# 3. From Traditional to Semantic Annotation

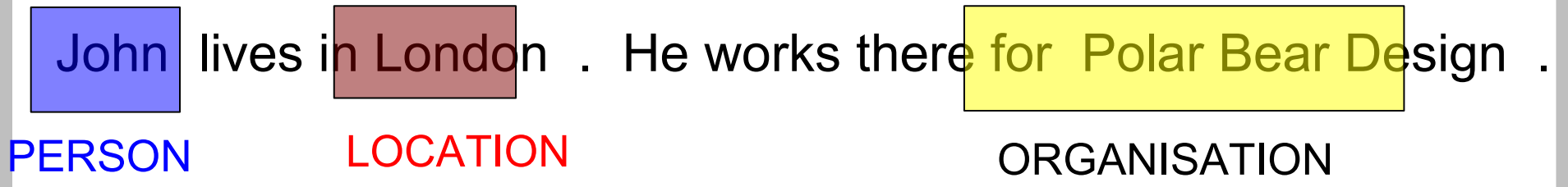
# Topics covered

- How do we move from traditional to semantic annotation, and what does that mean?
- Examples of semantic annotation in GATE
- Evaluation: moving beyond traditional IE measures

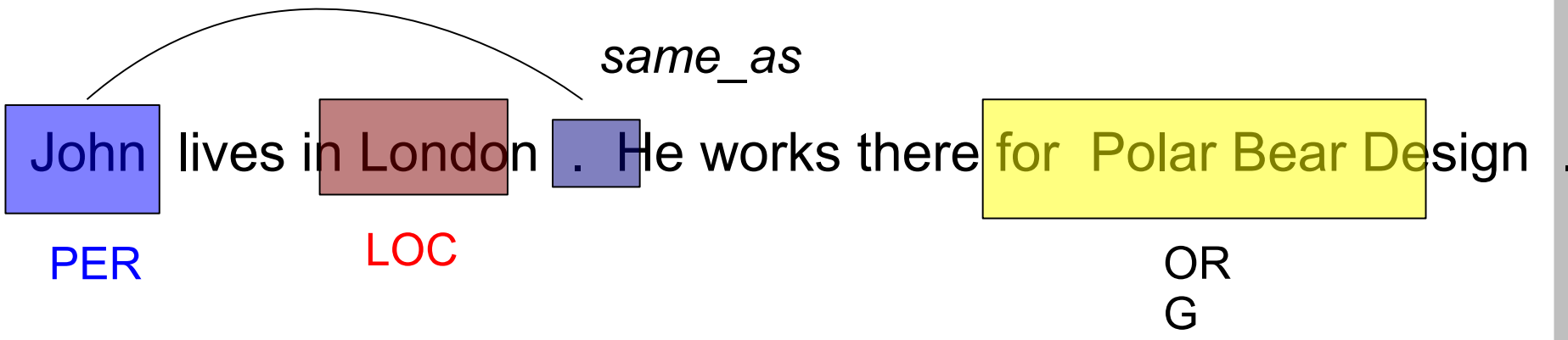
# Information Extraction for the Semantic Web

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

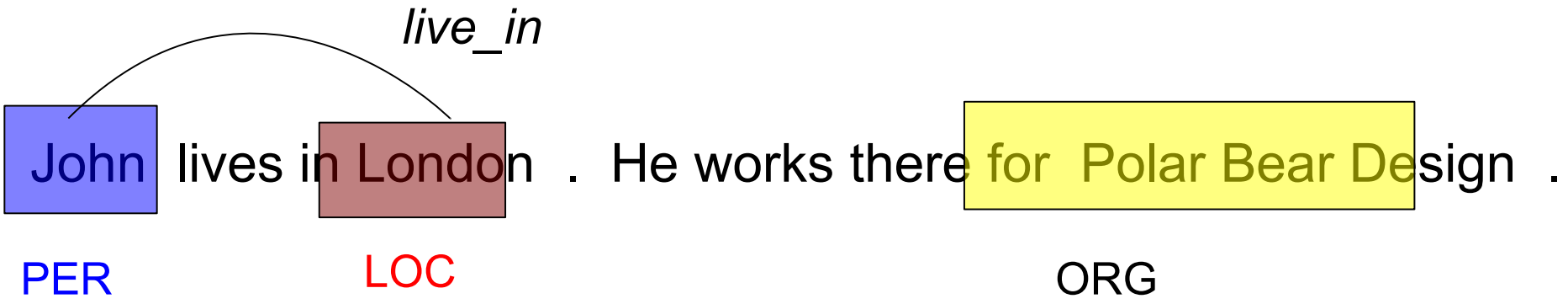
# Traditional NE Recognition



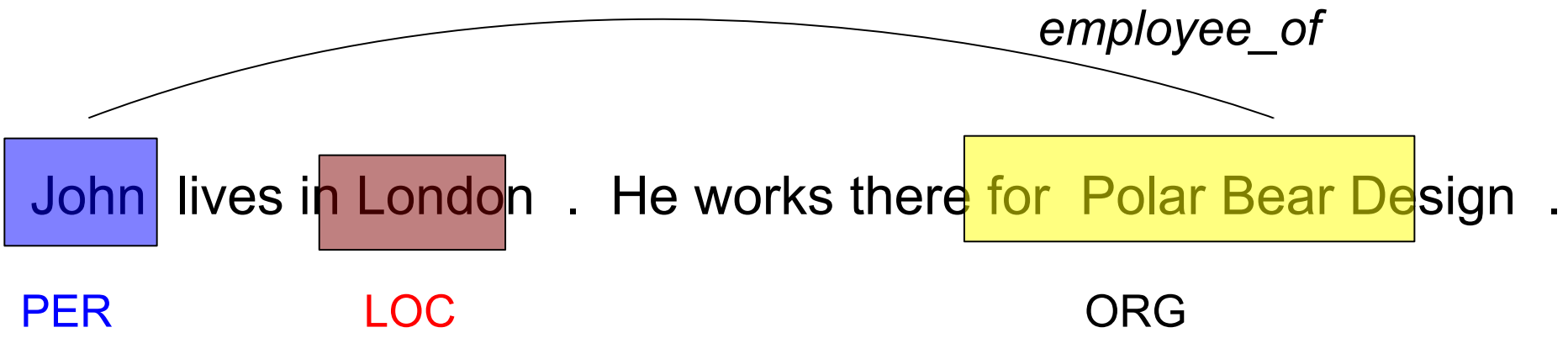
# Co-reference



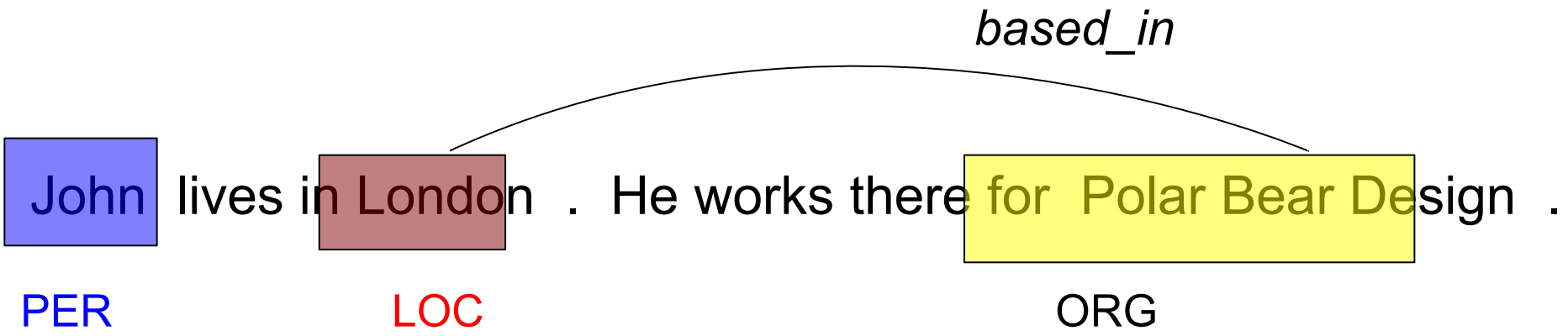
# Relations


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

# Relations (2)

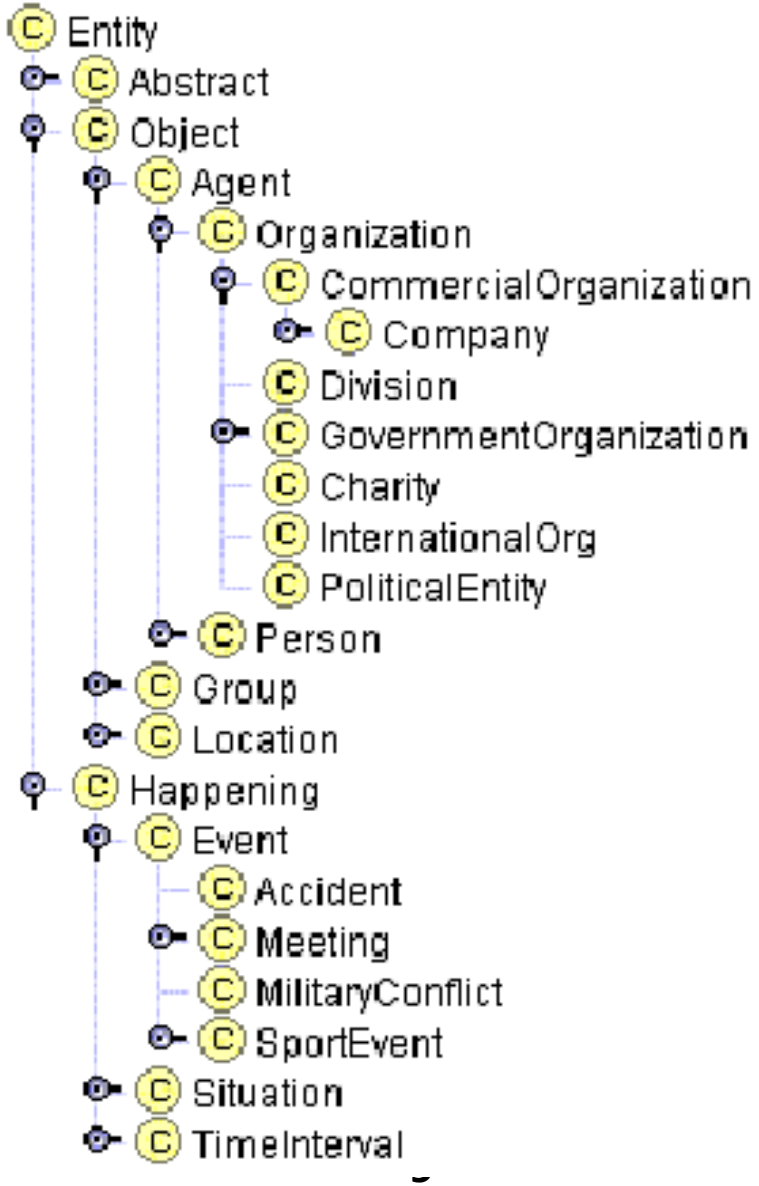

employee_of
John lives in London . He works there for Polar Bear Design .
PER
LOC
ORG

