Introduction to Text Mining

Module 2: Semantic Annotation

Aims of this module

- Introduce the essential concepts of ontologies and the semantic web
- Show how traditional NLP techniques such as information extraction can be “made semantic”
- Show how ontologies can be used to fulfil more complex information needs
- Use GATE as an example NLP toolkit to demonstrate real life applications for semantic web development

Motivation

- In module 1, we learnt that NLP techniques such as information extraction, disambiguation, term recognition etc. give meaning to unstructured text
- When this meaning is linked to an ontology, it
  - becomes reusable across the Semantic Web
  - enables processes such as reasoning to be carried out
- Ultimately, Linked Open Data enables this information to be exposed, shared and connected via dereferencable URIs
- All this leads eventually to the Holy Grail of knowledge understanding

"I got it on eBay."

Why do we need ontologies?

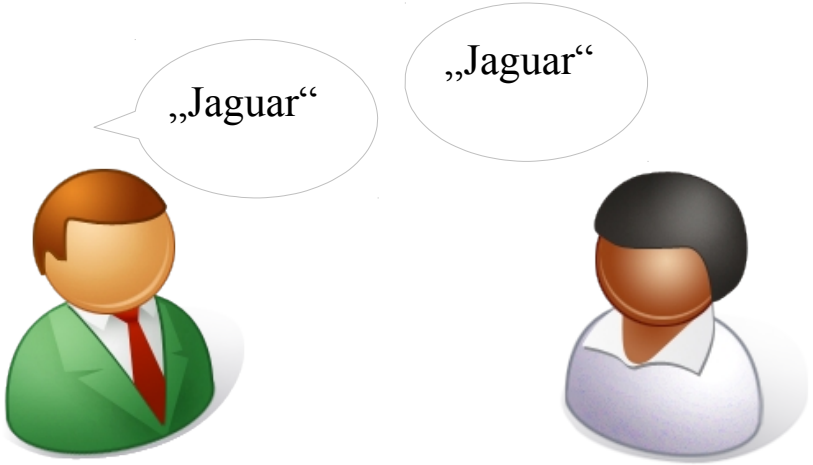
- share common understanding of the structure of information among people or software agents
- enable reuse of domain knowledge
- make domain assumptions explicit
- separate domain knowledge from the operational knowledge
- analyze domain knowledge

[Noy and McGuinness 2001]

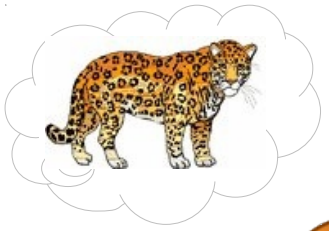
What is an Ontology?

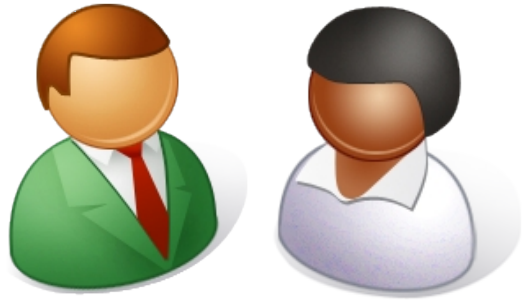
- Set of concepts (instances and classes)
- Relationships between them (is-a, part-of, located-in)
- Multiple inheritance
  - Classes can have more than one parent
  - Instances can have more than one class
- Ontologies are graphs, not trees

Example: Common Understanding

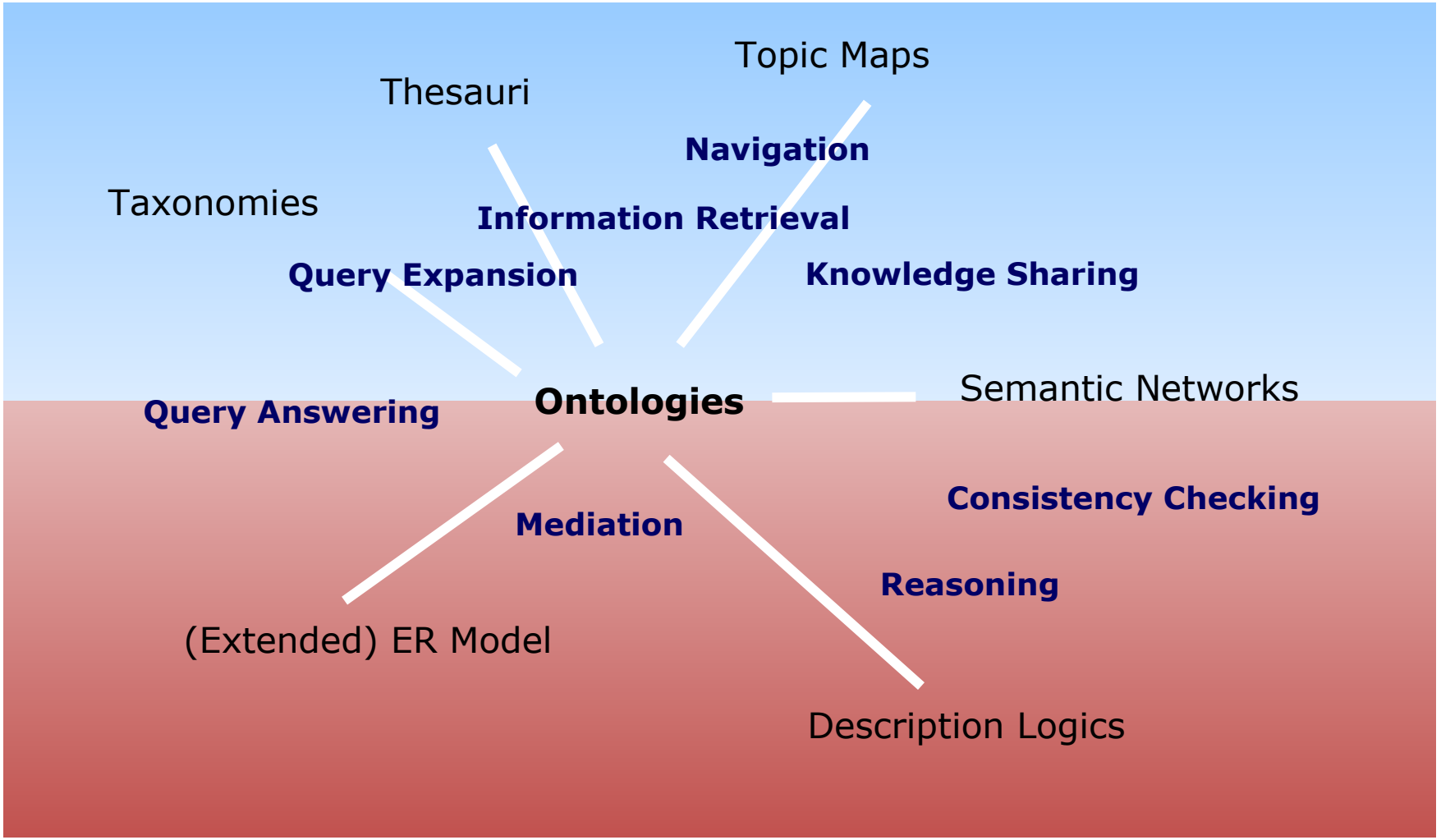


http://en.wikipedia.org/wiki/SEPECAT_Jaguar





Ontologies on the web

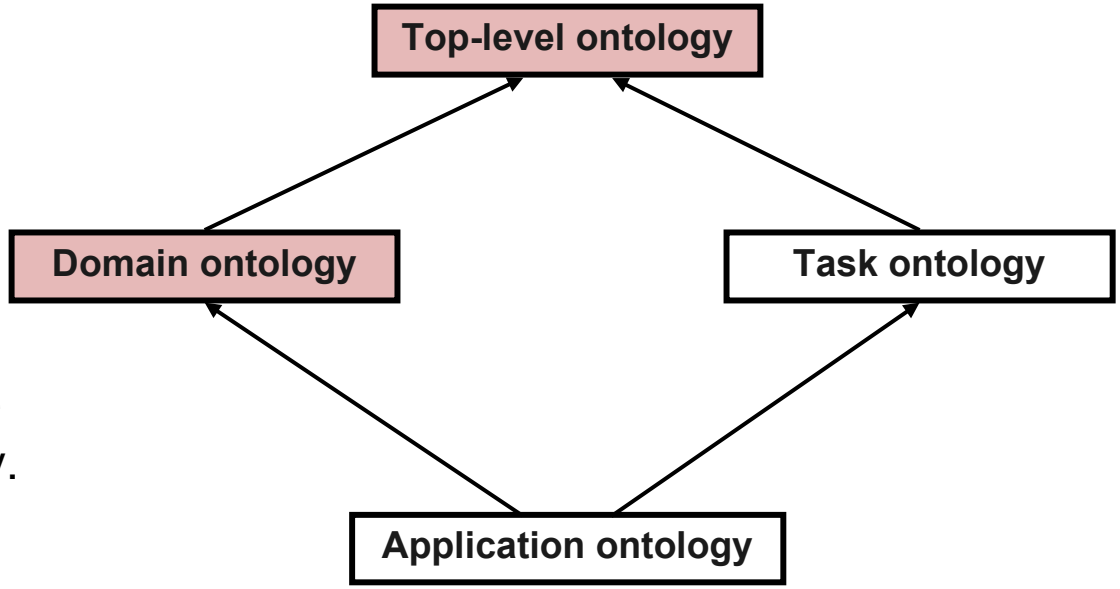


[Guarino, 1998]

Types of ontologies

Describe **very general concepts** like space, time, event, which are independent of a particular problem or domain. It seems reasonable to have unified top-level ontologies for large communities of users.

Describe the vocabulary related to a **generic domain** by specializing the concepts introduced in the top-level ontology.

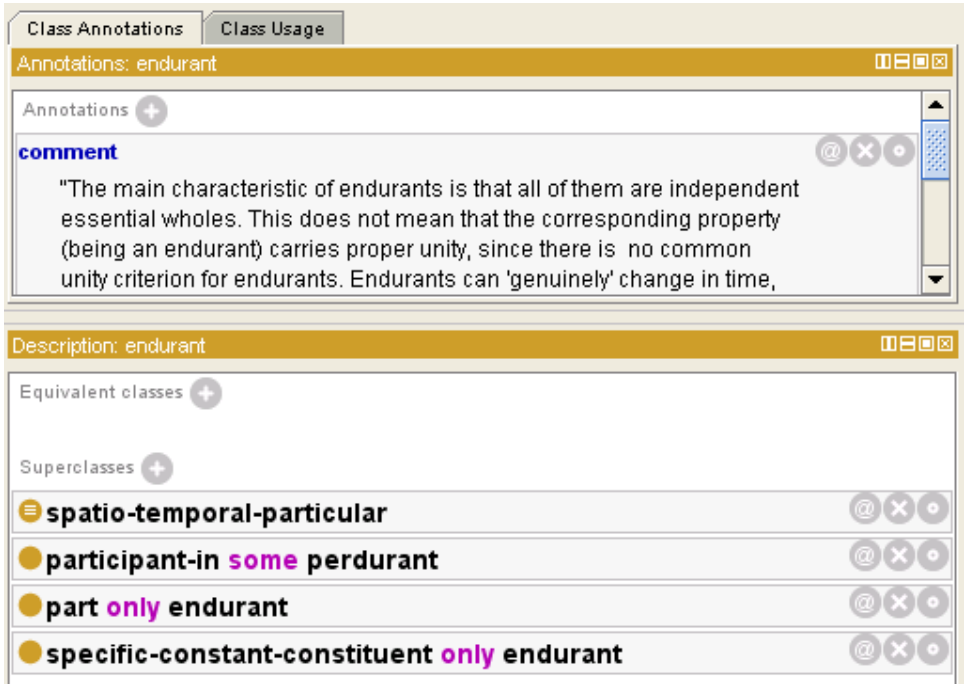


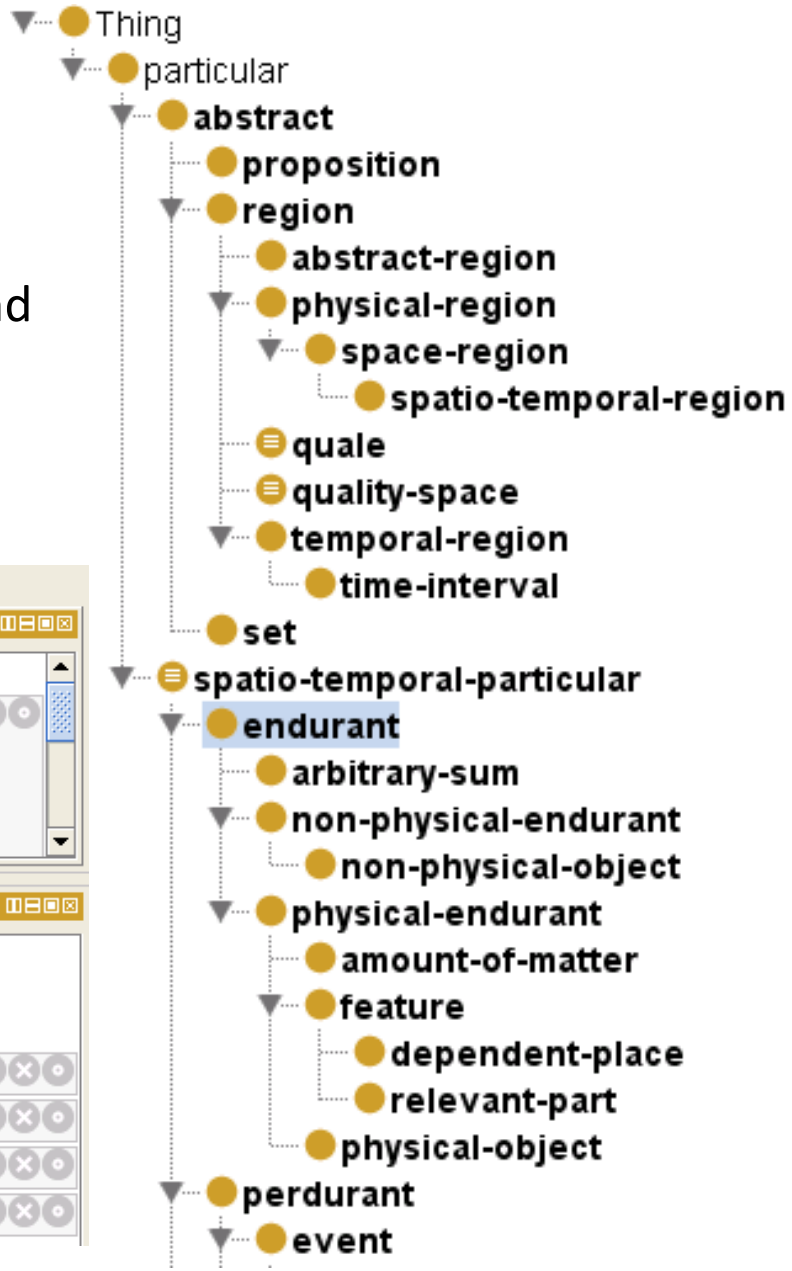
Describe the vocabulary related to a **generic task or activity** by specializing the top-level ontologies.

These are the most specific ontologies. Concepts in application ontologies often correspond to **roles played by domain entities while performing a certain activity**.

