

Infrastructure for Human Language Technology and the Semantic Web

<http://gate.ac.uk>

Open Source
Scalable
Robust

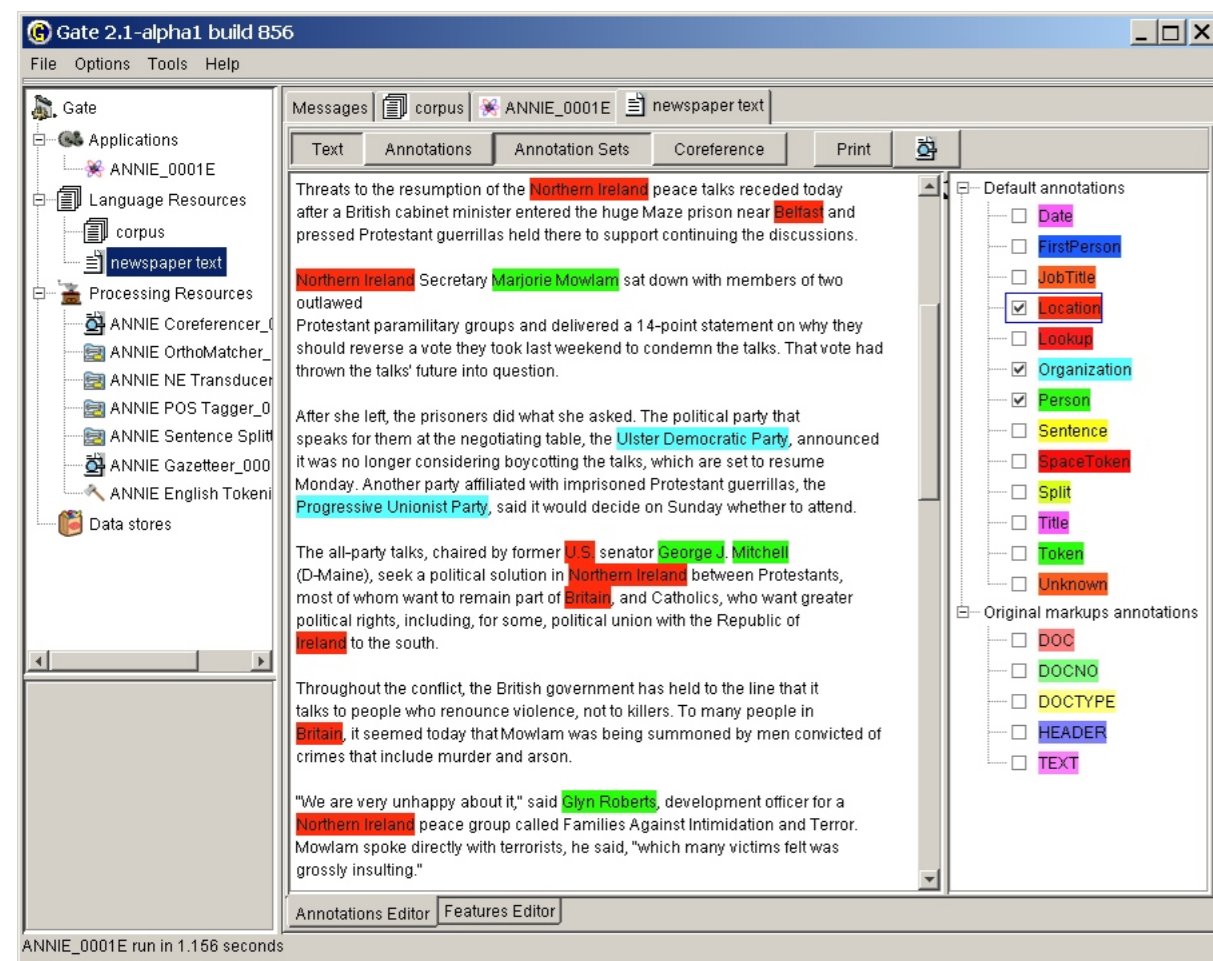
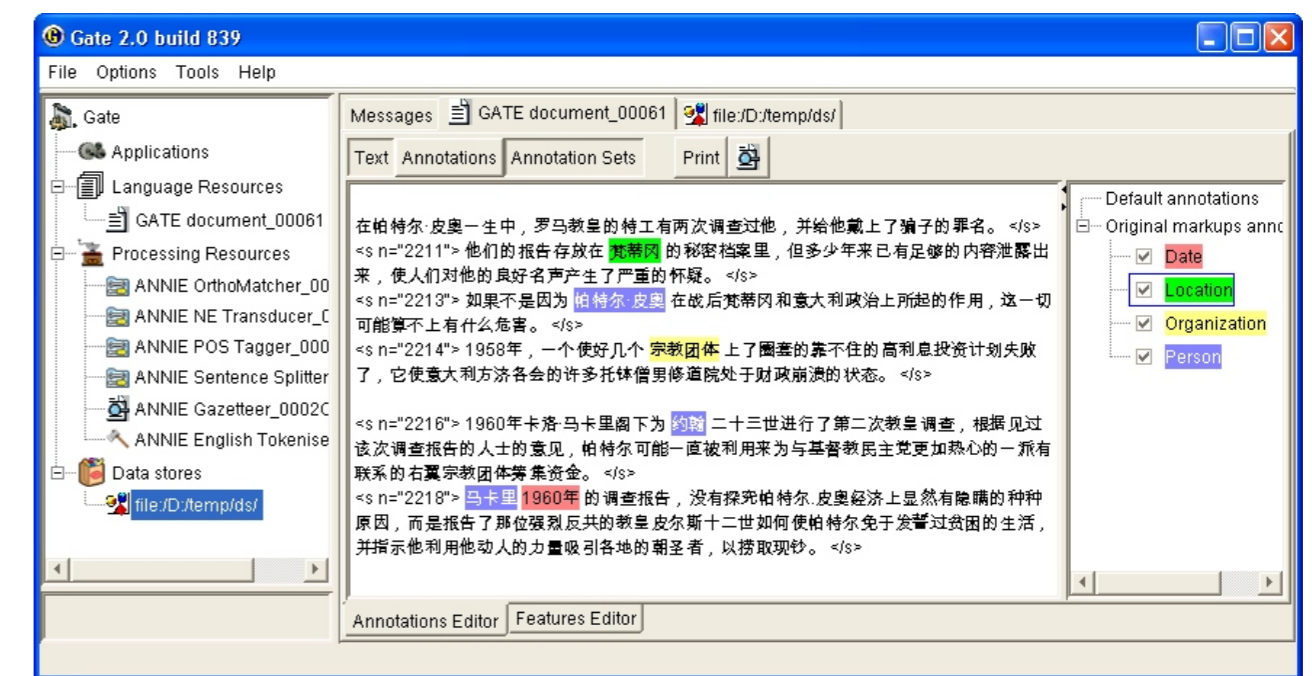
Stable
Customisable
Reusable Components

