greeks will be postponed to next week, as the group will have to stagger off default by honouring bonds.

The EU and the IMF have agreed to provide Greece with a €100 billion loan facility.

Type	Set	Start	End	Text
Location		1222	1228	Greece
Location		1222	1228	Argentina
Location		1222	1228	Greece
Location		1222	1228	Argentina
Location		1222	1228	Greece
Location		1222	1228	Argentina
Location		1222	1228	Greece
Location		1222	1228	Argentina
Location		1233	1241	Athens
Organization		1558	1558	EU
Organization		1558	1558	IMF

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

- 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



Ontology Learning / Population

- Ontology Population: add new facts to a given ontology. The ontology structure and many classes and individuals are already there:

“Westerwelle visits Ghana”

→ :GWesterwelle01 :actorOf :Event001 .

:Event001 a :VisitingEvent .

:Event001 :destination :Ghana .

...

- Ontology Learning: also create or extend the structure of the ontology.



Semantic Annotation: How

- Manually
GATE: ontology based annotation using OAT/RAT or through crowdsourcing
- Automatically
 - Gazetteer/rule/pattern based
GATE: OntoRoot gazetteer, LKB gazetteer, JAPE, ...
 - Classifier (ML) based – see the LODIE lecture later
 - Combination of the two



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



OAT

As well as picking MPs for Westminster, voters will elect councillors in 164 local authorities across England.

Voting in the general election will take place in 649 constituencies, with nearly 4,150 candidates 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.

Labour leader Gordon Brown went to vote shortly after 1100 BST at a community centre close to his home in North Queensferry, Fife. His wife Sarah was with him.

Nick Clegg, leader of the Liberal Democrats, arrived at a polling station in Sheffield Hallam at 1120 BST. His wife Miriam is unable to vote in the general election because she is a Spanish citizen.

The leader of the Scottish National Party, **Alex Salmond**, cast his vote shortly before noon, at Macduff in Banffshire. He is a member of the SNP in the constituency of Ynys Mon in north Wales.

Polling in one constituency - Thirsk and Marches - will be held on 5 May because of the death of one of the candidates.

ELECTION 2010 ON THE BBC

Type	Set	Start	End	Id	
Mention		1277	1289	55	{class=http

Ontology Tree(s) Options

test-ontology-instances.owl_00018

test-ontology-instances.owl_00018

- Entity
 - Location
 - Person
 - Leader
 - Nick_Clegg
 - Gordon_Brown
 - David_Cameron
 - Leader_0002A
 - Leader_0002B
 - Leader_0002C
 - Organization

