

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

© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com



“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

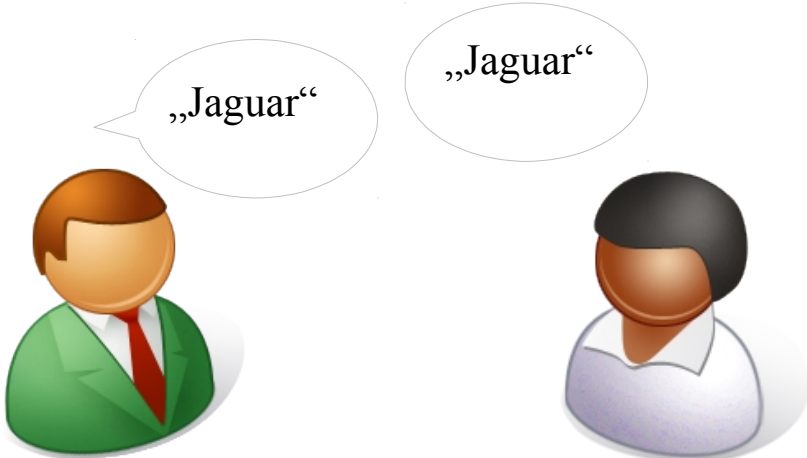
[Noy and McGuinness 2001]

What is an Ontology?

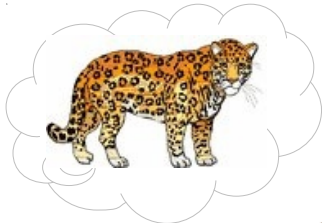
- Set of concepts (instances and classes)
- Relationships between them (is-a, part-of, located-in)
- Multiple inheritance
 - Classes can have more than one parent
 - Instances can have more than one class
- Ontologies are graphs, not trees



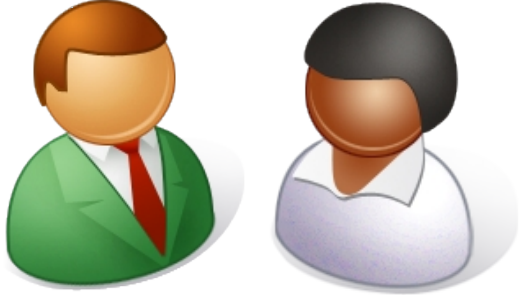
Example: Common Understanding



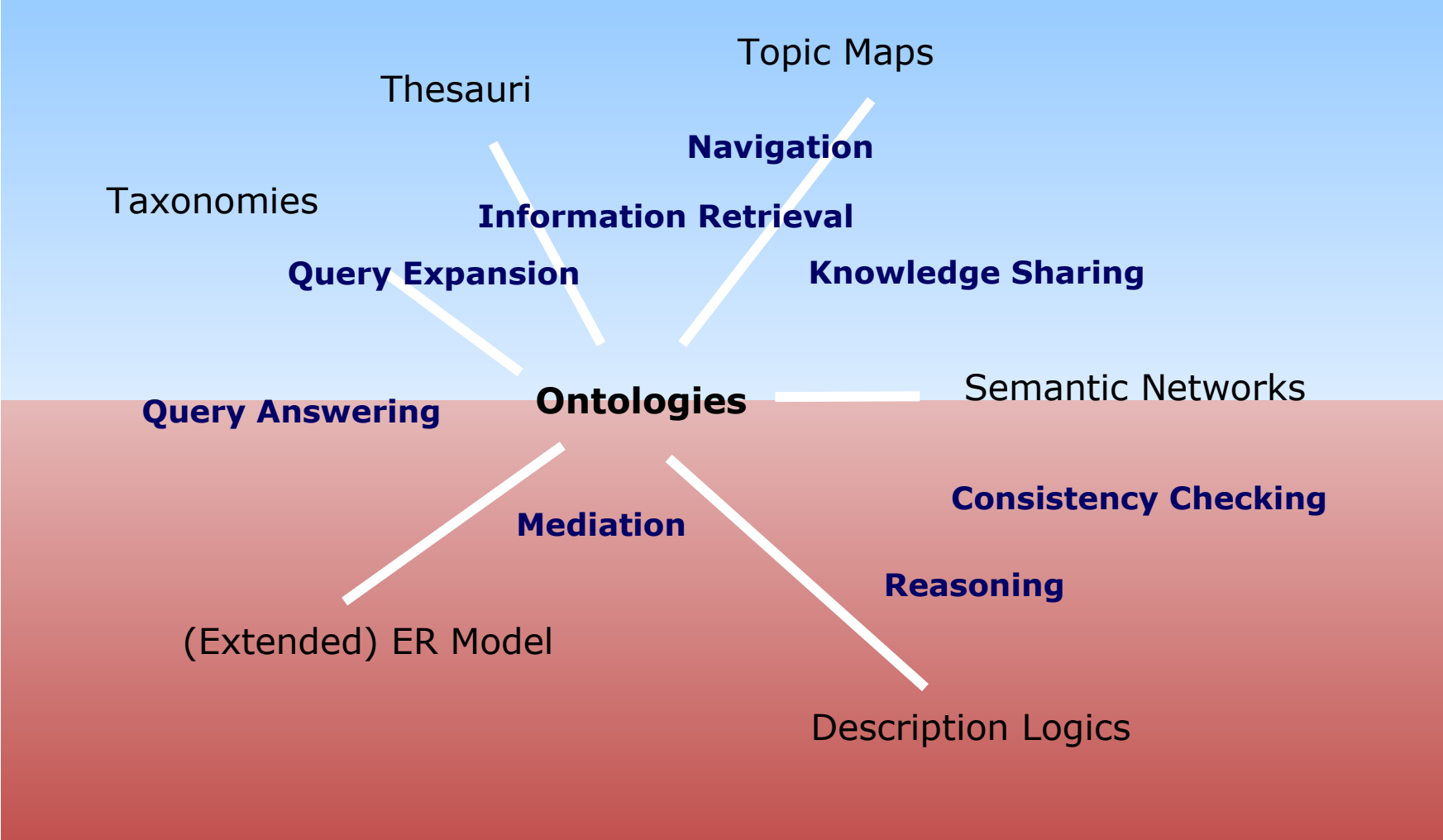
http://en.wikipedia.org/wiki/SEPECAT_Jaguar



Jaguar	
	
<small>A French Air Force Jaguar A during a vertical climb over the Ad-Elle Sea, in support of Operation Joint Forge</small>	
Role	Ground attack
Manufacturer	SEPECAT (Breguet/DAC)
First flight	8 September 1968
Introduced	1973
Retired	2005 (France) / 2007 (UK)
Status	Active
Primary users	Royal Air Force Armée de l'Air Indes Air Force
Builder built	543 (?)
Unit cost	US\$15.5-16.5 million in 1997(?)

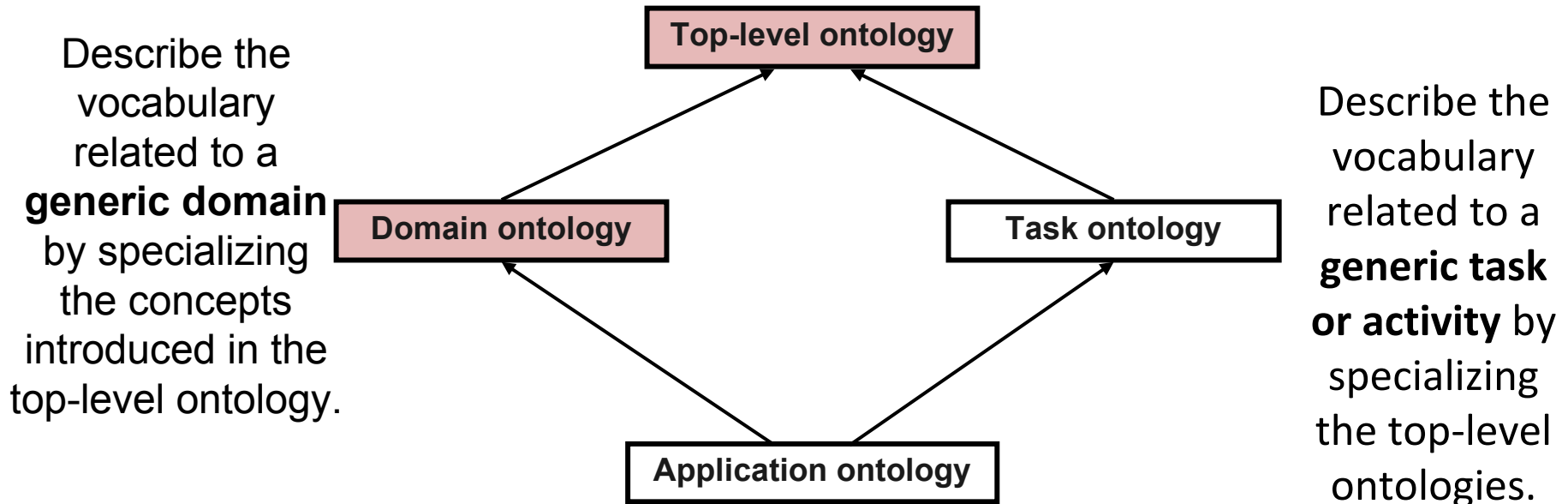


Ontologies on the web



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.