100% Java
XML, HTML, RTF, ...
Oracle, PostgreSQL backends

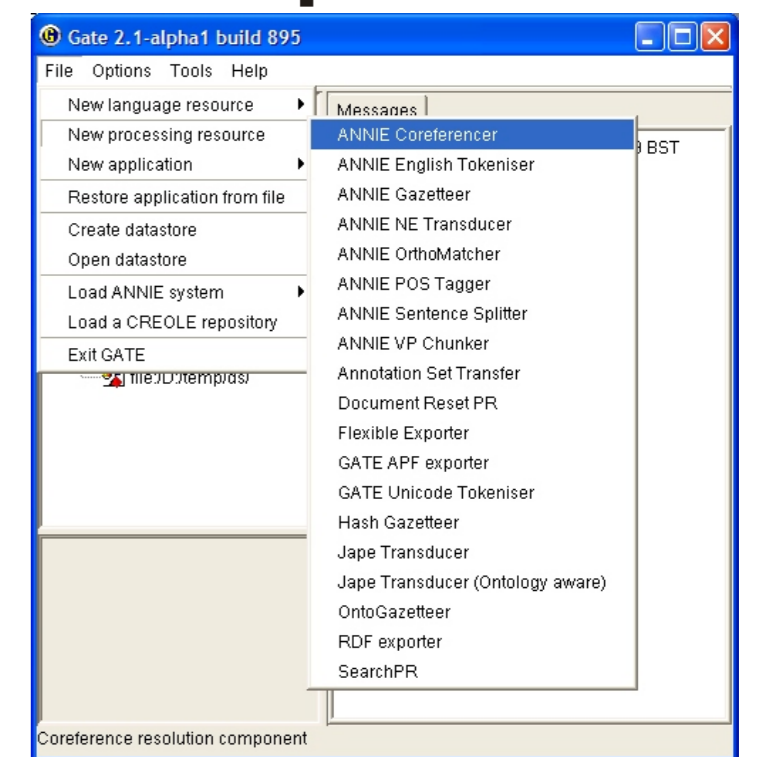
Contact: Hamish Cunningham (hamish@dcs.shef.ac.uk)