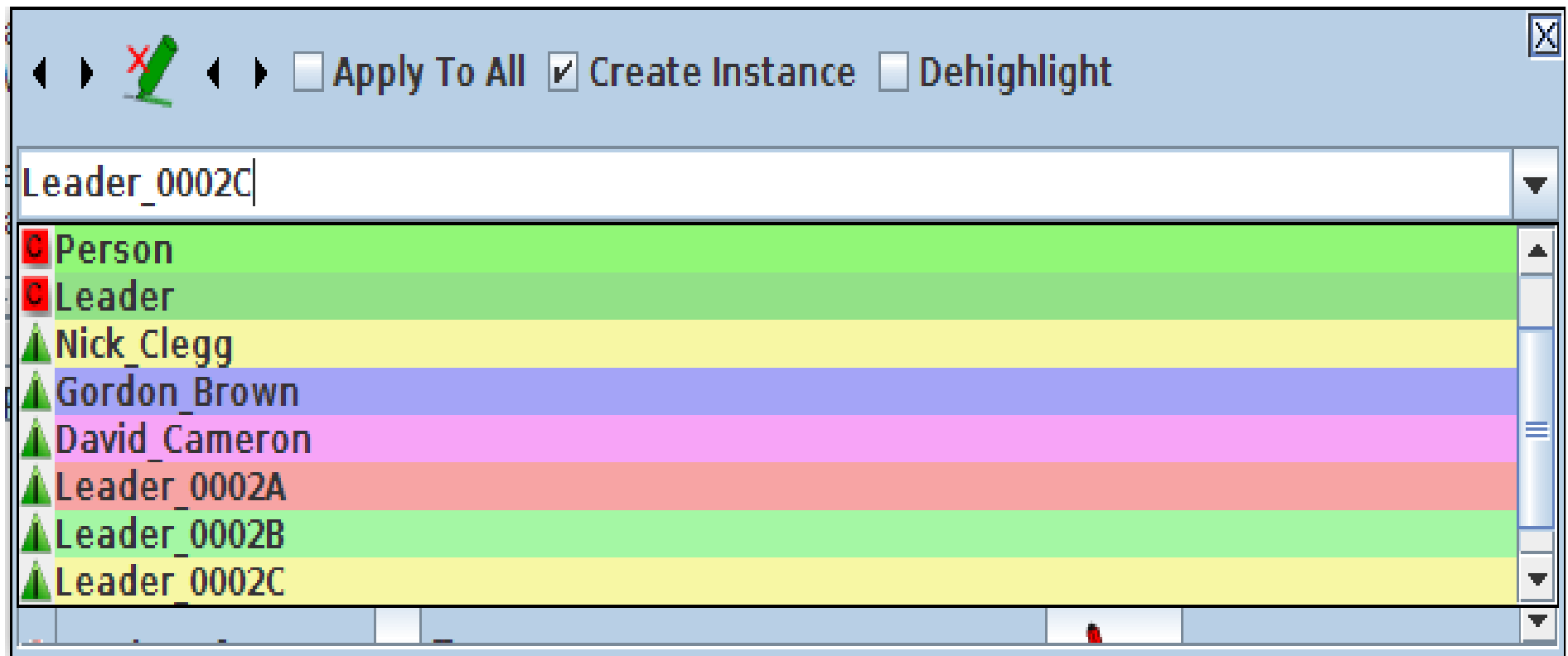
Apply To All Create Instance Dehighlight

Leader_0002C

- Person
- Leader
- Nick_Clegg
- Gordon_Brown
- David_Cameron
- Leader_0002A
- Leader_0002B
- Leader_0002C



OAT: The Editor Pop-up





Relation Annotation Tool (RAT)

- RAT annotates a document with ontology instances and creates relations between annotations by means of ontology object properties.
- It is compatible with OAT, but focuses on relations between annotations modelled as object properties
- Plugin `Ontology_Tools`
- It is comprised of 2 viewers: **RATC** (RAT-Concept) and **RATI** (Rat-Instance).
- Buttons **RATC** and **RATI** in document editor work in tandem
- The RATC pane (on the RHS) looks similar to OAT. Click the checkbox beside a class to display the relevant instances.



Adding a property value

close to his home in North Queensberry, Fife. His wife Sarah was with him.

Nick Clegg, leader of the Liberal Democrats, arrived at a polling station in Sheffield Hallam at 1120 BST. His wife Miriam is unable to vote in the general election because she is a Spanish citizen.