DOLCE

Descriptive Ontology for Linguistic and Cognitive Engineering
http://www.loa-cnr.it/DOLCE.html

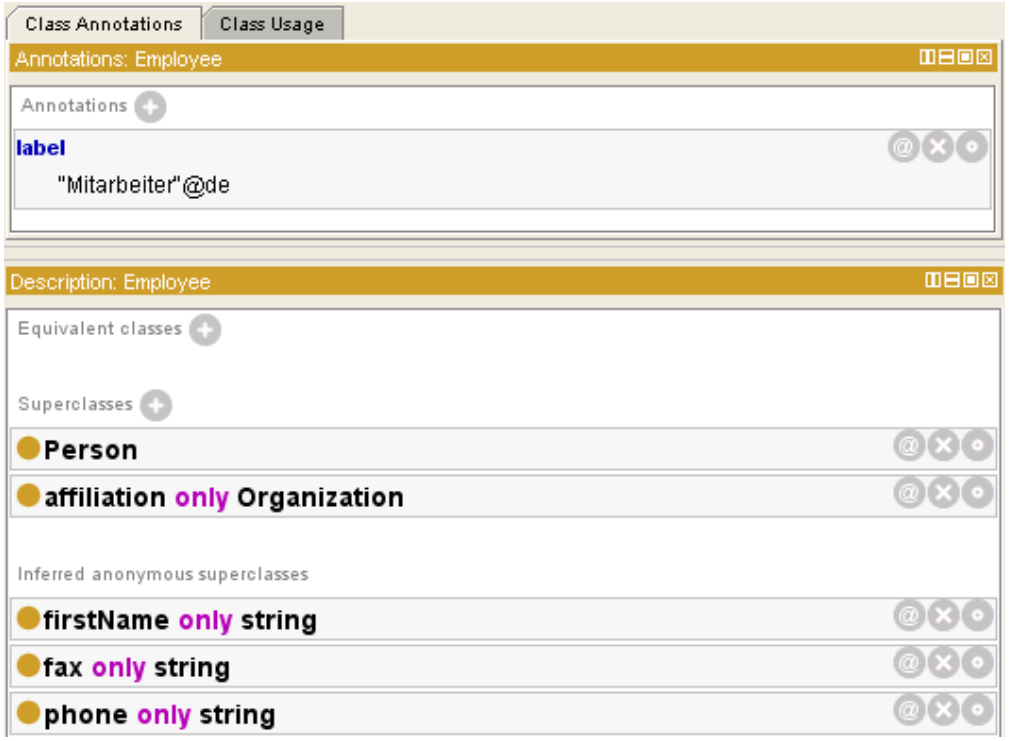




SWRC

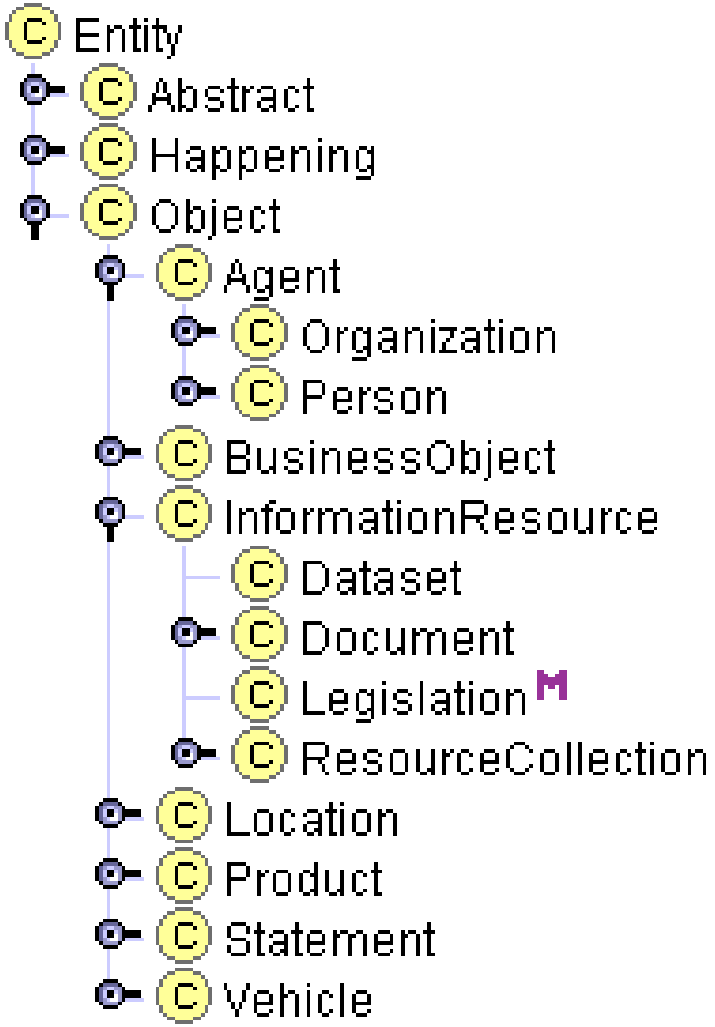
Semantic Web for Research Communities

http://ontoware.org/projects/swrc/



PROTON

- a lightweight upper-level ontology developed by Ontotext
- 250 NE classes
- 100 relations and attributes
- covers mostly NE classes, and ignores general concepts

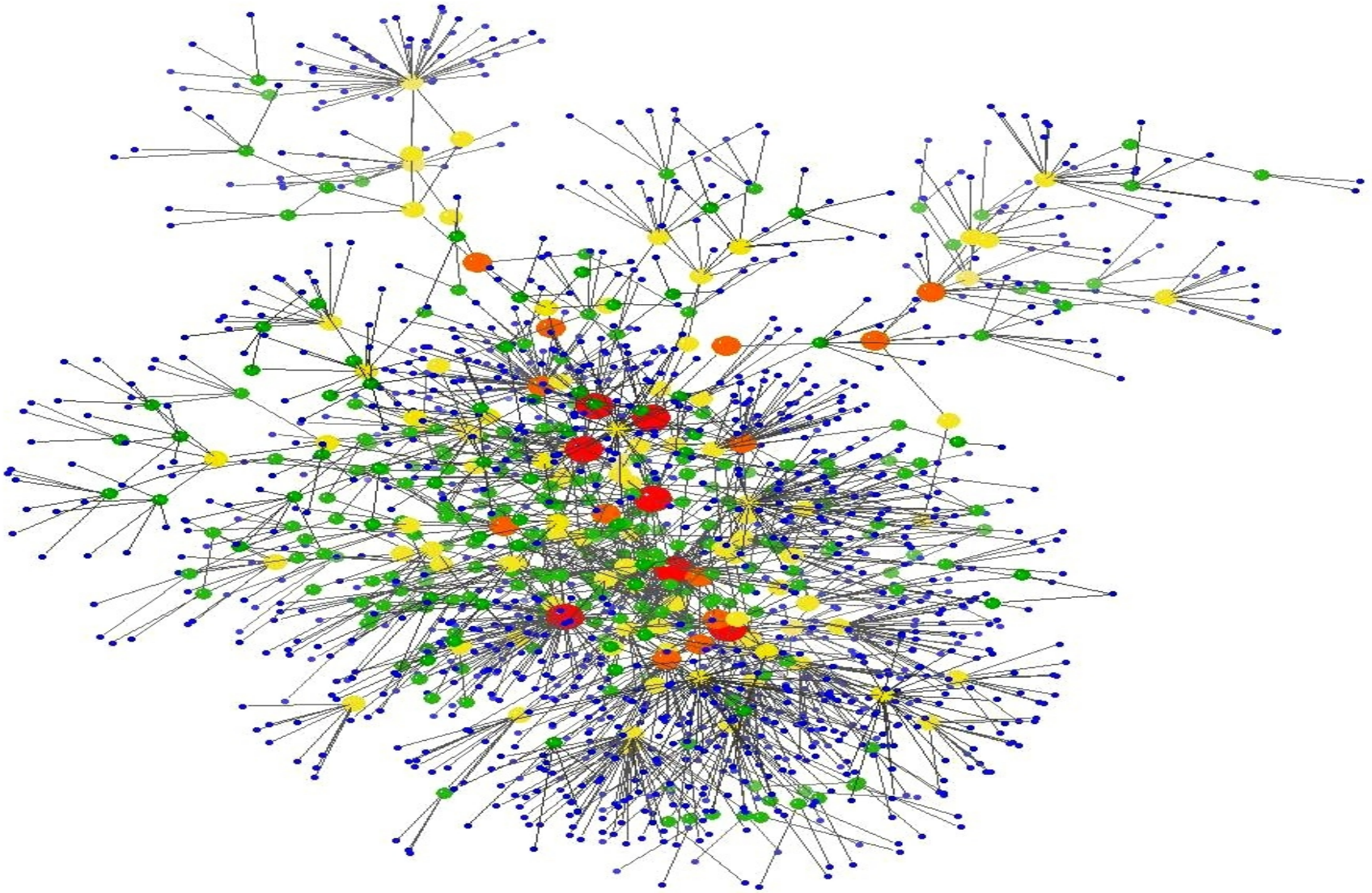


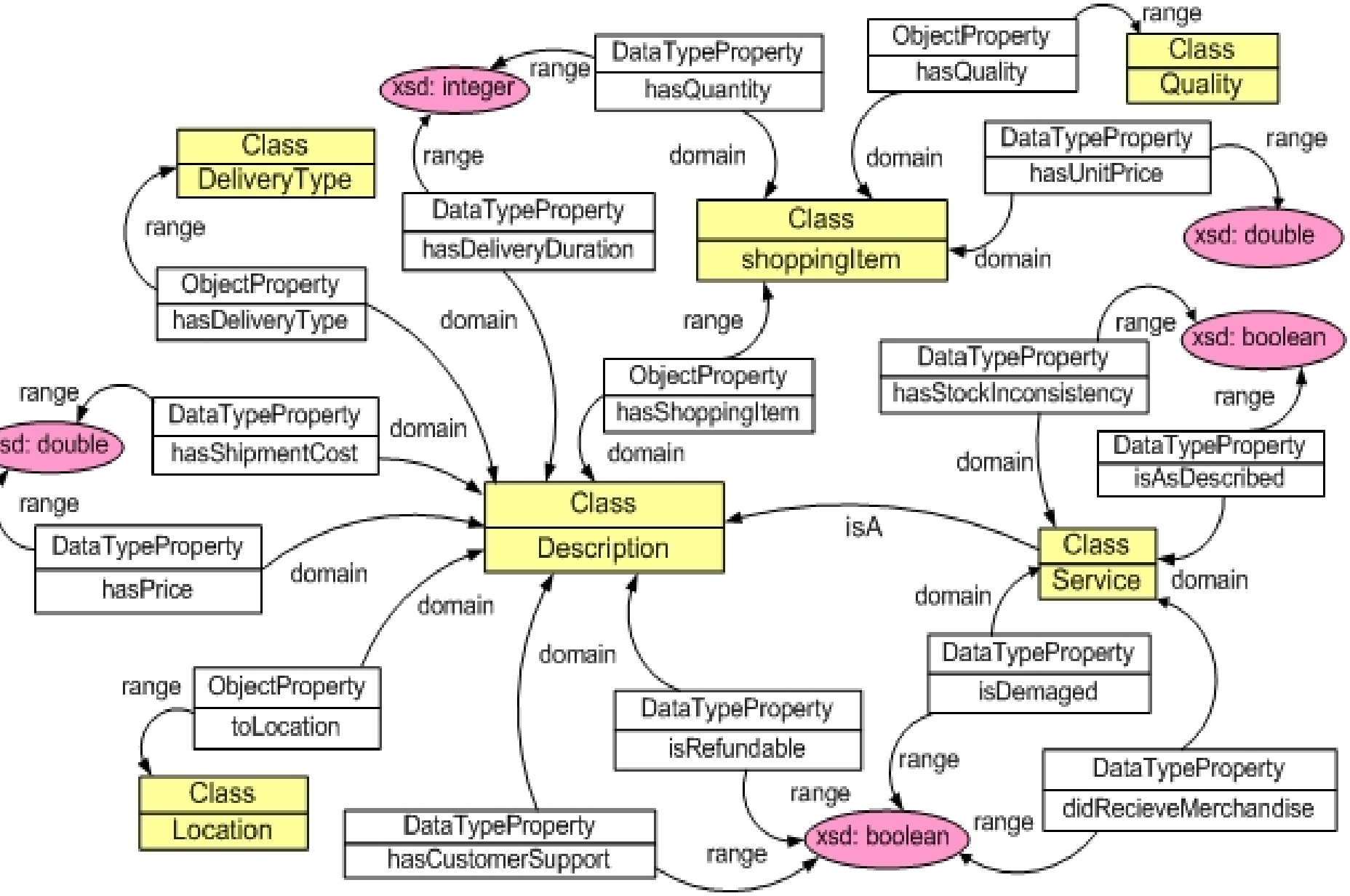
http://proton.semanticweb.org

Other Ontologies

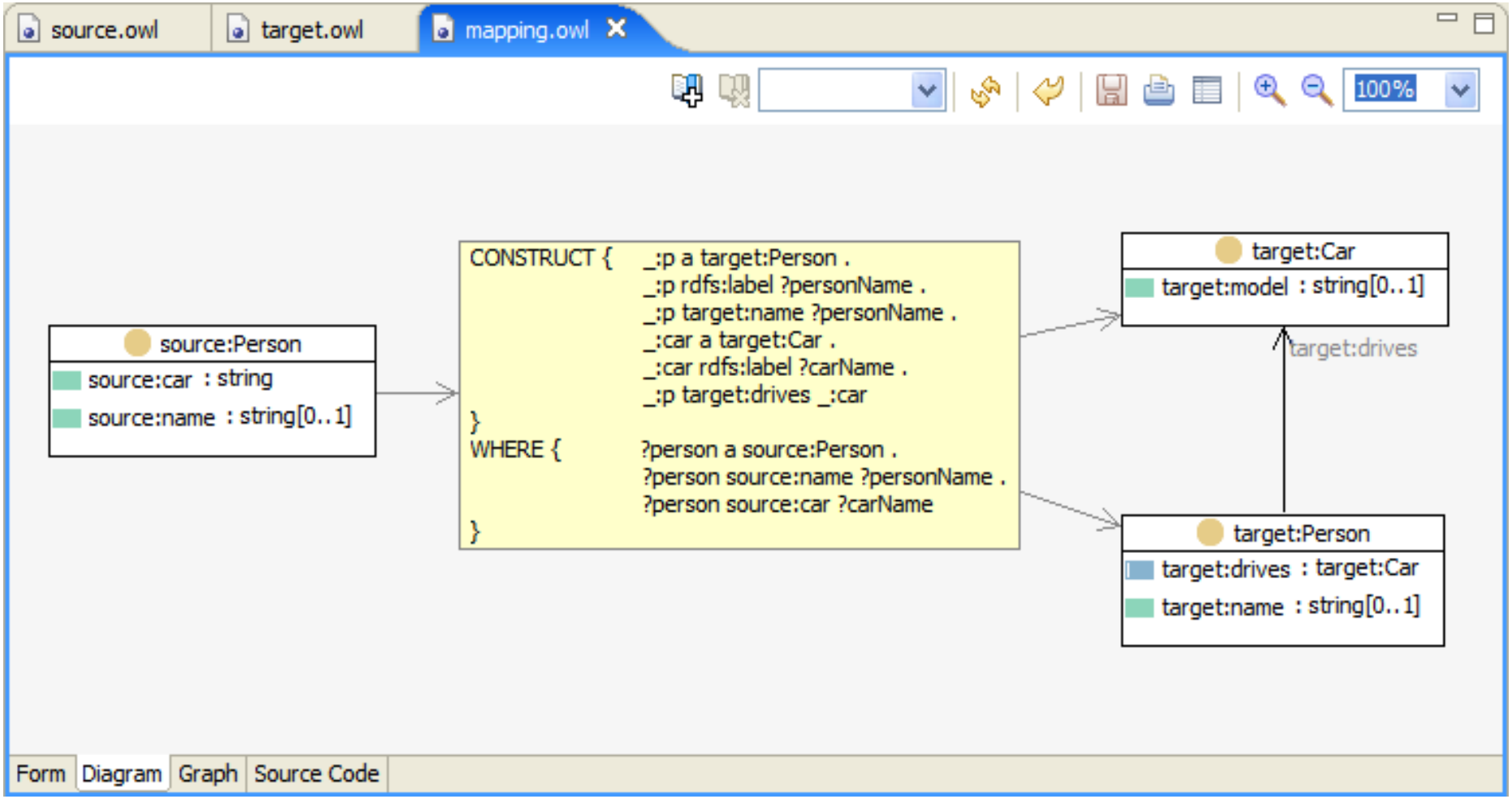
SUMO: http://www.ontologyportal.org
CYC: http://www.cyc.com
WordNet: http://wordnet.princeton.edu
Dublin Core: http://dublincore.org
Pizza: http://www.co-ode.org/ontologies/pizza/2007/02/12/
Time: http://www.w3.org/TR/2006/WD-owl-time-20060927/
Gene Ontology: http://www.geneontology.org
GALEN: http://www.openclinical.org/prj_galen.html
NCI Thesaurus: http://nciterms.nci.nih.gov
…

Ontologies can be scary monsters




DataTypeProperty
hasQuantity
range
xsd: integer
ObjectProperty
hasQuality
range
Class
Quality
domain
range
Class
DeliveryType
DataTypeProperty
hasUnitPrice
range
xsd: double
DataTypeProperty
hasDeliveryDuration
range
domain
Class
shoppingItem
domain
range
ObjectProperty
hasDeliveryType
domain
range
xsd: boolean
DataTypeProperty
hasStockInconsistency
range
ObjectProperty
hasShoppingItem
domain
range
sd: double
DataTypeProperty
hasShipmentCost
domain
range
DataTypeProperty
isAsDescribed
domain
range
Class
Description
isA
Class
Service
DataTypeProperty
hasPrice
domain
domain
domain
domain
domain
range
ObjectProperty
toLocation
DataTypeProperty
isRefundable
DataTypeProperty
isDemaged
range
range
range
Class
Location
DataTypeProperty
hasCustomerSupport
range
xsd: boolean
range
DataTypeProperty
didRecieveMerchandise

Querying them can look scary too....



You may now be wondering if you're in the right place...

