2. GATE:
Providing NLP Tools for the Semantic Web

Topics covered

- Introduction to GATE
- Semantic Annotation Example: Business Intelligence
- Manual ontology creation: CLONE
- Representing linguistic representation in ontologies: Localisation

GATE: a vision for text mining

- It is difficult to access **unstructured** information efficiently
- IE automates **extraction of facts** from text at reasonable accuracy and cost, increasing the value and utility of unstructured content
- **Interlinking** of text and data enables more efficient search, navigation and querying
- GATE is an architecture for language engineering, containing numerous plugins and resources for all kinds of NLP tasks
- One of the main things GATE is used for is information extraction

GATE is...

- open source software capable of solving almost any text processing problem
- a mature and extensive community of developers, users, educators, students and scientists
- a defined and repeatable process for creating robust and maintainable text processing workflows
- in active use for all sorts of language processing tasks and applications, including voice of the customer (opinion mining); cancer research; drug research; decision support; recruitment; web mining; information extraction; semantic annotation
- the result of a €multi-million R&D programme running since 1995, funded by commercial users, the EC, BBSRC, EPSRC, AHRC, JISC, etc.
- used by thousands of corporations, SMEs, research labs and Universities worldwide

In short…

GATE includes:

- **components and plugins** for language processing, e.g. parsers, machine learning, summarisation, stemmers, IR tools, IE components for various languages...
- tools for **visualising** and manipulating text, annotations, ontologies, parse trees, etc.
- various **information extraction** tools
- **evaluation** and **benchmarking** tools

We have used it to develop further **applications** for:

- opinion mining, summarisation, terminology extraction, ontology generation, relation finding, and so on.

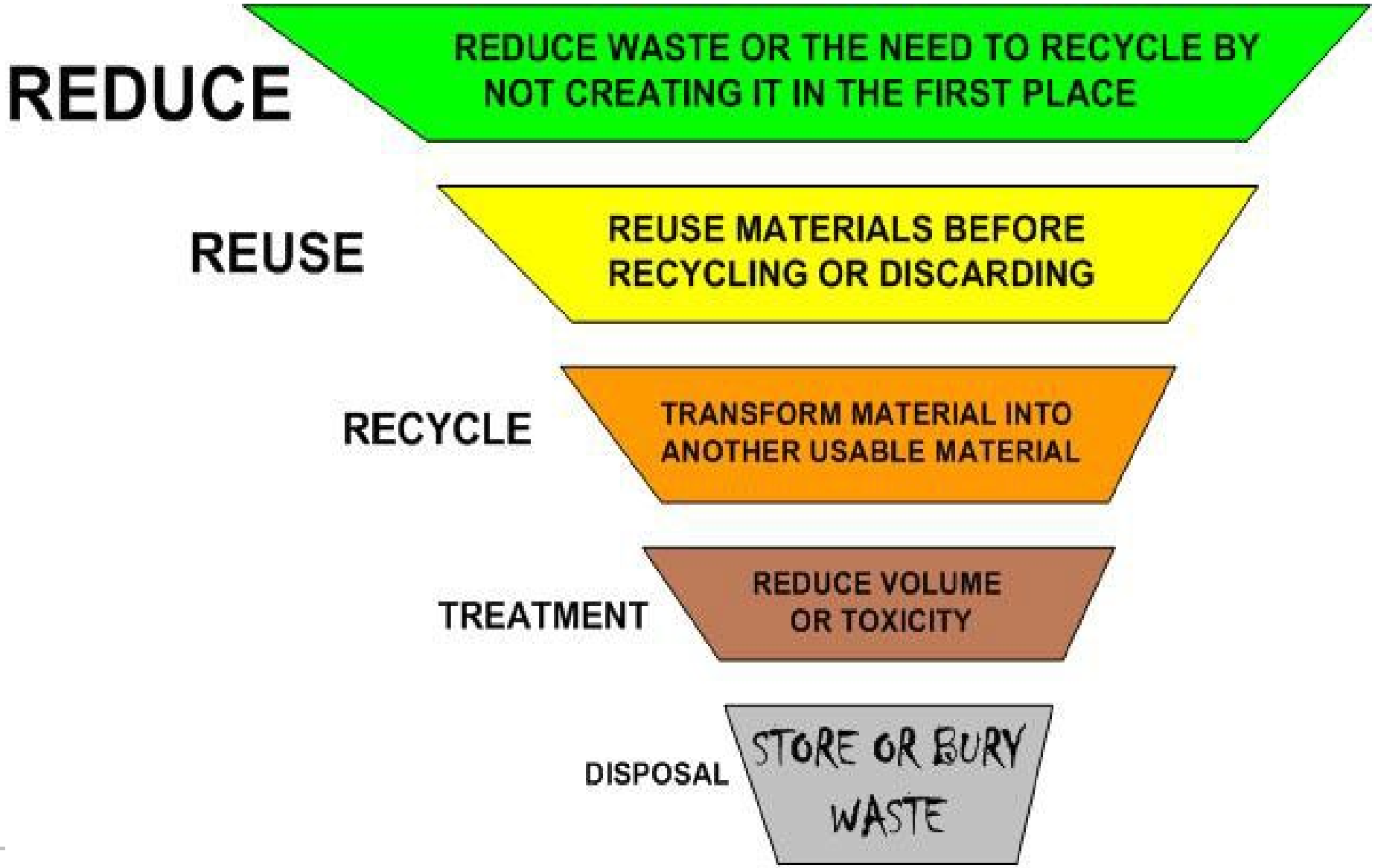
GATE: the Swiss Army Knife of NLP

- Has an attachment for almost every eventuality
- Some are hard to prise open
- Some are useful, but you might have to put up with a bit of clunkiness in practice
- Some will only be useful once in a lifetime, but you're glad to have them just in case.
- There are many imitations, but nothing like the real thing.

History of GATE

- **early 1990s**: you want me to write that all over again?
- **1995-7**: first GATE (and "large-scale IE") project
- **1996**: GATE 1: Tcl/Tk, Perl, C++, ...
- **2002**: release of completely rewritten version 2, 100% Java
- **2009**: mature ecosystem with established community
  - Tens of thousands of research users
  - 25,000 downloads per year
  - commercial users getting serious

GATE is very eco-friendly!



GATE commercial users

Typical commercial **uses**:

- dynamic search and indexing of repositories
- finding relations between elements in distributed repositories
- aggregating information from different text sources
- populating repositories
- fact finding from distributed knowledge sources

Typical **users**:

- Pharmaceutics, news, intelligence (business, competitor, government, etc.), manufacturing, telecommunications

AstraZeneca Websites | Search Go

AstraZeneca

AstraZeneca International

- Home
- About us
- Medicines
- Research
- Careers
- Responsibility
- Partnering
- Media
- Investors

LIFE INSPIRING IDEAS

Our business is focused on turning good ideas into innovative, effective medicines that make a real difference in important areas of healthcare.