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

DOLCE

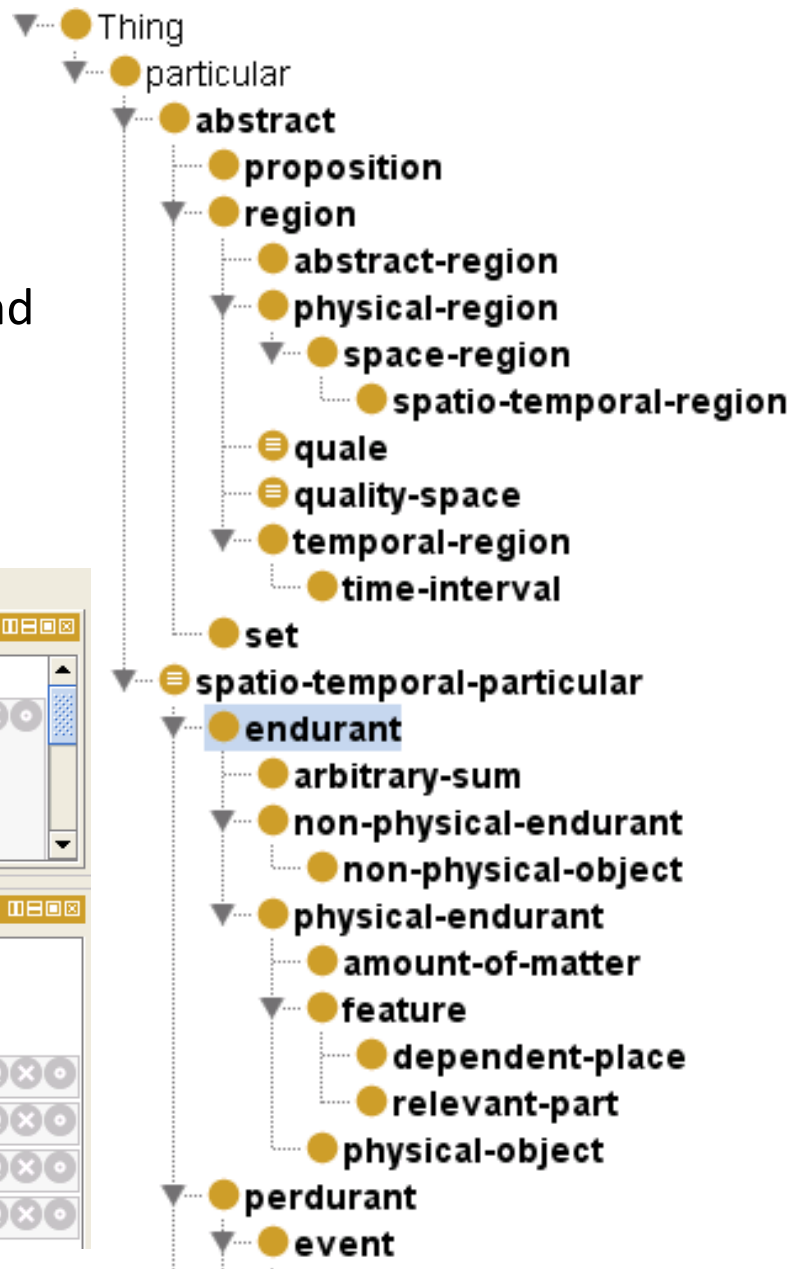
Descriptive Ontology for Linguistic and
Cognitive Engineering
<http://www.loa-cnr.it/DOLCE.html>

The screenshot shows a software interface with two main sections: "Class Annotations" and "Class Usage".

Class Annotations: The "Annotations: endurant" section shows a list of annotations. The "comment" annotation is expanded, displaying the text: "The main characteristic of endurants is that all of them are independent essential wholes. This does not mean that the corresponding property (being an endurant) carries proper unity, since there is no common unity criterion for endurants. Endurants can 'genuinely' change in time,".

Class Usage: The "Description: endurant" section shows a list of equivalent classes and superclasses. The superclasses are:

- spatio-temporal-particular
- participant-in some perdurant
- part only endurant
- specific-constant-constituent pink endurant



SWRC

Semantic Web for Research
Communities

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

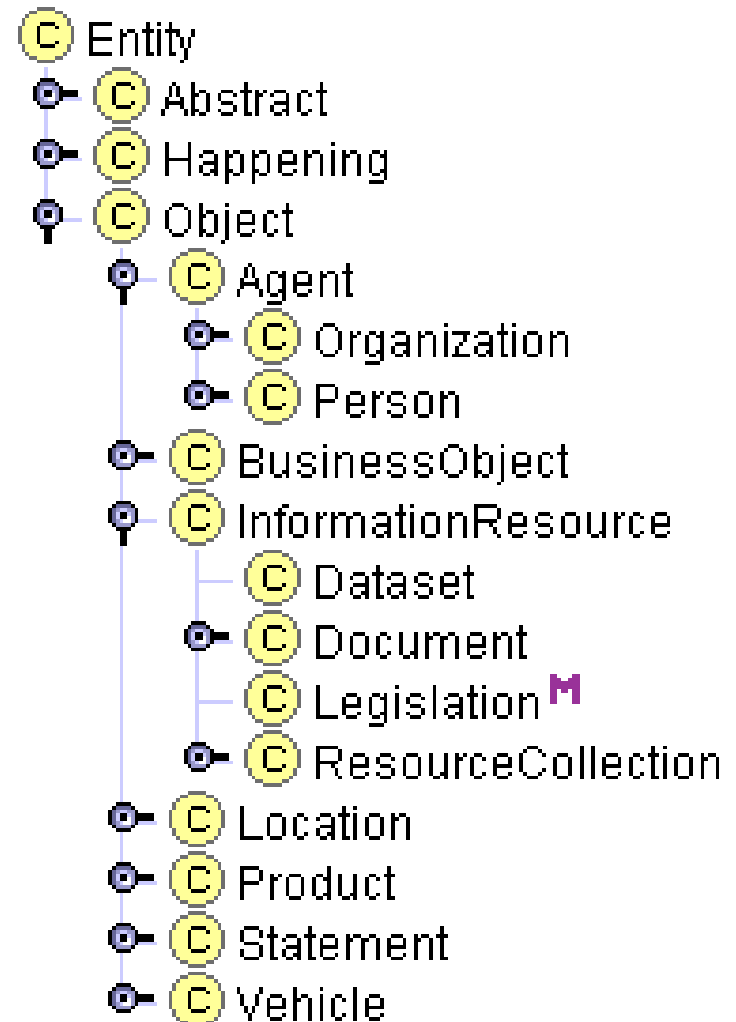
The screenshot shows two panels from an ontology editor:

- Class Annotations: Employee**
 - Annotations: +
 - label: "Mitarbeiter"@de
- Description: Employee**
 - Equivalent classes: +
 - Superclasses: +
 - Person
 - affiliation only Organization
 - Inferred anonymous superclasses:
 - firstName only string
 - fax only string
 - phone only string