# Relations (3)

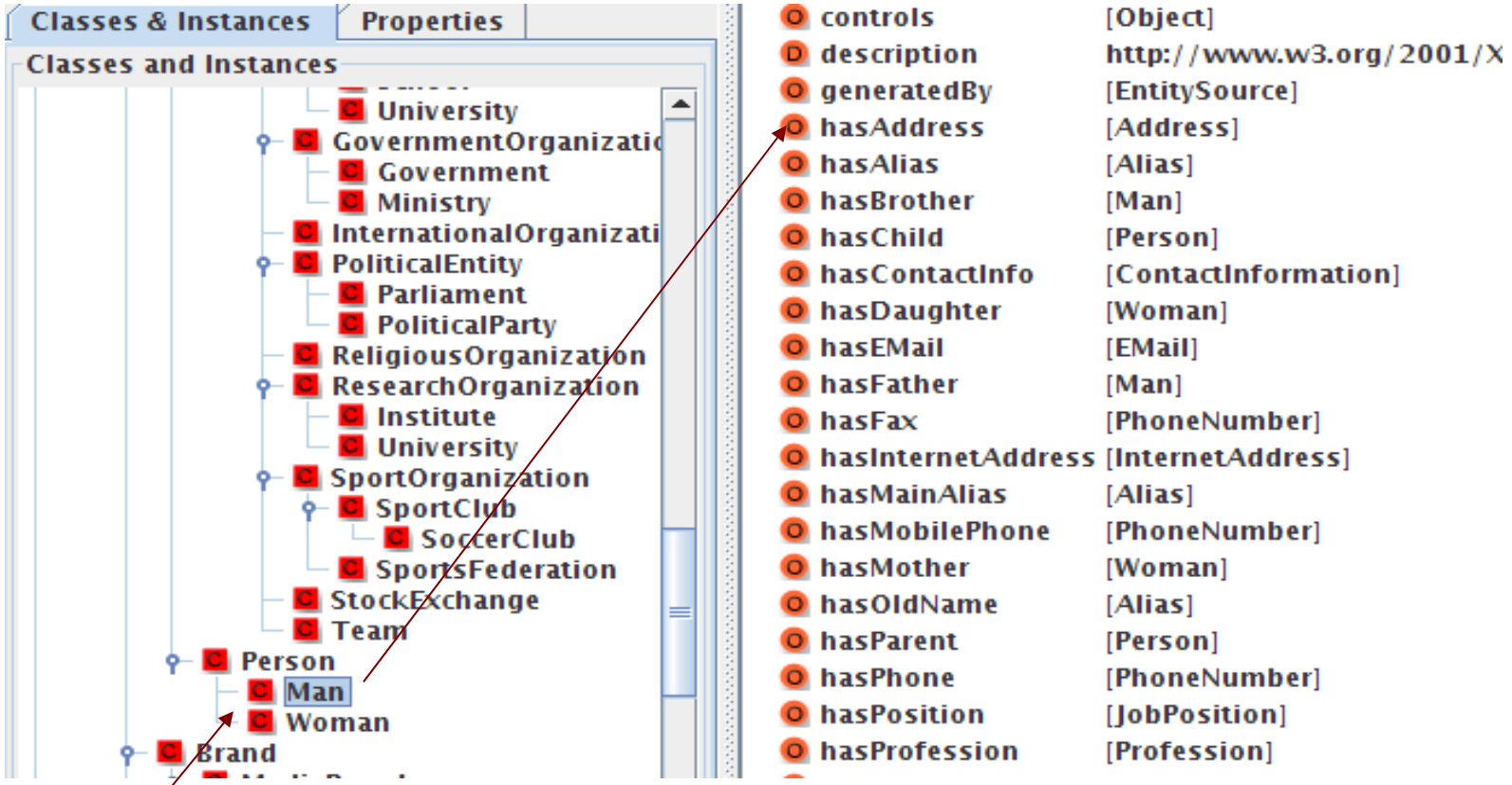

based_in
John lives in London . He works there for Polar Bear Design .
PER
LOC
ORG

# Richer NE Tagging

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

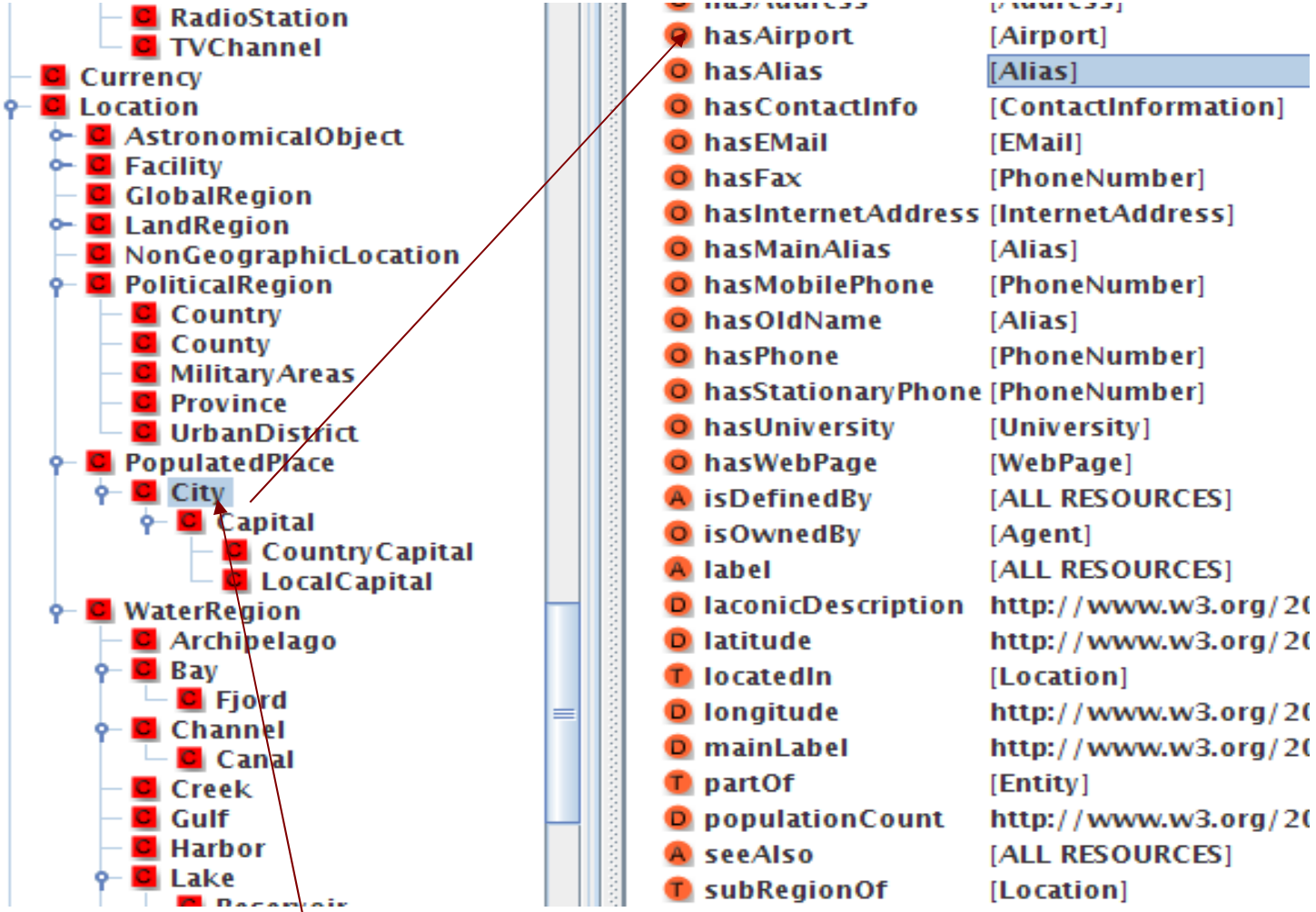


# Ontology-based IE




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

# Ontology-based IE (2)




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

# 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 John lives in 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 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 risks and the actors involved can help to improve system performance
- Backbone for other processes, for example visualising results on a timeline

# Text mining and semantic annotation

- Extract **structured data** from text by
  - Linking references to entities
  - Linking entities to their semantic descriptions
- Automatic **semantic annotation** based on IE technology
- Attaches **metadata** to documents, which makes them more useful and more easily processable
- They can then be used for searching and hyperlinking, categorising, and monitoring
- Adds value to content of libraries, enabling user **interaction** with content
- Enhanced capability for cross-referencing and **dynamic** document classification

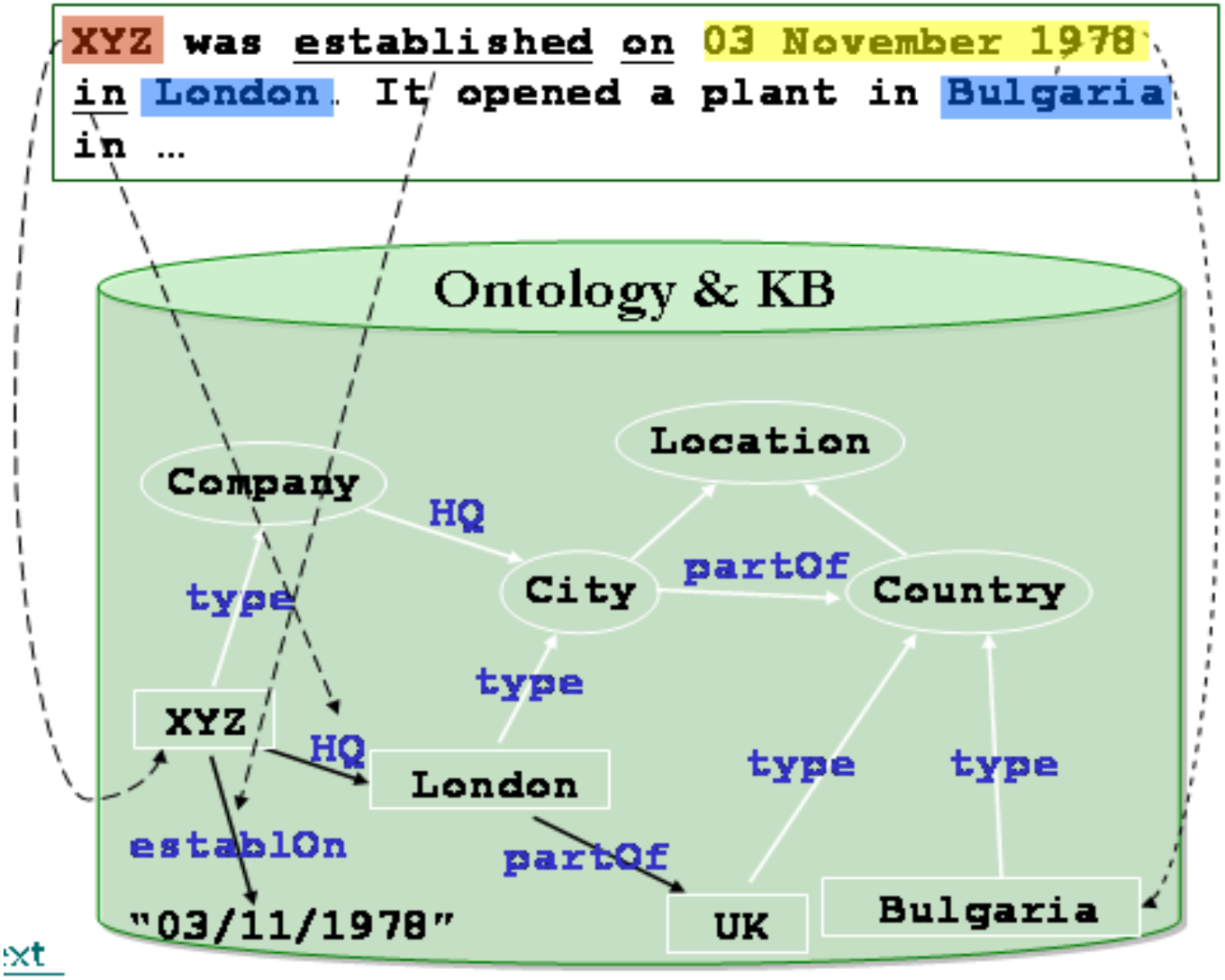
# Some Terminology

- **Semantic annotation** – annotate in the texts all mentions of instances relating to concepts in the ontology
- **Ontology learning** – automatically derive an ontology from texts
- **Ontology population** – given an ontology, populate the concepts with instances derived automatically from a text