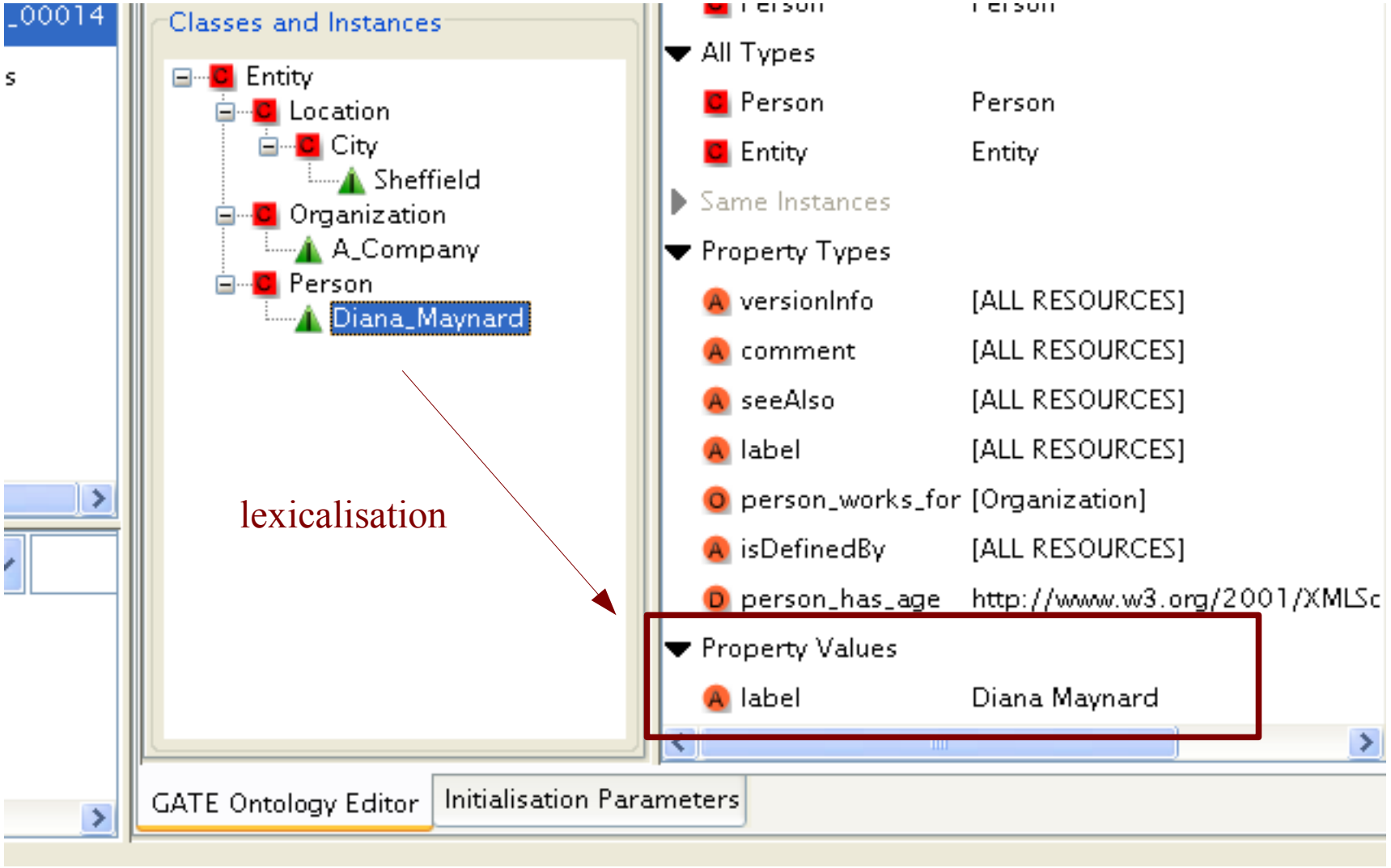
Some basics about ontologies

- For those who don't know anything about the Semantic Web, the whole concept can be rather confusing
- Luckily, you don't actually need to know much about ontologies to use them for semantic annotation
- But you do need to understand a few basics:
  - class, instance, property
  - how you can relate annotations to ontologies
  - what you can do with a semantically annotated corpus

URIs, Labels, Comments

- The name of a class or instance is a URI (Universal Resource Identifier), e.g. http://gate.ac.uk/example#Person
- URIs cannot have spaces: use underscores instead
- 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**
- These apply to all classes in the ontology

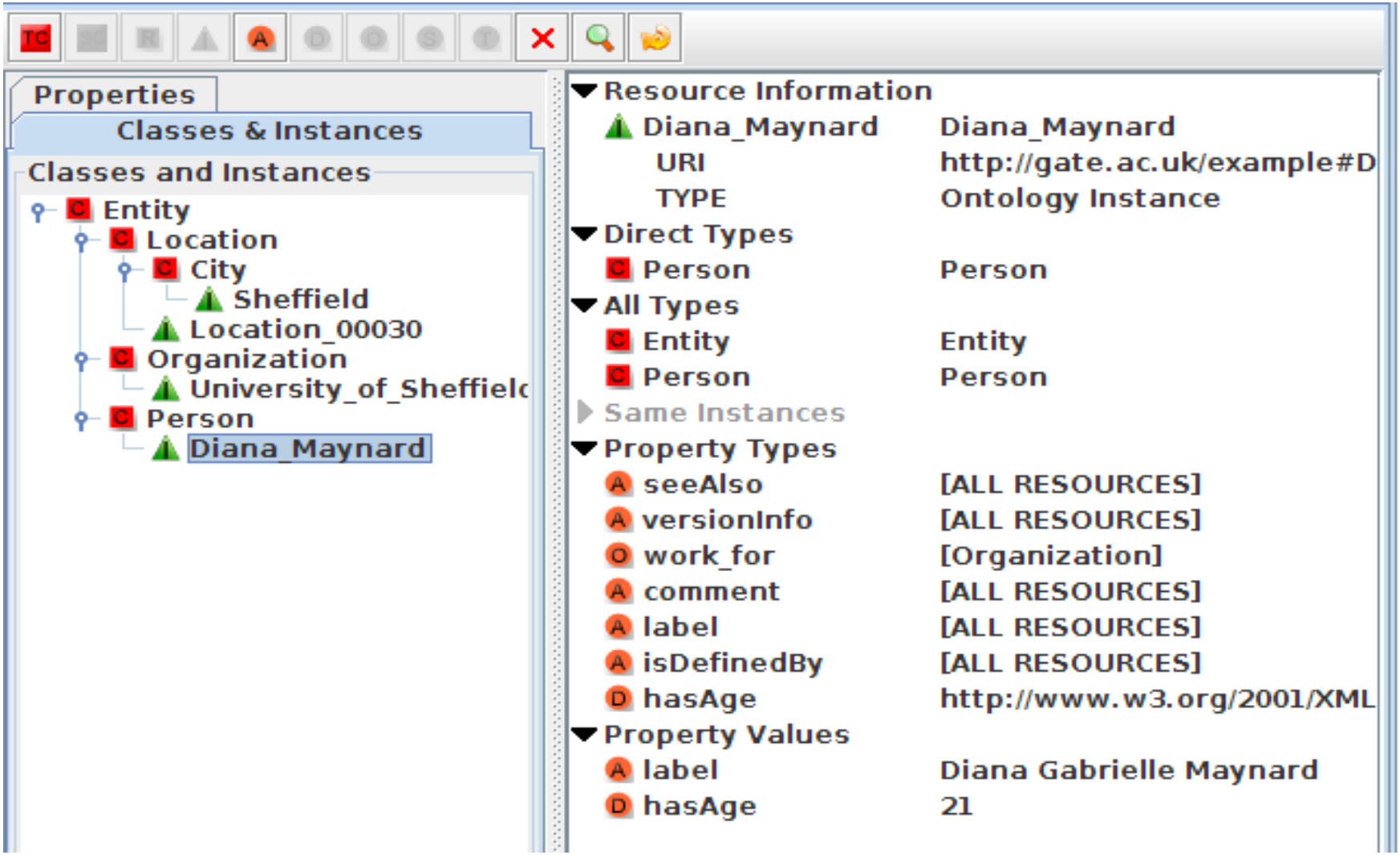
New label


Classes and Instances
Entity
Location
City
Sheffield
Organization
A_Company
Person
Diana_Maynard
lexicalisation
All Types
Person Person
Entity Entity
Same Instances
Property Types
versionInfo [ALL RESOURCES]
comment [ALL RESOURCES]
seeAlso [ALL RESOURCES]
label [ALL RESOURCES]
person_works_for [Organization]
isDefinedBy [ALL RESOURCES]
person_has_age http://www.w3.org/2001/XMLSc
Property Values
label Diana Maynard
GATE Ontology Editor
Initialisation Parameters

Datatype Properties

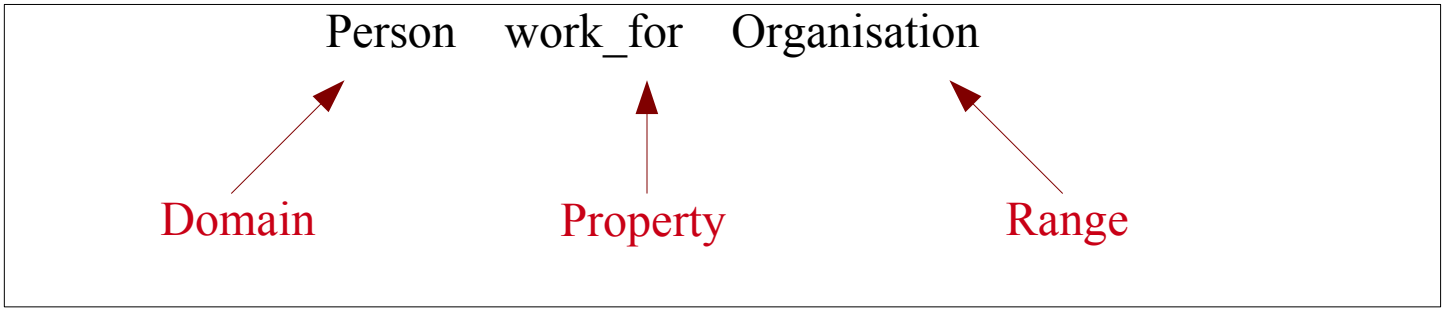
- Datatype properties link instances with data values
- Datatype properties can be of type: boolean, date, integer
- For example, a person could have an age property, e.g. “hasAge”
- The **domain** is the name of the class to whose instances the property applies
- In this case, the domain would be the “Person” class
- If more than one class is listed as a domain, these restrict the property to those individuals that belong to the intersection of the class descriptions
- The **value** of the age property would be the integer denoting the age of the person in question
- Basically, this equates to saying that everything that's a Person can have an age, and (in our example) the age of “Diana Maynard” is 21.

Datatype property

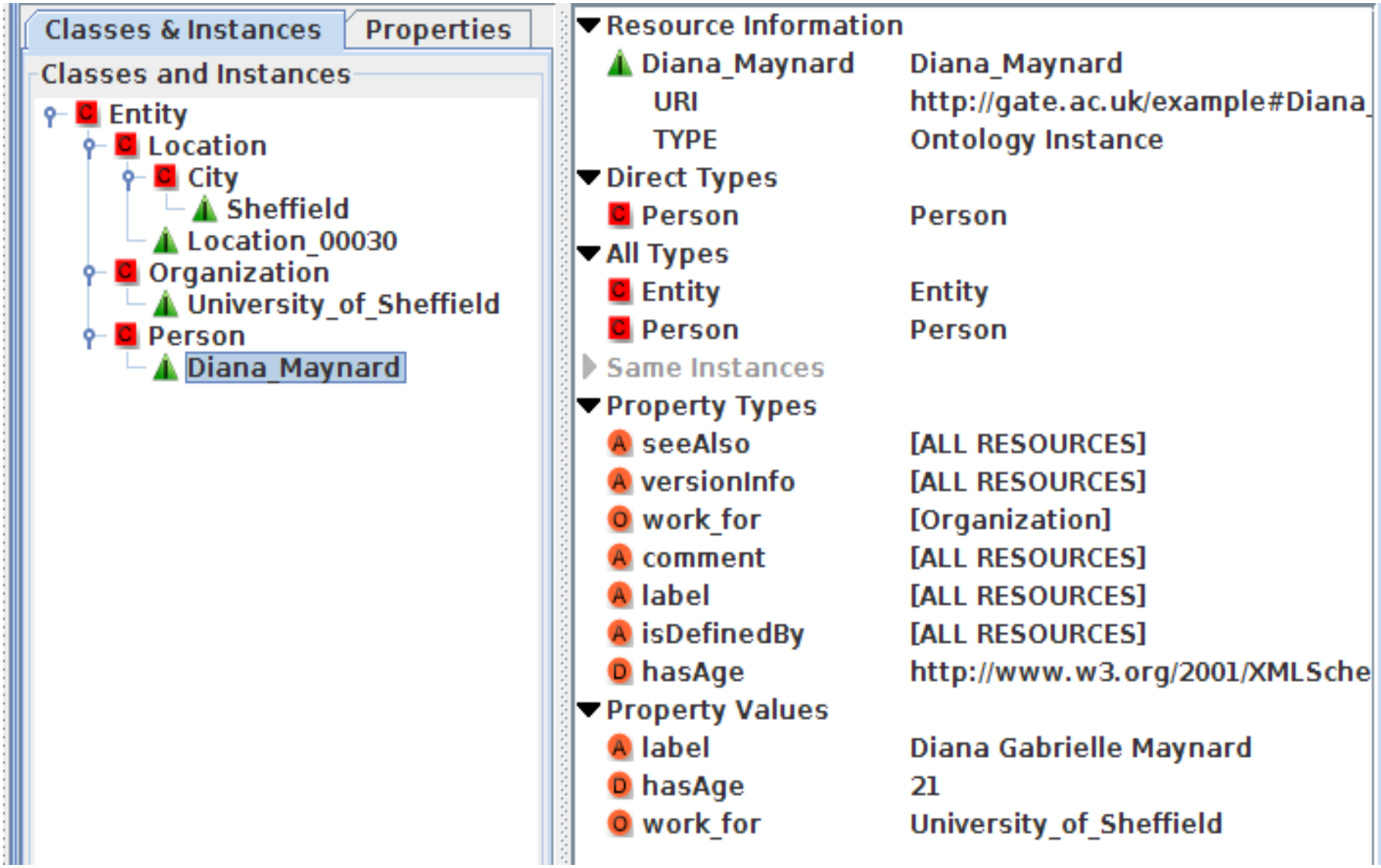


Object Properties

- 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”)



Object Property



Ontology Design Principles

- There are many ways to encode a domain in an ontology – use your application needs as a guide
- Ontology authoring is often iterative and evolves with your text analysis application
- Classes vs instances: this can vary, but as a rough guide, proper nouns are usually instances, common nouns are usually classes
- Level of granularity: what subclasses do you need?
    - e.g do organisations need subclasses such as government, education, charity, etc?
- Domains and ranges:
    - Make sure they are generic enough, but not too generic
    - Information is propagated downwards, so don't add both a class and its subclasses as domain/range
- Avoid using unions of classes for domains or ranges when a common superclass is available

Semantic Annotation

Information Extraction for the Semantic Web

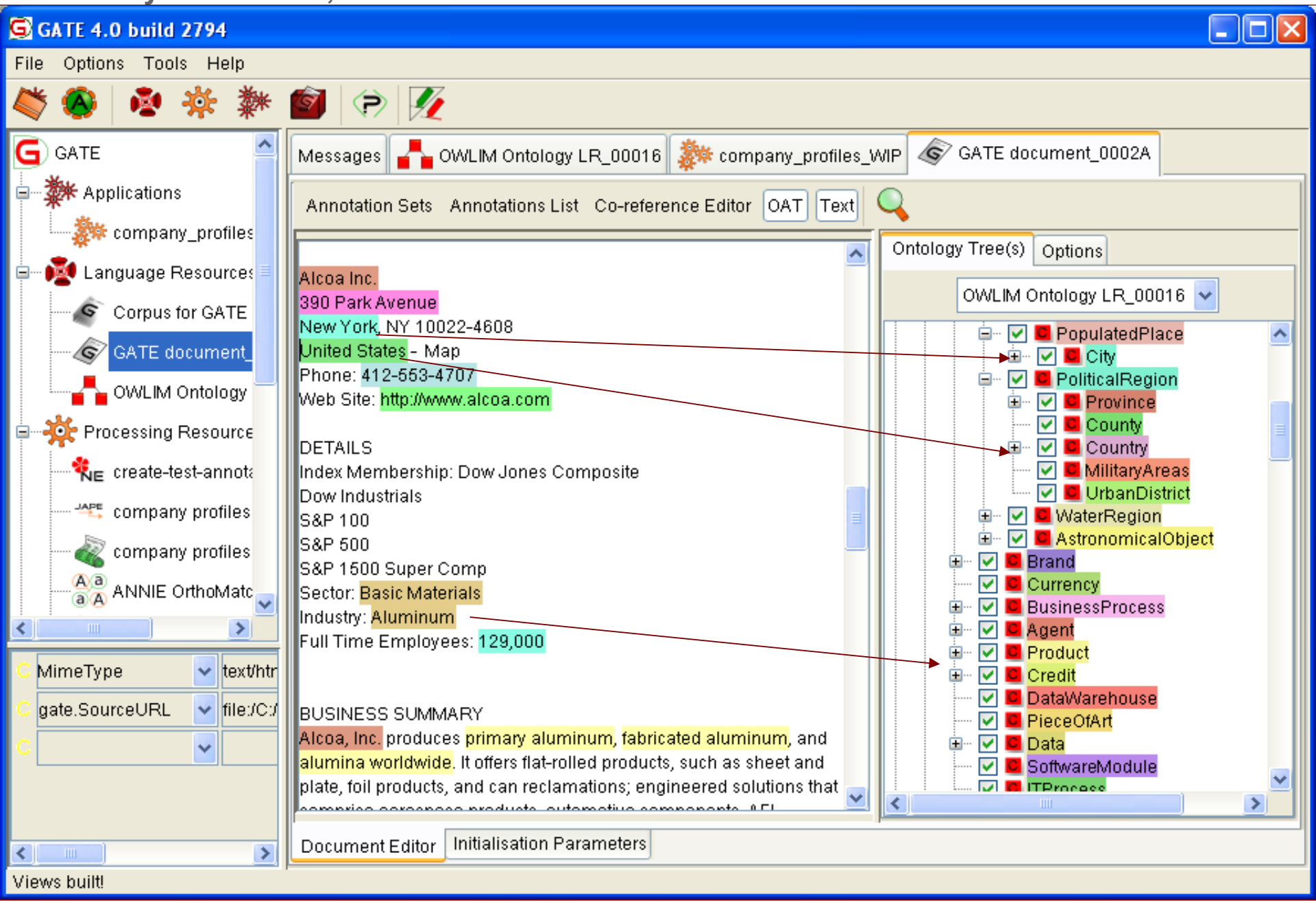
- Traditional IE is based on a flat structure, e.g. recognising Person, Location, Organisation, Date, Time etc.
- For the Semantic Web, we need information in a hierarchical structure
- Idea is that we attach semantic metadata to the documents, pointing to concepts in an ontology
- Information can be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

Semantic Annotation for Business Intelligence

- This application from the EU Musing project demonstrates how we can make use of ontology-based information extraction for real-life business intelligence
- Risk analysis, e.g. which companies make good investments, mergers etc.
- Region selection, e.g. where is the best place for internationalisation efforts (import and export, setting up a call centre, company mergers etc)

Extracting Company Information

- Identify Company Name, Address, Parent Organization, Shareholders..
- These associated pieces of information should be asserted as property values of the company instance
- Statements for populating the ontology need to be created
  - “Alcoa Inc” hasAlias “Alcoa”
  - “Alcoa Inc” hasWebPage “http://www.alcoa.com”


GATE 4.0 build 2794
File Options Tools Help
GATE
Applications
company_profiles
Language Resources
Corpus for GATE
GATE document_
OWLIM Ontology
Processing Resource
create-test-annota
company profiles
company profiles
ANNIE OrthoMatc
MimeType
text/htr
gate.SourceURL
file:/C:/
Messages
OWLIM Ontology LR_00016
company_profiles_WIP
GATE document_0002A
Annotation Sets
Annotations List
Co-reference Editor
OAT
Text
Alcoa Inc.
390 Park Avenue
New York, NY 10022-4608
United States - Map
Phone: 412-553-4707
Web Site: http://www.alcoa.com
DETAILS
Index Membership: Dow Jones Composite
Dow Industrials
S&P 100
S&P 500
S&P 1500 Super Comp
Sector: Basic Materials
Industry: Aluminum
Full Time Employees: 129,000
BUSINESS SUMMARY
Alcoa, Inc. produces primary aluminum, fabricated aluminum, and alumina worldwide. It offers flat-rolled products, such as sheet and plate, foil products, and can reclamations; engineered solutions that
Ontology Tree(s)
Options
OWLIM Ontology LR_00016
PopulatedPlace
City
PoliticalRegion
Province
County
Country
MilitaryAreas
UrbanDistrict
WaterRegion
AstronomicalObject
Brand
Currency
BusinessProcess
Agent
Product
Credit
DataWarehouse
PieceOfArt
Data
SoftwareModule
Document Editor
Initialisation Parameters
Views built!