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/>

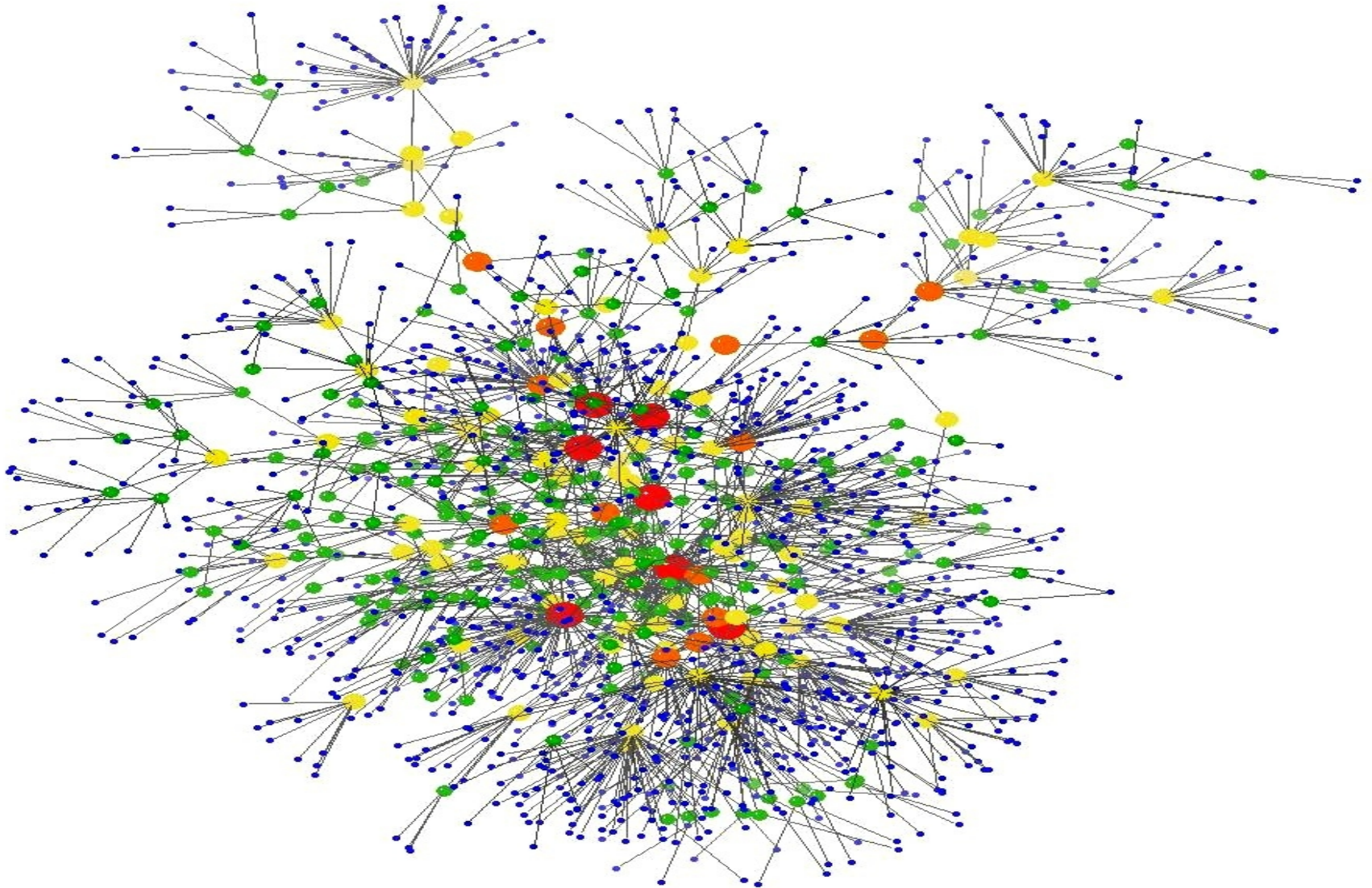
Gene Ontology: <http://www.geneontology.org>

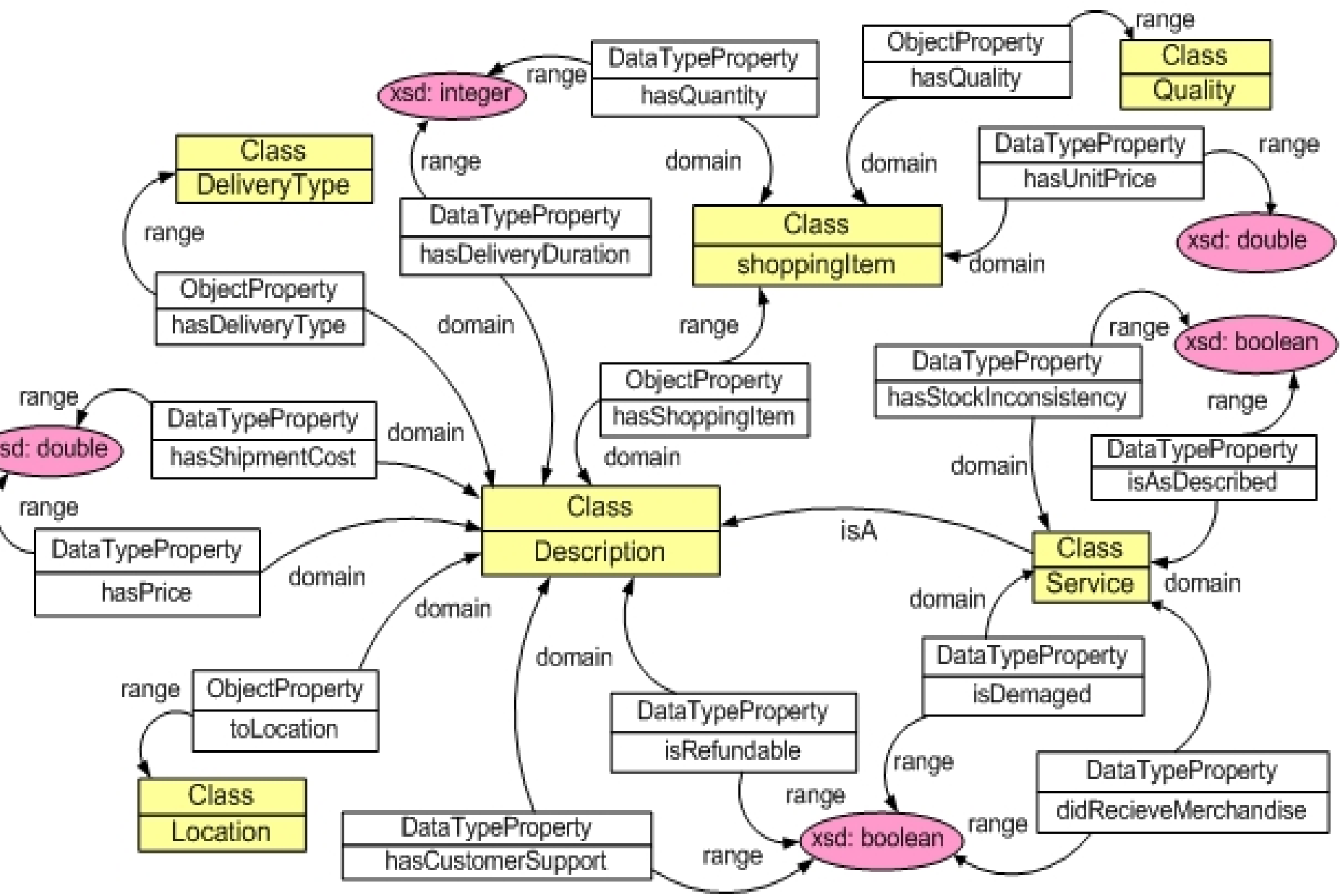
GALEN: http://www.openclinical.org/prj_galen.html

NCI Thesaurus: <http://nciterns.nci.nih.gov>

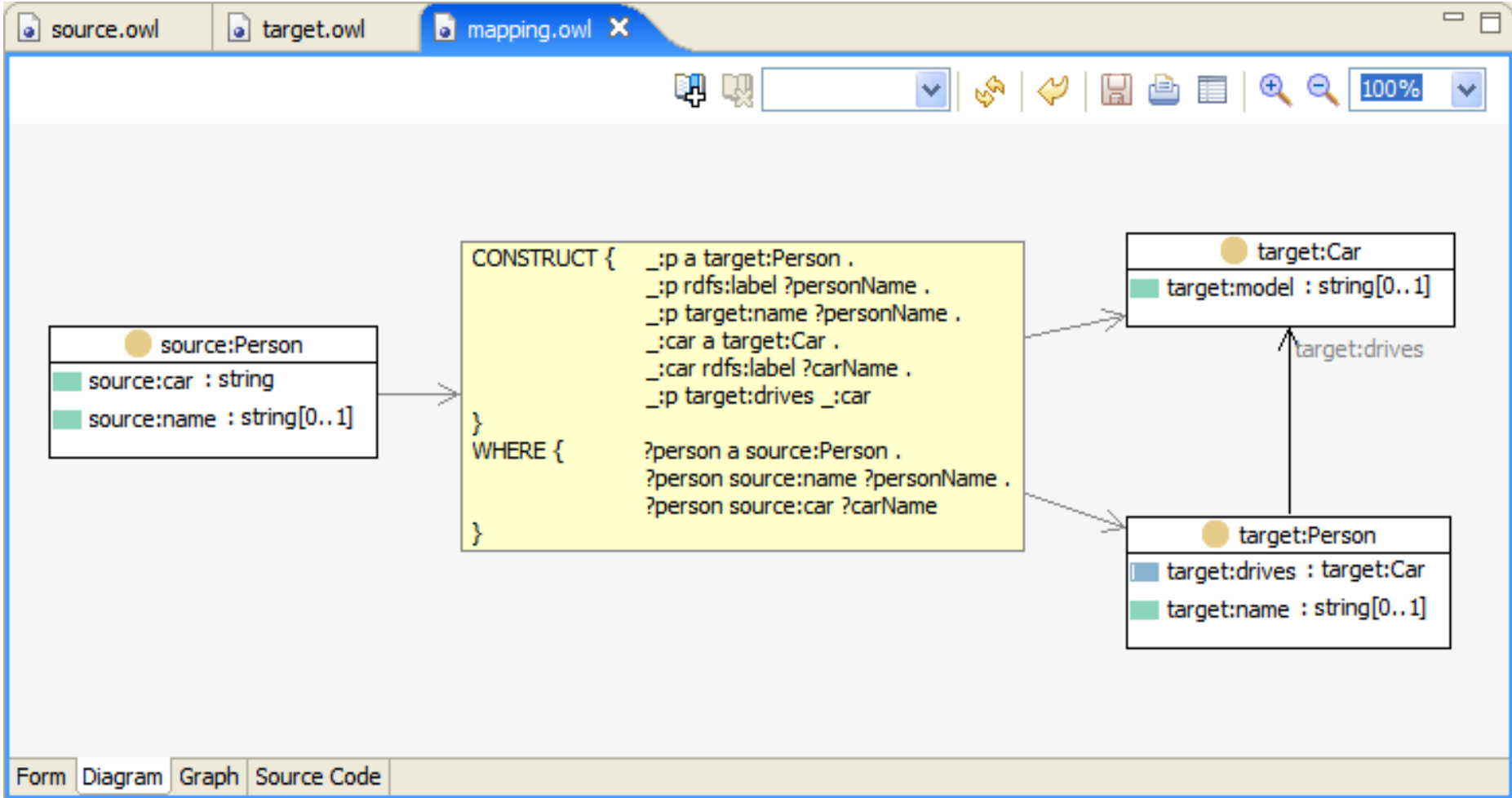
...