READ MORE

1 2 3 4

Our medicines

- Arimidex
- Crestor
- Nexium

Our research

In the fight against human disease, we focus on six disease areas where we believe our skills and experience can make the most difference:

- Cancer
- Cardiovascular
- Gastrointestinal
- Infection
- Neuroscience

Our responsibility

- Animal research
- Clinical trials
- Safety of medicines

Latest news

01 June 2009

- AstraZeneca and Merck & Co., Inc. Form Pioneering Collaboration to Investigate Novel Combination Anticancer Regimen

- more news

1 2 3 4 5

Global Web Site | Roche Worldwide | Deutsch | Site Map | Contact | + Text Size –

Roche

Home | About Roche | **Products** | Corporate Responsibility | Research & Development | Media | Investors | Careers

Search for...

Roche Product Portfolio

Focus on unsolved medical problems

Our aim is to develop new and improved drugs, and diagnostic tests and services that offer significant benefits over existing options.

Explore our Portfolio

- Products A-Z
- Diseases
- Products for Researchers
- Solutions for Diagnostics

Pharmaceuticals

Roche has brought many highly effective drugs onto the market and is a world leader in innovative cancer drugs. Other areas include viral infections, metabolic, central nervous system disorders and inflammatory diseases. > More

Solutions for Diagnostics

As the world leader in in-vitro diagnostics, we supply a wide range of rapid, reliable instruments and tests for disease screening and diagnosis in laboratories, at the point of care, and for patient self-management. > More

Products for Researchers

Roche Applied Science supplies a broad array of instruments and highly specific reagents and test kits for use in the diverse research market.

The portfolio is especially strong in genomics and proteomics. > More

fizzback

Engaging Customers

Search: GO

Home | Customer Engagement | Clients | Resources | About us

Industries | Solutions

+ Telecommunications & Media
+ Retail & Consumer Goods
+ Financial Services & Insurance
+ Transportation & Travel
+ Leisure & Entertainment

Fizzback is an **on-demand solution** that drives **customer engagement** at the **point-of-experience** » more

Spotlight

T-Mobile selects Fizzback

T-Mobile UK has partnered with Fizzback to drive satisfaction aross all touch-points...read more

Resources

Case Studies

Video Testimonials

Latest News » more

T-Mobile to present at ECEW

T-Mobile will present how Fizzback has allowed them to drive the customer...more

GENERIC

We're

Home | Niche Sectors | Careers | About Us | Contact Us

Business Intelligence

Business Intelligence offers a filter to identify the most us
data to the management team within a business...

As the universe continuously expands, so the demand for specialist skills increases; without the right experience and talent, no busines

Innovantage
ADVANTAGE THROUGH INNOVATION

e-mail us | 02920 022 600

- Home
- About Us
- Product
- Careers
- News

Revolutionising recruitment practices for the HR & Staffing sector

Find. Analyse. Connect. With Insight 3.0

Enjoy sophisticated access to information and market intelligence on companies who advertise online

Insight 3.0 gives you a detailed understanding of the online job advertising activity of hundreds of thousands of UK employers. It enables you to monitor and assess the recruitment needs and business activity of clients, prospects and competitors. It also allows you to audit your own recruitment and identify the most appropriate media for your ads.

simple ingenuity | recruitment news | register

garlik™ The online identity experts

e-mail address

D.O.B. 01 Jan 1945

forgotten your details? Log in

Home | About Garlik | Garlik products | What our customers say | News | Sign up

Garlik's DataPatrol helps people take control of their personal information and protect themselves against identity theft and financial fraud...

DataPatrol for businesses

DataPatrol for individuals

Latest news

all news

DataPatrol service review
5/15/09

Garlik Secures Further Funding in Battle Against Cybercrime
4/23/09

Welcome to Garlik

Garlik are leading technology innovators and identity experts set up by the founders of the online banks Egg and First Direct. With our range of products and services we aim to give individuals power over the use of their personal

Financial fraud soars

Economic crisis fuels new cybercrime wave. Check out our UK cybercrime report.

Homepage

Tailored text, pictures and video for use on any platform

Contact Us

PRESS ASSOCIATION Sport

New name, same unrivalled coverage

Wire Service

All the day's breaking news stories in words and pictures.

Digital

Content to power websites, widgets, mobile services, digital display screens and much more.

Images

Log in to see the latest news, sport and showbiz pictures, plus an archive of over 15 million images.

Video

News, sport and entertainment coverage from around the UK available as footage, clips and packages.

Specialist News

Tailored news feeds for corporate websites, in-depth news from Westminster and detailed financial coverage.

Pages

Newspaper and magazine production services from individual pages and supplements to entire cover-to-cover solutions. Bespoke services available.

PR Services

Media monitoring and training to press release distribution and broadcast consultancy - PR services from a journalistic point of view.

Business Information

Essential news and information services for public and private sector organisations, Government departments, financiers and PR companies.

Enhancing the Independent's website with video

The Independent has launched an enhanced UK video news offering following an innovative distribution deal with Press Association and Octopus Media Technology. Read the full video deal report.

2012 Olympic updates

TSO
information & publishing solutions

Providing fail-safe information and publishing solutions for more than 200 years

Part of the Williams Lea Group

Information and Publishing Solutions

TSO (The Stationery Office) provides expert management of all the printed and digital information organisations share with their internal and external audiences. Applying proven information management methods and experience, we deliver transformational solutions and rapid results.

We have been providing fail-safe information management and publishing solutions to private and public sector organisations for more than 200 years.

Find out how our **information and publishing solutions** can reduce costs, increase efficiency and improve interaction with citizens and stakeholders.

