

FREE

Open source, licensed under LGPL allowing unrestricted commercial use, hosted on SourceForge.

100% JAVA

Runs on **any platform** supporting Java 5 or later. Developed and tested daily on Linux, Windows, and Mac OS X.

MATURE AND ACTIVELY SUPPORTED

In development **since 1996**; now at version 5.0; around 20 active developers.

COMPREHENSIVE

Support for manual annotation, performance evaluation, information extraction, [semi-]automatic semantic annotation, and many other tasks. Over **50 plugins** included with the standard distribution, containing over 70 resource types. Many others available from independent sources.

STANDARDS-BASED

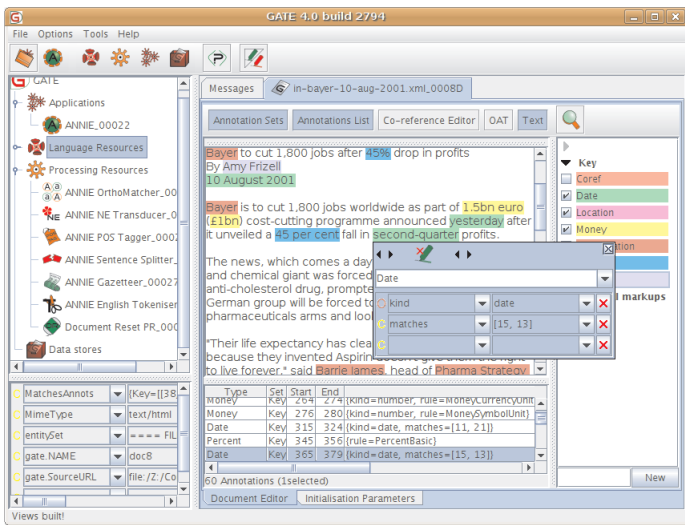
Reference implementation in **ISO TC37/SC4 LIRICS** project; supports XCES, ACE, TREC etc. formats; founder member of **OASIS/UIIMA** committee.

EFFICIENT

Optimisations included with the latest version provide a 20 to 40% speed and memory usage improvement. Highly efficient finite state text processing engine; many plugins with linear execution time.

POPULAR

Assessed as “outstanding” and “internationally leading” by an anonymous EPSRC peer review. Used at thousands of sites: companies, universities and research laboratories, all over the world. ~35,000 downloads/year. Rolling funding for more than 15 staff at the University of Sheffield.



DATA MANAGEMENT

Pluggable input filters with out of the box support for XML, HTML, PDF, MS Word, email, plain text, etc.

Common in-memory data model built around stand-off annotation, documents and corpora.

Persistent storage layer with support for XML, Oracle, PostgreSQL, or Java serialisation. I/O interoperation with many other systems.

STANDARD ALGORITHMS

Ready made implementations for many typical NLP tasks such as tokenisation, POS tagging, sentence splitting, named entity recognition, co-reference resolution, machine learning, etc.

USER INTERFACE

Comprehensive tool set for data editing and visualisation, rapid application development, manual annotation, ontology management.

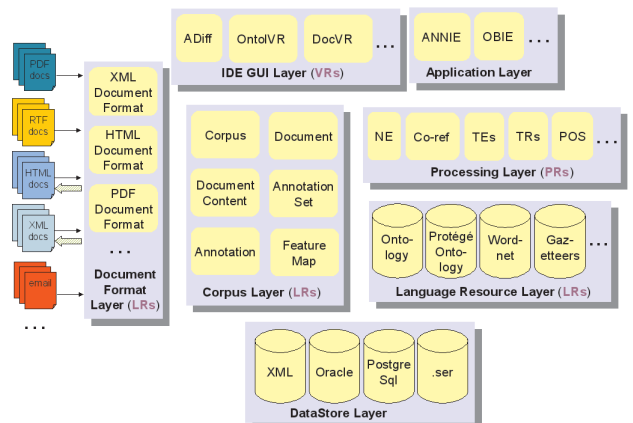
INTEGRATION

Leveraging the power of other projects such as:

- **Information Retrieval:** Lucene (Nutch, Solr), Google and Yahoo search APIs, MG4j;
- **Machine Learning:** Weka, MaxEnt, SVMLight, etc.;
- **Ontology Support:** Sesame and OWLIM;
- **Parsing:** RASP, Minipar, and SUPPLE;
- **Other:** UIMA, Wordnet, Snowball, etc.

COMMUNITY AND SUPPORT

Friendly and active community of developers and users offers efficient help. Commercial support available from Ontotext and Matrixware.



OVERVIEW

GATE, a General Architecture for Text Engineering, was first released in 1996, then completely re-designed, re-written, and re-released in 2002. The system is now one of the most widely-used systems of its type and is a relatively comprehensive infrastructure for language processing software development.