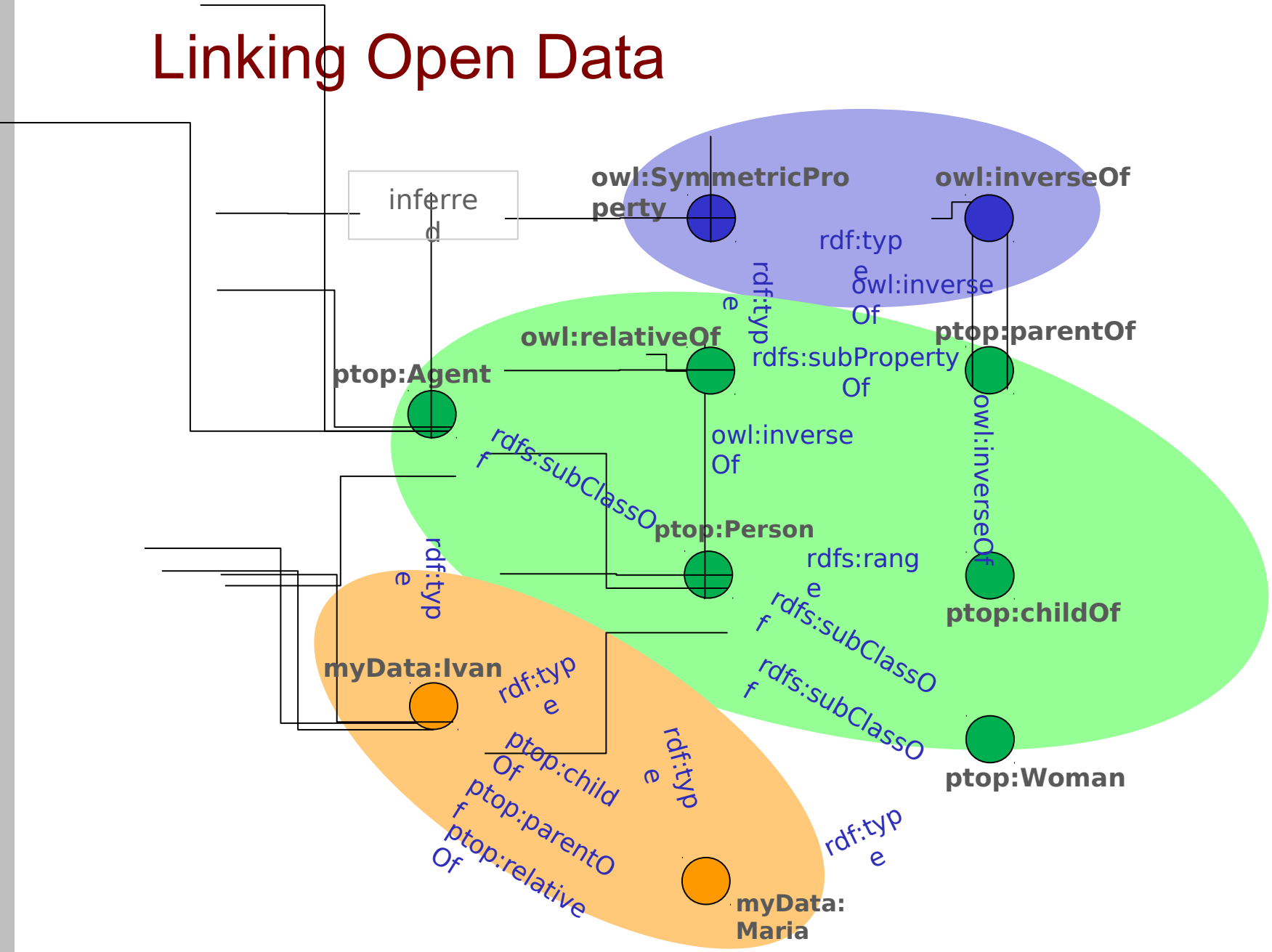
# Semantic Annotation vs Ontology Population

- **Semantic Annotation**
  - Mentions of instances in the text are annotated wrt concepts (classes) in the ontology.
  - Requires that instances are disambiguated.
  - It is the **document** which is modified.
- **Ontology Population**
  - Generates new instances in an ontology from a text.
  - Links unique mentions of instances in the text to instances of concepts in the ontology.
  - Instances must be not only disambiguated but also co-reference between them must be established.
  - It is the **ontology** which is modified.