Information Extraction

- ◆ High Performance
- ◆ Robust across genres and formats
- ◆ Multilingual
- ◆ Portable

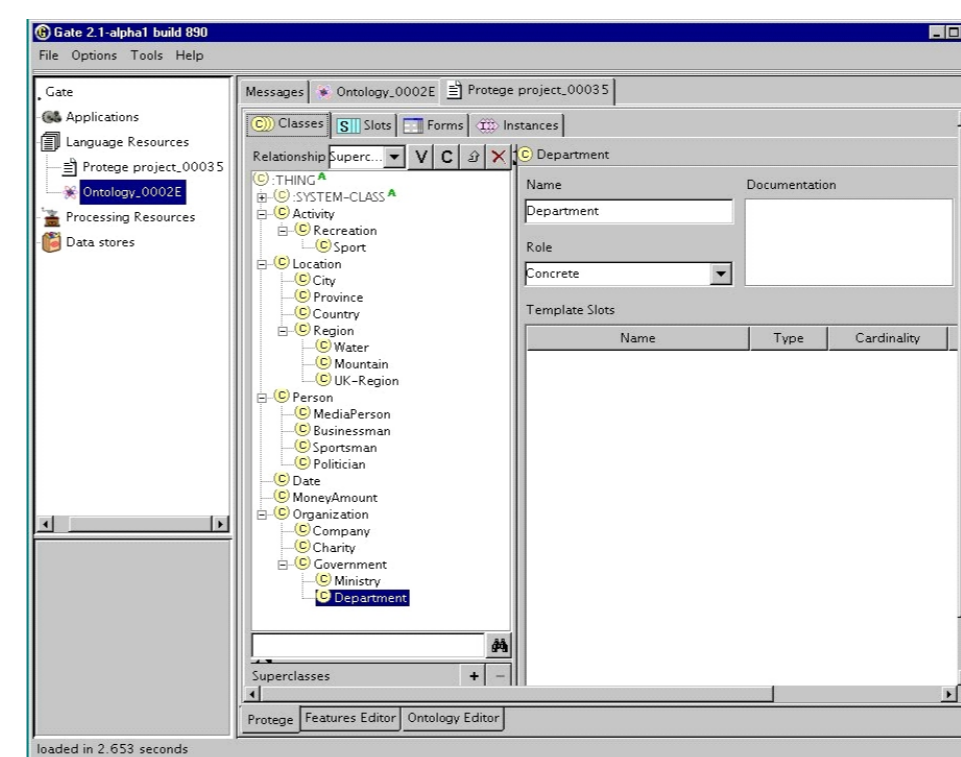


Components



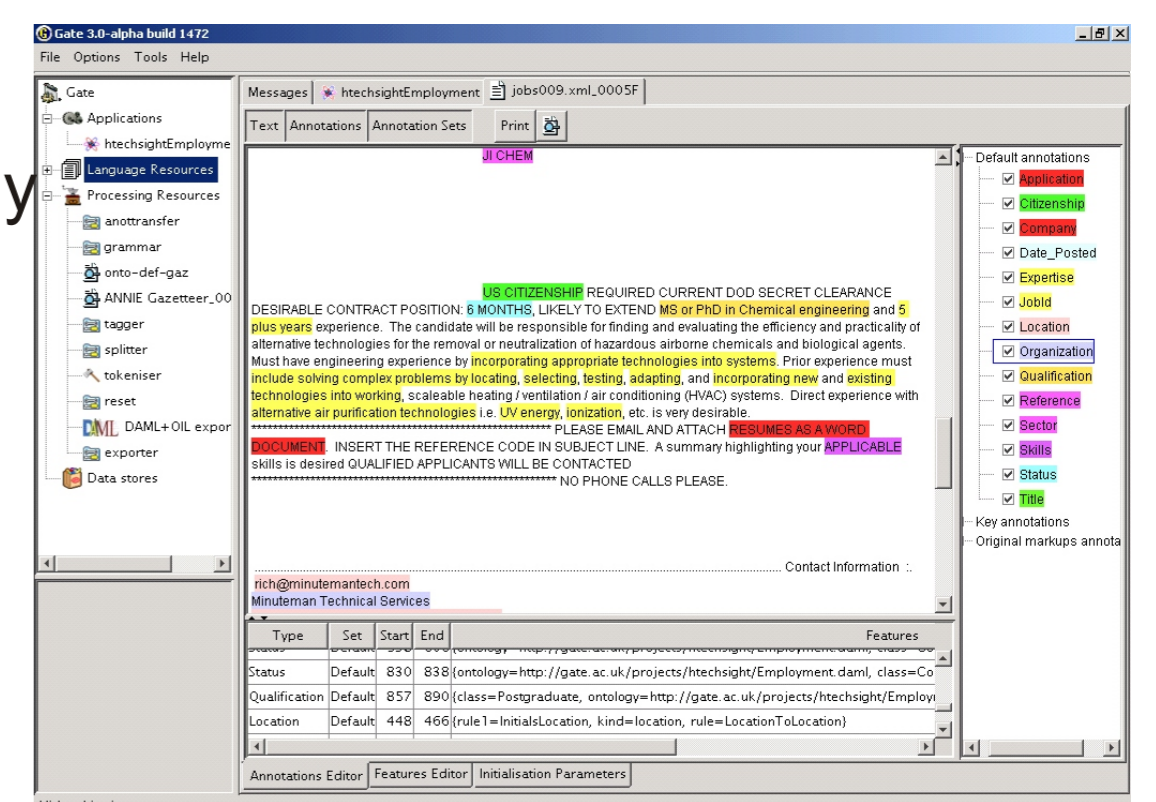
Ontologies