The new UIMA architecture from IBM/Apache has taken inspiration from GATE and IBM have paid the University of Sheffield to develop an interoperability layer between the two systems.

Key features of GATE are:

- Component-based development reduces the systems integration overhead in collaborative research.
- Automatic performance measurement of Language Engineering (LE) components promotes quantitative comparative evaluation.
- Distinction between low-level tasks such as data storage, data visualisation, discovery and loading of components and the high-level language processing tasks.
- Clean separation between between data structures and algorithms that process human language.
- Consistent use of standard mechanisms for components to communicate data about language, and use of open standards such as Unicode and XML.
- Insulation from idiosyncratic data formats (GATE performs automatic format conversion and enables uniform access to linguistic data).
- Provision of a baseline set of LE components that can be extended and/or replaced by users as required.

INFORMATION EXTRACTION

Information Extraction (IE) is a process which takes unseen texts as input and produces fixed-format, unambiguous data as output. This data may be used directly for display to users, or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in Information Retrieval (IR) applications.

IE covers a family of applications including named entity recognition, relation extraction, event detection.

GATE has been used for **IE applications** in domains including bioinformatics, health and safety, and 17th century court reports.

IE systems built on GATE have been evaluated among the top ones at **international competitions** (MUC, ACE, Pascal). A system built by the GATE team came top in two of three categories in the NTCIR 2007 patent classification competition.



THE GATE FAMILY

- **GATE Developer**: an integrated development environment for language processing components bundled with the most widely used Information Extraction system and a comprehensive set of other plug-ins
- **GATE Embedded**: an object library optimised for inclusion in diverse applications giving access to all the services used by GATE Developer and more
- **GATE Teamware** a collaborative annotation environment for high volume factory-style semantic annotation projects built around a workflow engine and the GATE Cloud backend web services
- **GATE Cloud**: a parallel distributed processing engine that combines GATE Embedded with a heavily optimised service infrastructure

FIRST COUSINS: THE ONTOTEXT FAMILY

- **Ontotext KIM**: UIs demonstrating our multiparadigm approach to information management, navigation and search
- **Ontotext Mimir**: (Multi-paradigm Information Management Index and Repository) a massively scaleable multiparadigm index built on Ontotext's semantic repository family, GATE's annotation structures database plus full-text indexing from MG4J

Working with Ontologies

ONTOLOGY ABSTRACTION LAYER

Based on Sesame RDF store (<http://openrdf.org>), with additional OWL support provided by OWLIM (<http://www.ontotext.com/owlim/>), leading to one of the fastest and most scalable triple stores. Ontologies can be loaded with storage in memory, on disk or on a dedicated server.

ONTOLOGIES IN GATE

Taxonomical relations can be used in annotation matching, thus enhancing JAPE's power of generalisation.

Graphic interface tools for ontology visualisation, ontology editing, and semantic annotation of text are included with GATE.

ONTOLOGY LEARNING

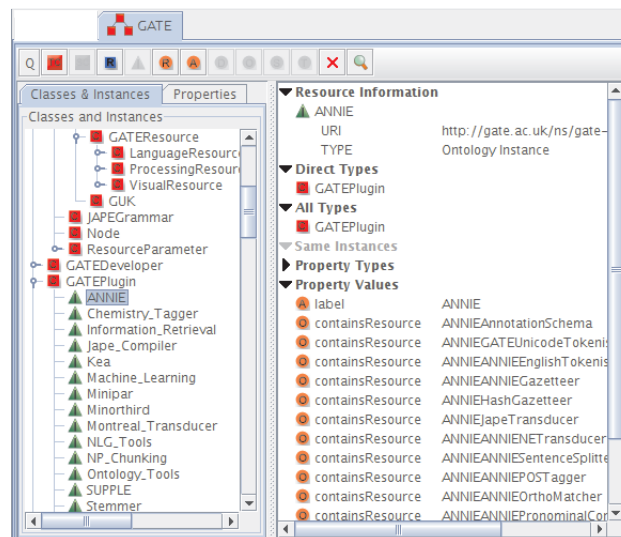
Automatically extending ontologies with knowledge extracted from text through Information Extraction.

KNOWLEDGE BASE POPULATION

Automatically populating knowledge bases with instance data extracted from text. This is related to Semantic Annotation.

AN EXAMPLE

The GATE ontology shown to the right has been automatically derived from source code and associated meta-data and automatically populated from the software and user documentation.



Research Projects

E-SCIENCE

- **AKT:** Advanced Knowledge Technologies;
- **MultiFlora:** support for biodiversity research;
- **MiAKT:** collaborative problem solving in medical informatics;
- **CLEF:** tools for integrating patient information from text and images.