# Ontology Population

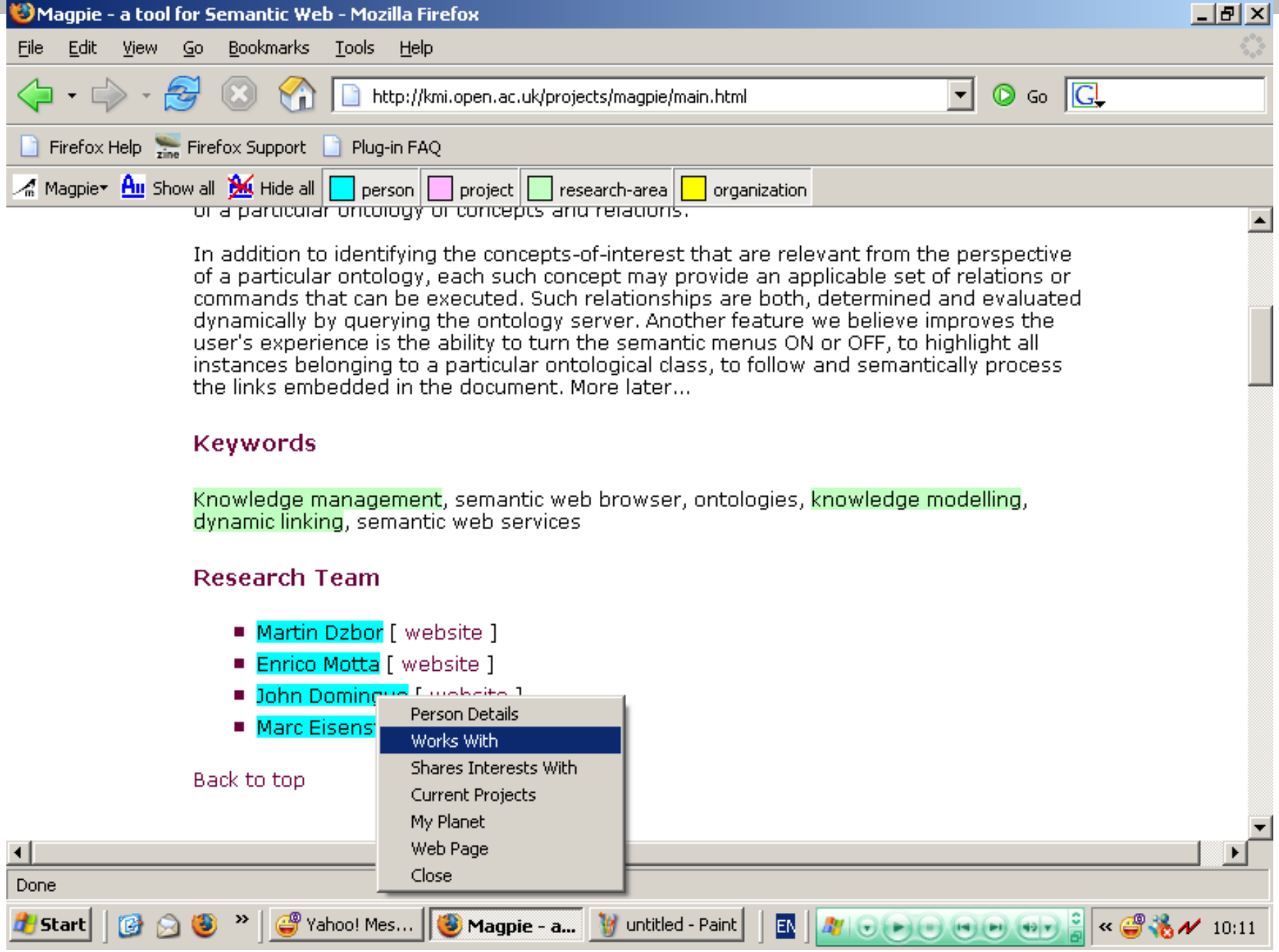


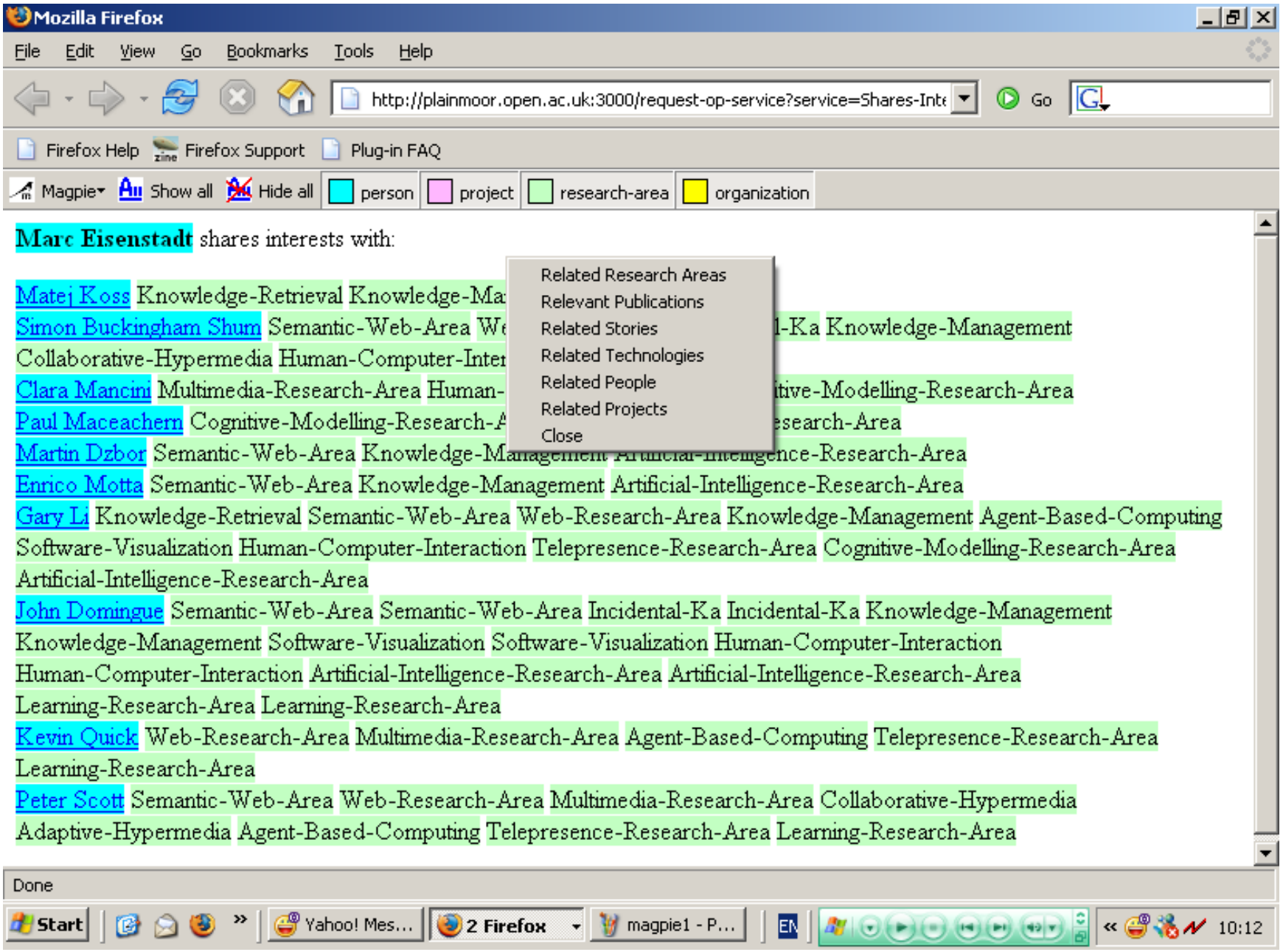
# Linking Open Data



# Magpie: semantic annotation in use

- Developed by the Open University
- Plugin for standard web browser
- Automatically associates an ontology-based semantic layer to web resources, allowing relevant services to be linked
- Provides means for a structured and informed exploration of the web resources
- e.g. looking at a list of publications, we can find information about an author such as projects they work on, other people they work with, etc.

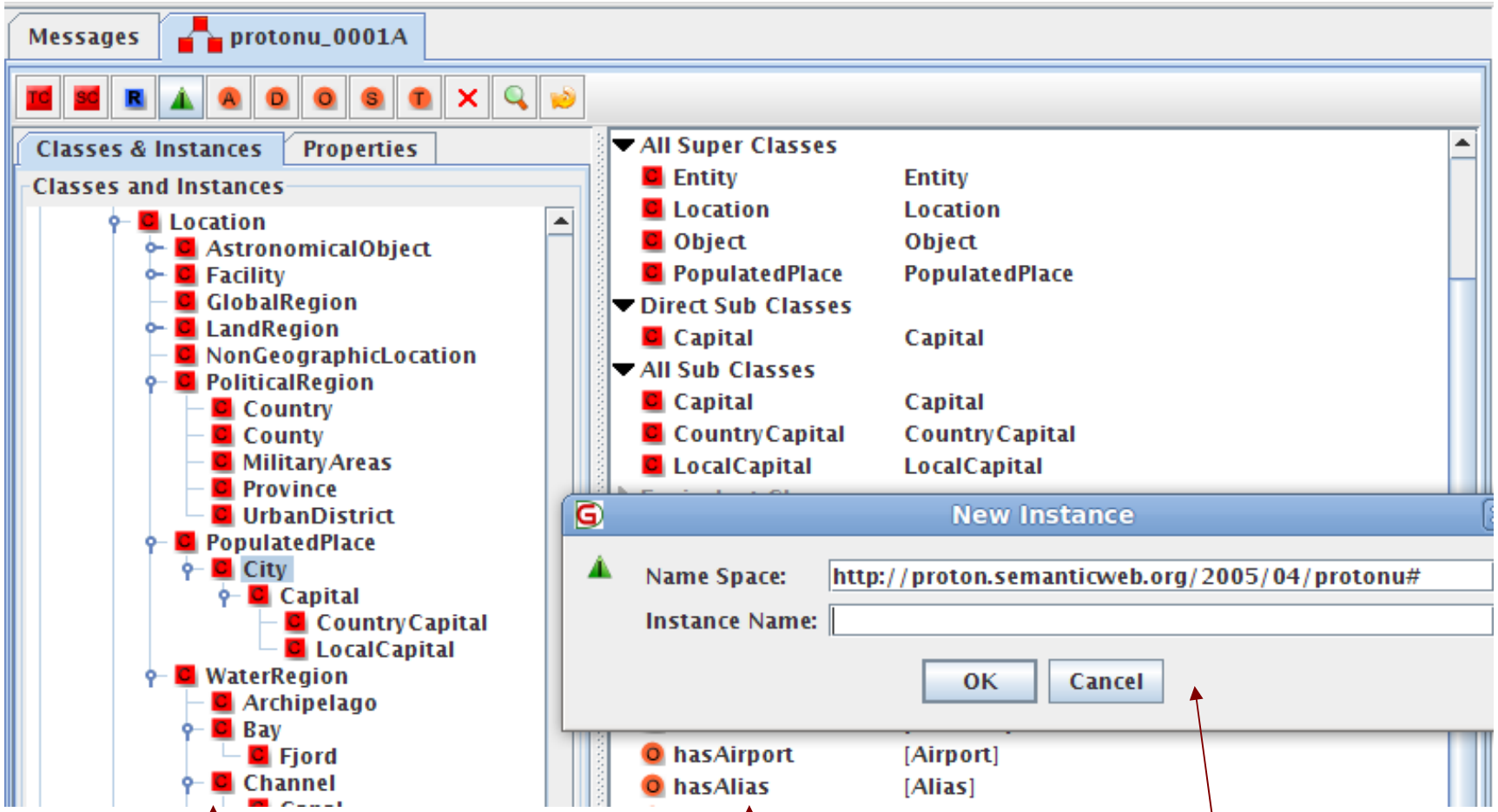

Magpie - a tool for Semantic Web - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://kmi.open.ac.uk/projects/magpie/main.html
Go
Firefox Help Firefox Support Plug-in FAQ
Magpie Show all Hide all person project research-area organization
of a particular ontology of concepts and relations.
In addition to identifying the concepts-of-interest that are relevant from the perspective of a particular ontology, each such concept may provide an applicable set of relations or commands that can be executed. Such relationships are both, determined and evaluated dynamically by querying the ontology server. Another feature we believe improves the user's experience is the ability to turn the semantic menus ON or OFF, to highlight all instances belonging to a particular ontological class, to follow and semantically process the links embedded in the document. More later...
Keywords
Knowledge management, semantic web browser, ontologies, knowledge modelling, dynamic linking, semantic web services
Research Team
Martin Dzbor [ website ]
Enrico Motta [ website ]
John Doming
Marc Eisens
Person Details
Works With
Shares Interests With
Current Projects
My Planet
Web Page
Close
Back to top
Done
Start
Yahoo! Mes...
Magpie - a...
untitled - Paint
EN
10:11


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://plainmoor.open.ac.uk:3000/request-op-service?service=Shares-Int
Go
Firefox Help
Firefox Support
Plug-in FAQ
Magpie
Show all
Hide all
person
project
research-area
organization
Marc Eisenstadt shares interests with:
Related Research Areas
Relevant Publications
Related Stories
Related Technologies
Related People
Related Projects
Close
Matej Koss Knowledge-Retrieval Knowledge-Ma
Simon Buckingham Shum Semantic-Web-Area We
l-Ka Knowledge-Management
Collaborative-Hypermedia Human-Computer-Inter
Clara Mancini Multimedia-Research-Area Human-
itive-Modelling-Research-Area
Paul Maceachern Cognitive-Modelling-Research-A
esearch-Area
Martin Dzbor Semantic-Web-Area Knowledge-Ma
nce-Research-Area
Enrico Motta Semantic-Web-Area Knowledge-Management Artificial-Intelligence-Research-Area
Gary Li Knowledge-Retrieval Semantic-Web-Area Web-Research-Area Knowledge-Management Agent-Based-Computing
Software-Visualization Human-Computer-Interaction Telepresence-Research-Area Cognitive-Modelling-Research-Area
Artificial-Intelligence-Research-Area
John Domingue Semantic-Web-Area Semantic-Web-Area Incidental-Ka Incidental-Ka Knowledge-Management
Knowledge-Management Software-Visualization Software-Visualization Human-Computer-Interaction
Human-Computer-Interaction Artificial-Intelligence-Research-Area Artificial-Intelligence-Research-Area
Learning-Research-Area Learning-Research-Area
Kevin Quick Web-Research-Area Multimedia-Research-Area Agent-Based-Computing Telepresence-Research-Area
Learning-Research-Area
Peter Scott Semantic-Web-Area Web-Research-Area Multimedia-Research-Area Collaborative-Hypermedia
Adaptive-Hypermedia Agent-Based-Computing Telepresence-Research-Area Learning-Research-Area
Done
Start
Yahoo! Mes...
2 Firefox
magpie1 - P...
EN
10:12