Region Selection Application

- Idea is to find automatically where the best location is for a particular type of business internationalisation
- The user specifies various facts about the business and goal, e.g. export, direct investment, alliance, company size and type
- A number of social, political, geographical and economic indicators or variables about the regions are collected by the system, e.g. surface area, labour costs, tax rates, population, literacy rates, etc.
- These then are fed into a statistical model which calculates a ranking of the most suitable regions for the business

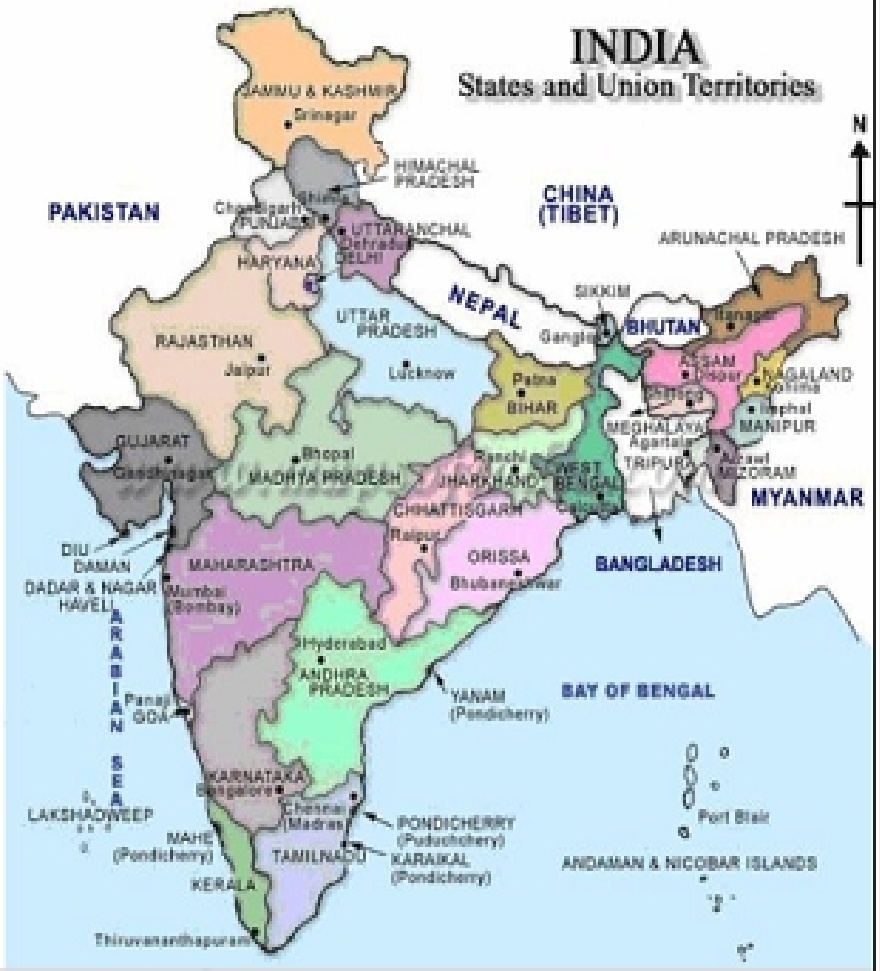
Region Selection Pilot

New search

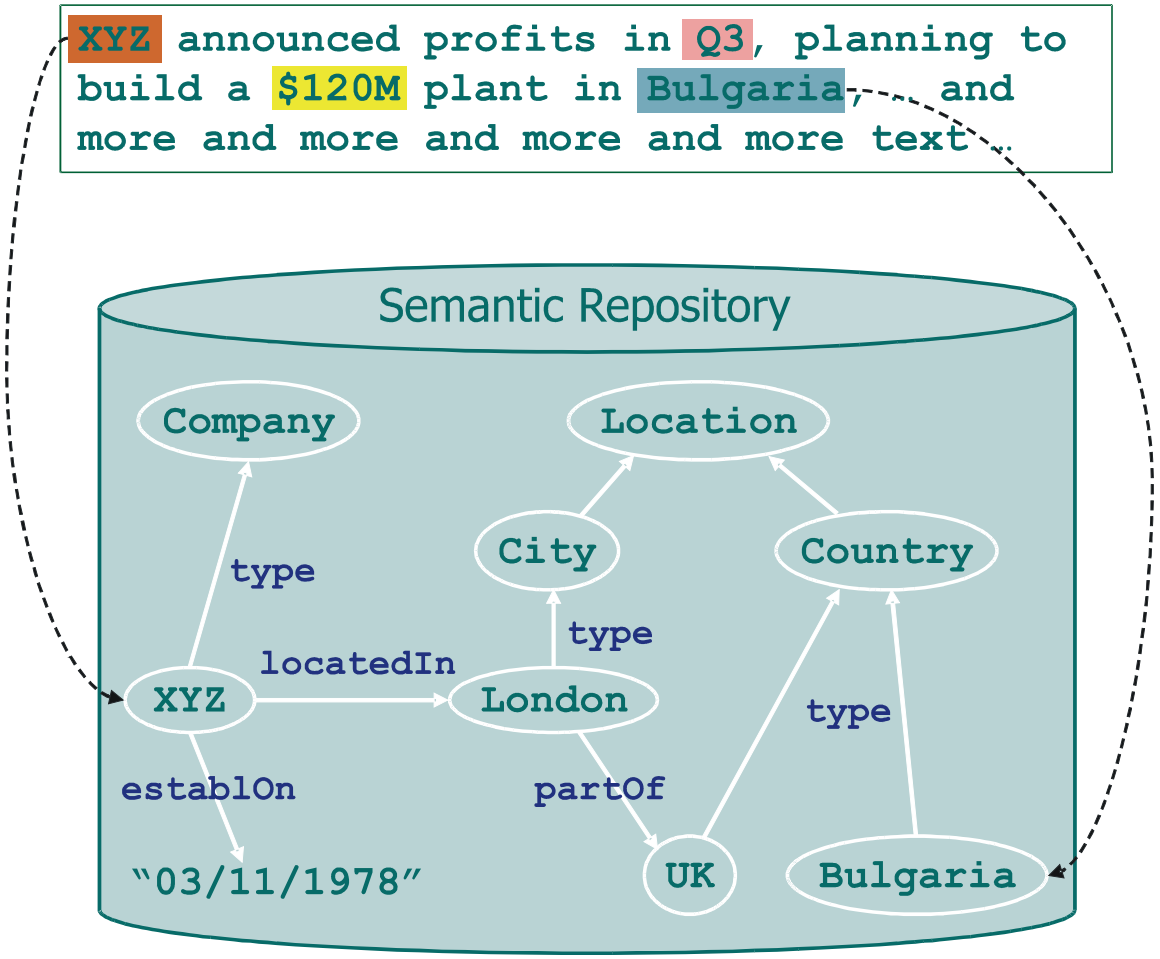
Results

The best 5 Indian regions for a company with the characteristics supplied are returned (best region first) -- available is also an indication of the most relevant region characteristics that contributed to its ranking (either positively or negatively depending on the score sign).

	Region	Global Score	1st (most influent) factor	Partial score	2nd factor	Partial score	3rd factor	Partial score
1	**Delhi**	174.68	Density of population per sq. km.	128.43	Literacy	51.38	Labour cost	-50.77
2	**Daman&Diu**	134.34	Decadal growth of Population	68.78	Literacy	47.10	Population	-23.75
3	**Chandigarh**	113.60	Density of population per sq. km.	107.13	Labour cost	-68.15	Literacy	51.38
4	**Lakshadweep**	48.76	Literacy	77.07	Population	-23.83	Decadal growth of Population	-21.40
5	**Mizoram**	45.92	Literacy	77.07	Population	-23.15	Density of population per sq. km.	-13.27