Contact us for more information

Bookshop

Order and purchase any UK book in print from TSO's online bookshop

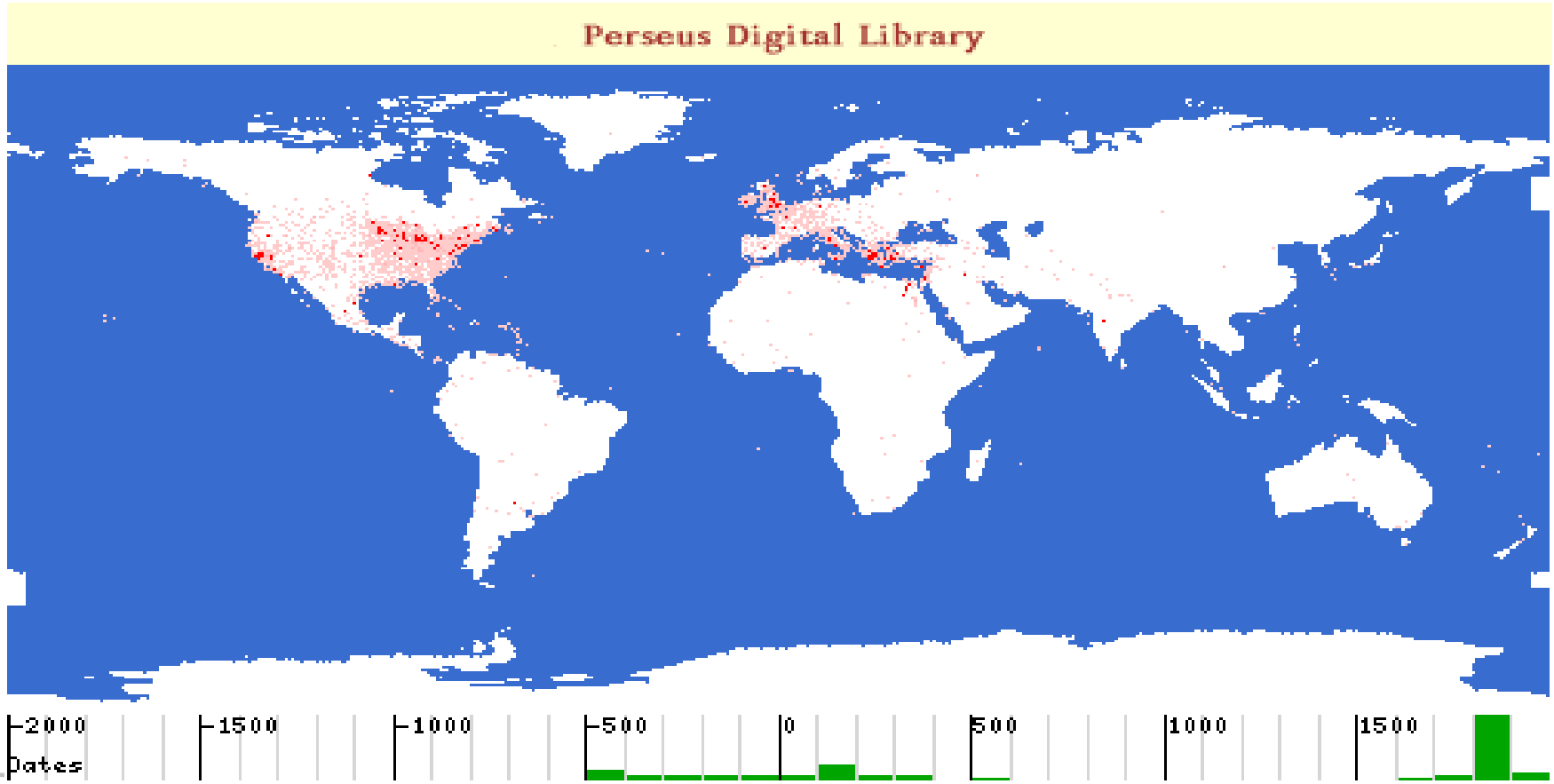
Enter tsoshop.co.uk

Latest News

The London Gazette used to create Dick Turpin e-fit . . . more

Semantic Annotation

- Adding information to documents that is usable by machines to enable better presentation, navigation or searching, e.g. Perseus:



A graph of the places and dates mentioned in this collection

Perseus Named Entity Search (Beta)

Search for places: [ ] matching this name [Browse]

Enter the name of a place, like "Springfield" or "Athens", to find all locations matching the name, or enter a state ("Illinois") or country ("Canada") to find all places within that state or nation. You may also enter more than one of these to narrow your search ("Athens, Greece" or "Springfield, Illinois, United States").

Note that abbreviations ("USA", "Ill.") do *not* work at present--please stick to full names!

Search for a person: [ ] [Browse]

In: ☑ Forenames ☑ Surnames ☐ Full name

Searching for "Washington" in "Forenames" and "Surnames" will return all people with Washington as a first or last name, respectively. A full-name search will find anyone who matches the entire search string ("Washington Irving").

Search for dates: [any] [ ], [ ] A.D. [Browse]

Month Day Year

Enter a month, day and/or year to search for references to that date. You do not need to fill out every field: searching only for "1863" will find all references to the year 1863, while searching for "July 4" will find all references to the 4th of July, regardless of year.

Search for dates from: [any] [ ], [ ] A.D.

Month Day Year

To: [any] [ ], [ ] A.D. [Browse]