# 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

# GATE Ontology Editor

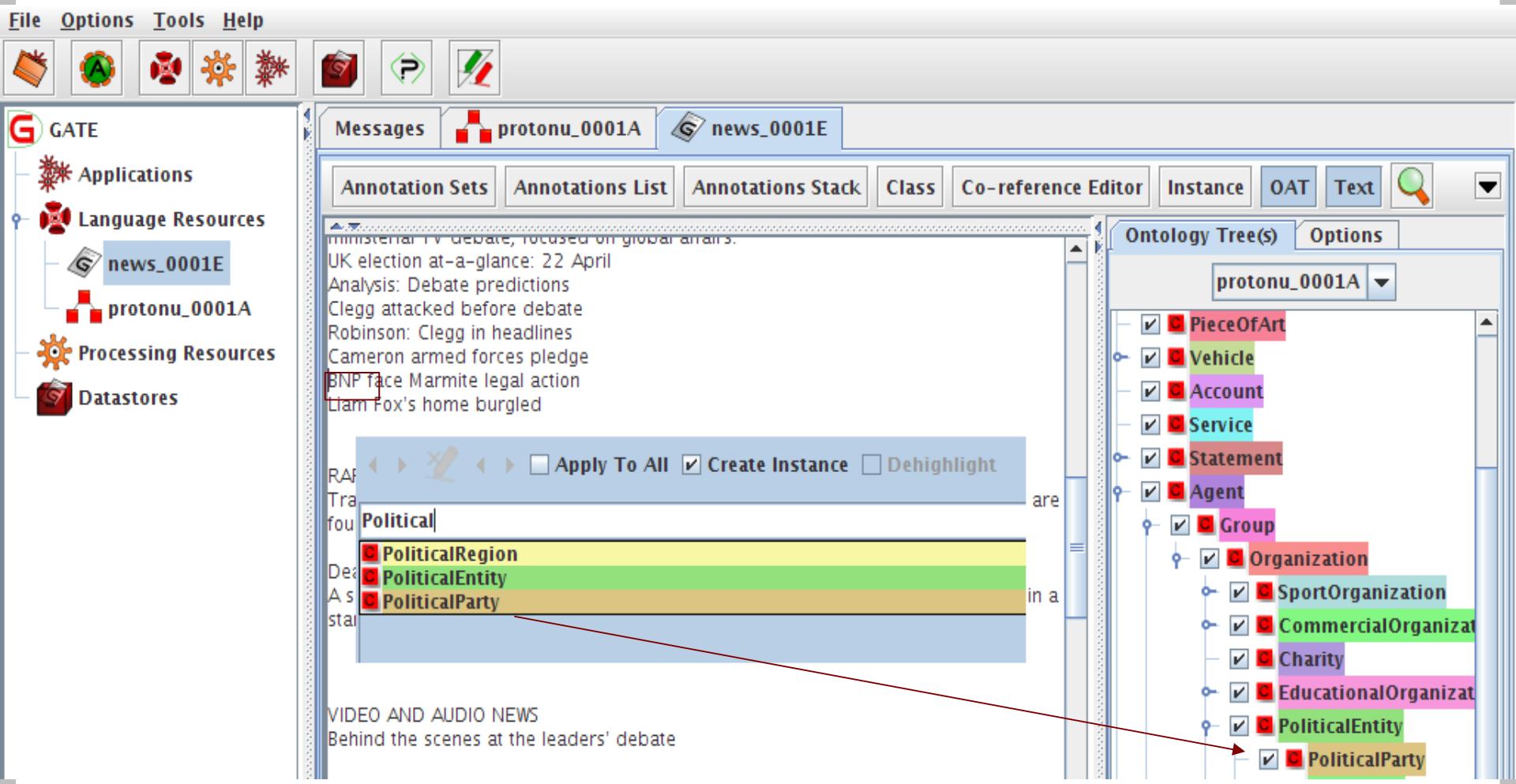



Classes and Instances

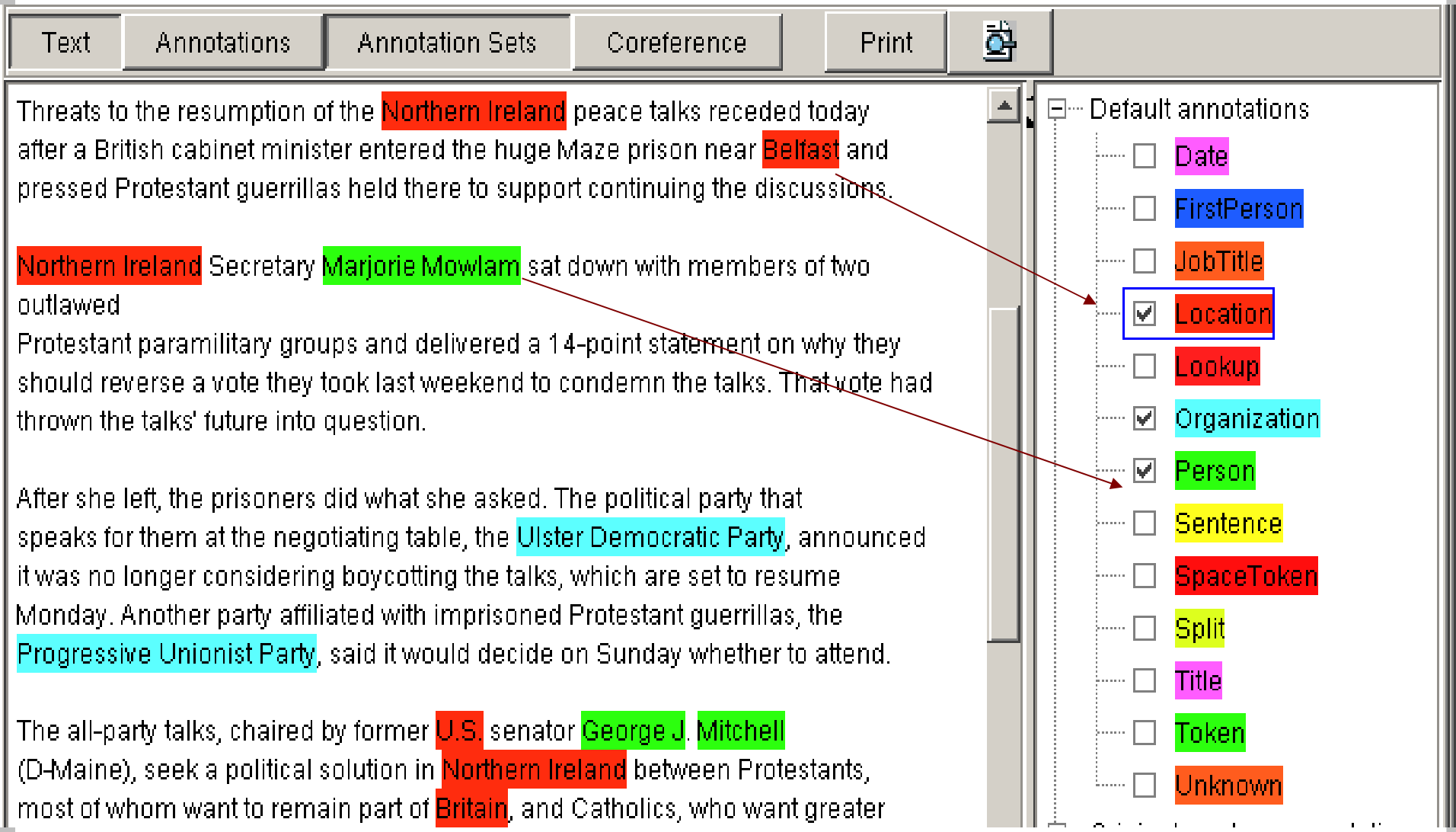
Properties

Adding a new instance

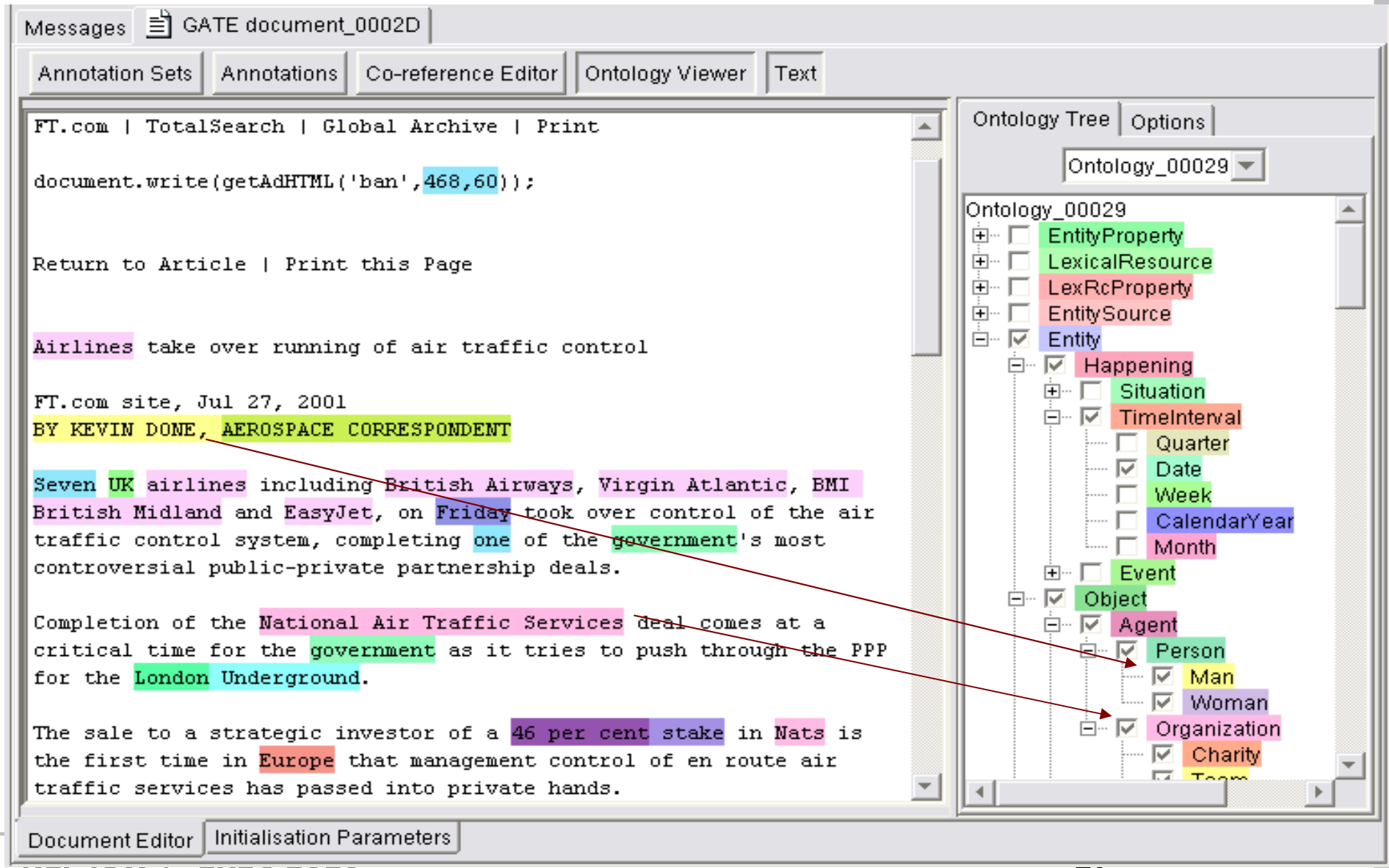
# Manual annotation with OAT



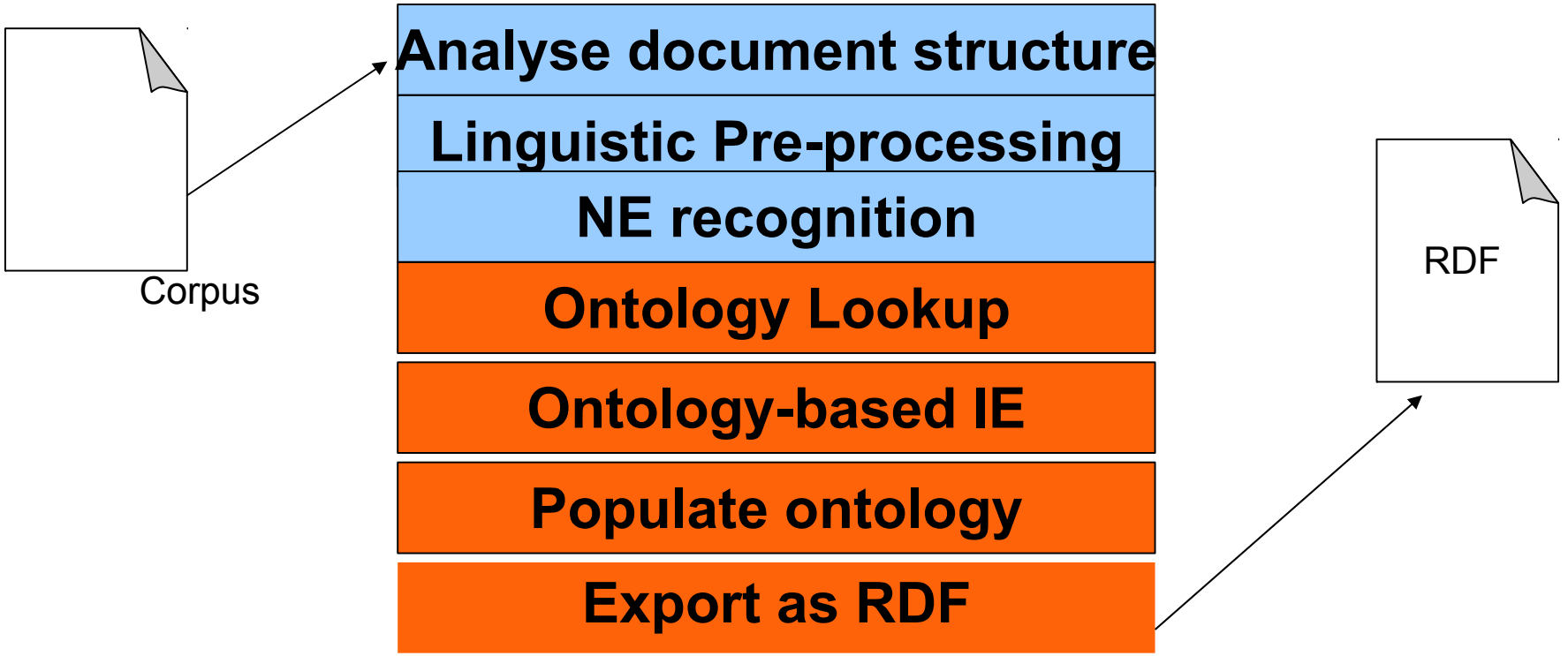
# Traditional IE in GATE



# Semantic IE in GATE



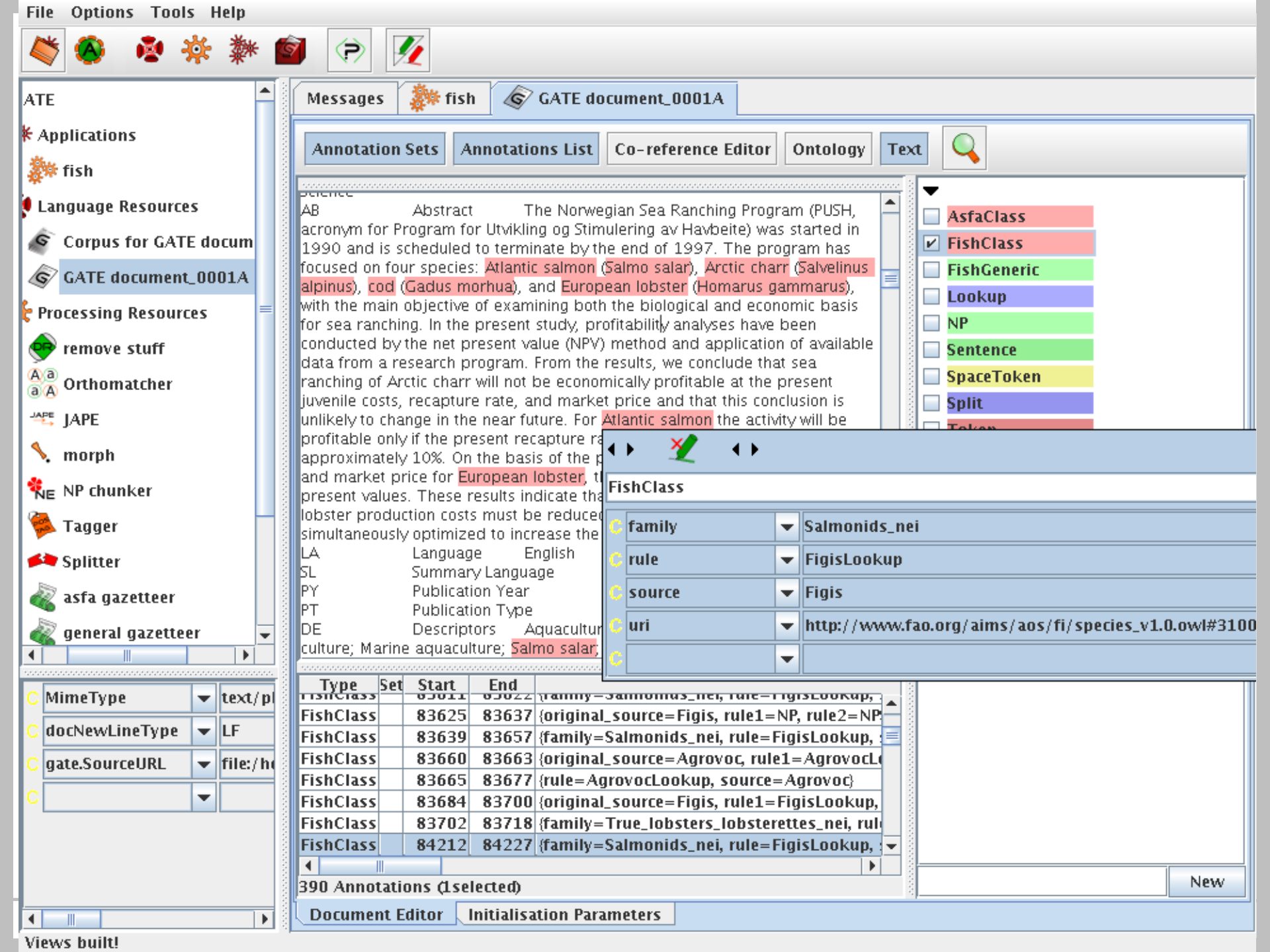
# Typical Semantic Annotation pipeline

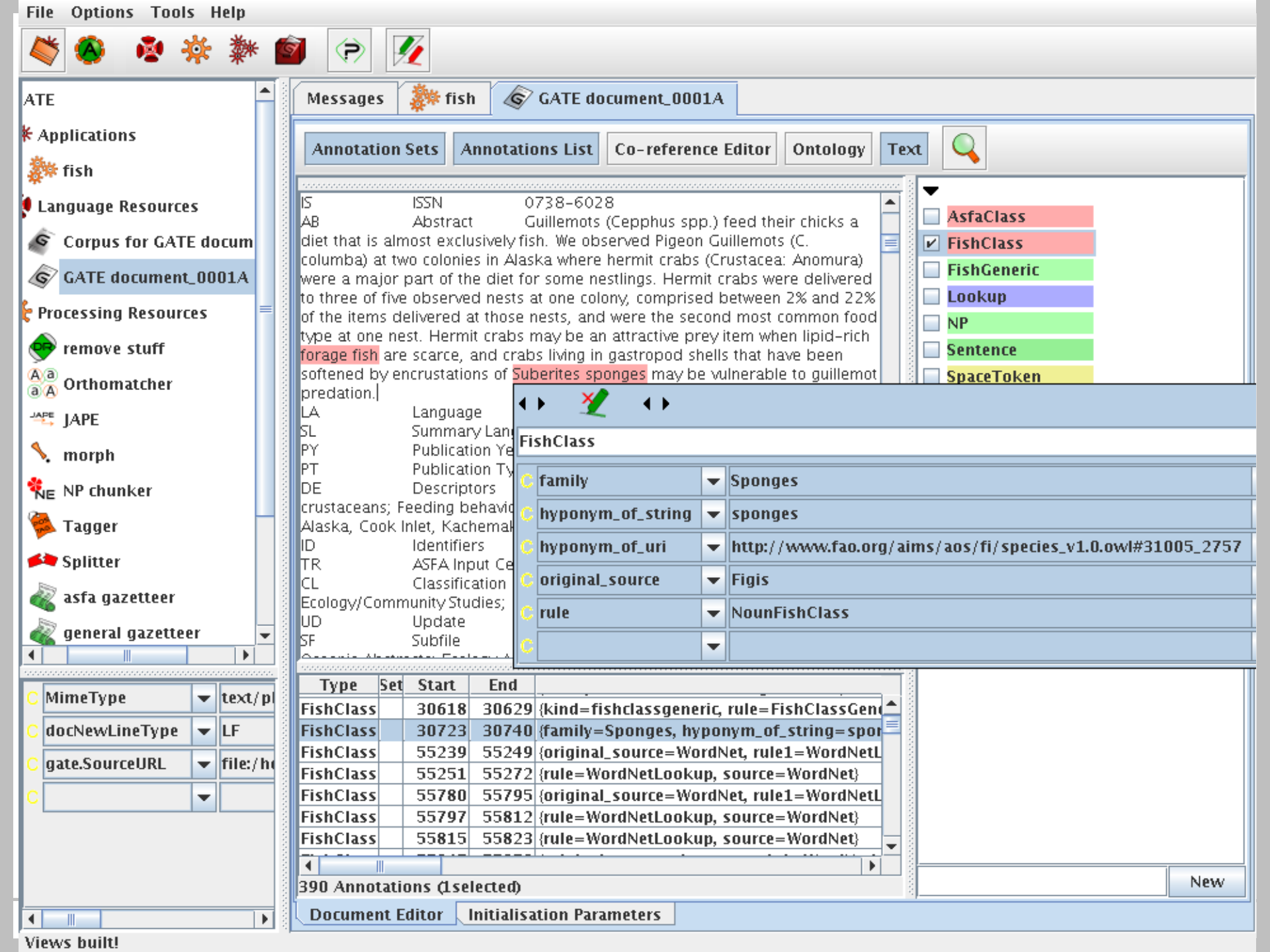


# SPRAT

## Semantic Pattern Recognition and Annotation Tool

- Combination of semantic annotation, ontology population and learning
- Identifies
  - existing classes and instances listed in the ontology and their morphological variants
  - potential new classes and instances not listed in the ontology
  - potential relations between the classes and instances, which can be encoded as properties in the ontology
- For the new entities found, it attempts to classify them in the ontology, based on linguistic information such as synonyms and hyponyms of existing entities, or to add new properties


File Options Tools Help
ATE
Applications
fish
Language Resources
Corpus for GATE docum
GATE document_0001A
Processing Resources
remove stuff
Orthomatcher
JAPE
morph
NP chunker
Tagger
Splitter
asfa gazetteer
general gazetteer
MimeType
text/pl
docNewLineType
LF
gate.SourceURL
file:/h
Messages
fish
GATE document_0001A
Annotation Sets
Annotations List
Co-reference Editor
Ontology
Text