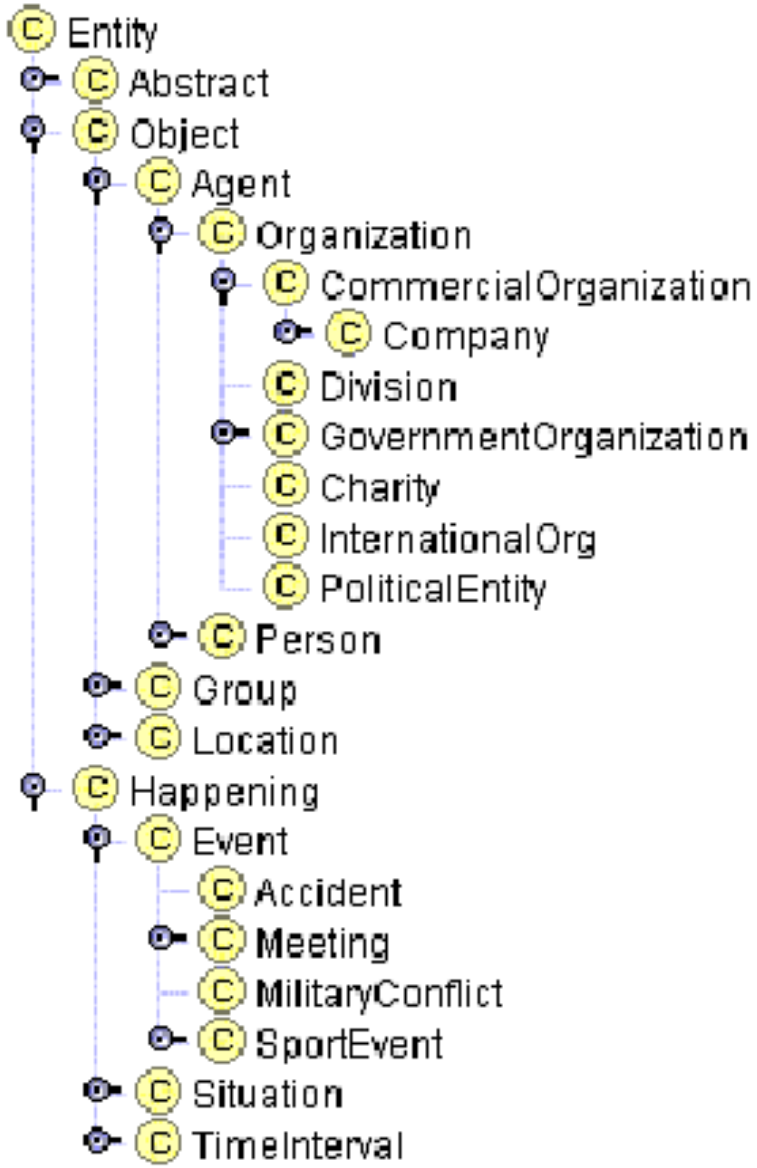


Semantic Annotation of Documents

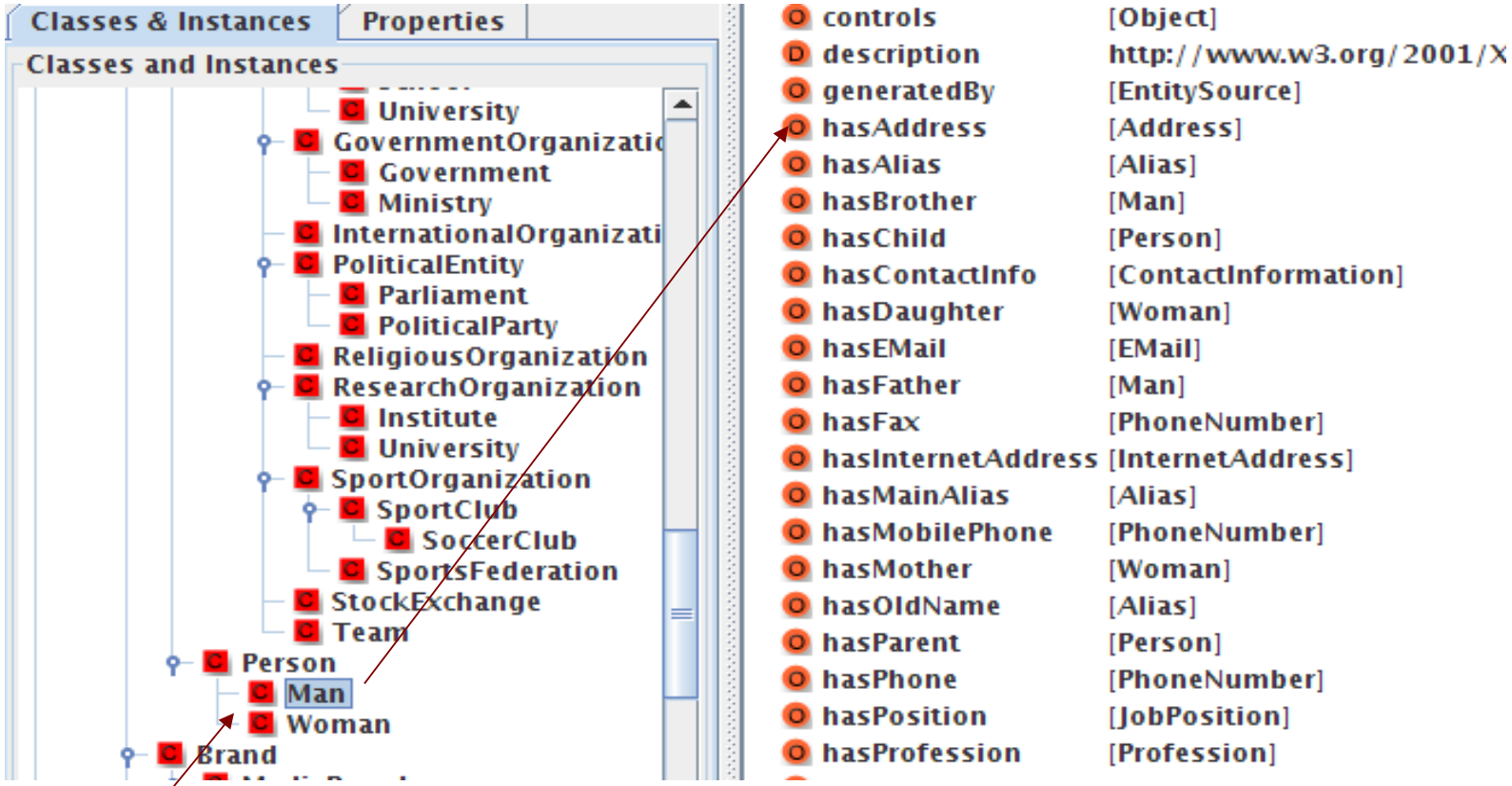


Richer NE Tagging

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

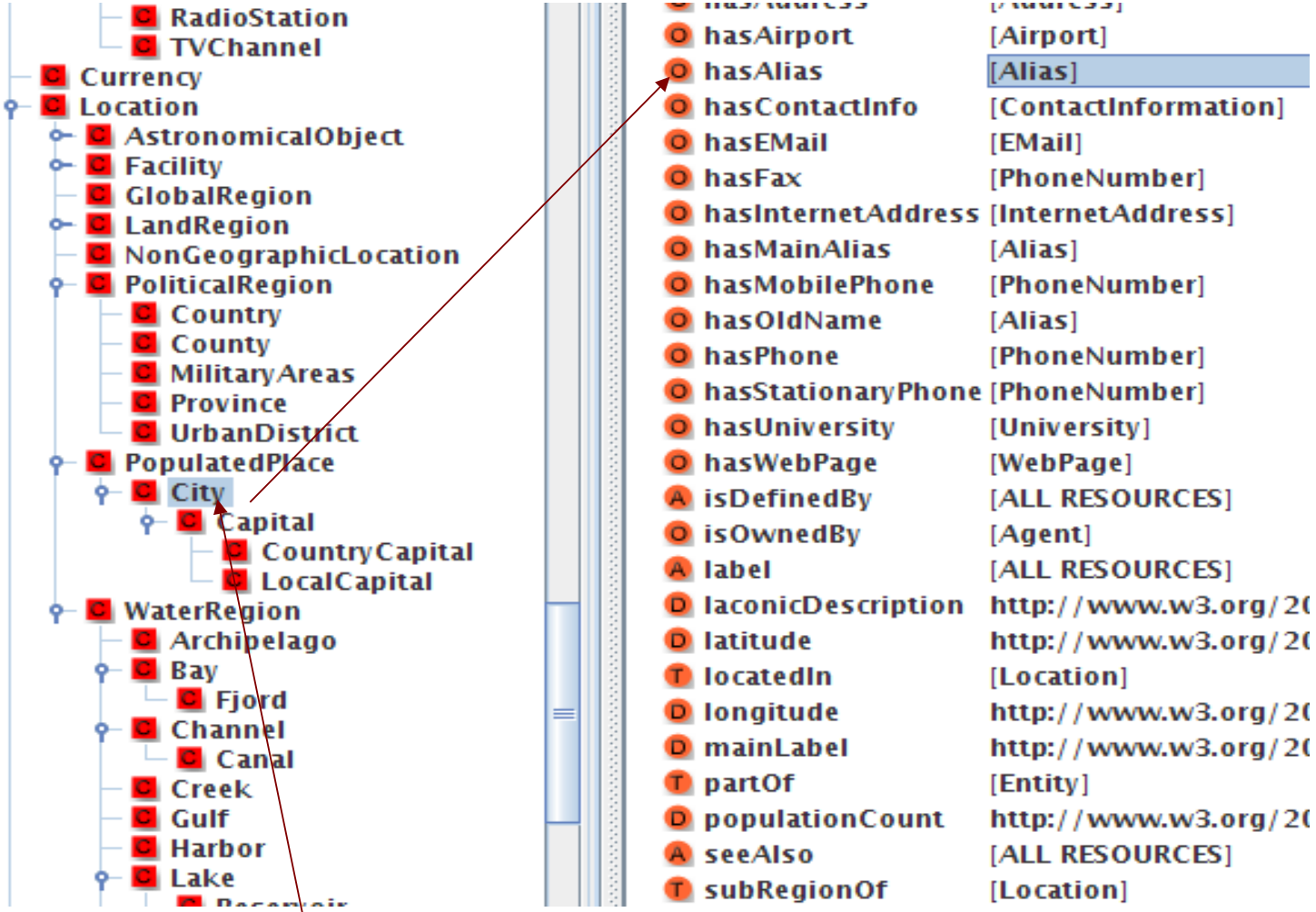


Ontology-based IE



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.

How does ontology-based IE help with IE?

- We can make inferences about all kinds of things once we have the annotations linked to an ontology
- We now know that cities can have airports, and people have phone numbers
- Since John is a man, we know he can have a wife
- If we know that the London, where John lives, is in England, we know that Polar Bear Design is also in England and not Ontario

Ontologies are useful for encoding the information found

- Enable us to define the concepts we're trying to find in texts
  - e.g., *aircraft accident, industrial action*
- As well as particular instances of these
  - e.g., *Qantas flight XYZ crashed on ..., BA cabin crew were on strike between March 20-23, 2010*
- And the relationships between them
  - e.g., *the plane that crashed belonged to Qantas and crashed on a specific date*

Using knowledge from the ontology

- The ontology tells us that
  - *Industrial action involves airport or airline staff and has a start and end date*
- It gives a clearly defined schema to annotate against
  - *if you annotate an instance of a strike, then you know this also requires you to annotate the airport/airline affected and the staff on strike*
- Extra knowledge about the different kinds of properties and the actors involved can help to improve system performance
- Backbone for other processes, for example visualising results on a timeline

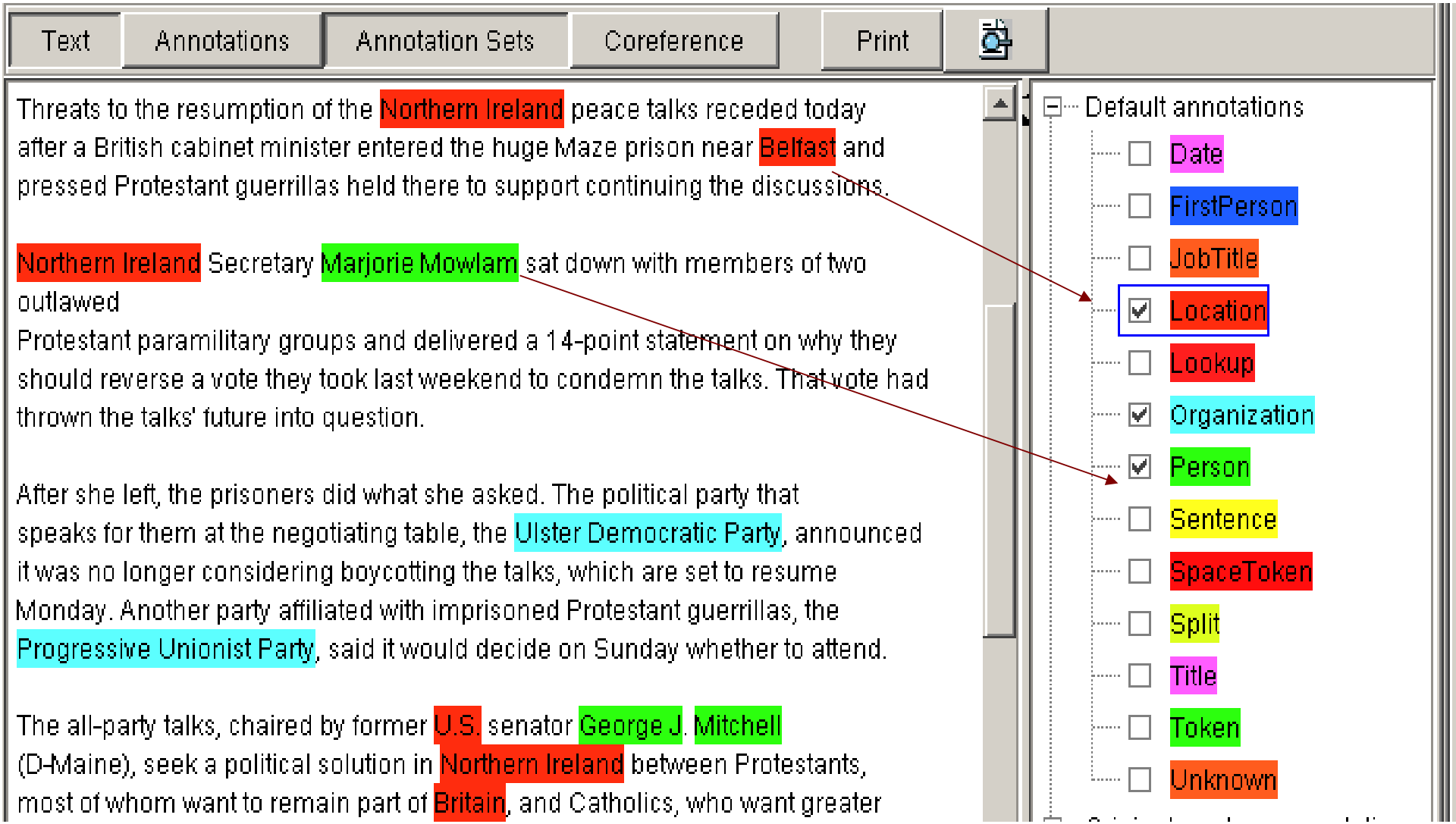
Semantic Annotation lets us do semantic search

- Semantic Search could match
  - a query: *Documents concerning a telecom company in Europe, John Smith as a director, and a date in the first half of 2002*.
  - with a document containing: *“At its meeting on the 10th of May, the board of Vodafone appointed John G. Smith as CTO"*
- Traditional search engines cannot do the required reasoning:
  - Vodafone is a mobile operator, which is a kind of telecom company;
  - Vodafone is in the UK, which is a part of Europe;
  - CTO is a type of director
  - 5th of May is a "date in first half of 2002”;
  - “John G. Smith” matches “John Smith”.

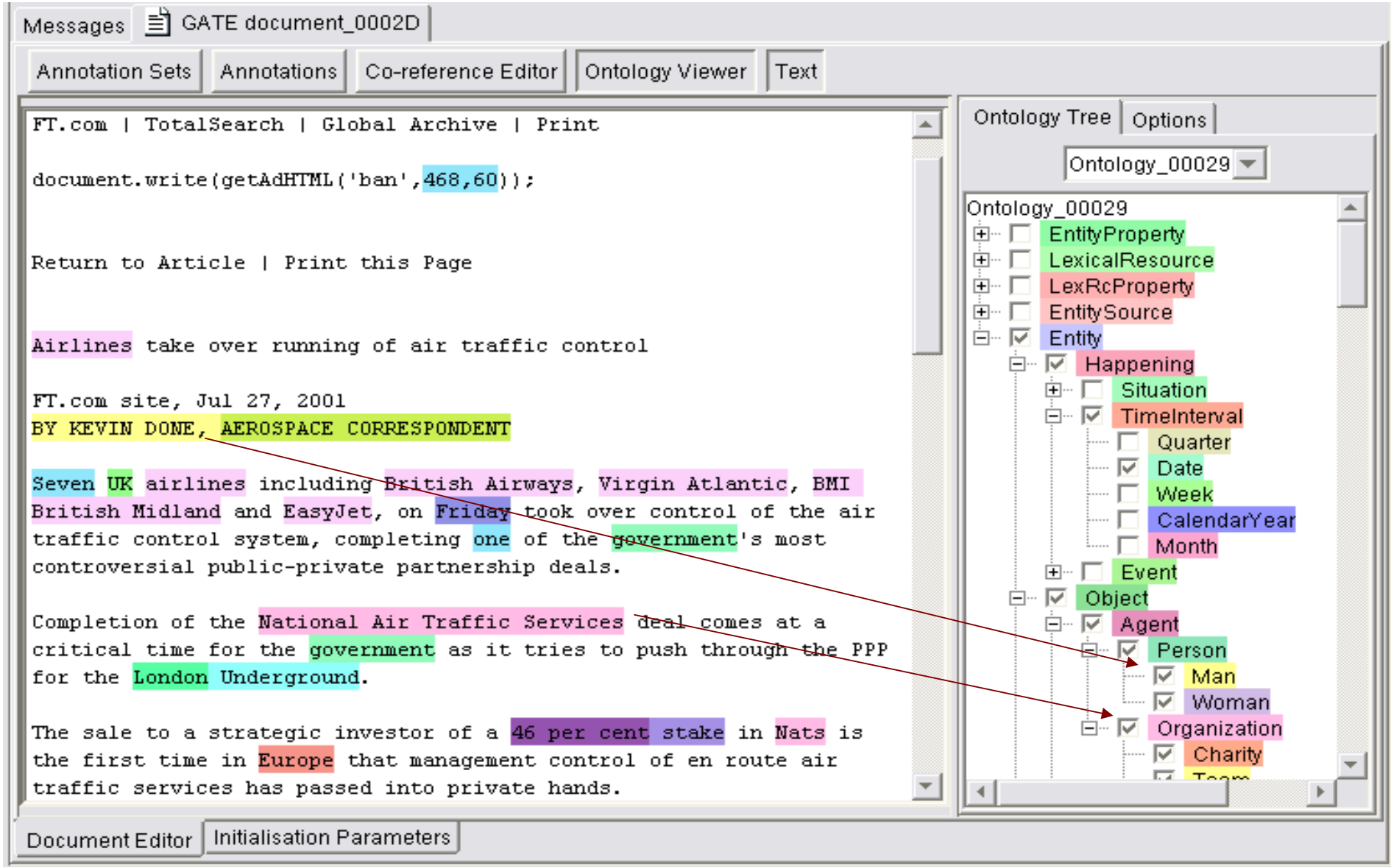
Automatic Semantic Annotation in GATE

- GATE supports ontologies as part of IE applications - Ontology-Based IE (OBIE)
- Supports semantic annotation and ontology population
- GATE has its own ontology API based on Sesame 2 and OWLIM 3
- Semantic annotation can combine learning and rule-based methods
- Enables use of large-scale linguistic resources for IE, such as WordNet

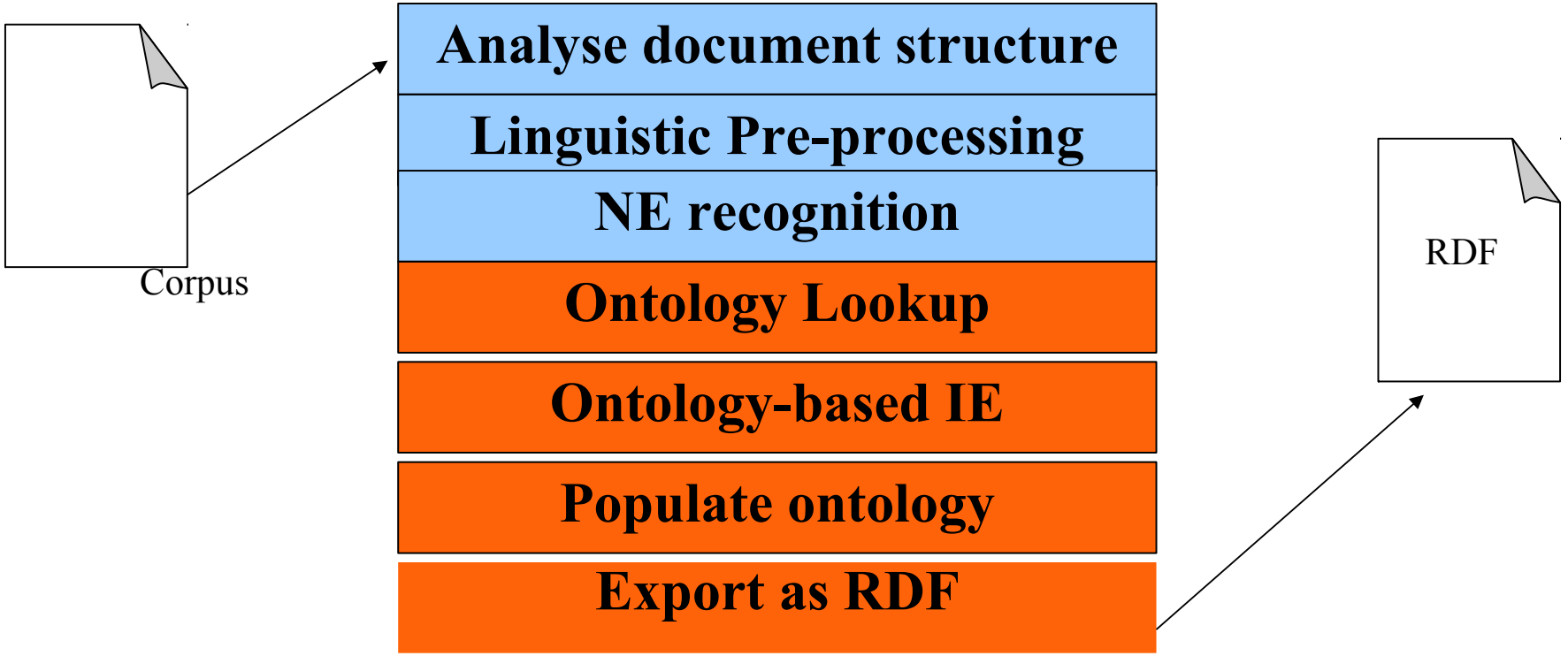
Traditional IE in GATE



Semantic IE in GATE



Typical Semantic Annotation pipeline



Semantic Annotation in the real world

- We can annotate with respect to an ontology via a lookup process, but usually, this isn't enough for real applications
- In GATE, you can map the ANNIE results to ontology classes
- Rules can also combine Lookups from traditional gazetteers with Lookups from ontologies and other clues, in order to detect Mentions
- You can also link co-referring items to the ontology
  - e.g. Mr. Brown = Gordon Brown = he
- Disambiguation: if several instances with label “John Smith”, pick the correct one
  - Context from the text can be matched against the ontology

Evaluation of Information Extraction

"We didn't underperform. You overexpected."

OBIE Demo

- OBIE (Ontology-Based Information Extraction) allows the user to inspect a document that has been automatically annotated with respect to an ontology and manually improve it.
- The user adds, deletes and changes annotations, then sends the corrected document back to the machine learning tool's trainer, so that further automatic annotation will be better.