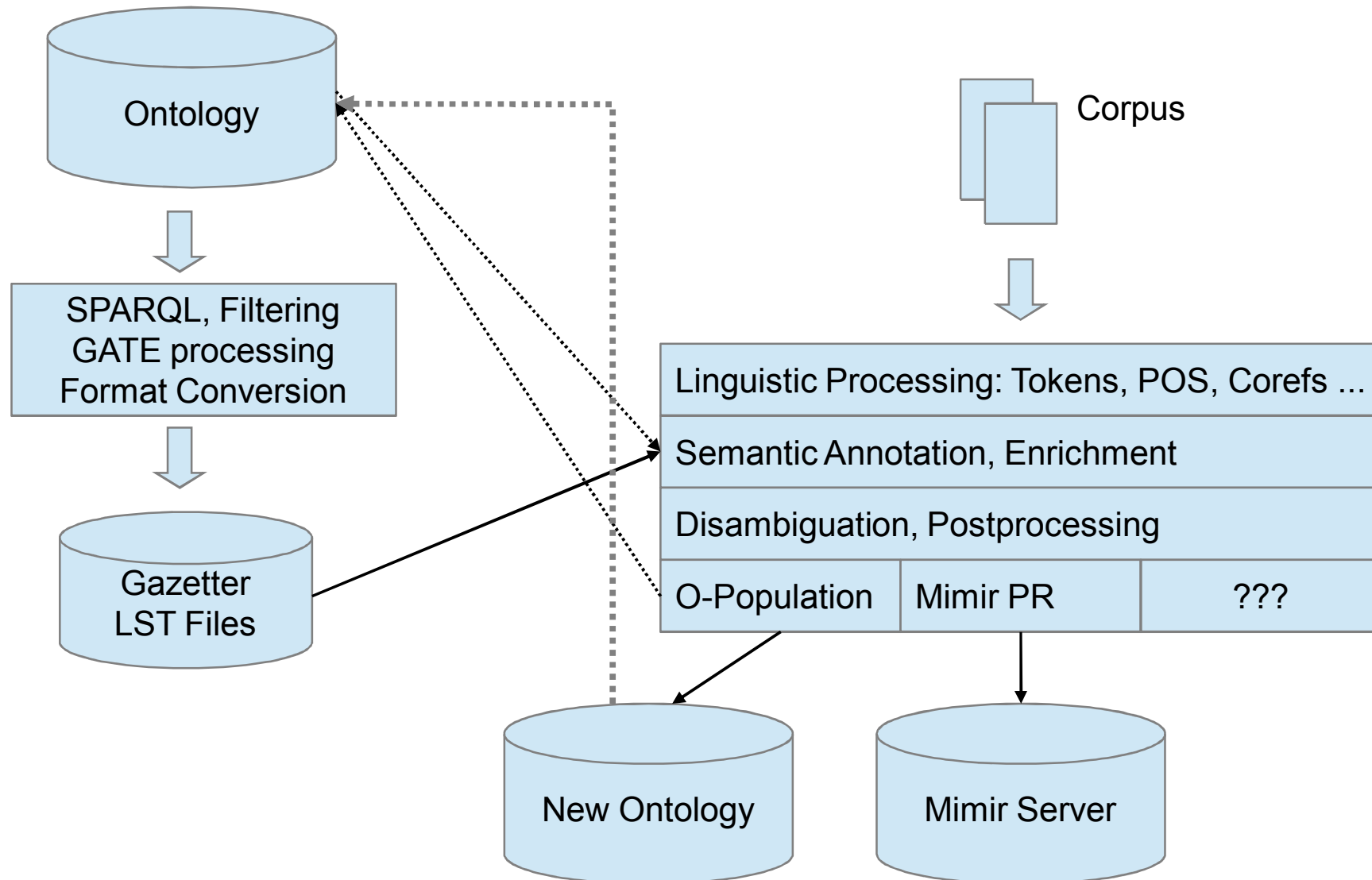
The leader of the Scottish National Party, Alex Salmond, cast his vote shortly before noon, at Macduff in Banffshire. Ieuan Wyn Jones of Plaid Cymru voted in the constituency of Ynys Mon in north Wales at lunchtime.

Filter: X

Instance	Label	Property	Value
Nick_Clegg_0	[Nick_Clegg]	person_works_for	[Organization]
		person_works_for	Liberal_Democrats



Semantic Annotation: The Big Picture



GATE: Automatic Semantic Annotation



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



LKB Gazetteer

- The LKB gazetteer is used to do ontology-based gazetteer lookup against very large ontologies, e.g. DBPedia, GeoNames and other Open Linked Data ontologies
- Uses a SPARQL query to create a gazetteer list from the ontology

```
SELECT DISTINCT ?label ?inst ?class
WHERE {
    ?inst rdf:type dbp:Country .
    ?inst foaf:name ?label .
    FILTER (lang(?label) = "en")
}
```

- Internally retrieves the result rows and converts them to gazetteer entries with inst and class features
- Creates a cache file that will load fast subsequently