AB Abstract The Norwegian Sea Ranching Program (PUSH, acronym for Program for Utvikling og Stimulering av Havbeite) was started in 1990 and is scheduled to terminate by the end of 1997. The program has focused on four species: Atlantic salmon (Salmo salar), Arctic charr (Salvelinus alpinus), cod (Gadus morhua), and European lobster (Homarus gammarus), with the main objective of examining both the biological and economic basis for sea ranching. In the present study, profitability analyses have been conducted by the net present value (NPV) method and application of available data from a research program. From the results, we conclude that sea ranching of Arctic charr will not be economically profitable at the present juvenile costs, recapture rate, and market price and that this conclusion is unlikely to change in the near future. For Atlantic salmon the activity will be
LA Language English
SL Summary Language
PY Publication Year
PT Publication Type
DE Descriptors
culture; Marine aquaculture; Salmo salar;
AsfaClass
FishClass
FishGeneric
Lookup
NP
Sentence
SpaceToken
Split
FishClass
family Salmonids_nei
rule FigisLookup
source Figis
uri http://www.fao.org/aims/aos/fi/species_v1.0.owl#3100
Type Set Start End
FishClass 83625 83637 {original_source=Figis, rule1=NP, rule2=NP
FishClass 83639 83657 {family=Salmonids_nei, rule=FigisLookup,
FishClass 83660 83663 {original_source=Agrovoc, rule1=AgrovocL
FishClass 83665 83677 {rule=AgrovocLookup, source=Agrovoc}
FishClass 83684 83700 {original_source=Figis, rule1=FigisLookup,
FishClass 83702 83718 {family=True_lobsters_lobsterettes_nei, rul
FishClass 84212 84227 {family=Salmonids_nei, rule=FigisLookup,
390 Annotations (1selected)
New
Document Editor
Initialisation Parameters
Views built!