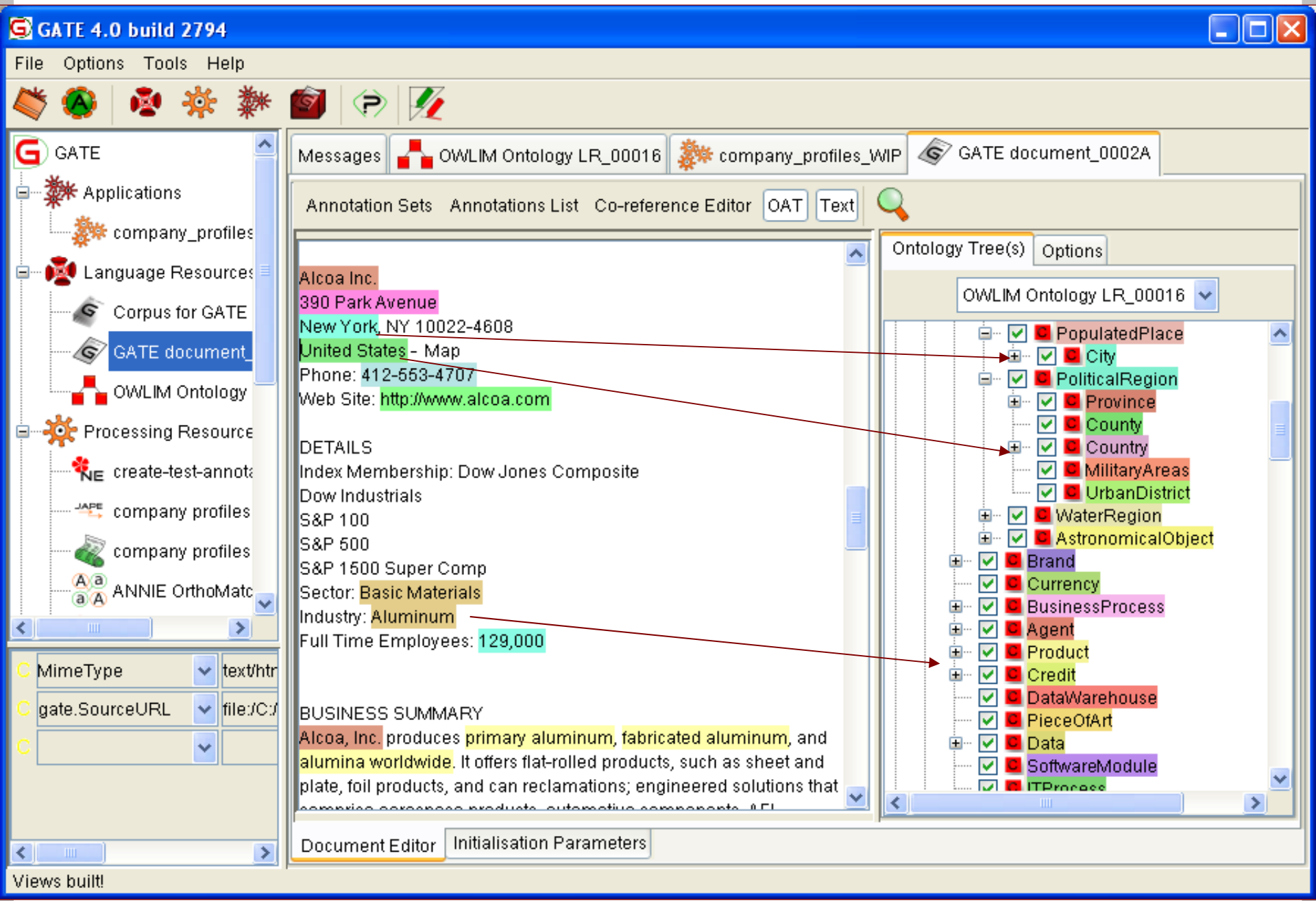
Enter a starting date and an ending date to find all occurrences of dates in

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

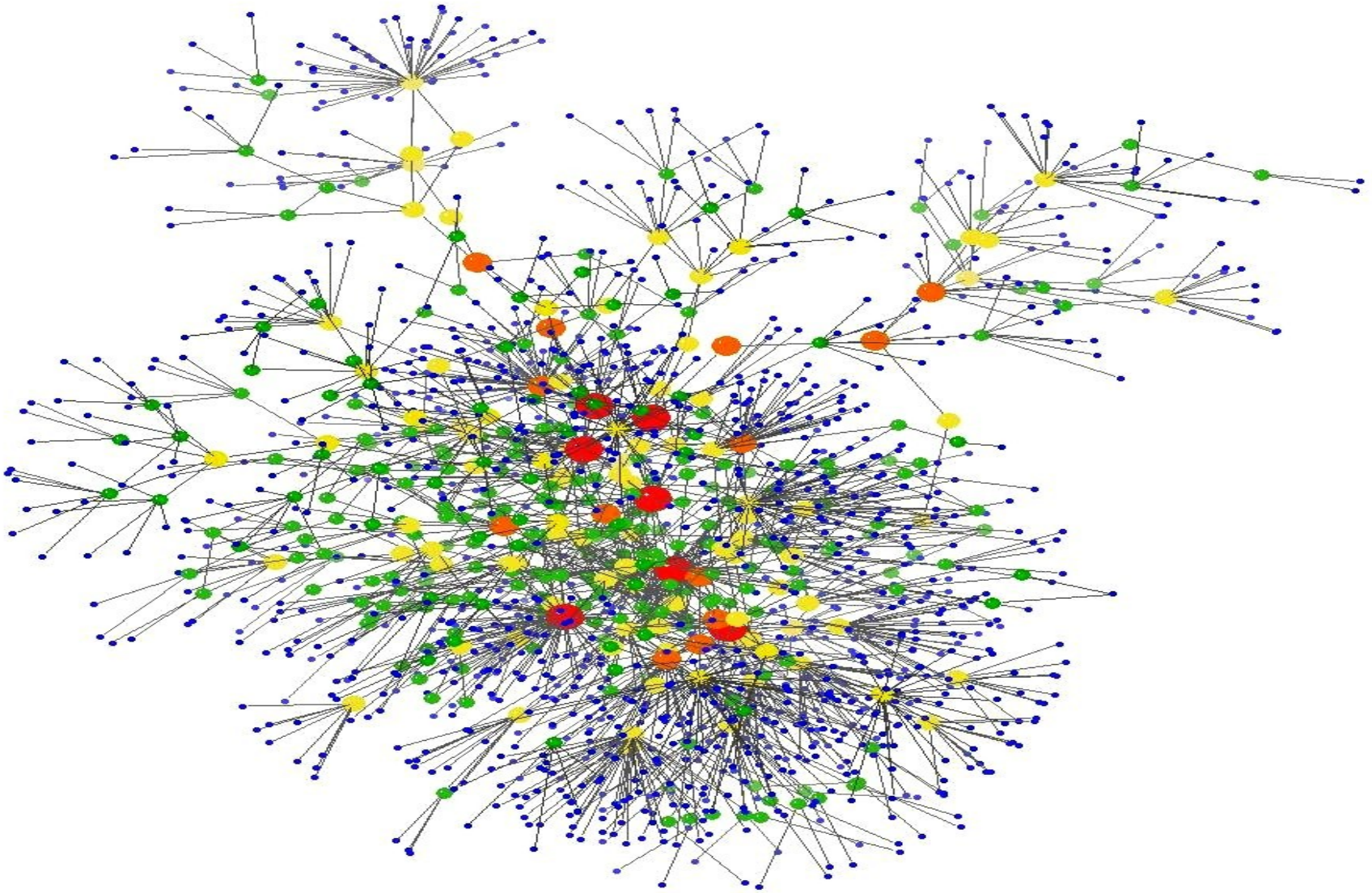
The best 5 Indian regions for a company with the characteristics supplied are returned (best region first) -- available is also an indication of the most relevant region characteristics that contributed to its ranking (either positively or negatively depending on the score sign).