Ontologies can be scary monsters





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

The screenshot shows the GATE Ontology Editor interface. On the left, a tree view under 'Classes and Instances' shows a hierarchy: Entity (red square icon) -> Location (red square icon) -> City (red square icon) -> Sheffield (green triangle icon). Entity -> Organization (red square icon) -> A_Company (green triangle icon). Entity -> Person (red square icon) -> Diana_Maynard (green triangle icon, highlighted with a blue dashed border). A red arrow points from the text 'lexicalisation' to the Diana_Maynard instance.

On the right, a list view shows the ontology structure:

- Person
- 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/XMLSchema>
- Property Values
 - label Diana Maynard

The 'Property Values' section is highlighted with a red box.

At the bottom, the window title is 'GATE Ontology Editor' and the current tab is '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

The screenshot shows an ontology editor interface. On the left, a tree view displays the class hierarchy under 'Classes & Instances':

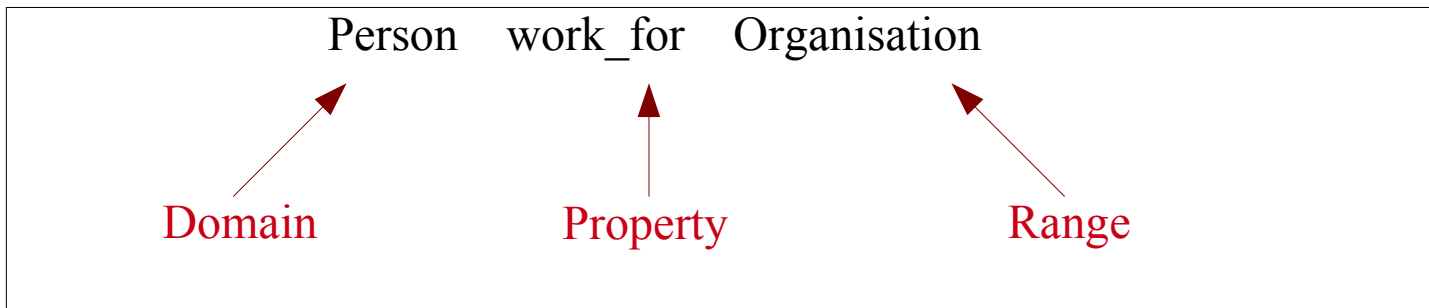
- Entity
 - Location
 - City
 - Sheffield
 - Location_00030
 - Organization
 - University_of_Sheffield
 - Person
 - Diana_Maynard

The 'Diana_Maynard' instance is selected. On the right, the 'Properties' panel shows details for this resource:

- Resource Information**
 - Diana_Maynard** Diana_Maynard
 - URI** <http://gate.ac.uk/example#D>
 - TYPE** Ontology Instance
- Direct Types**
 - Person Person
- All Types**
 - Entity Entity
 - Person Person
- Same Instances**
- Property Types**
 - seeAlso [ALL RESOURCES]
 - versionInfo [ALL RESOURCES]
 - work_for [Organization]
 - comment [ALL RESOURCES]
 - label [ALL RESOURCES]
 - isDefinedBy [ALL RESOURCES]
 - hasAge <http://www.w3.org/2001/XML>
- Property Values**
 - label Diana Gabrielle Maynard
 - hasAge 21

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

The image shows a screenshot of an ontology editor interface. On the left, there are two tabs: 'Classes & Instances' and 'Properties'. The 'Classes & Instances' tab is active, showing a hierarchical tree of classes and instances. The tree starts with 'Entity' (a class), which has three subclasses: 'Location', 'Organization', and 'Person'. 'Location' has two subclasses: 'City' and 'Location_00030'. 'City' has two instances: 'Sheffield' and 'Location_00030'. 'Organization' has one instance: 'University_of_Sheffield'. 'Person' has one instance: 'Diana_Maynard'. The 'Diana_Maynard' instance is highlighted with a blue selection box.

On the right, the 'Properties' tab is active, showing a list of properties for the selected resource 'Diana_Maynard'. The properties are grouped into several categories:

- Resource Information:**
 - Diana_Maynard** (URI): Diana_Maynard
 - URI**: http://gate.ac.uk/example#Diana_Maynard
 - TYPE**: Ontology Instance
- Direct Types:**
 - Person**: Person
- All Types:**
 - Entity**: Entity
 - Person**: Person
- Same Instances:** (None listed)
- Property Types:**
 - seeAlso**: [ALL RESOURCES]
 - versionInfo**: [ALL RESOURCES]
 - work_for**: [Organization]
 - comment**: [ALL RESOURCES]
 - label**: [ALL RESOURCES]
 - isDefinedBy**: [ALL RESOURCES]
 - hasAge**: <http://www.w3.org/2001/XMLSchema#integer>
- Property Values:**
 - label**: Diana Gabrielle Maynard
 - hasAge**: 21
 - work_for**: University_of_Sheffield

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>”

The screenshot displays the GATE 4.0 build 2794 interface. The main window shows a document titled "GATE document_0002A" with the following 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 comprise aerospace products, automotive components, etc.

The right-hand pane shows the "Ontology Tree(s)" for "OWLIM Ontology LR_00016". The tree is expanded to show the following classes:

- PopulatedPlace
- City
- PoliticalRegion
- Province
- County
- Country
- MilitaryAreas
- UrbanDistrict
- WaterRegion
- AstronomicalObject
- Brand
- Currency
- BusinessProcess
- Agent
- Product
- Credit
- DataWarehouse
- PieceOfArt
- Data
- SoftwareModule
- ITProcess

Red arrows indicate the mapping of text elements to ontology classes: "Alcoa Inc." maps to "Brand", "United States" maps to "Country", and "129,000" maps to "Quantity".