File Options Tools Help
ATE
Applications
fish
Language Resources
Corpus for GATE docum
GATE document_0001A
Processing Resources
remove stuff
Orthomatcher
JAPE
morph
NP chunker
Tagger
Splitter
asfa gazetteer
general gazetteer
Messages
fish
GATE document_0001A
Annotation Sets
Annotations List
Co-reference Editor
Ontology
Text
IS ISSN 0738-6028
AB Abstract Guillemots (Cepphus spp.) feed their chicks a diet that is almost exclusively fish. We observed Pigeon Guillemots (C. columba) at two colonies in Alaska where hermit crabs (Crustacea: Anomura) were a major part of the diet for some nestlings. Hermit crabs were delivered to three of five observed nests at one colony, comprised between 2% and 22% of the items delivered at those nests, and were the second most common food type at one nest. Hermit crabs may be an attractive prey item when lipid-rich forage fish are scarce, and crabs living in gastropod shells that have been softened by encrustations of Suberites sponges may be vulnerable to guillemot predation.
LA Language
SL Summary Lan
PY Publication Ye
PT Publication Ty
DE Descriptors
crustaceans; Feeding behavic
Alaska, Cook Inlet, Kachemak
ID Identifiers
TR ASFA Input Ce
CL Classification
Ecology/Community Studies;
UD Update
SF Subfile
AsfaClass
FishClass
FishGeneric
Lookup
NP
Sentence
SpaceToken
FishClass
family Sponges
hyponym_of_string sponges
hyponym_of_uri http://www.fao.org/aims/aos/fi/species_v1.0.owl#31005_2757
original_source Figis
rule NounFishClass
MimeType text/pl
docNewLineType LF
gate.SourceURL file:/h
Type Set Start End
FishClass 30618 30629 {kind=fishclassgeneric, rule=FishClassGen
FishClass 30723 30740 {family=Sponges, hyponym_of_string=spor
FishClass 55239 55249 {original_source=WordNet, rule1=WordNetL
FishClass 55251 55272 {rule=WordNetLookup, source=WordNet}
FishClass 55780 55795 {original_source=WordNet, rule1=WordNetL
FishClass 55797 55812 {rule=WordNetLookup, source=WordNet}
FishClass 55815 55823 {rule=WordNetLookup, source=WordNet}
390 Annotations (1selected)
New
Document Editor
Initialisation Parameters
Views built!

# Augmenting the Ontology

- The new classes found are linked to existing classes in the ontology
- For existing classes, and new classes which we identified as a synonym or hyponym of an existing classes, the link is to an existing ontology class
- When we don't identify a link to any existing class, we create a new class
- The changes to the ontology are stored and can be verified later by human experts

# Recognising components from the ontology

- In addition to the standard IE components, we use some special ontology components.
- The OntoRootGazetteer enables us to match words or phrases in the text with classes, instances or properties in an ontology, as any morphological variant
- 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)
- When new elements are added to the ontology, these features can be used to regenerate alternative forms

# Modifying the ontology

- We use technology taken from CLONE, called NEBONE (Named Entity Based ONtology Editing)
- CLONE is designed to create new classes, instances etc from raw (controlled) text generated by the user
- NEBONE enables changes to be made to the ontology based on IE from input texts (e.g. web pages) in natural language
- Morphological analysis enables both root forms and variants to be added to the ontology (as properties), along with other variants (e.g. capitalisation)

# Examples

“species of shark, such as hornsharks, leopard sharks and catfish”

-->

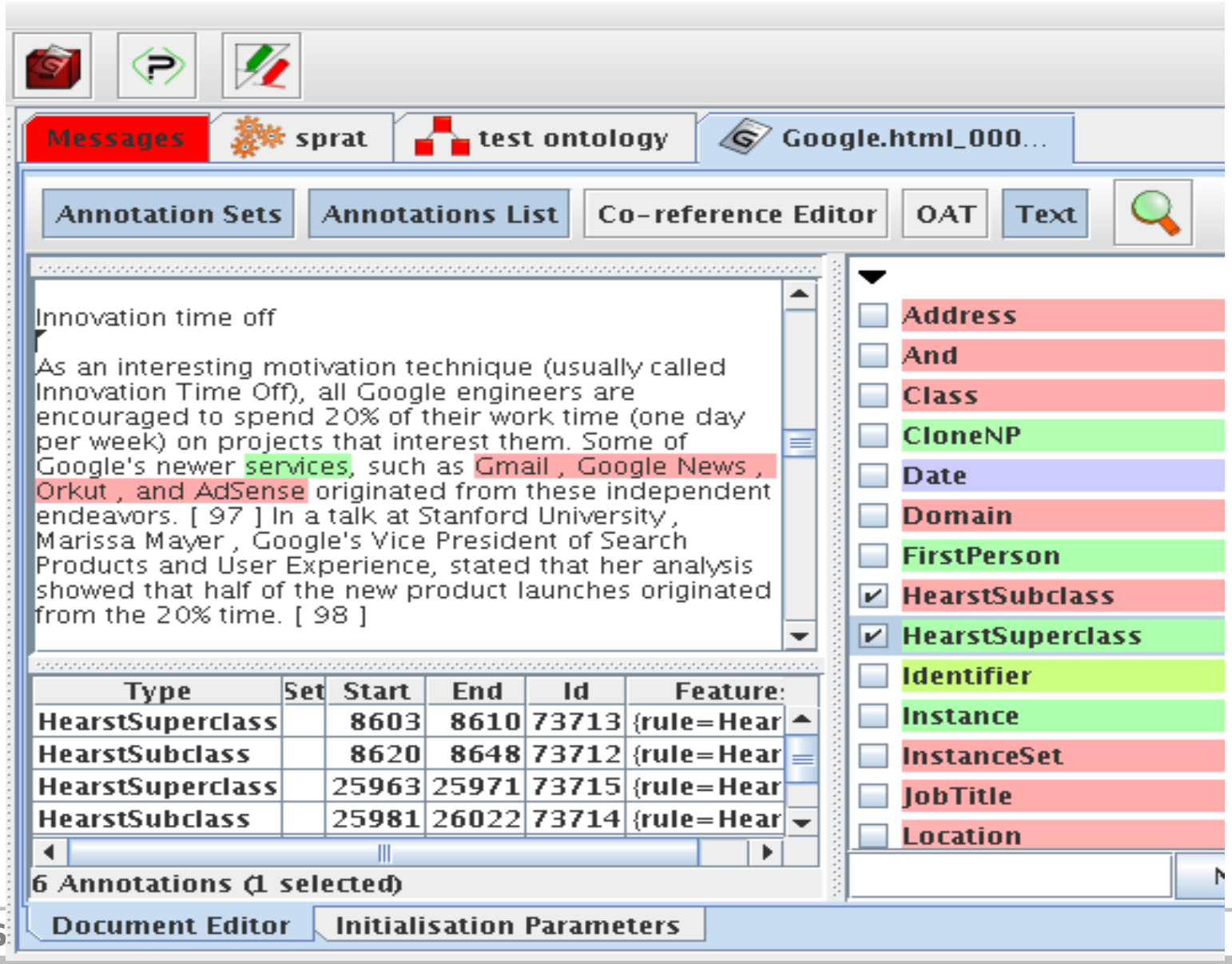
“hornshark”, “leopard shark” and "catfish” are added as subclasses of “shark”

“Google has begun experimenting with other markets, such as radio and print publications”

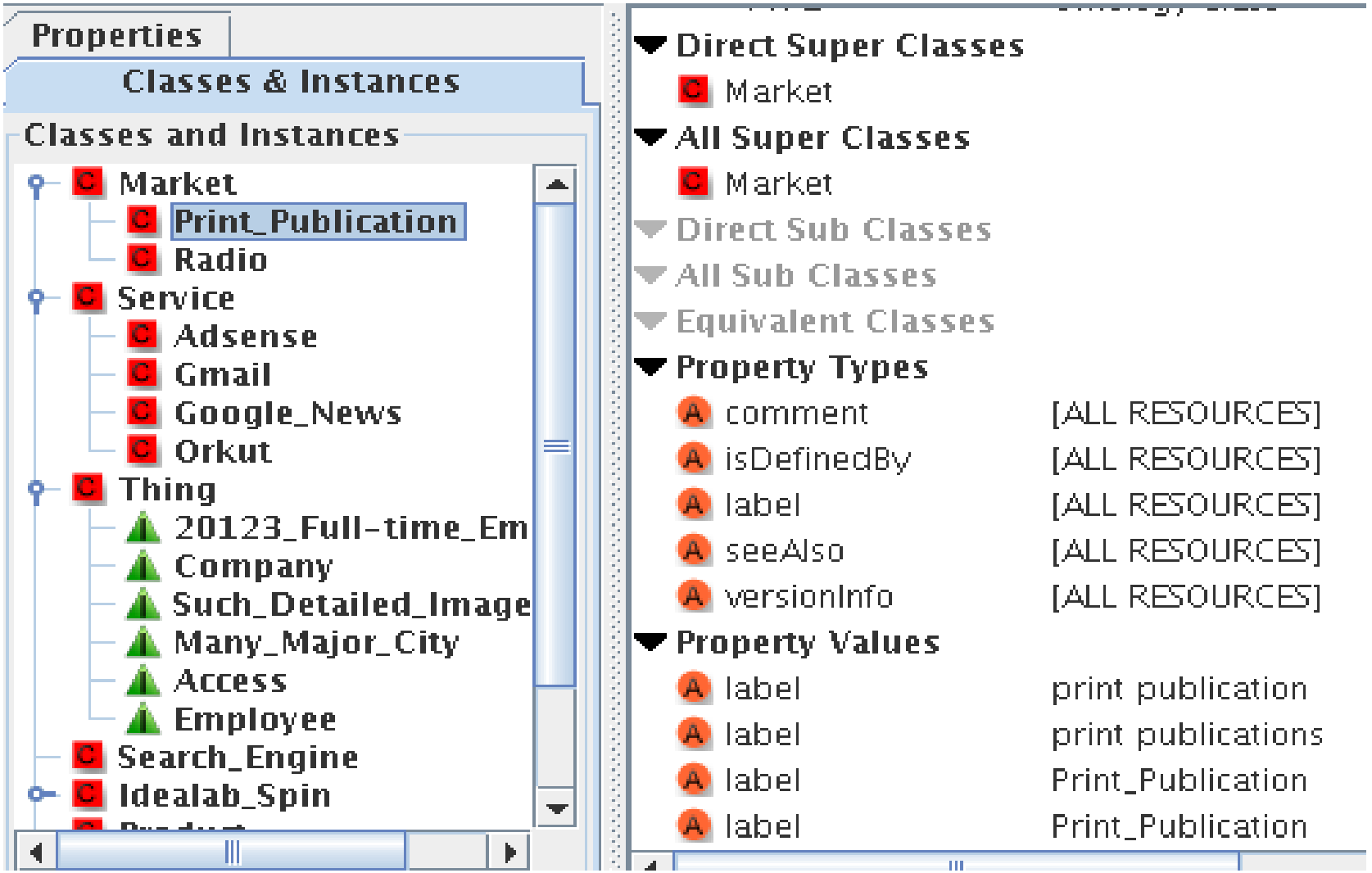
-->

“radio” and “print publication” added as subclasses of “market”