- ◆ Ontology management through Protégé & Sesame integration.
- ◆ RDF, OWL, DAML+OIL support.



Ontology Population using IE

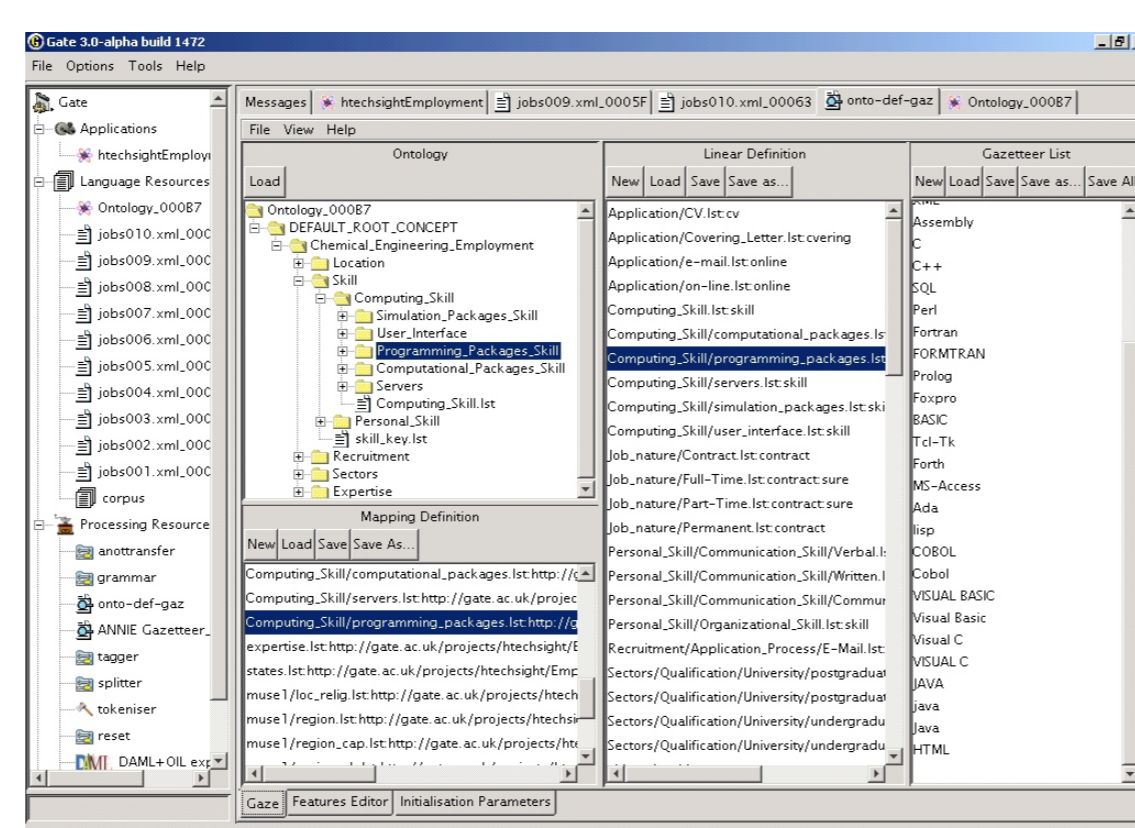
- ◆ Given a domain ontology
- ◆ Identify instances in text
- ◆ Generate semantic metadata