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

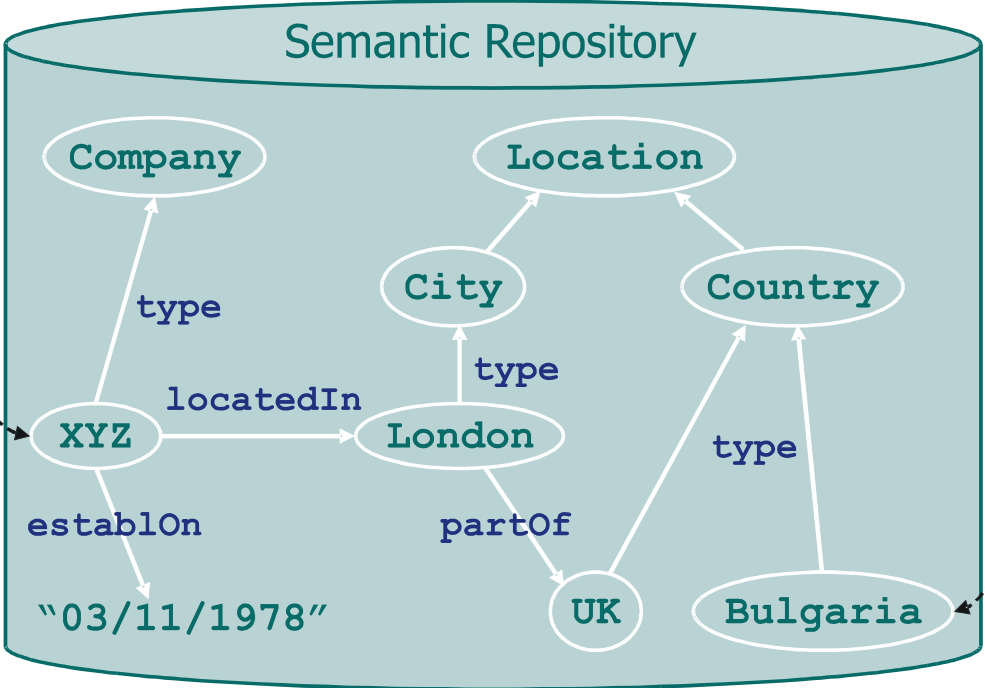
The best 5 Indian regions for a company with the characteristics supplied are returned (best region first) -- available is also an indication of the most relevant region characteristics that contributed to its ranking (either positively or negatively depending on the score sign).

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



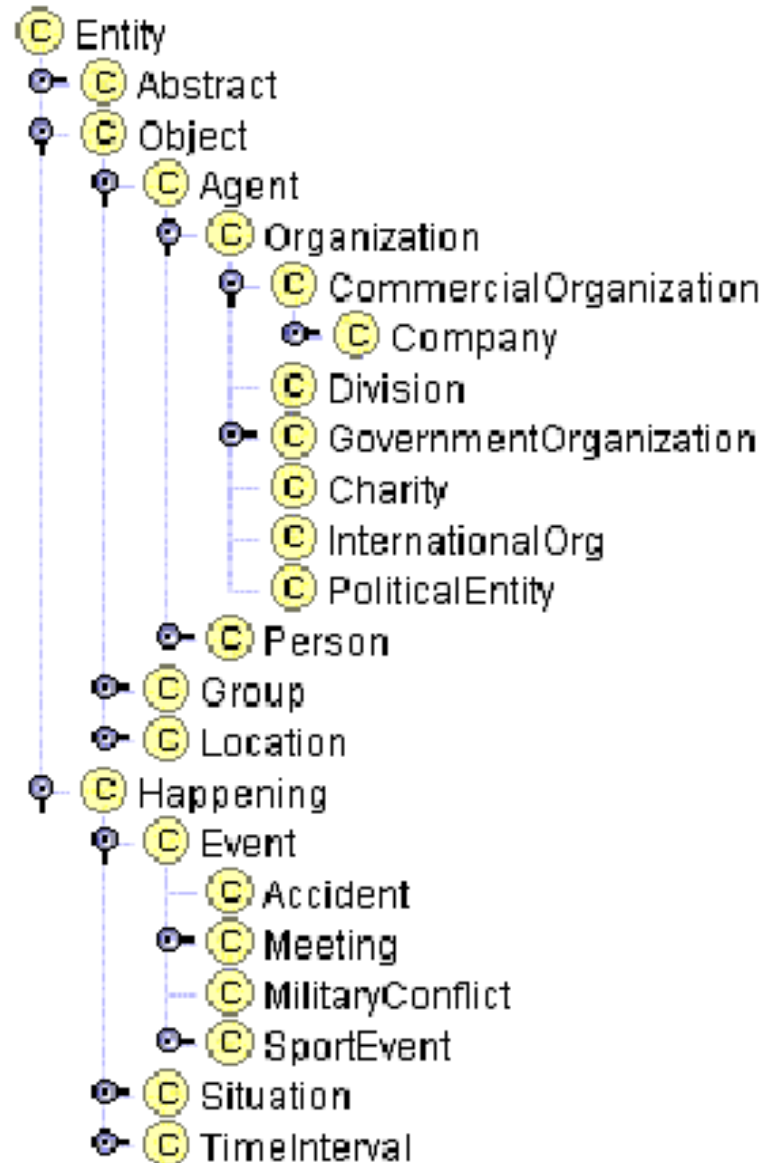
Semantic Annotation of Documents

XYZ announced profits in Q3, planning to build a \$120M plant in Bulgaria, and more and more and more and more text...