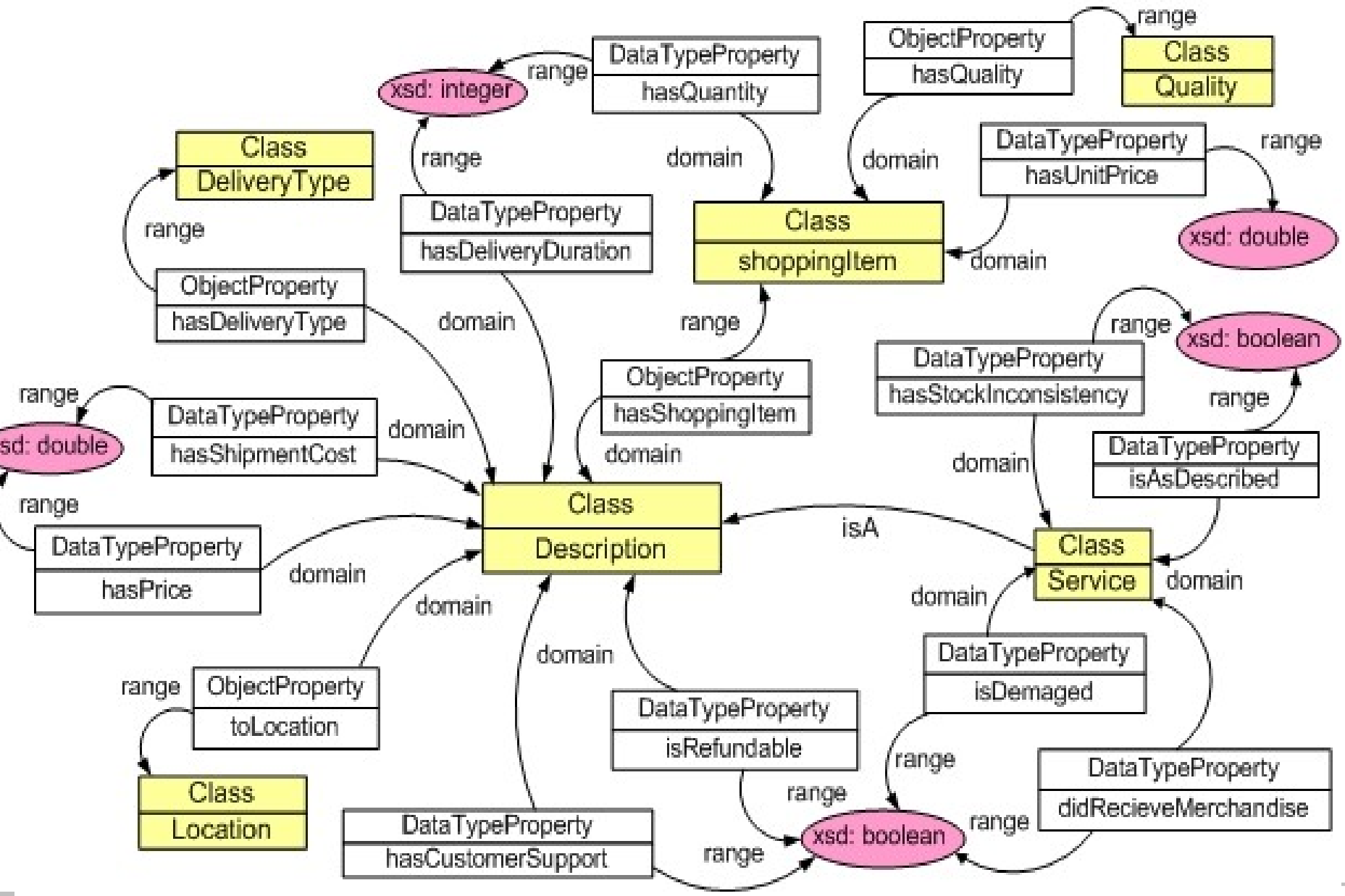
	Region	Global Score	1st (most influent) factor	Partial score	2nd factor	Partial score	3rd factor	Partial score
1	**Delhi**	174.68	Density of population per sq. km.	128.43	Literacy	51.38	Labour cost	-50.77
2	**Daman&Diu**	134.34	Decadal growth of Population	68.78	Literacy	47.10	Population	-23.75
3	**Chandigarh**	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

From NLP to the Semantic Web

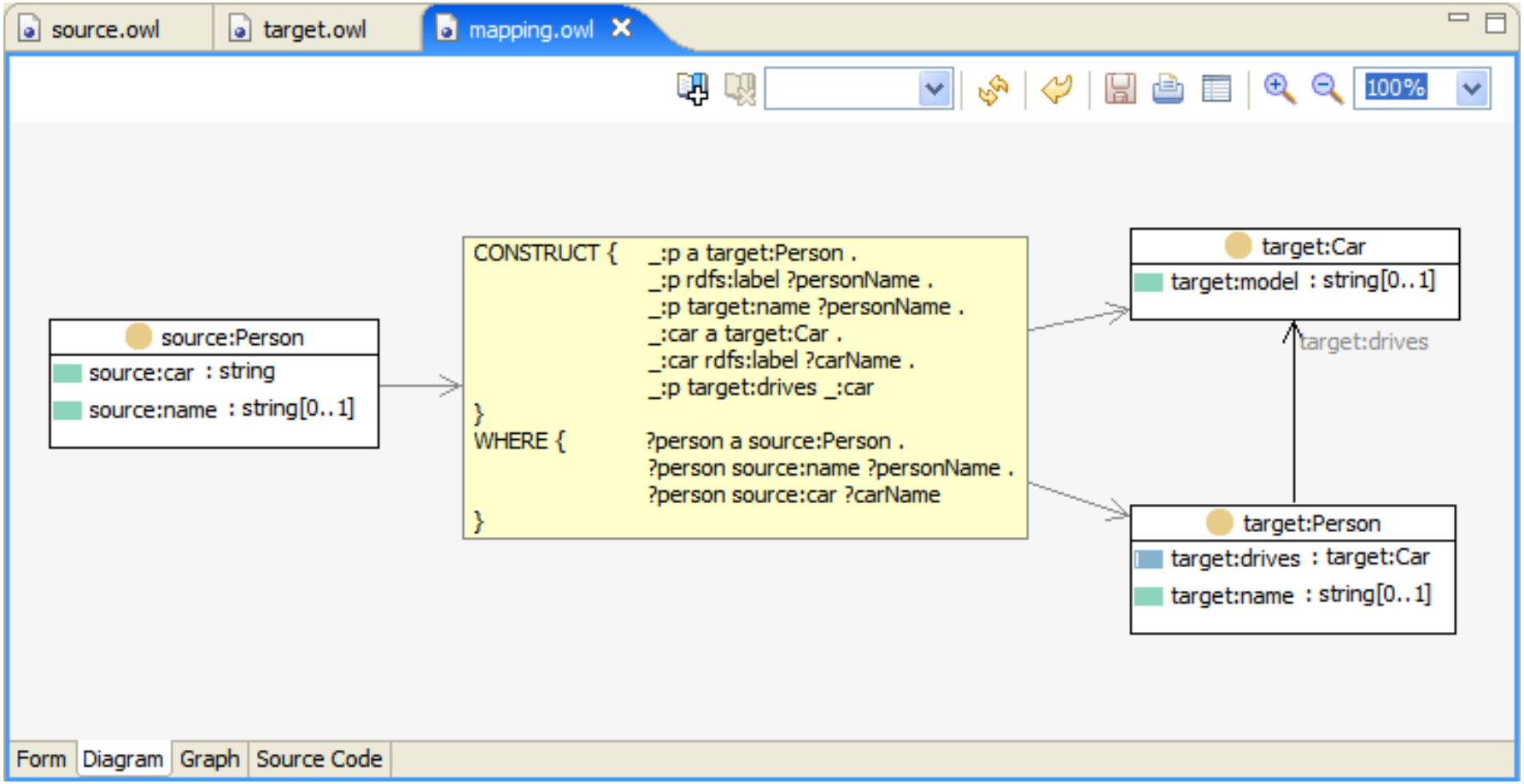
- By now you should be able to see the usefulness of the Semantic Web for NLP, and vice versa
- But how do we get from one to the other?
- Traditionally the two fields have been worlds apart
- They use their own language
- And even use our terms to mean different things!
- But fear not, help is at hand!