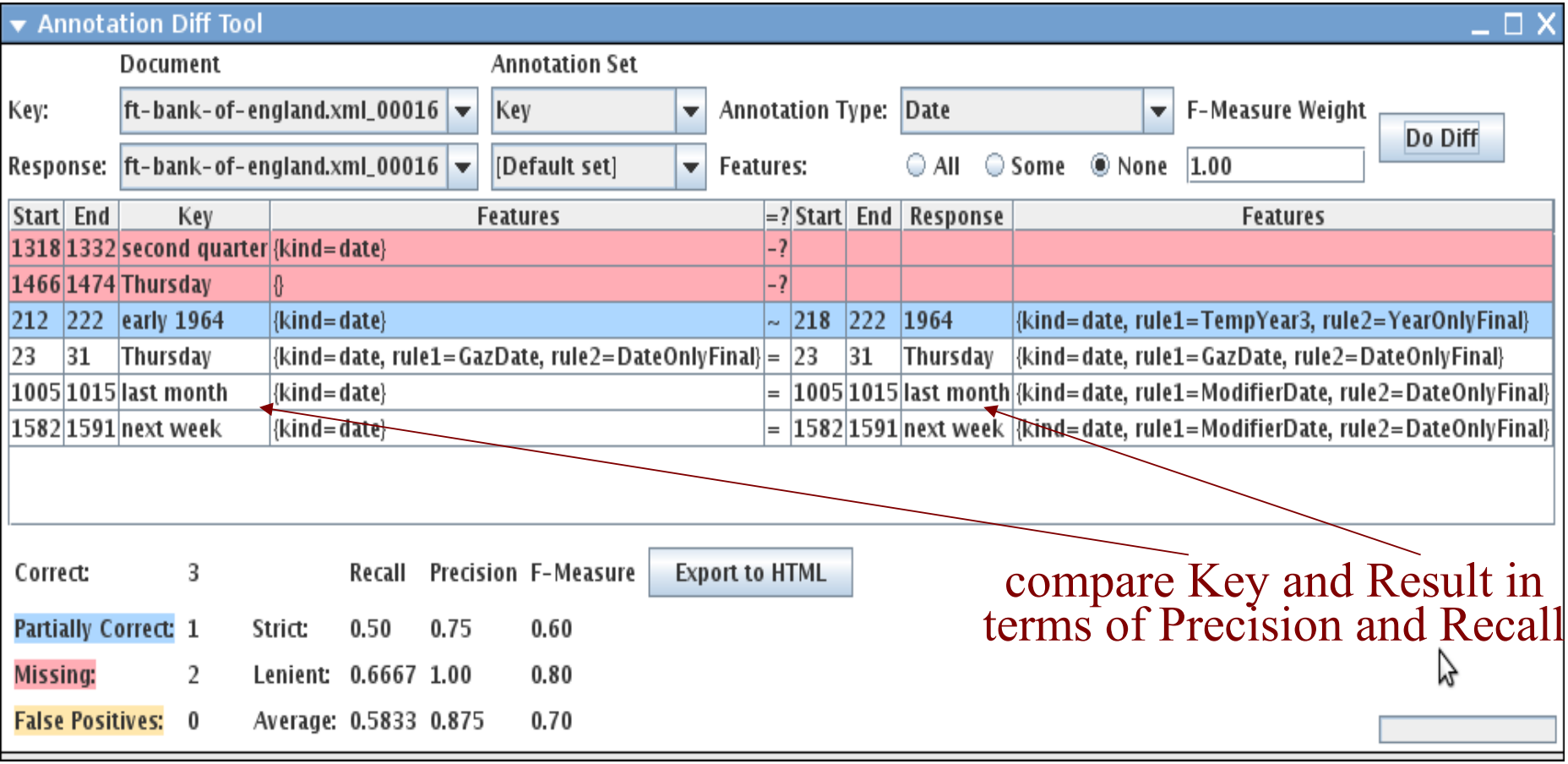
  http://gate.ac.uk/demos/obie/obie.html

Performance Evaluation

2 main requirements:

- **Evaluation metric**: mathematically defines how to measure the system's performance against human-annotated gold standard
- **Scoring program**: implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of annotation

Evaluation of Traditional IE



Evaluation tools in GATE

- GATE has various tools for evaluation
- Annotation Diff - compares annotations on a single document (Precision, Recall and F)
- Corpus Quality Assurance - compares annotations on a corpus using a variety of measures
- IAA - compares different sets of annotations (e.g. different manual annotators) on a corpus, using a variety of measures
- BDM tool calculates BDM between every class in an ontology
- IAA tool can use BDM scores as part of its evaluation

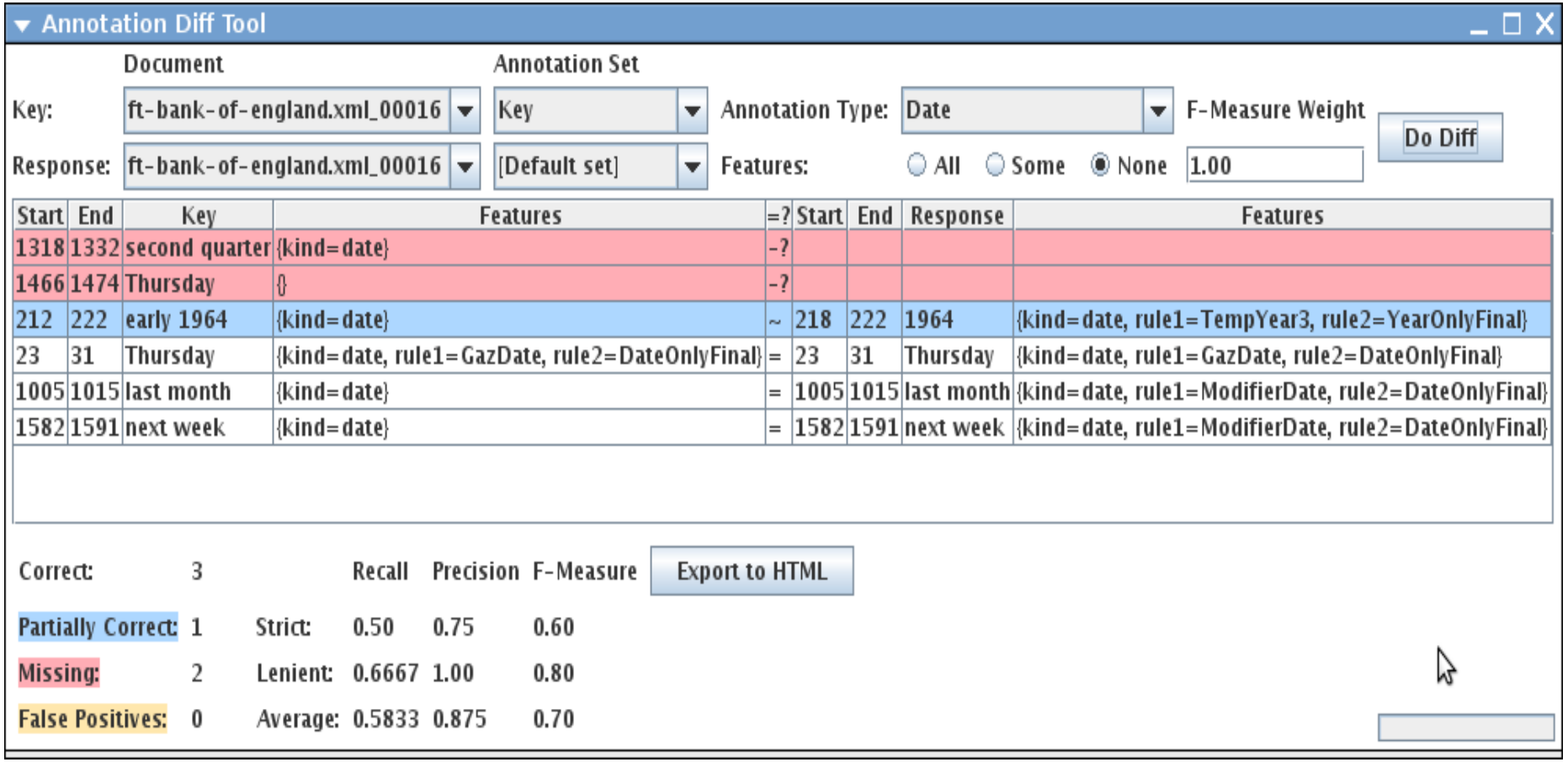
AnnotationDiff

- Graphical comparison of 2 sets of annotations
- Visual diff representation, like tkdiff
- Compares one document at a time, one annotation type at a time

Annotations are like squirrels…

Annotation Diff helps with “spot the difference”

Annotation Diff



A Word about Terminology

- Different communities use different terms when talking about evaluation, because the tasks are a bit different.
- The IE community usually talks about "correct", "spurious" and "missing"
- The IR community usually talks about "true positives", "false positives" and "negatives". They also talk about "false negatives", but you can ignore those.
- Some terminologies assume that one set of annotations is correct ("gold standard")
- Other terminologies do not assume one annotation set is correct
- When measuring inter-annotator agreement, there is no reason to assume one annotator is more correct than the other

Measuring success

- In IE, we classify the annotations produced in one of 4 ways:
- **Correct** = things annotated correctly

e.g. annotating “Hamish Cunningham” as a Person

- **Missing** = things not annotated that should have been

e.g. not annotating “Sheffield” as a Location

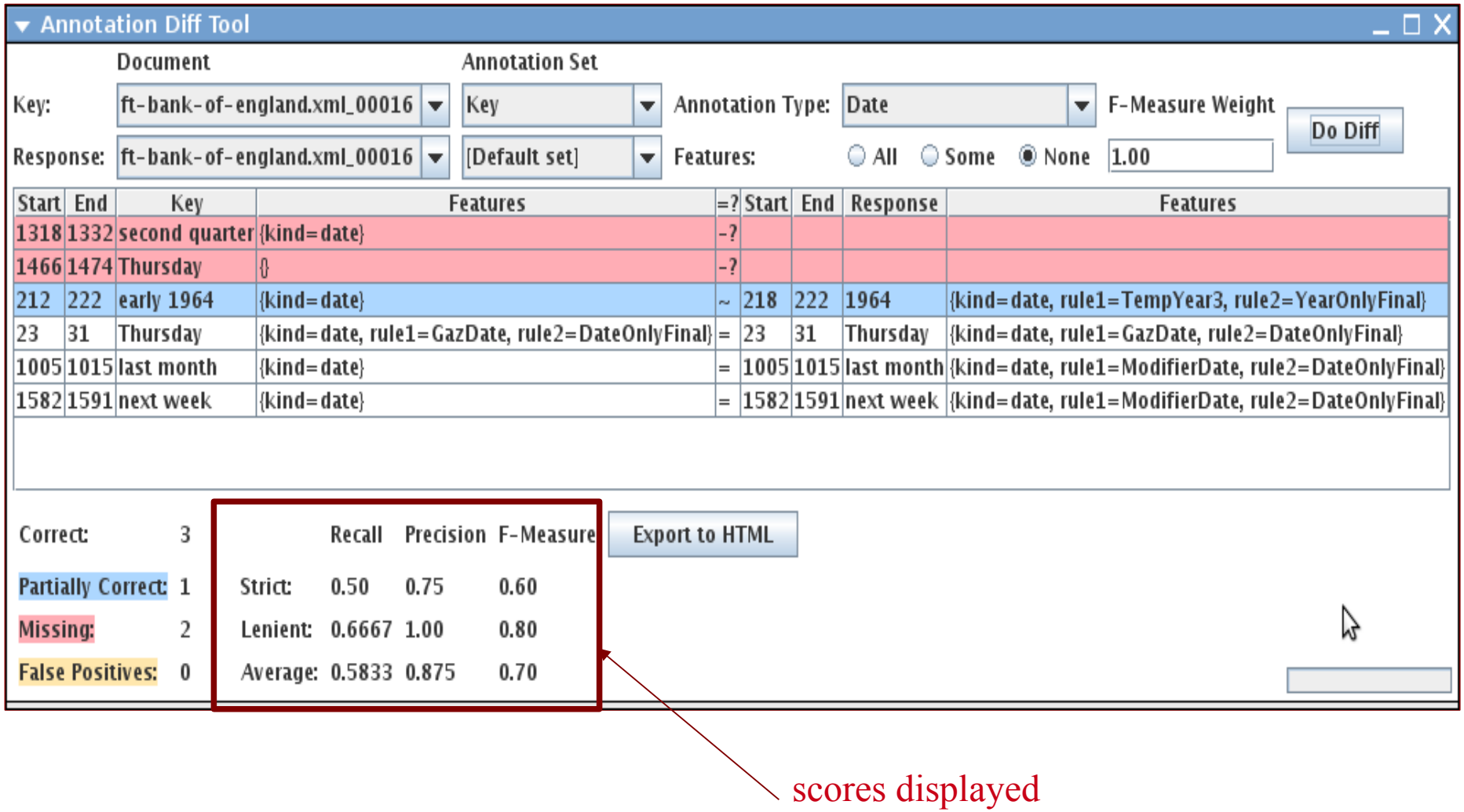
- **Spurious** = things annotated wrongly

e.g. annotating “Hamish Cunningham” as a Location

- **Partially correct** = the annotation type is correct, but the span is wrong

e,g, annotating just “Cunningham” as a Person (too short) or annotating “Unfortunately Hamish Cunningham” as a Person (too long)

Finding Precision, Recall and F-measure



scores displayed

Precision

- How many of the entities your application found were correct?
- Sometimes precision is called **accuracy**

$$Precision = \frac{Correct}{Correct + Spurious}$$

Recall

- How many of the entities that exist did your application find?
- Sometimes recall is called **coverage**

$$Recall = \frac{Correct}{Correct + Missing}$$

F-Measure

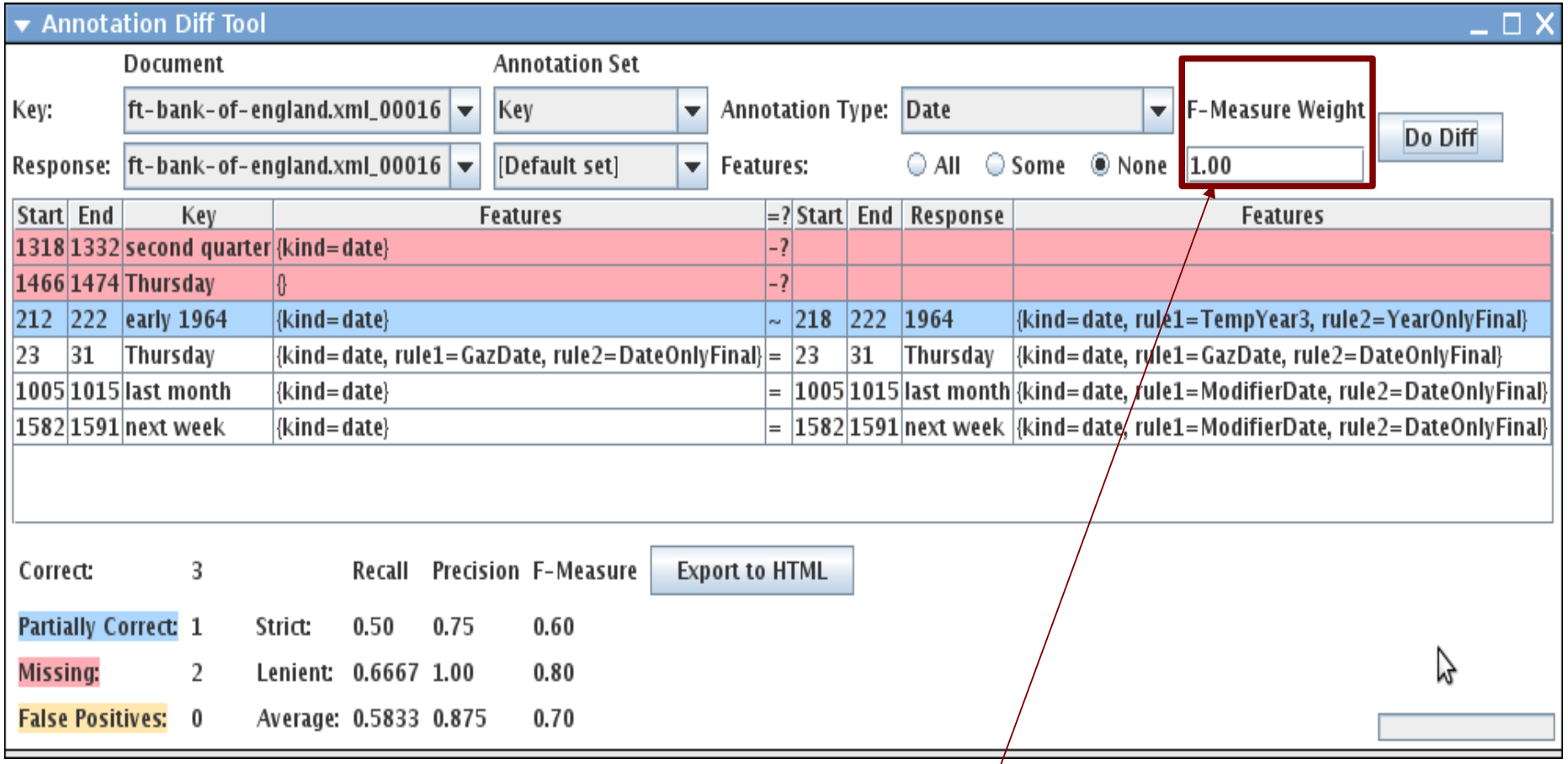
- Precision and recall tend to trade off against one another
  - If you specify your rules precisely to improve precision, you may get a lower recall
- If you make your rules very general, you get good recall, but low precision
- This makes it difficult to compare applications, or to check whether a change has improved or worsened the results overall
- F-measure combines precision and recall into one measure