LKB: Continued

- Lives in plugin Gazetteer_LKB
- LKB does not use the GATE ontology language resources. Instead, it uses its own mechanism to load and process ontologies.
- Set up your dictionary first. The dictionary is a folder with some configuration files. Use the samples at `GATE_HOME/plugins/Gazetteer_LKB/samples` as a guide or download a pre-built dictionary from ontotext.com/kim/lkb_gazetteer/dictionaries.
- The dictionary directory defines which repository to connect to, which SPARQL queries to use to initialise the gazetteer, etc.
- For details see

<http://gate.ac.uk/userguide/sec:gazetteers:lkb-gazetteer>



LKB: Example

- Samples in [gate/plugins/Gazetteer_LKB/samples/dictionary_from_remote_repository](#)
- An ontology-based gazetteer of actors from Dbpedia

Query:

```
1 SELECT ?Name ?Person ?Cls
2 FROM <http://www.ontotext.com/disable-sameAs>
3 WHERE {
4     ?Person a ?Cls ; rdfs:label ?Name .
5     FILTER (lang(?Name) = "en")
6     FILTER (?Cls = <http://dbpedia.org/ontology/Actor>)
7 }
```

- Test this query against <http://ldsr.ontotext.com/sparql>

SPARQL Query Results

[Home](#) > SPARQL Query

Results for [PREFIX rdfs:...](#) (100 of 850)

View as [Exhibit](#) Download SPARQL |

Name	Person	Cls
Jet Li@en	dbpedia:Jet_Li	dbp-ont:Actor
Tom Cruise@en	dbpedia:Tom_Cruise	dbp-ont:Actor
Cruise, Tom@en	dbpedia:Tom_Cruise	dbp-ont:Actor
Bruce Lee@en	dbpedia:Bruce_Lee	dbp-ont:Actor
Lee Armstrong@en	dbpedia:Lee_Armstrong	dbp-ont:Actor
Johnny Depp@en	dbpedia:Johnny_Depp	dbp-ont:Actor
Depp, Johnny@en	dbpedia:Johnny_Depp	dbp-ont:Actor
Zhang Ziyi@en	dbpedia:Zhang_Ziyi	dbp-ont:Actor
Chow Yun-fat@en	dbpedia:Chow_Yun-fat	dbp-ont:Actor
Tsui Hark@en	dbpedia:Tsui_Hark	dbp-ont:Actor
Sammo Hung@en	dbpedia:Sammo_Hung	dbp-ont:Actor

LKB: Try it

- Samples in `gate/plugins/Gazetteer_LKB/samples/dictionary_from_remote_repository`
- Load the ready-made application `sample_linked_data_mashup.gapp`
- This should load the Movie stars pipeline application
- Temporarily move away the LDSR Enrichment PR from the pipeline, leaving just the documents reset and the entertainers gazetteer
 - that's pre-built from the SPARQL query shown on the previous page
- Run the pipeline on the sample corpus and inspect the Lookup annotations

ear found Ricky at Golden Harvest with a leading role in John Woo 's Money Crazy . In 1979 Games Gamblers Play was released in the Japanese market. For this edition Michael shot a new scene, a fight between Ricky and Sam on the beach, and replaced the original Sammo Hung vs Sam Hui fight with it. The next Hui brothers production where Ricky teamed up with his brothers again. He has also released seven albums, most of them on vinyl in the 1970s and 1980s. There are three Ricky albums on successful films featuring the Hui brothers production where Ricky teamed up with his brothers again. He later received Super Model and Forever Young Music [edit]

Michael became a producer in 1990. He later received Super Model and Forever Young Music [edit]

Ricky was most active in his film career in the 1970s and 1980s. There are three Ricky albums on successful films featuring the Hui brothers production where Ricky teamed up with his brothers again. He later received Super Model and Forever Young Music [edit]

Hui has also released seven albums, most of them on vinyl in the 1970s and 1980s. There are three Ricky albums on



class	inst	
http://dbpedia.org/ontology/Actor	http://dbpedia.org/resource/Sammo_Hung	X
		X