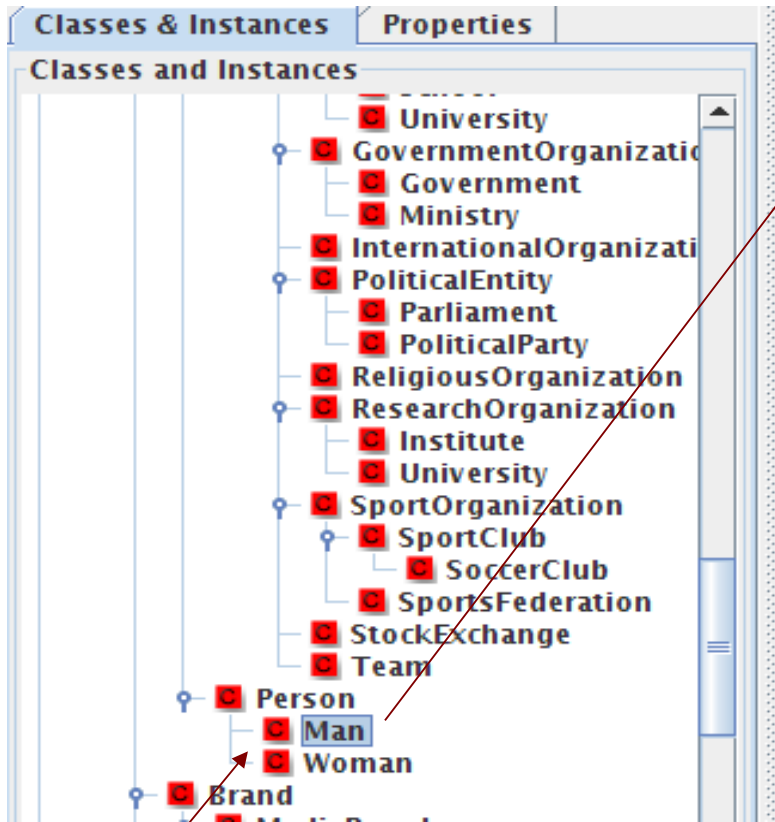


Richer NE Tagging

- 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

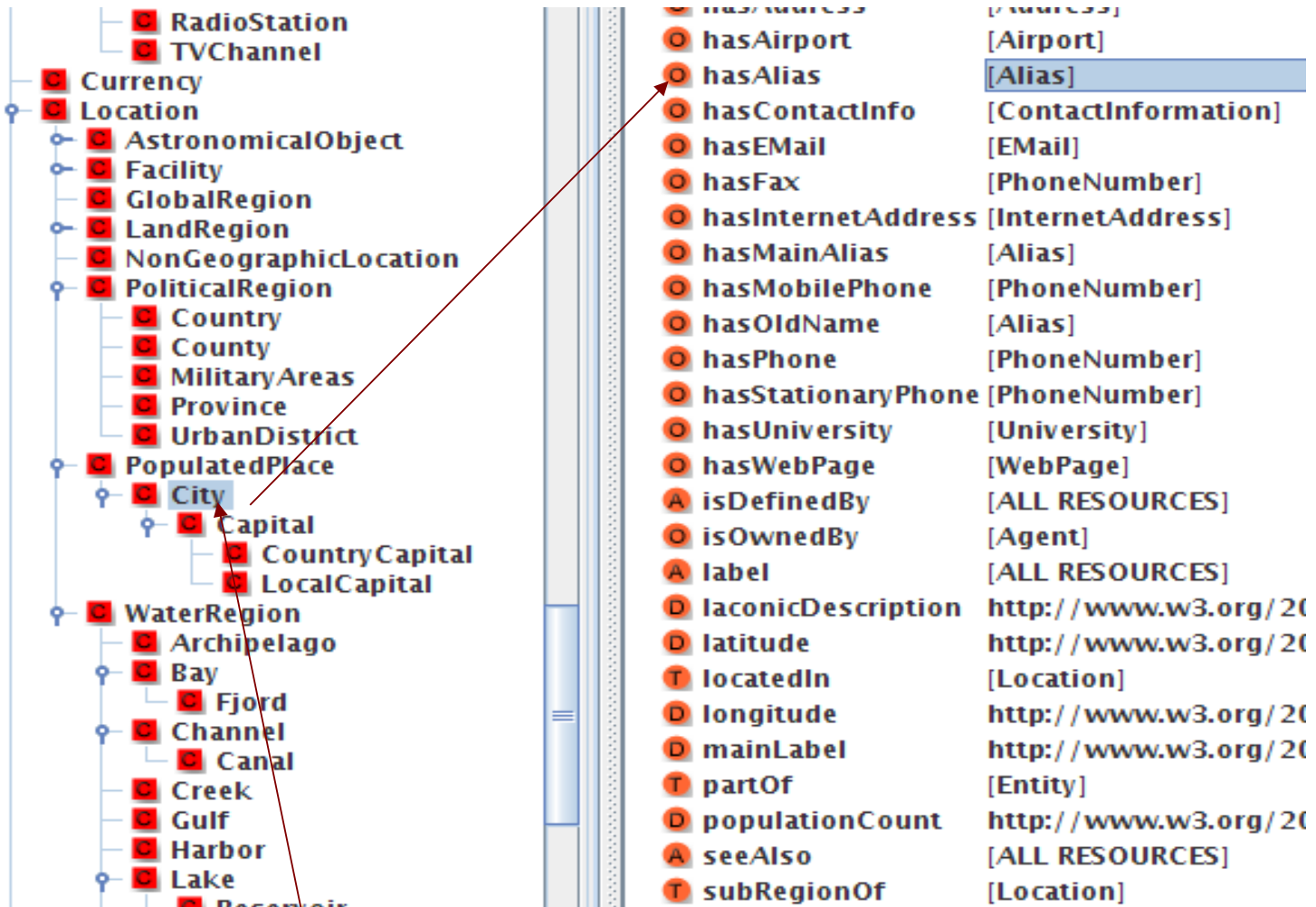


- controls [Object]
- description <http://www.w3.org/2001/X>
- generatedBy [EntitySource]
- hasAddress [Address]
- hasAlias [Alias]
- hasBrother [Man]
- hasChild [Person]
- hasContactInfo [ContactInformation]
- hasDaughter [Woman]
- hasEMail [EMail]
- hasFather [Man]
- hasFax [PhoneNumber]
- hasInternetAddress [InternetAddress]
- hasMainAlias [Alias]
- hasMobilePhone [PhoneNumber]
- hasMother [Woman]
- hasOldName [Alias]
- hasParent [Person]
- hasPhone [PhoneNumber]
- hasPosition [JobPosition]
- hasProfession [Profession]

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

The screenshot displays the GATE (General Architecture for Text Engineering) software interface. At the top, there are tabs for 'Text', 'Annotations', 'Annotation Sets', 'Coreference', and 'Print'. The main text area contains several paragraphs of news text with various words highlighted in colored boxes representing named entities. On the right side, there is a panel titled 'Default annotations' with a list of entity types, each with a checkbox. Two red arrows point from the 'Location' and 'Person' checkboxes in the list to the corresponding highlighted words in the text.

Text: Threats to the resumption of the Northern Ireland peace talks receded today after a British cabinet minister entered the huge Maze prison near Belfast and pressed Protestant guerrillas held there to support continuing the discussions.

Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.


After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.

The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater

Default annotations:

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Organization
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown

Semantic IE in GATE

Messages  GATE document_0002D

Annotation Sets Annotations Co-reference Editor Ontology Viewer Text

FT.com | TotalSearch | Global Archive | Print

```
document.write(getAdHTML('ban', 468, 60));
```

Return to Article | Print this Page

Airlines take over running of air traffic control

FT.com site, Jul 27, 2001
BY KEVIN DONE, AEROSPACE CORRESPONDENT

Seven UK airlines including British Airways, Virgin Atlantic, BMI British Midland and EasyJet, on Friday took over control of the air traffic control system, completing one of the government's most controversial public-private partnership deals.

Completion of the National Air Traffic Services deal comes at a critical time for the government as it tries to push through the PPP for the London Underground.

The sale to a strategic investor of a 46 per cent stake in Nats is the first time in Europe that management control of en route air traffic services has passed into private hands.

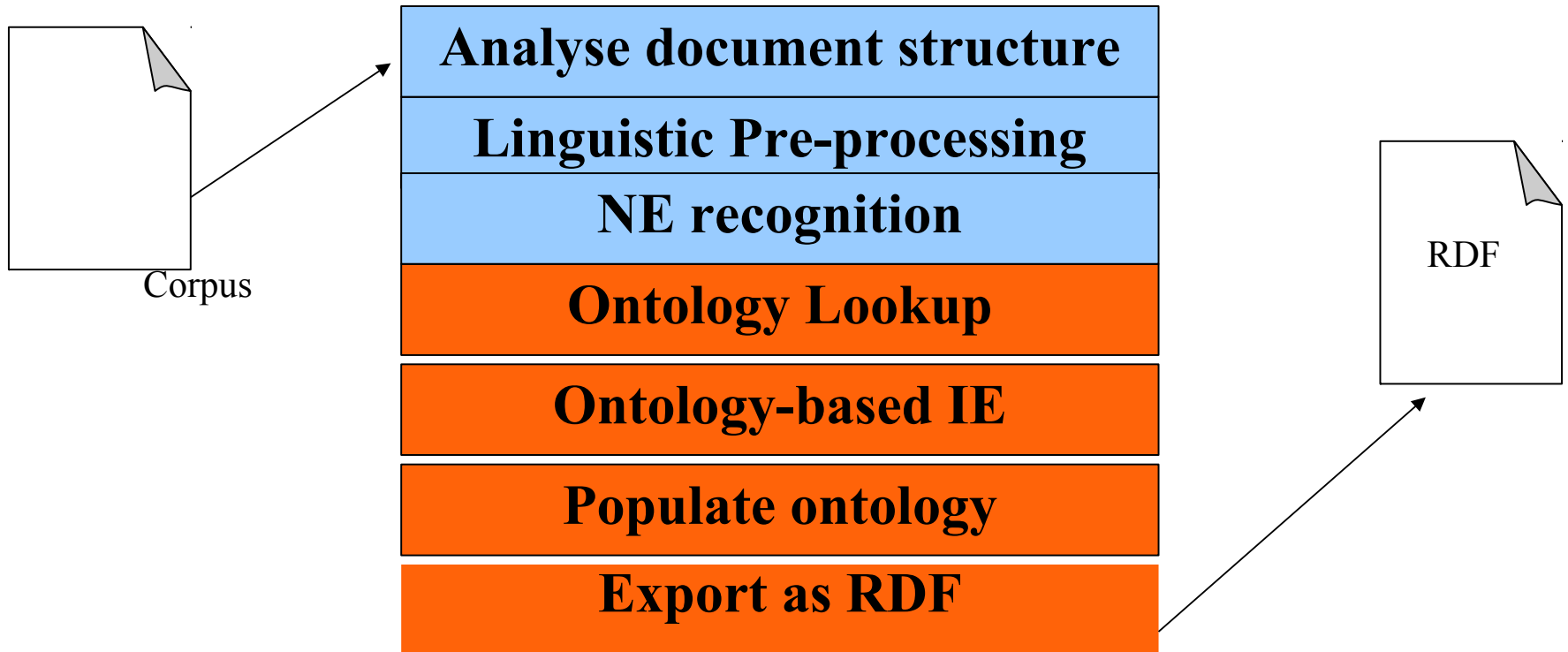
Document Editor Initialisation Parameters

Ontology Tree Options

Ontology_00029

- EntityProperty
- LexicalResource
- LexRcProperty
- EntitySource
- Entity
 - Happening
 - Situation
 - TimeInterval
 - Quarter
 - Date
 - Week
 - CalendarYear
 - Month
 - Event
 - Object
 - Agent
 - Person
 - Man
 - Woman
 - Organization
 - Charity
 - Team

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

Annotation Diff Tool

Document: ft-bank-of-england.xml_00016 Annotation Set: Key

Key: ft-bank-of-england.xml_00016 Annotation Type: Date F-Measure Weight: 1.00

Response: ft-bank-of-england.xml_00016 [Default set] Features: All Some None

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct: 3 Recall Precision F-Measure

Partially Correct: 1 Strict: 0.50 0.75 0.60

Missing: 2 Lenient: 0.6667 1.00 0.80

False Positives: 0 Average: 0.5833 0.875 0.70

compare Key and Result in terms of Precision and Recall

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

Annotation Diff Tool

Document: ft-bank-of-england.xml_00016 Annotation Set: Key

Key: ft-bank-of-england.xml_00016 Annotation Type: Date F-Measure Weight: 1.00

Response: ft-bank-of-england.xml_00016 [Default set] Features: All Some None

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct: 3 Recall Precision F-Measure

Partially Correct: 1 Strict: 0.50 0.75 0.60

Missing: 2 Lenient: 0.6667 1.00 0.80

False Positives: 0 Average: 0.5833 0.875 0.70

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

Annotation Diff Tool

Document: ft-bank-of-england.xml_00016 Annotation Set: Key

Key: ft-bank-of-england.xml_00016 Annotation Type: Date

Response: ft-bank-of-england.xml_00016 [Default set] Features: All Some None F-Measure Weight: 1.00