Ontologies can be scary monsters




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

Querying them can look scary too....



Making ontologies accessible to ordinary people

- For those who don't know anything about the Semantic Web, the whole concept can be rather confusing
- CLONE is designed to make the whole process of creating ontologies a bit less intimidating for the non-expert user

Controlled Language for Ontology Editing

- Aim to provide a controlled language for basic ontology-editing (and later, querying) functions:
  - easy to learn from examples and simple rules
  - relatively easy to deploy (Java, GATE)
  - unambiguous
  - compact (e.g., create many classes or instances with one sentence)
  - natural but grammatically lax

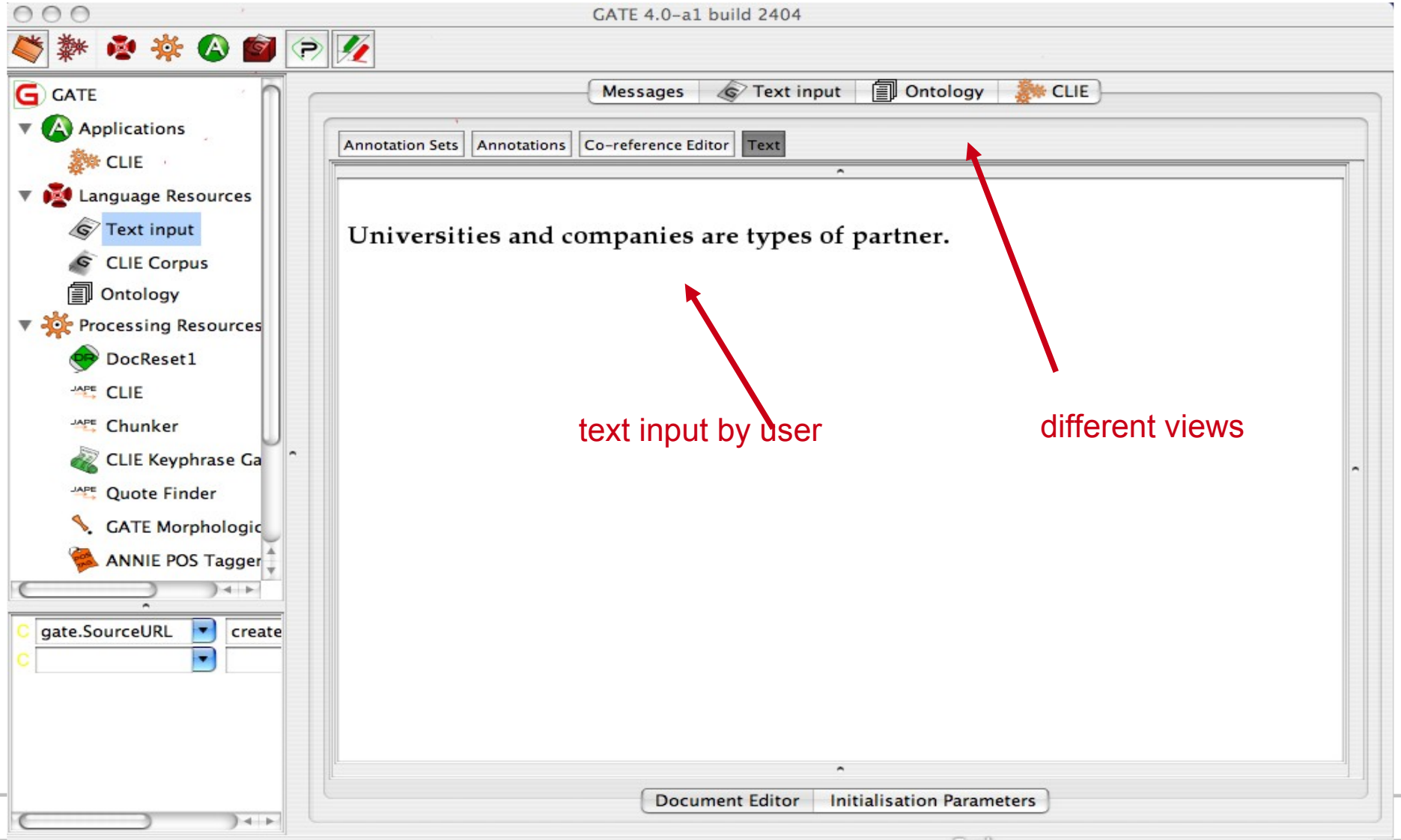
What is it for?