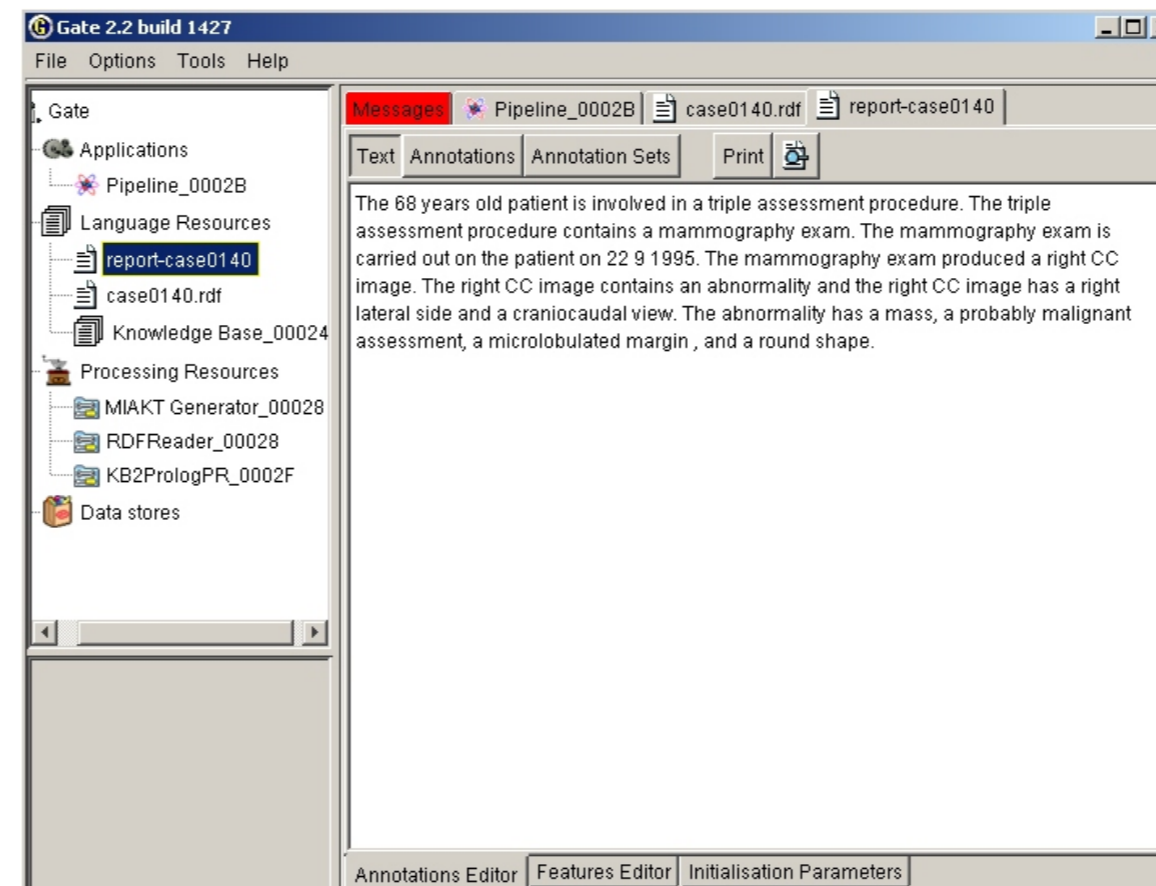
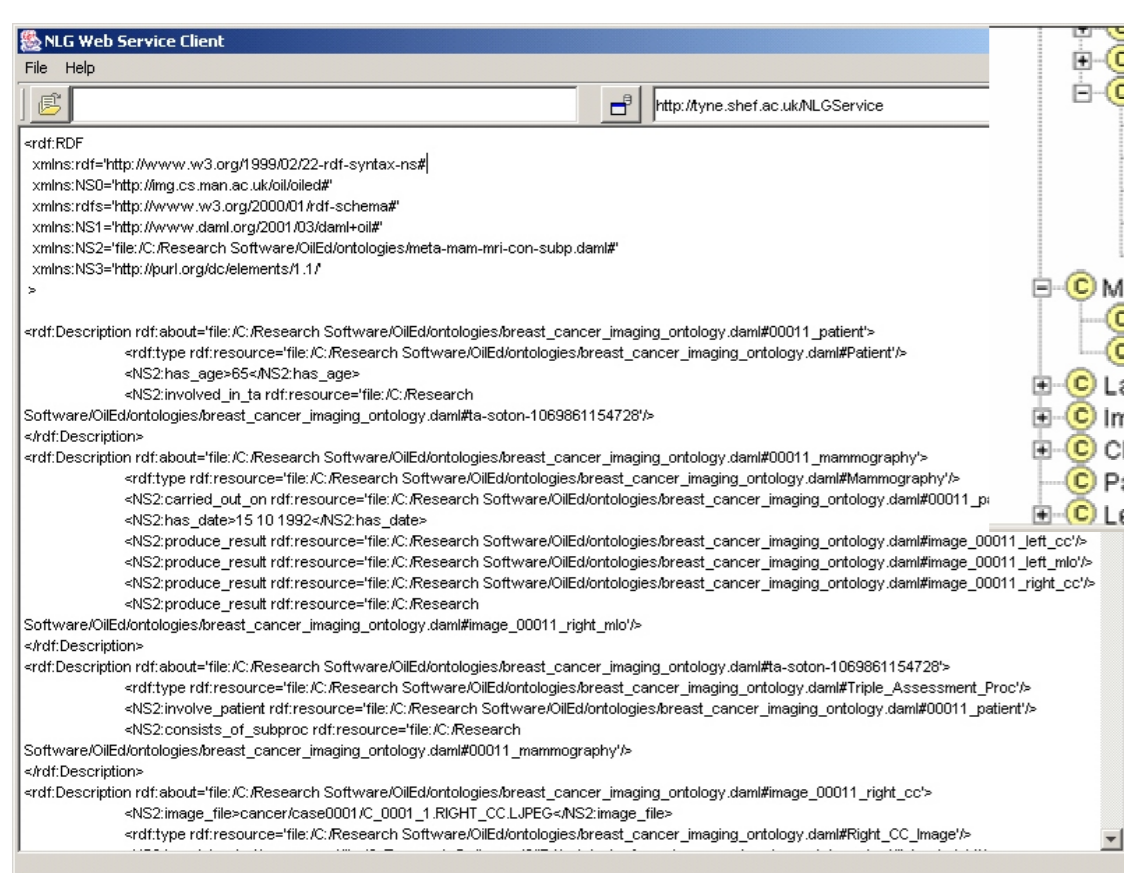
Machine Learning