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct: 3

Partially Correct: 1	Strict:	0.50	0.75	0.60
Missing: 2	Lenient:	0.6667	1.00	0.80
False Positives: 0	Average:	0.5833	0.875	0.70

scores displayed

Precision

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

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

Recall

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

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

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

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

Annotation Diff defaults to F1

Annotation Diff Tool

Document: ft-bank-of-england.xml_00016
Annotation Set: Key
Annotation Type: Date
F-Measure Weight: 1.00
Response: ft-bank-of-england.xml_00016
Features: All Some None

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct: 3
Partially Correct: 1
Missing: 2
False Positives: 0

Strict: 0.50 0.75 0.60
Lenient: 0.6667 1.00 0.80
Average: 0.5833 0.875 0.70

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

Annotation Diff Tool

Document: ft-bank-of-england.xml_00016 Annotation Set: Key Annotation Type: Date F-Measure Weight: 1.00

Response: ft-bank-of-england.xml_00016 [Default set] Features: All Some None

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct: 3 Recall Precision F-Measure

Partially Correct:	1	Strict:	0.50	0.75	0.60
Missing:	2	Lenient:	0.6667	1.00	0.80
False Positives:	0	Average:	0.5833	0.875	0.70

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

Annotation Diff Tool

Document: ft-bank-of-england.xml_00016
Annotation Set: Key
Annotation Type: Date
F-Measure Weight: 1.00
Response: ft-bank-of-england.xml_00016
Features: All Some None

Do Diff

Start	End	Key	Features	? Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?			
1466	1474	Thursday	{}	-?			
212	222	early 1964	{kind=date}	~	218	222	1964 {kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday {kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month {kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week {kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

Correct: 3 Recall Precision F-Measure Export to HTML

Partially Correct: 1 Strict: 0.50 0.75 0.60

Missing: 2 Lenient: 0.6667 1.00 0.80

False Positives: 0 Average: 0.5833 0.875 0.70

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.

The screenshot shows the GATE Developer 5.2-snapshot build 3518 interface. The main window displays a list of documents in a corpus. The 'Corpus Quality Assurance' tab is highlighted at the bottom of the Display pane.

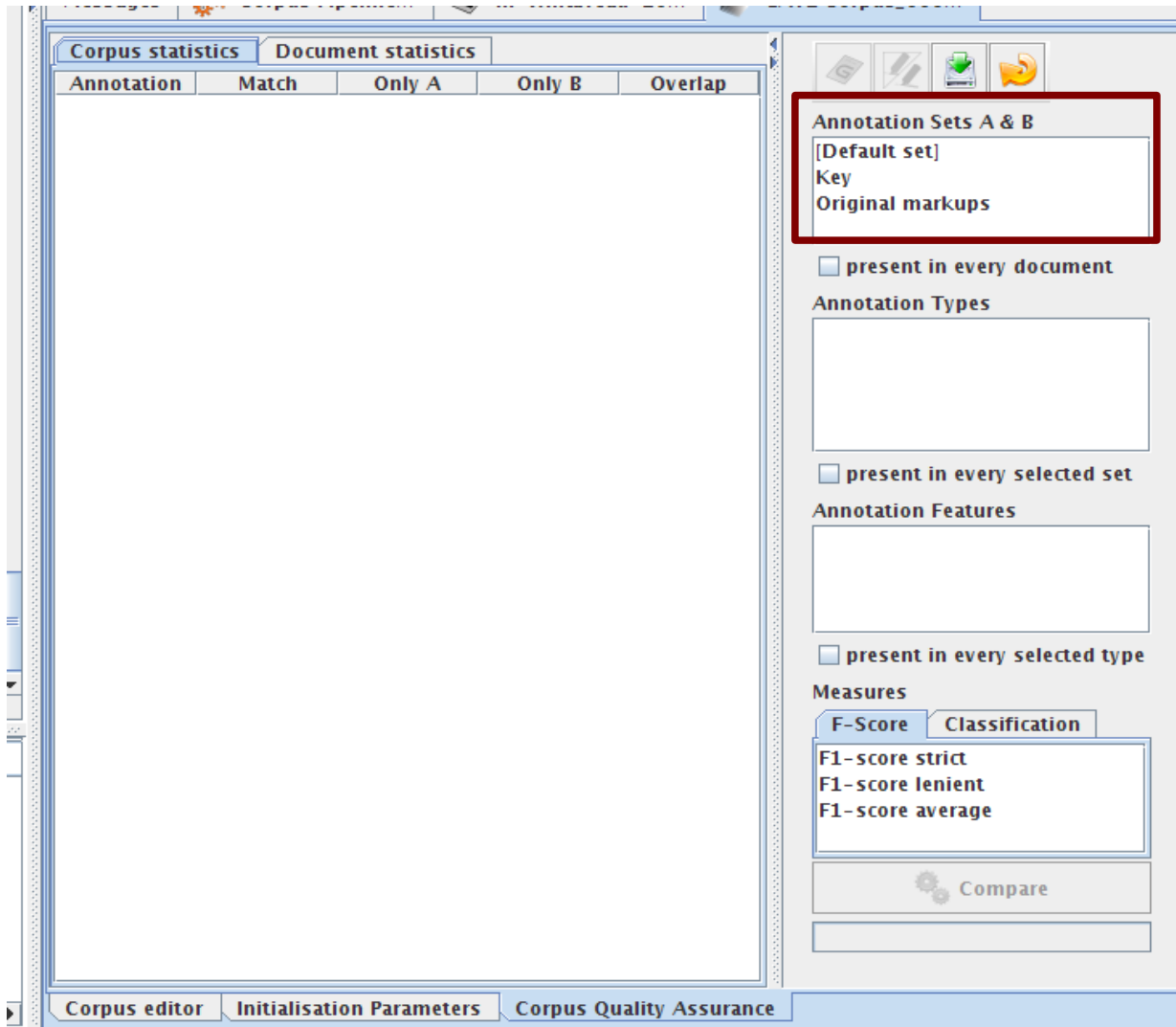
Messages Corpus Pipeline... in-whitbread-10... GATE Corpus_000...

All the documents loaded in the system are in this corpus.

Index	Document name
0	ft-BT-07-aug-2001.xml_0001B
1	ft-BT-briefing-02-aug-2001.xml_0001C
2	ft-BT-loop-01-aug-2001.xml_0001D
3	ft-GKN-09-aug-2001.xml_0001E
4	ft-SSL-10-aug-2001.xml_0001F
5	ft-WestLB-BT-05-aug-2001.xml_00020
6	ft-airlines-27-jul-2001.xml_00021
7	ft-airtours-08-aug-2001.xml_00022
8	ft-bank-of-england-02-aug-2001.xml_00023
9	ft-bank-of-uk-08-Aug-2001.xml_00024
10	ft-bmi-09-may-2001.xml_00025
11	ft-bmi-25-feb-2001.xml_00026
12	ft-bmi-airline-07-aug-2001.xml_00027
13	ft-bt-03-aug-2001.xml_00028
14	ft-bt-26-jul-2001.xml_00029