F-Measure

- Also known as the “harmonic mean”
- Usually, precision and recall are equally weighted
- This is known as F1
- To use F1, set the value of the F-measure weight to 1
- This is the default setting

$$F = 2 \cdot \left(\frac{precision \cdot recall}{precision + recall} \right)$$

Annotation Diff defaults to F1

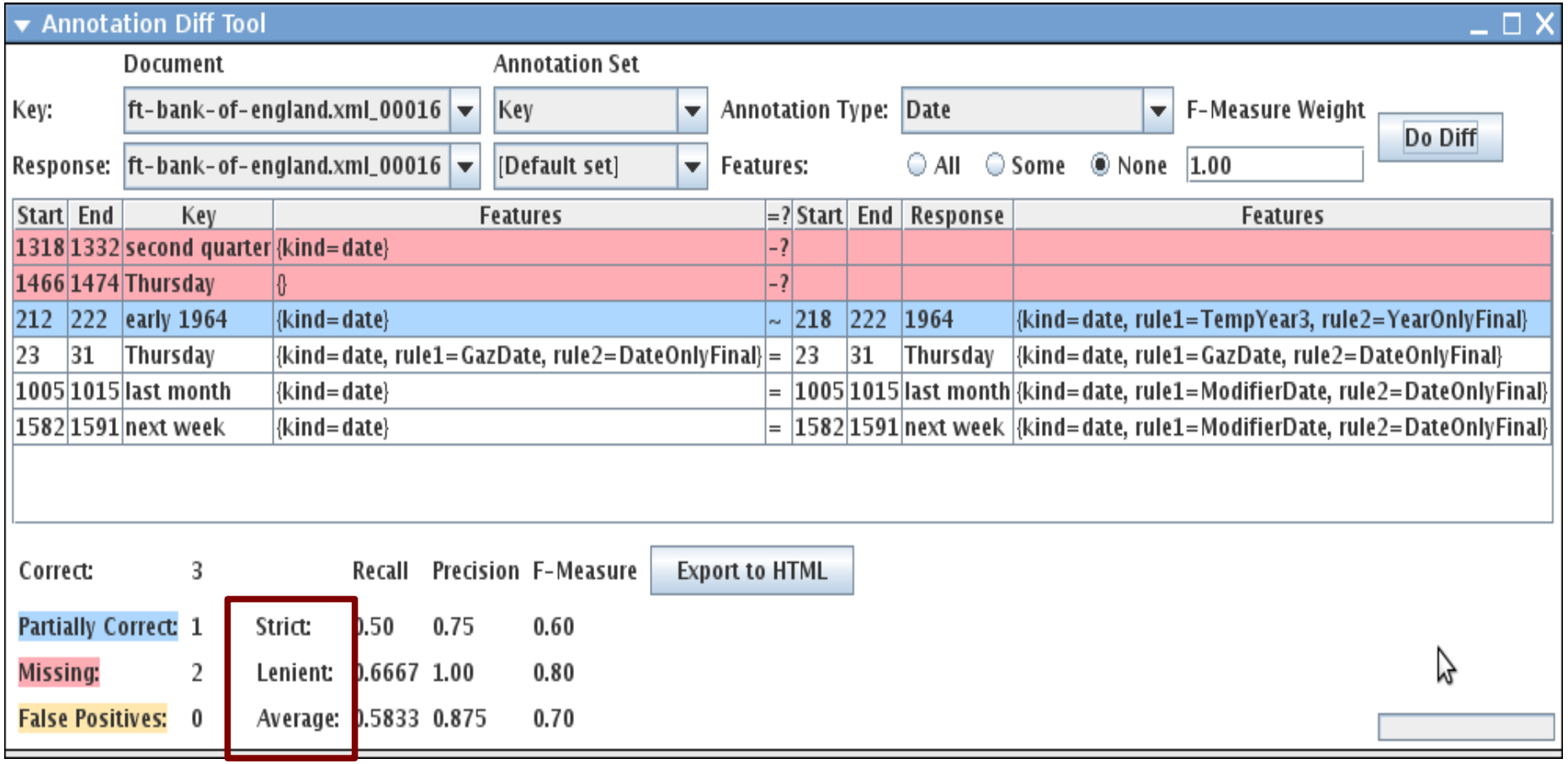


F-measure weight set to 1

Statistics can mean what you want them to....

- How we want to measure partially correct annotations may differ, depending on our goal
- In GATE, there are 3 different ways to measure them
- The most usual way is to consider them to be “half right”
- Average: Strict and lenient scores are averaged (this is the same as counting a half weight for every partially correct annotation)
- Strict: Only perfectly matching annotations are counted as correct
- Lenient: Partially matching annotations are counted as correct. This makes your scores look better :-)

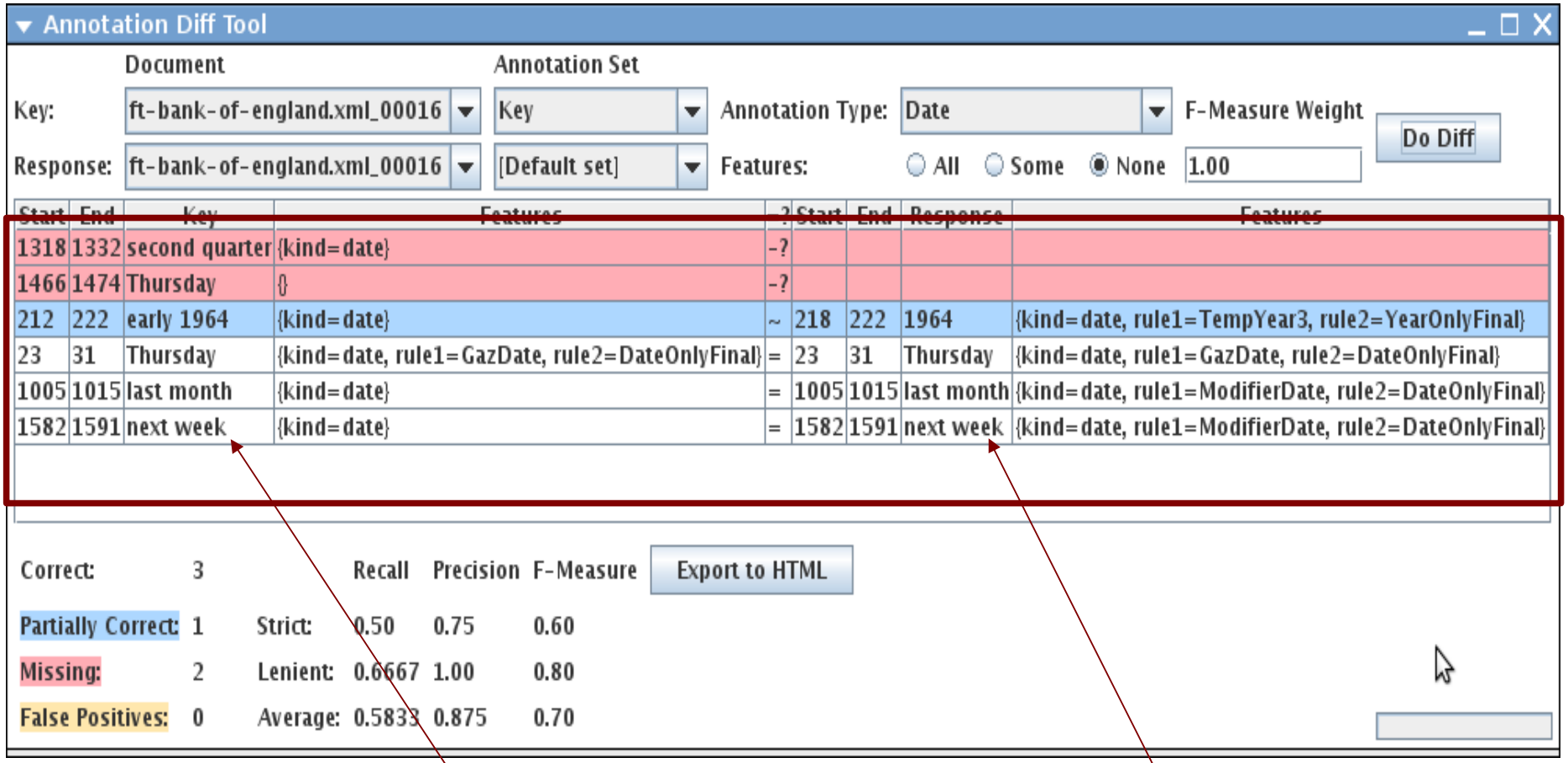
Strict, Lenient and Average



Comparing the individual annotations

- In the AnnotationDiff, colour codes indicate whether the annotation pair shown are correct, partially correct, missing (false negative) or spurious (false positive)
- You can sort the columns however you like

Comparing the annotations



Key annotations

Response annotations

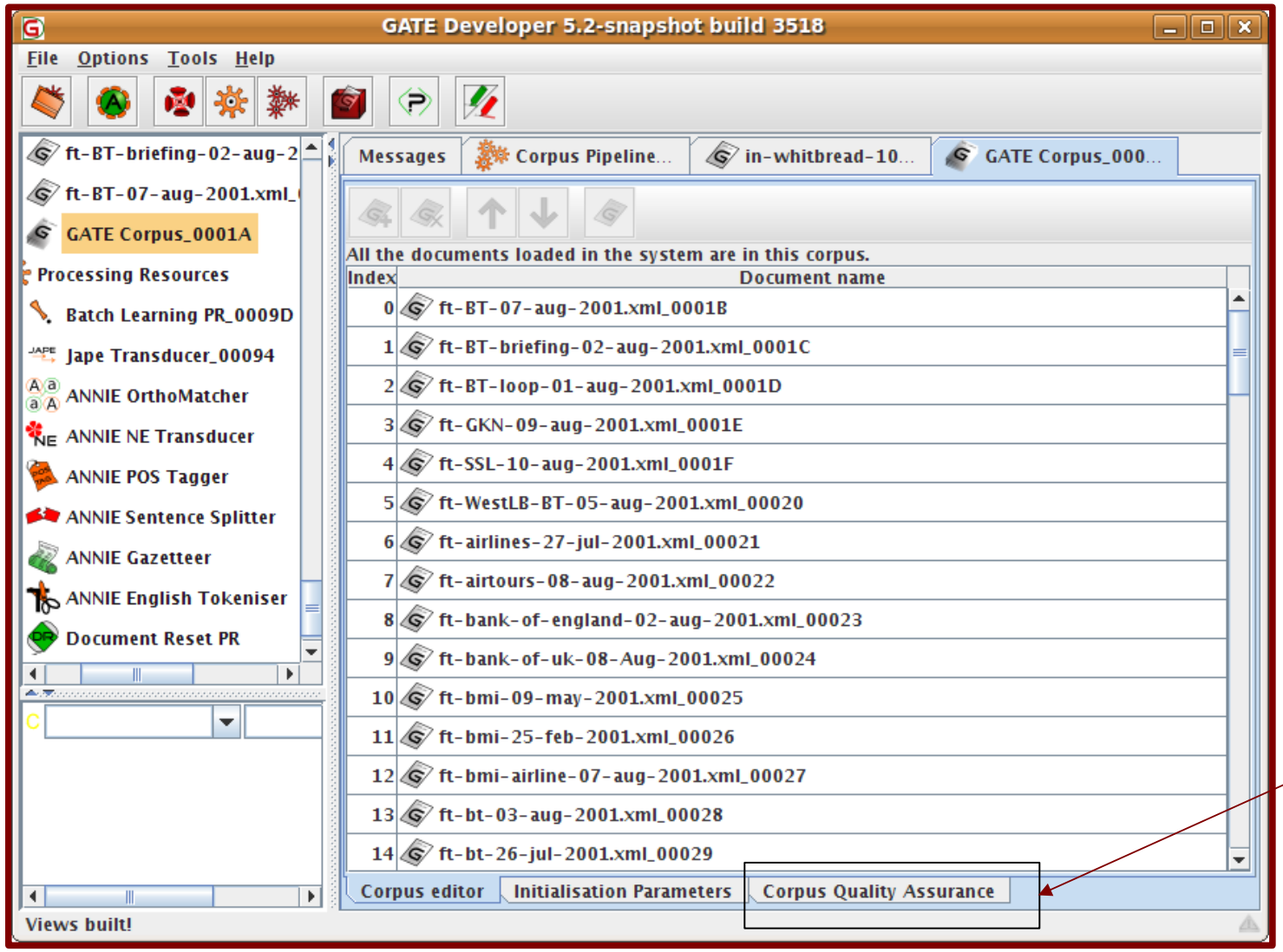
Annotation Diff Hands-On

- Open the hands-on news-texts corpus in GATE as before
- Open the AnnotationDiff tool (Tools → Annotation Diff)
- For the Key set (containing the manual annotations) select **Key** annotation set
- For the Response set (containing annotations from ANNIE) select **Default** annotation set
- Select an annotation type (e.g. **Organization**)
- Click on "Compare"
- Scroll down the list, to see correct, partially correct, missing and spurious annotations
- Try changing the various settings

Corpus Quality Assurance

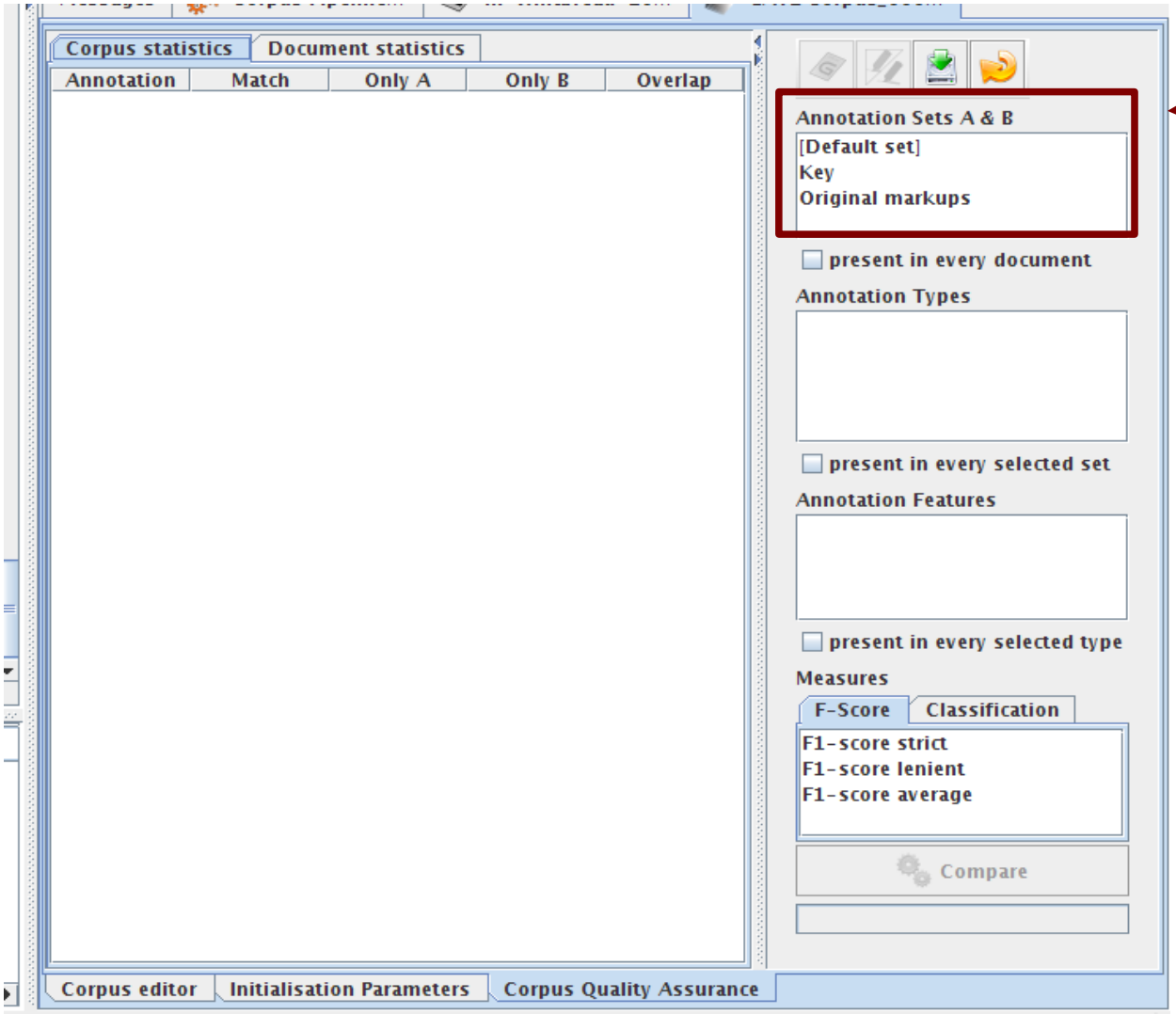
- Corpus Quality Assurance tool extends the Annotation Diff functionality to the entire corpus, rather than on a single document at a time
- It produces statistics both for the corpus as a whole (Corpus statistics tab) and for each document separately (Document statistics tab)
- It compares two annotation sets, but makes no assumptions about which (if either) set is the gold standard. It just labels them A and B.
- This is because it can be used to measure Inter Annotator Agreement (IAA) where there is no concept of “correct” set

Try out Corpus Quality Assurance

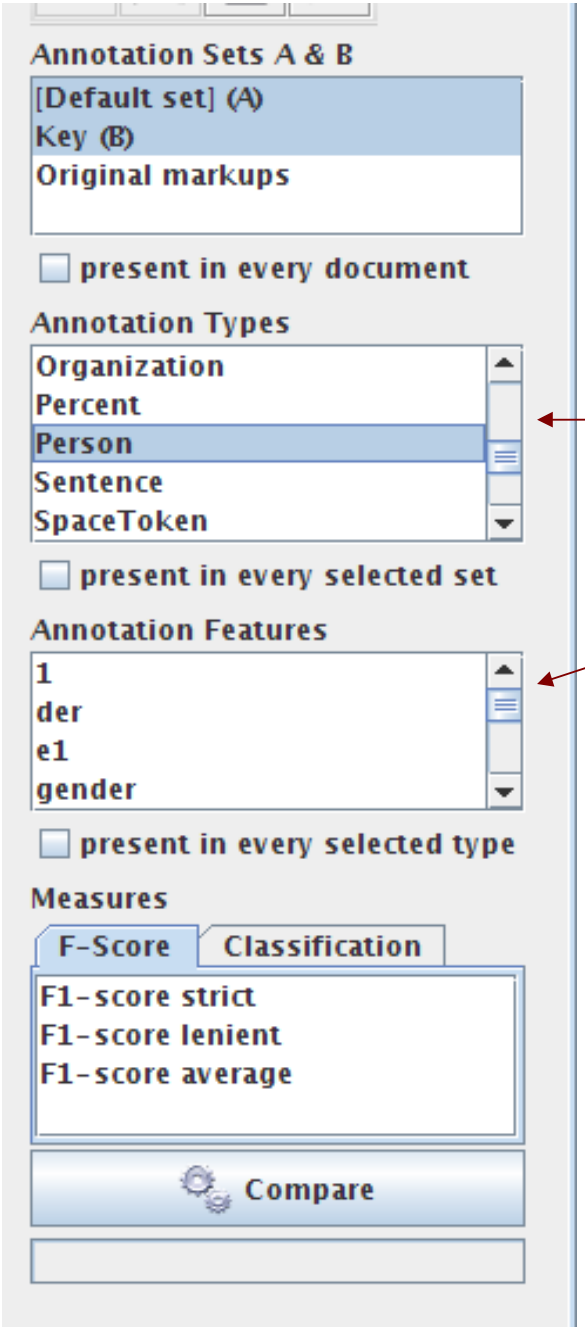


- Double click your hands-on corpus and click the Corpus Quality Assurance tab at the bottom of the Display pane.