DIGITAL LIBRARIES

- **GATE/ETCSL:** automatic morphological analysis of literary Sumerian texts;
- **EMILLE:** building a 63M words corpus of South Asian languages;
- **OldBaileyIE:** named entity recognition on 17th century Old Bailey court reports;
- **PrestoSpace:** automatic creation of meta-data for news broadcasts used for advanced indexing and conceptual search.

SEMANTIC WEB AND KNOWLEDGE TECHNOLOGIES

- **SEKT:** Next Generation Knowledge Management;
- **hTechSight:** building a KM platform for the chemical industry;
- **TAO:** migrating legacy applications to open, semantics-based, service oriented architectures;

```
<rdf:Description rdf:about="file:/breast_cancer_ontology.daml#01401_patient">
  <rdf:type rdf:resource="file:/breast_cancer_ontology.daml#Patient"/>
  <NS2:has_age>68</NS2:has_age>
  <NS2:involved_in_ta
rdf:resource="file:/breast_cancer_ontology.daml#ta-soton-1069861276136"/>
</rdf:Description>
<rdf:Description rdf:about="file:/breast_cancer_ontology.daml#01401_mammography">
  <rdf:type rdf:resource="file:/breast_cancer_ontology.daml#Mammography"/>
  <NS2:carried_out_on rdf:resource="file:/breast_cancer_ontology.daml#01401_patient"/>
  <NS2:has_date>22 9 1995</NS2:has_date>
  <NS2:produce_result
rdf:resource="file:/breast_cancer_ontology.daml#image_01401_left_cc"/>
  <NS2:produce_result
rdf:resource="file:/br
  <NS2:produce_re
rdf:resource="file:/br
  <NS2:produce_re
rdf:resource="file:/br
</rdf:Description>
```

The 68 years old patient is involved in a triple assessment procedure. The triple assessment procedure contains a mammography exam. The mammography exam is carried out on the patient on 22 9 1995. The mammography exam produced a right CC image. The right CC image contains an abnormality and the right CC image has a right lateral side and a craniocaudal view. The abnormality has a mass, a probably malignant assessment, a microlobulated margin , and a round shape.

- **NeON:** shaping the future infrastructure for semantic applications;
- **MediaCampaign:** automating the detection and tracking of media campaigns on television, Internet and in the press .

HUMAN LANGUAGE TECHNOLOGY

- **MUSE:** Named entity recognition from diverse text types;
- **SAFE:** collaborative mixed initiative semantic annotation;
- **LIRICS:** definition of an ISO standard for language technology with a reference implementation.



Keys: **Person** **Location** **Organization**

date and that his seriously ill in hospital on the people of the buses travelling on this why stop [SIL] now is to be met [s] the lib dems' continues to eat his hat [SIL] into the break [SIL] and [SIL] be mature [SIL] enough to use [SIL] them [s] no welcome to the bbc's news at one o'clock [s] the trial of another war but has collapsed at the opening [SIL] hour brown face charges of stealing valuables including an ornate what we're able in sailing ship [SIL] worth more than half a million pounds from diana princess of wales [s] the case against him the society jeweller edge and having collapsed before a jury was sworn in [s] the prosecution said there was no realistic prospect of conviction [s] so if the collapse of the trial of paul but last month [s] now this to be an internal review of the metropolitan police's investigations [SIL] are royal correspondent nicholas witchell [SIL] is at the old bailey [SIL] make [SIL] and this was a decision taken [SIL] personally by v. director of public prosecutions david calvert smith [s] it will i think kirby the cause of a huge relief within [SIL] the royal palaces have been [SIL] those of the royal palaces to be in desperate really is a kick [SIL] this whole further into touch and the crown prosecution service [SIL] has now obliged [s] the metropolitan police though are not happy they believe that in this [SIL] instance then it came out [SIL] a four and thorough investigation [SIL] and most unusually in explaining the decision to drop the case [SIL] the prosecution went through [SIL]