Corpus editor Initialisation Parameters **Corpus Quality Assurance**

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

Annotation Sets A & B

[Default set] (A)
Key (B)
Original markups

present in every document

Annotation Types

Organization
Percent
Person
Sentence
SpaceToken

present in every selected set

Annotation Features


1
der
e1
gender

present in every selected type

Measures

F-Score Classification

F1-score strict
F1-score lenient
F1-score average

 Compare

- 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

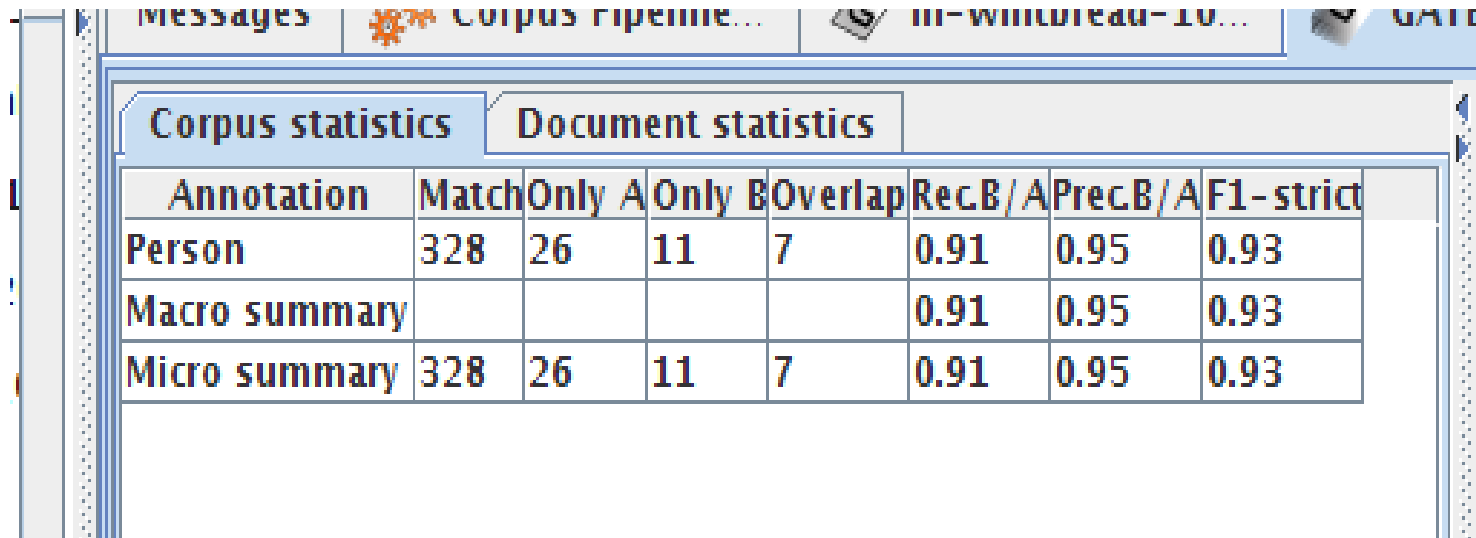
- In the “Measures” box, select the kind of F1.0 score you want “Strict, Lenient, Average” or any combination of them. Suggestion: try just “lenient” at first

- Select Compare

The screenshot shows a software interface with several sections:

- Annotation Sets A & B:** A list containing "[Default set] (A)", "Key (B)", and "Original markups". Below the list is a checkbox labeled "present in every document".
- Annotation Types:** A list containing "Organization", "Percent", "Person", "Sentence", and "SpaceToken". Below the list is a checkbox labeled "present in every selected set".
- Annotation Features:** A list containing "1", "der", "e1", and "gender". Below the list is a checkbox labeled "present in every selected type".
- Measures:** A section with two tabs: "F-Score" (selected) and "Classification". Under the "F-Score" tab, there is a list containing "F1-score strict", "F1-score lenient", and "F1-score average". This section is highlighted with a red box.
- Compare:** A button with a gear icon and the text "Compare".

Corpus Statistics Tab

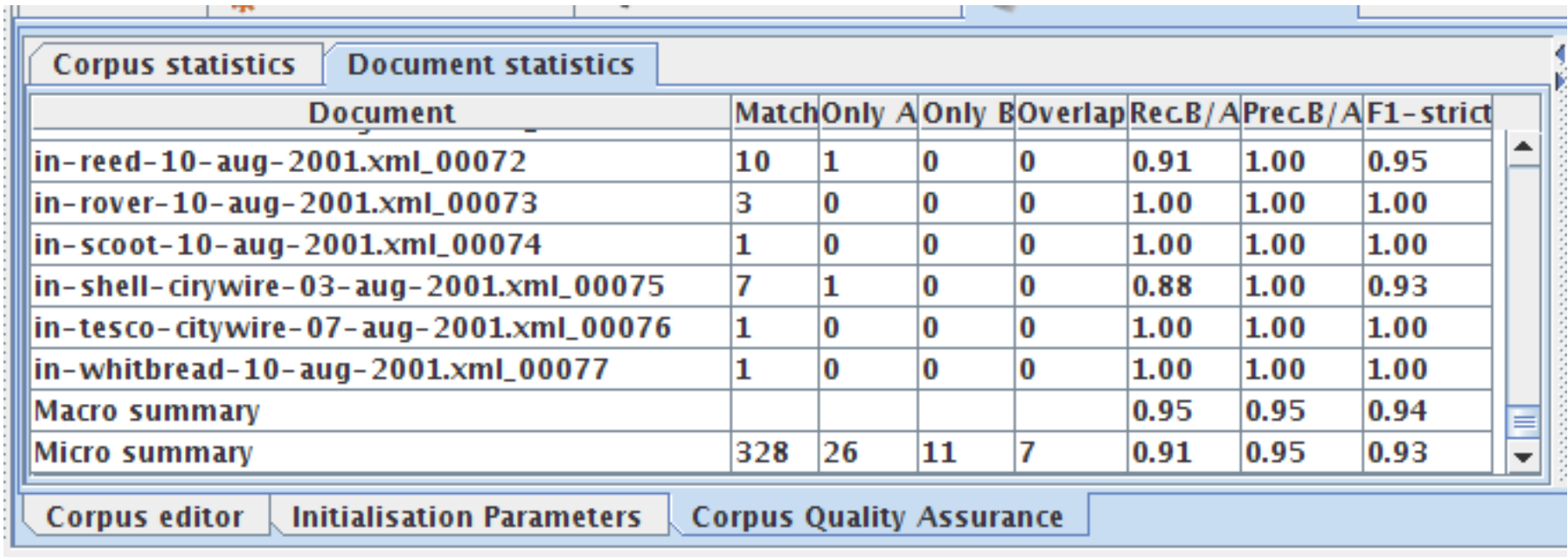


The screenshot shows a software window with a 'Corpus statistics' tab selected. The window title bar includes 'messages', 'Corpus ripenne...', 'in-windread-10...', and 'GATE'. The 'Corpus statistics' tab is active, and the 'Document statistics' tab is also visible. The table below displays the following data:

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



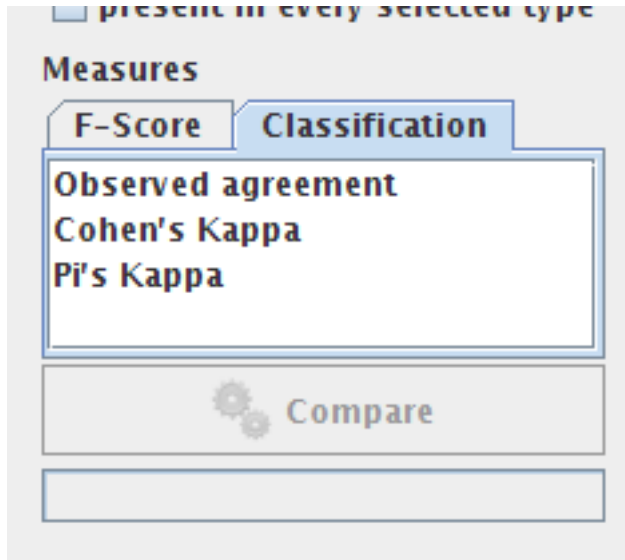
Document	Match	Only A	Only B	Overlap	Rec.B/A	Prec.B/A	F1-strict
in-reed-10-aug-2001.xml_00072	10	1	0	0	0.91	1.00	0.95
in-rover-10-aug-2001.xml_00073	3	0	0	0	1.00	1.00	1.00
in-scoot-10-aug-2001.xml_00074	1	0	0	0	1.00	1.00	1.00
in-shell-citywire-03-aug-2001.xml_00075	7	1	0	0	0.88	1.00	0.93
in-tesco-citywire-07-aug-2001.xml_00076	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

- 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:

$$AP = \frac{BDM}{BDM + \textit{Spurious}}$$

$$AR = \frac{BDM}{BDM + \textit{Missing}}$$

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

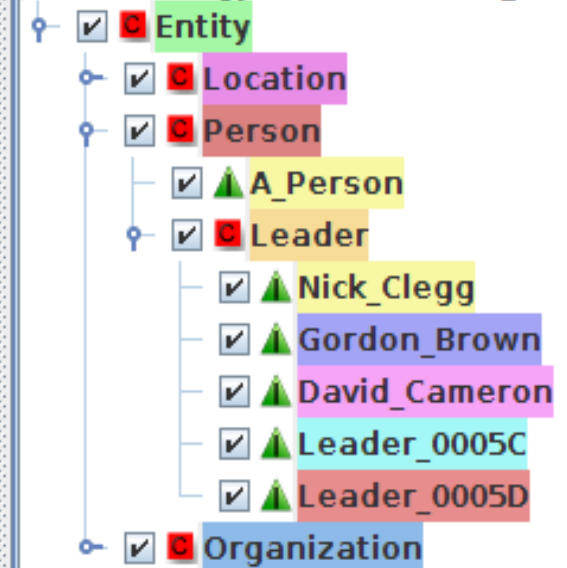
David Cameron was the first of the main UK party leaders to cast their vote. The Tory leader went to a community hall in Witney, Oxfordshire, shortly after 1030 BST, accompanied by his wife Samantha.

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.

test-ontology-instances.owl_022:



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.