Select Annotation Sets

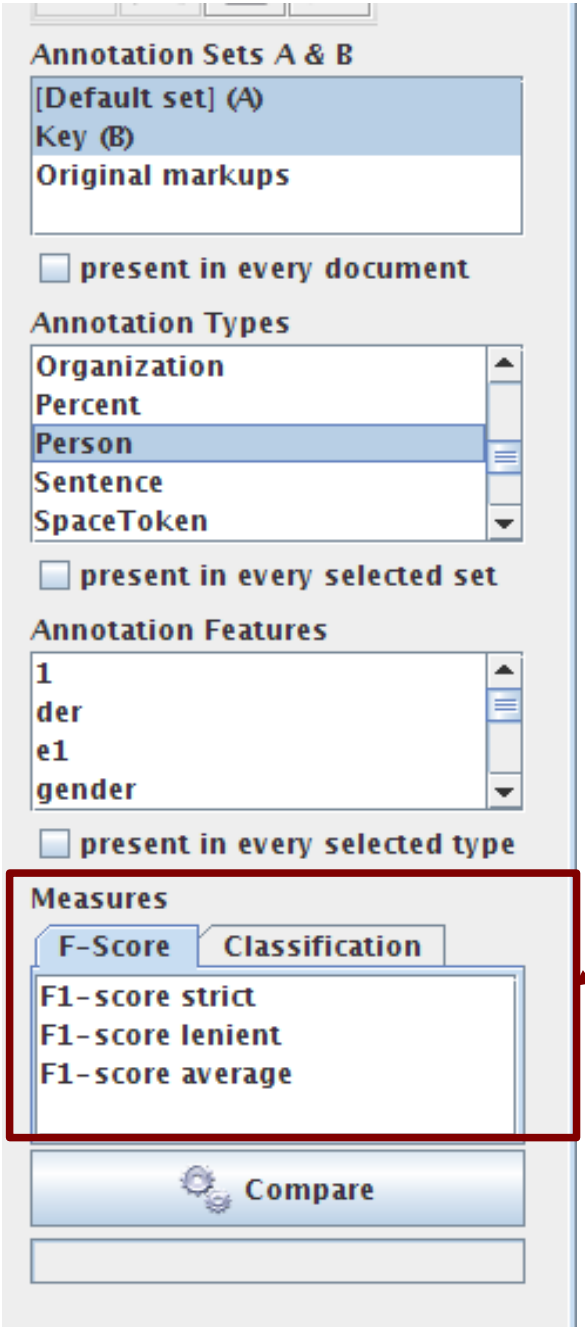


- Select the annotation sets you wish to compare.
- Click on the Key annotation set – this will label it set A.
- Now click on the default annotation set - this will label it set B.



Select Type

- Select the annotation type to compare (suggestion: select Organisation, Person and Location for now)
- Select the features to include (if any – leave unselected for now)
- You can select as many types and features as you want.

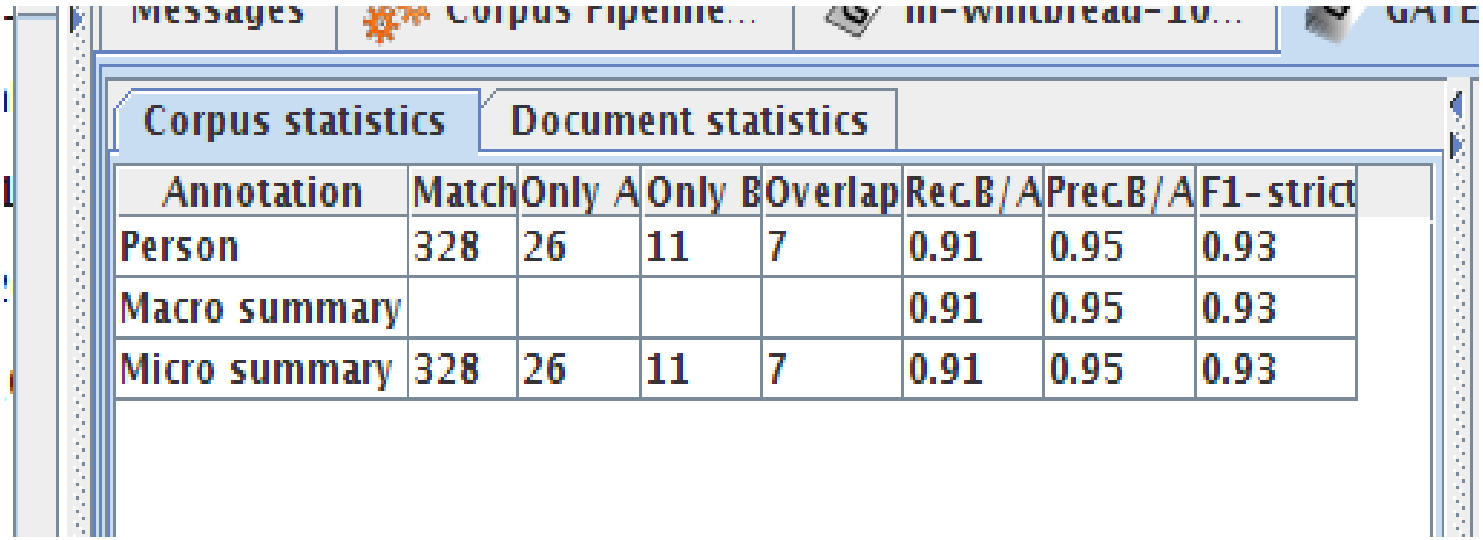


Select measure

- In the "Measures" box, select the kind of F1.0 score you want "Strict, Lenient, Average" or any combination of them. Suggestion: try just "lenient" at first

- Select Compare

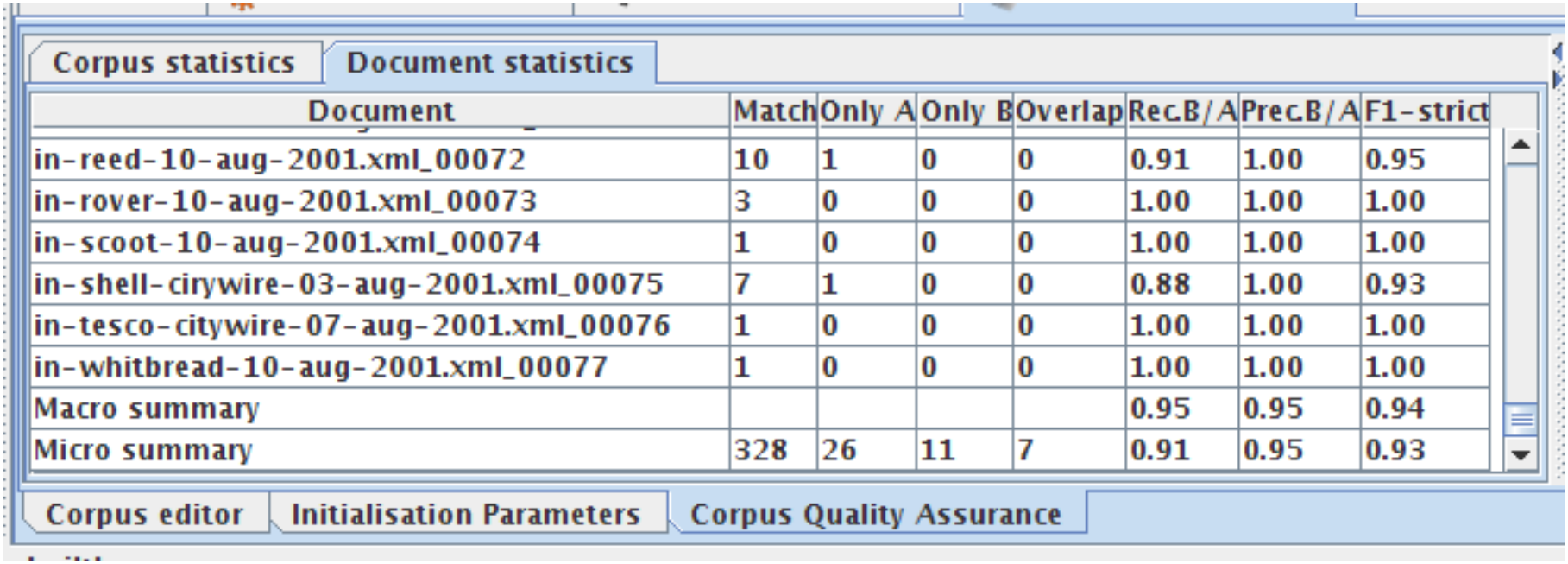
Corpus Statistics Tab



Annotation	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
Person	328	26	11	7	0.91	0.95	0.93
Macro summary					0.91	0.95	0.93
Micro summary	328	26	11	7	0.91	0.95	0.93

- Each annotation type is listed separately
- Precision, recall and F measure are given for each
- Two summary rows provide micro and macro averages

Document Statistics Tab



Corpus statistics | Document statistics

Document	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
in-reed-10-aug-2001.xml_00072	10	1	0	0	0.91	1.00	0.95
in-rover-10-aug-2001.xml_00073	3	0	0	0	1.00	1.00	1.00
in-scoot-10-aug-2001.xml_00074	1	0	0	0	1.00	1.00	1.00
in-shell-cirywire-03-aug-2001.xml_00075	7	1	0	0	0.88	1.00	0.93
in-tesco-citywire-07-aug-2001.xml_00076	1	0	0	0	1.00	1.00	1.00
in-whitbread-10-aug-2001.xml_00077	1	0	0	0	1.00	1.00	1.00
Macro summary					0.95	0.95	0.94
Micro summary	328	26	11	7	0.91	0.95	0.93

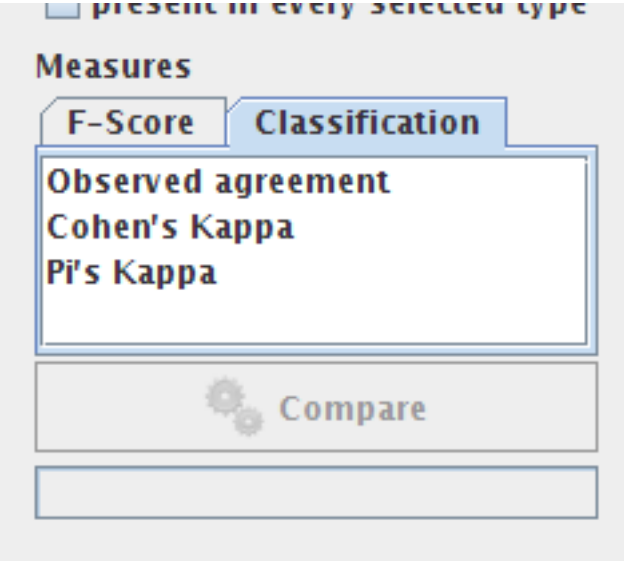
Corpus editor | Initialisation Parameters | Corpus Quality Assurance

- Each document is listed separately
- Precision, recall and F measure are given for each
- Two summary rows provide micro and macro averages

Micro and Macro Averaging

- Micro averaging treats the entire corpus as one big document, for the purposes of calculating precision, recall and F
- Macro averaging takes the average of the rows

Classification Measures



- By default, Corpus Quality Assurance presents the F-measures
- However, classification measures are also available
- These are not suitable for entity extraction tasks

Evaluation for Semantic IE

- Traditional IE is evaluated in terms of Precision, Recall and F-measure.
- But these are not sufficient for ontology-based IE, because the distinction between right and wrong is less obvious
- Some mistakes can be "more wrong" than others
  - Nick Clegg → Person (not Leader) – still logically correct
  - Nick Clegg → Location – wrong
- Similarity metrics need to be integrated so that items closer together in the hierarchy are given a higher score, if wrong

Balanced Distance Metric

- BDM measures the closeness of two concepts in an ontology or taxonomy
- It produces a real number between 0 and 1
- The more closely related the two concepts are in an ontology, the greater their BDM score is
- It is dependent on a number of features:
    - the length of the shortest path connecting the two concepts
    - the depth of the two concepts in the ontology
    - size and density of the ontology

Augmented Precision and Recall

BDM is integrated with traditional Precision and Recall in the following way to produce a score at the corpus level:

$$AP = \frac{\text{BDM}}{BDM + Spurious} \qquad AR = \frac{\text{BDM}}{BDM + Missing}$$

Examples of misclassification

Entity	Response	Key	BDM
Sochi	Location	City	0.724
FBI	Org	GovOrg	0.959
Al-Jazeera	Org	TVCompany	0.783
Islamic Jihad	Company	ReligiousOrg	0.816
Brazil	Object	Country	0.587
Senate	Company	Political Entity	0.826

Ontology-sensitive F-measure in GATE

- The Corpus QA tool can also be set to calculate an ontology-sensitive F-measure, using the BDM score to replace the number of correct matches
- Load the document “voting-example-bdm.xml”
- Right click on the document and select “New corpus with this document”
- Double click on the document and select the Corpus QA tool
- Select **Key** and **Test** for sets A and B; select **Mention** for the Annotation Type and **class** for the feature
- Select **F1-average** and **F1-average-BDM** as classification types, and compare results

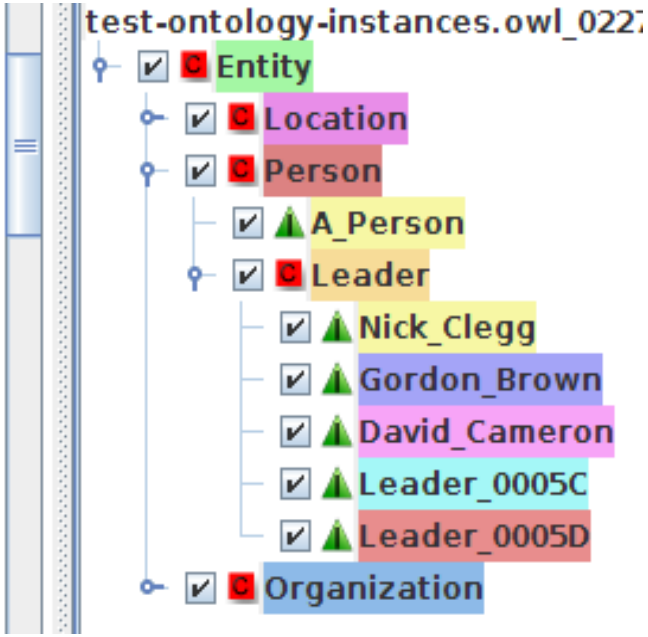
Our example text again

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. Ieuan Wyn Jones of Plaid Cymru voted in the constituency of Ynys Mon in north Wales at lunchtime.



Clegg is marked as a Person, instead of Leader

Salmond is missing

Results

- The traditional scores:
  - Match = 2, Only A (missing) = 2, Only B (spurious) = 1, Overlap (Partial) = 0
  - Recall = 0.50, Precision = 0.67, F1 = 0.57
- BDM-sensitive scores:
  - Recall = 0.60, Precision = 0.81, F1 = 0.69
    (The BDM Match = 2.418, not shown)

Summary

- Introduced the idea of semantic annotation and ontologies
- Had a look at why you might want to use them
- Shown how to evaluate both traditional and semantic information extraction in different ways
- Module 3 tomorrow will look at opinion mining as a specific kind of text mining - in particular, with respect to social media

Further materials

Ontology design principles:

http://lsdis.cs.uga.edu/SemWebCourse/OntologyDesign.ppt

BDM:

http://gate.ac.uk/userguide/sec:eval:bdmplugin

Semantic Annotation:

K. Bontcheva, B. Davis, A. Funk, Y. Li and T. Wang. Human Language Technologies. Semantic Knowledge Management, John Davies, Marko Grobelnik, and Dunja Mladenic (Eds.), Springer, 37-49, 2009.

K.Bontcheva, H.Cunningham, A.Kiryakov and V.Tablan. Semantic Annotation and Human Language Technology. Semantic Web Technology: Trends and Research. John Wiley and Sons Ltd. 2006.

D. Maynard, Y. Li and W. Peters. NLP Techniques for Term Extraction and Ontology Population. Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text, P. Buitelaar and P. Cimiano (editors). IOS Press, 2007.