- Creating formal data is a high initial barrier for small organisations and individuals wanting to create ontologies to benefit from semantic knowledge technologies
- CLOnE is aimed at ordinary people who don't know much about ontologies (and don't want to...)
- Allows user to create simple ontologies by typing plain text in a window, and seeing the output ontoloy in another window
- Tests showed people found it easier to use than complex ontology authoring tools such as Protege
- Starting point for more complex applications, e.g. access control policy management system (University of Kent) - users start from a template and can then easily see how to modify it

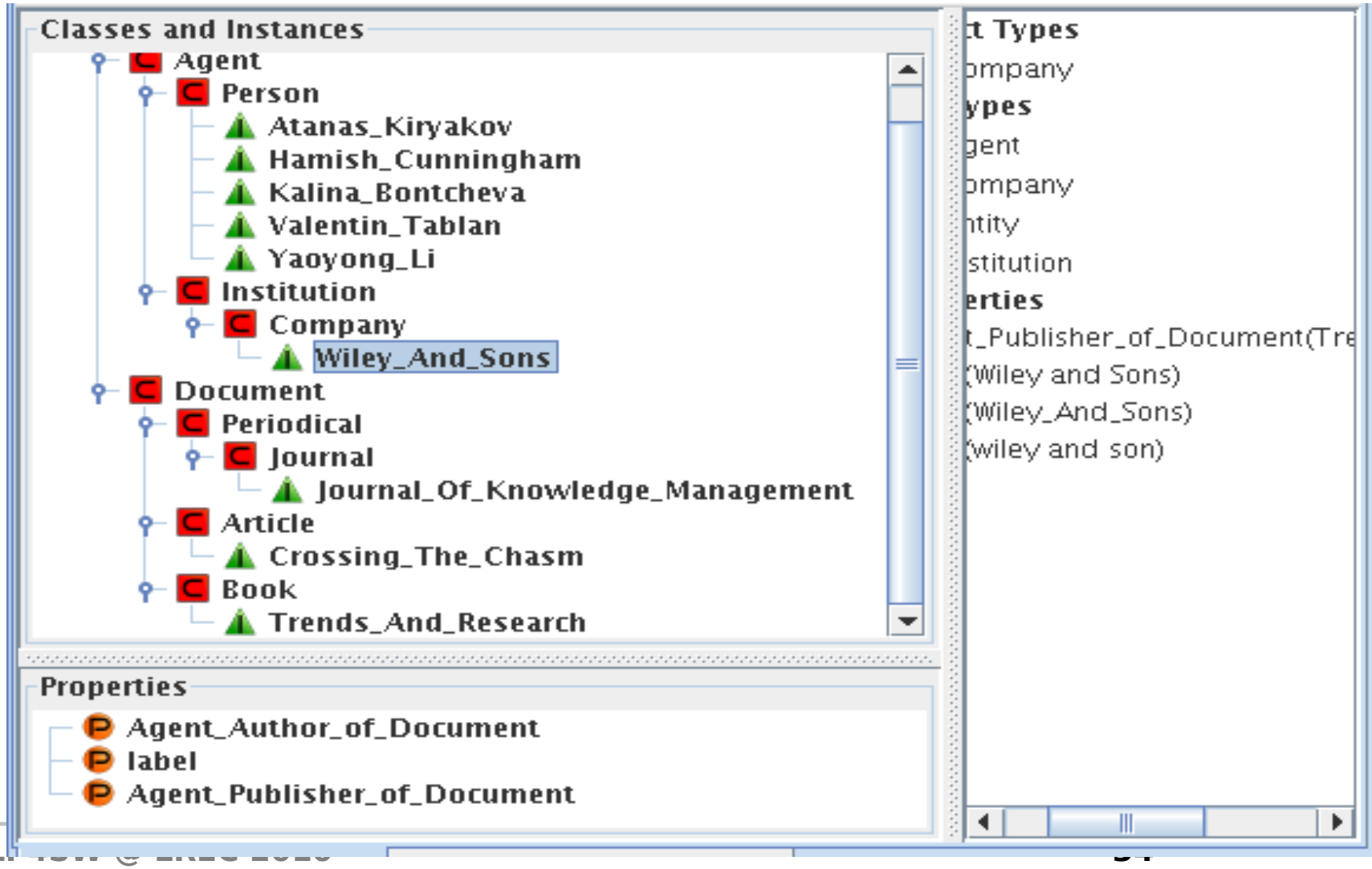
Implementation

- Developed and tested in the GATE GUI, but deployable as a service
- GATE application using text as input to modify an ontology
- Based partly on standard NLP components and modified IE components, with manipulation of the GATE ontology API
- Formed the basis for later tools for more flexible ontology generation from text (SPRAT and SARDINE)