▶ Open Search & Annotate tool



Other Gazetteers

- ExtendedGazetteer from StringAnnotation plugin (<http://code.google.com/p/gateplugin-stringannotation/>)
 - can handle arbitrary features
- it can work both on the document text or on the value of features without wrapping a flexible gazetteer around it
- Requires tokenized text because it uses the Tokens (or some other annotation) to find word boundaries
- Default field separator: the tab character. Gazetteer lists which have entries and then features separated by tabs work by default. You can override that default separator
- Two runtime parameters to choose if a match should occur at a word boundary or only at the beginning and at the end



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
```

Matches the class Person
or any of its subclasses

```
Rule: PersonLookup  
(  
  {Lookup.class == Person}
```

```
):person
```

```
-->
```

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

Adds class and instance information
as features on the Mention annotation



Ontology-aware JAPE example

Ontology-aware JAPE applies only to a feature named “class” and only if the PR's ontology parameter is set.

```
{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!

These rules apply equally to the string in the JAPE rule and in the value of the annotation's class feature.



Templates to simplify namespaces

Template declarations can be used to simplify namespaces.

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

If you switch to a newer version of PROTON, you only need to change the Template declarations, not every JAPE LHS. (See the GATE User Guide <http://gate.ac.uk/userguide/sec:jape:templates> for more details and examples.)

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



Matching subclasses

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

Lookup			
C	URI	http://gate.ac.uk/example#David_Cameron	X
C	class	http://gate.ac.uk/example#Leader	X
C	classURI	http://gate.ac.uk/example#Leader	X
C	classURIList	[http://gate.ac.uk/example#Leader]	X
C	heuristic_level	0	X
C	inst	http://gate.ac.uk/example#David_Cameron	X
C	majorType		X
C	type	instance	X



The rule matches because Leader is a subclass of Person



Semantic Enrichment

- Add additional knowledge to semantically annotated mentions
- Simplest: add features
e.g. add the name of the country, zip code for a city
→ if we have city names to disambiguate, may use zip code to disambiguate!
- Use Java API in JAPE RHS, Groovy or own PR
- SemanticEnrichment PR from the Gazetteer_LKB plugin
 - SPARQL Endpoint (not GATE Ontology LR)
 - Run SPARQL query for each URI in inst
 - add query result to 'connections' feature



Semantic Enrichment PR

- Adding new data to semantic annotations by querying external RDF (Linked Data) repositories
- A semantic annotation is an annotation that is linked to an RDF entity by having the URI of the entity in the 'inst' feature of the annotation
- This PR runs a SPARQL query against a given repository and puts a comma-separated list of the values mentioned in the query output in the 'connections' feature of the annotation
- Run-time parameters:
 - List of annotation types to enrich and input AS
 - Delete on no relations (**true/false**)
 - Query



Hands On: Semantic Enrichment

- Add the LDSR Enrichment PR back into your pipeline, making sure it is last
- Run the pipeline on the sample corpus and inspect again the Lookup annotations, especially their **connections** feature
- You will need internet connection for this to work

A screenshot of the GATE software interface. The top part shows a configuration window for a Lookup annotation. It includes fields for "Previous boundary", "Next boundary", an "Overlapping" checkbox, and a "Target set: Undefined" dropdown. Below these are "Context" and "Lookup" fields. The "Context" field contains the text "Hui - Wikipedia, the free encyclopediaRicky HuiFrom Wikipedia, the free". The "Lookup" field contains a yellow box with a mouse cursor pointing to it. Below the configuration window, a list of connections is displayed in a blue background, including "a.org/ontology/Actor", "a.org/resource/Guangdong,http://rdf.freebase.com/ns/en.guangdong_province_china,http://data.nytimes.com/guangdong_province_china_geo,", and "a.org/resource/Ricky_Hui".

- How do results change, if you modify the query to say LIMIT 1, instead of LIMIT 10?



QUESTIONS?