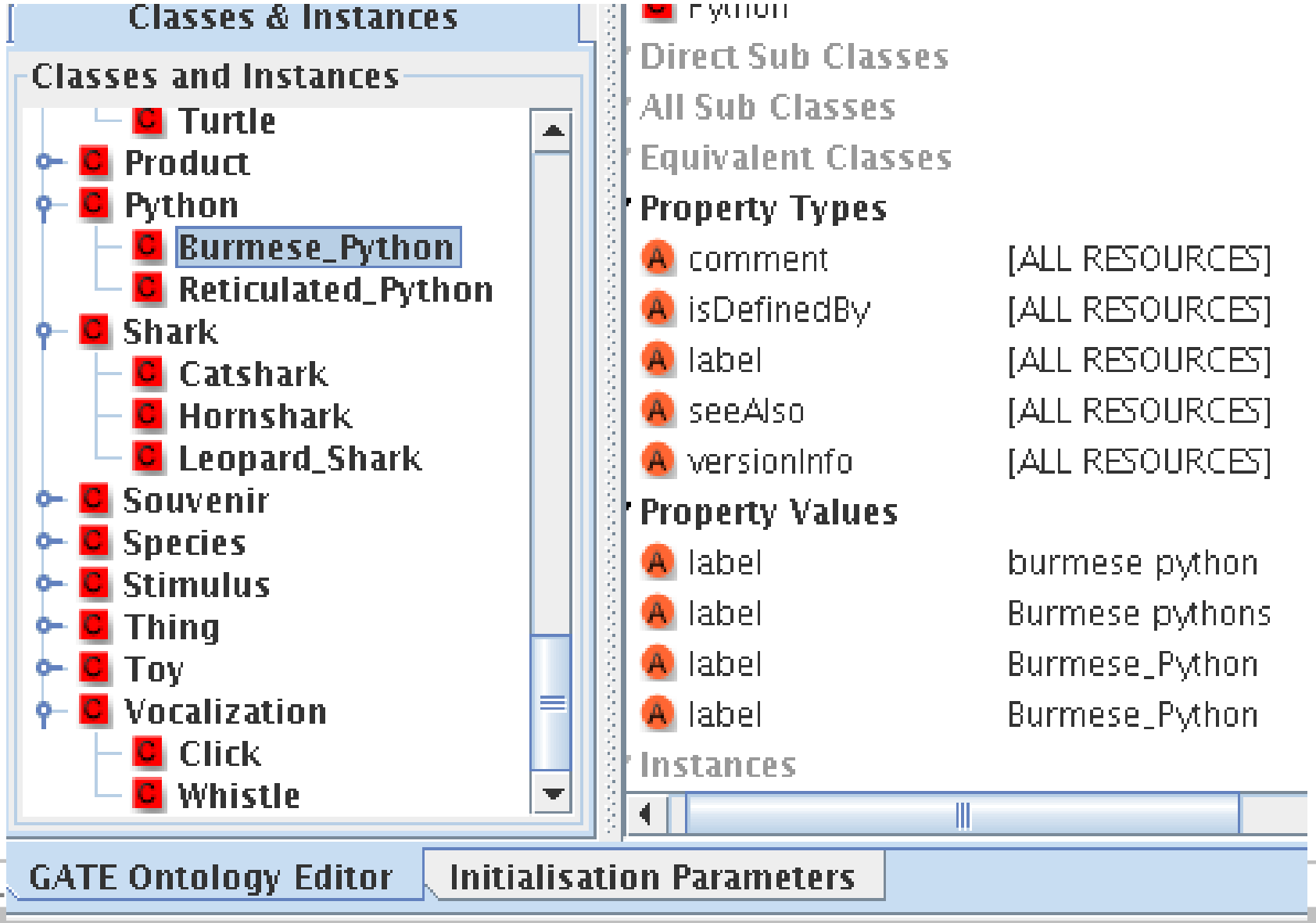
# Annotated Google wikipedia page



# Ontology from Google wikipedia page



# Generated “animal” ontology



# Evaluation of Semantic Annotation

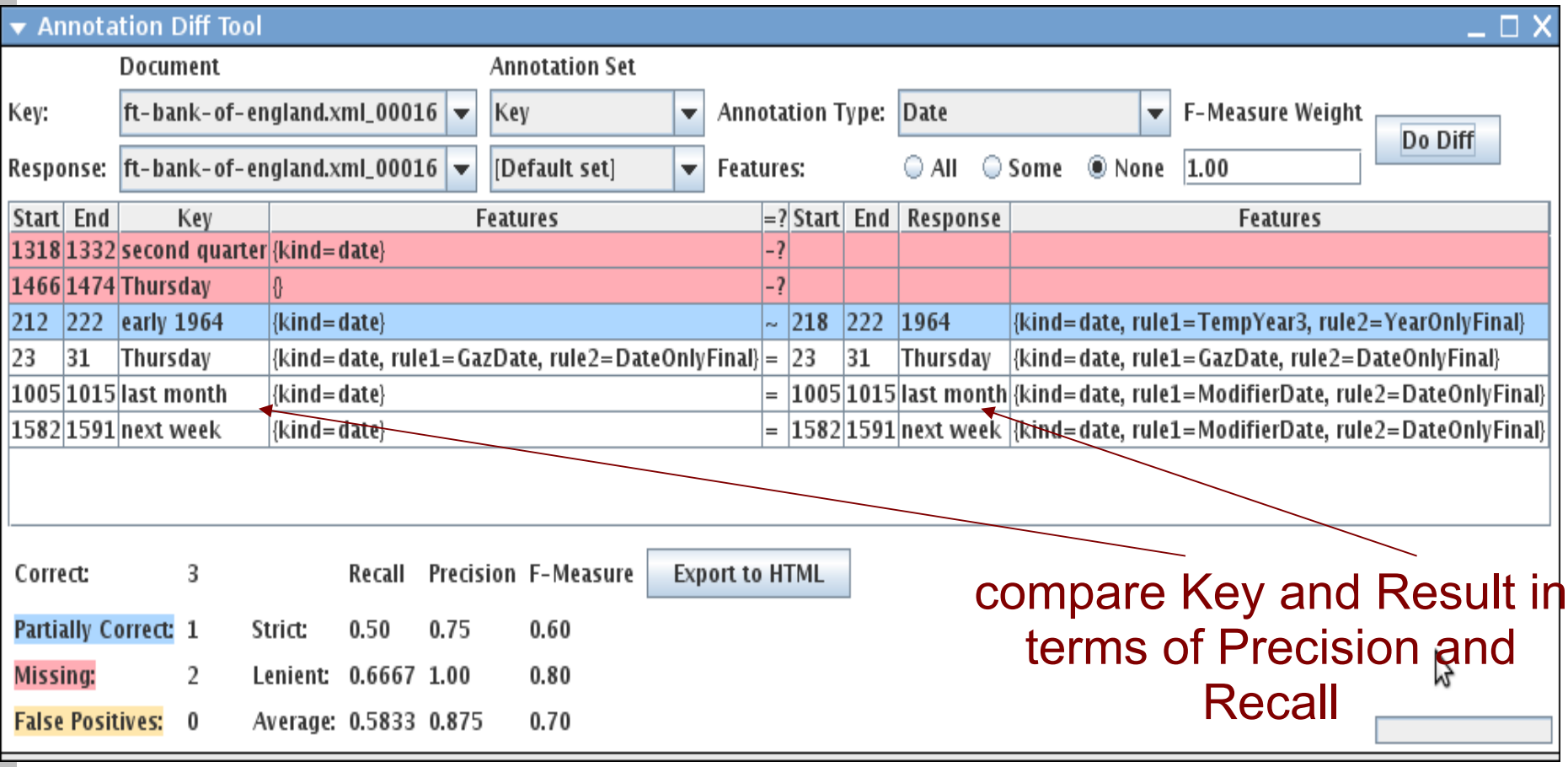
“We didn’t underperform. You overexpected.”

# 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




compare Key and Result in terms of Precision and Recall

# 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
- Recognising a *Person* as a *Location* is clearly wrong, but recognising a *Research Assistant* as a *Lecturer* is not so wrong
- Similarity metrics need to be integrated so that items closer together in the hierarchy are given a higher score, if wrong

# Augmented Precision and Recall

- BDM (**Balanced Distance Metric**) compares key and response concepts wrt a given ontology
- In the case of ontological mismatch, provides an indication of how serious the error is, and weights it accordingly
- BDM provides a score between 0 and 1 for each key/response match instead of a binary measure

# Augmented Precision and Recall

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

$$AP = \frac{BDM}{BDM + Spurio} \qquad AR = \frac{BDM}{BDM + Missi}$$

# 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 |

# 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

Messages | sara-marked | IAA

```
(0.0, 0.15555556, 0.6888889, 0.6);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.18421052,
0.20588236, 0.19444445)

pairAnns=(Key1,Output1), type=Mention, label=OperationalRisk:   (correct, partial, spurious, missing)=
(0.0, 0.0, 0.0, 0.044444446);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.0, 0.0, 0.0)

pairAnns=(Key1,Output1), type=Mention, label=PandemicThreat:   (correct, partial, spurious, missing)=
(0.0, 0.022222223, 0.022222223, 0.0);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.5, 1.0,
0.6666667)

pairAnns=(Key1,Output1), type=Mention, label=SafetyIssue:   (correct, partial, spurious, missing)=
(0.0, 0.022222223, 0.2, 0.022222223);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.1, 0.5,
0.16666666)

pairAnns=(Key1,Output1), type=Mention, label=SecurityConcern:   (correct, partial, spurious, missing)=
(0.0, 0.044444446, 0.22222222, 0.06666667);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient:
(0.16666666, 0.39999998, 0.2352941)

Overall pairs and types:   (correct, partial, spurious, missing)= (0.0, 0.37777779, 1.4888889,
0.8666667);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.20238096, 0.30357143, 0.24285716)


********  The F-measure based on the BDM scores specified in the following:

Macro averaged over 45 documents:

For each pair of annotators, each type and each label:
Annotation type *Mention*
For each pair of annotators
For pair (Key1,Output1):   (correct, partial, spurious, missing)= (0.0, 0.7111111, 1.1555556,
0.8888889);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.29259259, 0.3057239, 0.28395063)

Overall pairs:   (correct, partial, spurious, missing)= (0.0, 0.7111111, 1.1555556, 0.8888889);
(precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.29259259, 0.3057239, 0.28395063)

Overall pairs and types:   (correct, partial, spurious, missing)= (0.0, 0.7111111, 1.1555556,
0.8888889);  (precision, recall, F1)= (0.0, 0.0, 0.0);  Lenient: (0.29259259, 0.3057239, 0.28395063)


 ********
```

# Summary

- Why semantic annotation?
  - Motivation for moving from traditional to semantic annotation
- How to do it?
  - Examples of practical applications in GATE
- Why evaluate semantic annotation differently?
  - BDM measure
- How to do it?
  - Example of GATE IAA tool