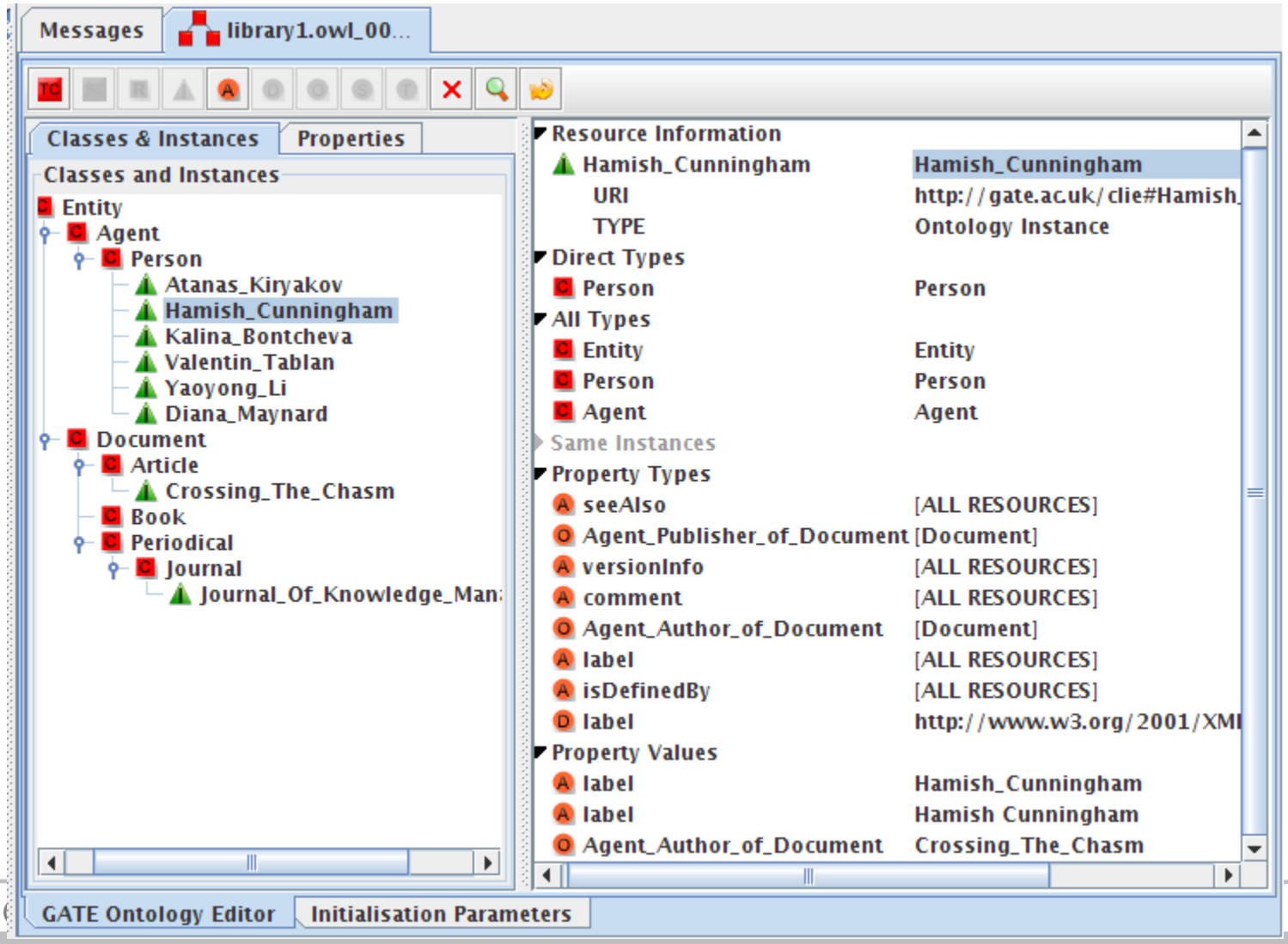
CLONE interface



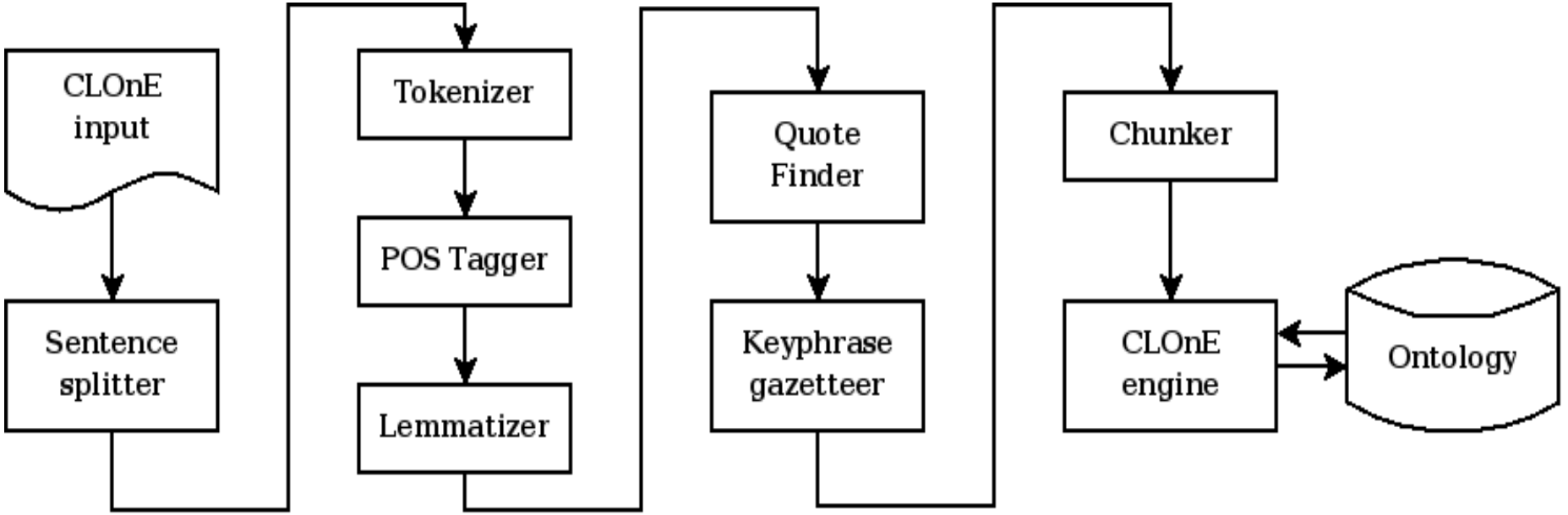
Generated ontology



Ontology in more detail



Implementation



How does it work?

- User inputs text in the controlled language
- Text is analysed in GATE and either accepted as valid or user is informed about errors
- A valid sentence is one that is matched by a rule in GATE
- If the text is valid, the ontology is modified according to the rules matched:
    - Create and delete classes, subclass relations and instances
    - Create and instantiate datatype and object properties
- Ssentences consist of **keyphrases** (fixed pre-determined expressions and punctuation) and **chunks** (free text)
- Rules are implemented in JAPE, the pattern-matching rule-action language used in GATE

Rules and Actions

- There are agents and documents.

[Create new classes for “agent” and “document”]

- Universities and persons are types of agent.

[Create subclasses of “agent” called “university” and “person”]

- “The University of Sheffield” is a university.

[Create an instance of “university” called “The University of Sheffield”]

- Deliverables and conferences have dates as deadlines.

[Create appropriate properties and values]

- Forget that Hamish Cunningham is a person.

[Delete the instance “Hamish Cunningham” from the ontology]

Example JAPE rule

- Create an instance of a class, eg. “Spot is a dog”.

```
(
  (ChunkList):instances
  ({Lookup.majorType == "CLIE-InstanceOf"}):keyphrase
  ({Chunk}):class
)
```

Matches the following sequence:

- a list of one or more NPs
- an entry in a list of “is a” terms
- a single NP

Usability

- Evaluation on different groups of users (naive and expert ontologists)
- Both groups found it easy to learn the syntax of the controlled language
- Most people found it easier to use than more complex ontology editors, especially for simple tasks
- Limitations in functionality
- Most NLP tasks (semantic annotation etc) do not require complex ontological structure

Representing linguistic information

- Ontology Localization is a part of the ontology development process
- Involves adapting an ontology to a concrete language or culture community
- Results in some kind of multilingual ontological system
- Users requiring ontology-based applications don't always speak the same language or have the same cultural background

Representing the label information

- Ontologies are conceptual constructs without linguistics
- Concepts are abstract notions whose labels are arbitrary
- Lexicalizations that function as labels for this concepts are only considered to be evocative of the ontological meaning of the concepts
- Implicit mapping assumption between lexical and conceptual knowledge: intensional senses from a lexical model are mapped to extensional interpretations on ontology elements

Summary

- Introduced GATE for NLP
- Some examples of how it can be used for basic Semantic Web applications, for non-experts in SW technology
- Introduced CLONE: a tool for generating simple ontologies for the non-expert
- Representation of linguistic information in ontologies for e.g. localisation
- Next session will introduce more complex technology: making the transition from traditional to semantic annotation in GATE, and how to evaluate the results