many of the facets of what would of been the prosecution case [s] now they say [SIL] kept changing his story how initially said [SIL] that these items had been disposed off [SIL] before [SIL] princess of wales's death of the money had been given to her [s] in fact the evidence was clearly they were disposed off [s] after [SIL] her death [SIL] and then [SIL] [SIL] apparently saying that he be passed to dispose of these items by [SIL] another [SIL] man [SIL] [SIL] butler and it was that [SIL] of course that gave [SIL] the crown their way out of this [SIL] the second instalment of the saga [SIL] of the crown [SIL] and the butler's [SIL] [SIL] [SIL] and other buffalo who work for the prince and princess of wales [SIL] a mother of a clue who's experienced the curious way in which some royal households [SIL] have dealt with gifts and their disposal [s] his co accused and have any [SIL] society you learned was charged with receiving items from [SIL] [SIL] among them this model of an arab dow which would be a wedding present the prince and princess [SIL] from the emir of [SIL] [SIL] the two men stood in the dock of the number one court [SIL] and pleaded not guilty to all the charges against them [s] that was the cue [SIL] for the crown to begin its retreat [s] crown counsel william boyce qc explained how mr [SIL] had changed his story in relation to the dow several times [SIL] in the end claiming his butler colleague [SIL] [SIL] bar all had told him to sell it [SIL] and that was the crown's escape route [SIL] because mr powell [SIL] had had that conversation with [SIL] [SIL] this is what william boyce told the court [SIL] narrowed [SIL] final position was that his actions in relation to the dow were effectively [SIL] authorized by poor bowel [s] if however [SIL] barrow was willing to inform [SIL] [SIL] that he was holding on to property belonged to [SIL] princess of wales [SIL] without discernible disapproval from her [s] it enhances the credibility of hell [SIL] s account [s] but he reasonably accepted that poor barrel [SIL] was acting in good faith [s] that was the prosecution's way [SIL] of throwing in the town [s] both men were acquitted [s] mr [SIL] s lawyers said his client was delighted at [SIL] our ground [SIL] many of the

Ontotext
Knowledge and Language Engineering Lab of Sima

KIM
Knowledge and Information Management Platform

John Prescott, a Man

Property	Value
hasMainAlias	John Prescott
hasPosition	manager
hasAlias	John Prescott
hasAlias	Brown

Related Entities

Resource	Link to John Prescott
manager	holder

Copyright © 2004 Ontotext Lab, Sima AI, Bulgaria

Headline: Another right royal embarrassment

Description: The collapse of a second trial involving a royal butler raises more awkward questions.

SEMANTIC ANNOTATION

Automatic and semi-automatic production of **semantic meta-data** for text and multimedia. GATE identifies mentions of known concepts and instances from an ontology. This type of meta-data enables a search by meaning paradigm to enhance traditional retrieval methods. Searches like "find companies located in Western Europe involved in the high tech sector" become possible.

RICHNEWS

RichNews is an example of applying semantic annotation to **multimedia broadcasts**. It combines timing information from automatic speech recognition and semantics extracted from news web pages. The meta-data produced is then used for advanced retrieval facilities.

SCIENCE

GATE is a platform for experimental repeatability, quantitative evaluation, collaborative development and integration.

EDUCATION

GATE and the ANNIE IE system are used in classrooms across the world and in many postgraduate projects.

BUSINESS

GATE has been engineered to a high standard in order to be suitable for deployment in commercial applications software, and is based on components, mobile code and internet-based distribution. A serious effort has been made to achieve a very high level of quality; unit and regression tests are run nightly on three different platforms.

Our IE software is quality-controlled and Sheffield has applied IE in very many domains, and developed World-leading expertise in producing robust systems

Annotation Diff tool

Key: Document_00016 | Annotation Set: | Key: | Annotation Type: Date | F-Measure Weight: 1.00

Response: Document_00016 | [Default set] | Features: All | Some | None | 1.00

Start	End	Key	Features	Start	End	Response	Features
1192	1204	last quarter	(kind=date)	=	1192	1204	last quarter (rule2=DateOnlyFinal, kind=date, rul
2431	2447	previous quarter	(kind=date)	=	2431	2447	previous quarter (rule2=DateOnlyFinal, kind=date, rul
2022	2028	Friday	(rule2=DateOnlyFinal, kind=date, rule1=GazDate)	=	2022	2028	Friday (rule2=DateOnlyFinal, kind=date, rul
1877	1888	end of June	(kind=date)	~	1884	1888	June (rule2=DateOnlyFinal, kind=date, rul
643	654	this autumn	()	?~	50	54	fall (kind=date)
1906	1922	three months ago	(kind=date)	-?			
225	231	Monday	(rule2=DateOnlyFinal, kind=date, rule1=GazDate)	<	225	231	Monday (kind=time)

Correct: 3 | Recall Precision F-Measure | Export to HTML

Partially Correct: 1 | Strict: 0.4286 0.50 0.4615

Missing: 3 | Lenient: 0.5714 0.6667 0.6154

False Positives: 2 | Average: 0.50 0.5833 0.5385

for diverse applications.

Commercial users include: Glaxo Smith Kline PLC, AT&T, Master Foods NV, British Gas PLC, Syntalex Ltd., Lernout Hauspie GmbH., Thompson Corp, Innovantage, Garlik, Fizzback, Spock.

ACKNOWLEDGEMENTS

GATE research is or has been funded by the EU, EPSRC, AHRB, BBSRC, and commercial grants.

Sponsored by: **Ontotext.com**, **Matrixware.com**

CONTACT:

Prof. Hamish Cunningham

<http://www.dcs.shef.ac.uk/~hamish/>