- Algorithms:
- ◆ WEKA - decision trees, Bayesian methods
 - ◆ Maximum Entropy
 - ◆ SVM Light

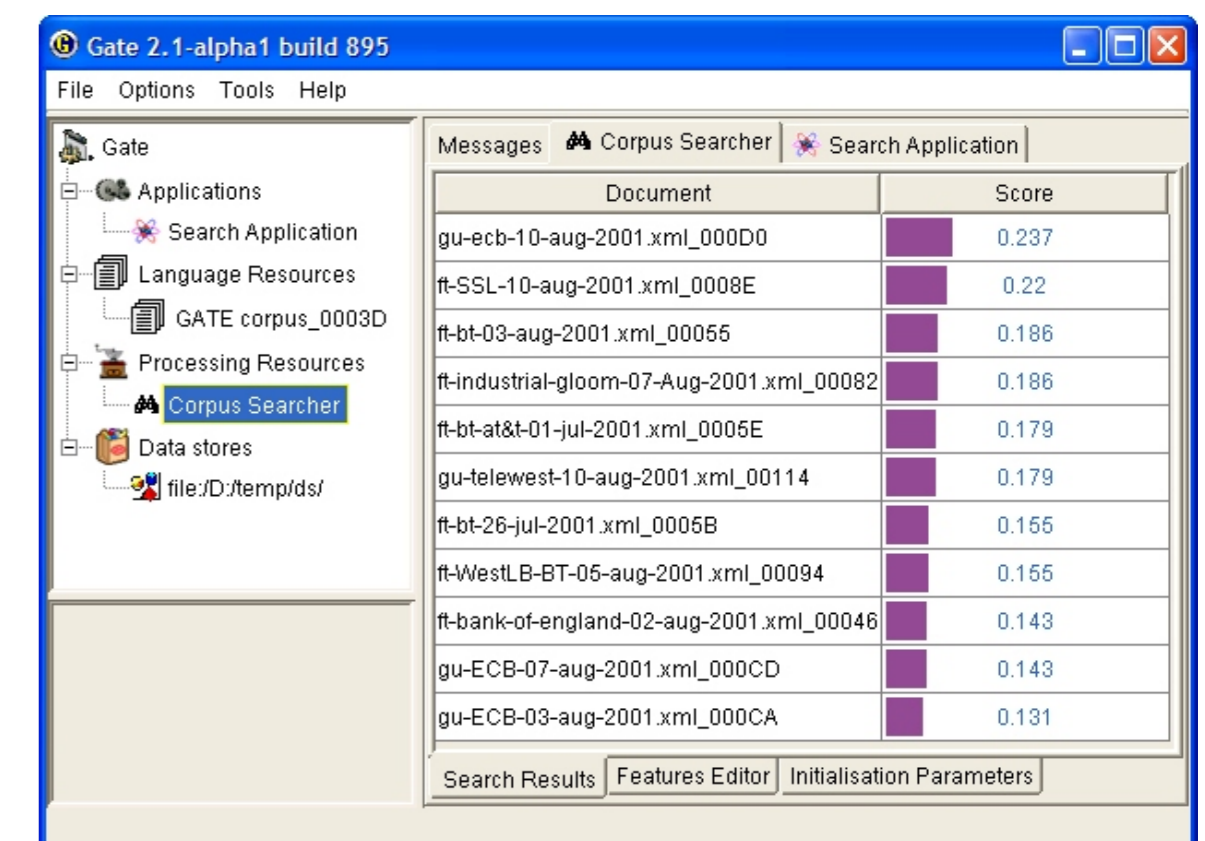
- Used for:
- ◆ Information Extraction
 - ◆ IE-based ontology population



Automatic Text Generation from Ontologies



Information Retrieval



Full featured implementation based on

Research

- ◆ Repeatability
 - ◆ Collaboration
 - ◆ Reuse vs. Reinvention
 - ◆ Quantitative evaluation
- Semantic Web projects:
- ◆ SEKT - <http://sekt.semanticweb.org>
 - ◆ KnowledgeWeb - knowledgeweb.semanticweb.org
 - ◆ hTechSight - <http://www.h-techsight.org>
 - ◆ AKT - <http://www.akt.org>
- Digital library projects:
- ◆ PrestoSpace - <http://www.prestospace.org>
 - ◆ ETC SL - <http://www-etcsl.orient.ox.ac.uk>

Commercial

- Used at:
- ◆ BT Exact
 - ◆ Reuters plc.
 - ◆ GlaxoSmithKline plc.
 - ◆ Merck KgAa
 - ◆ British Gas Plc.

Performance Evaluation

String	KeyStart	KeyEnd	Key	String	ResponseStart	ResponseEnd	Response
England	2358	2365	England	2358	2365	2365	
UK	258	260	UK	258	260	260	
Hampshire	2638	2647					
Swanwick	2886	2894					
Europe	746	752	Europe	746	752	752	
Wales	2370	2375	Wales	2370	2375	2375	
UK	2801	2803	UK	2801	2803	2803	
Swanwick	2628	2636					
UK	931	933	UK	931	933	933	

Precision strict: 1.0000 Recall strict: 0.6667 F-Measure strict: 0.8000
 Precision average: 1.0000 Recall average: 0.6667 F-Measure average: 0.8000
 Precision lenient: 1.0000 Recall lenient: 0.6667 F-Measure lenient: 0.8000

GATE research is funded by EU & EPSRC grants

Integrated tools for performance evaluation.