



GATE

General Architecture for Text Engineering

Presented by Ahmed Magdy Ezzeldin

What is Text Engineering?

- Text or Language Engineering means applying **scientific** principles to the design, construction and maintenance of **tools** to help deal with **information** that has been **expressed in natural languages** (the languages that people use for communicating with one another).

Applications

- Automatic summarization
- Co-reference resolution
- Discourse analysis (elaboration, explanation, contrast, question, statement, assertion)
- Machine translation
- Morphological segmentation
- Named entity recognition
- Natural language generation
- Natural language understanding
- Optical character recognition (OCR)
- Part-of-speech tagging
- Parsing
- Question answering
- Relationship extraction
- Sentiment analysis (Polarity)
- Speech recognition (Speech segmentation)
- Sentence breaking, Word segmentation, Topic segmentation
- Word sense disambiguation



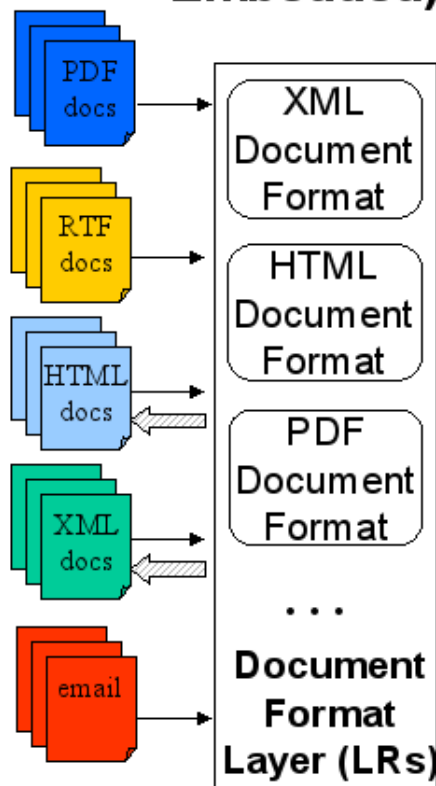
What is GATE?

- General Architecture for Text Engineering
- **Java** suite of NLP tools
- University of **Sheffield**
- Initial Release **1995** (17 years ago)
- Last Stable Release **6.1** May 6, **2011**
- **Languages** : English, Spanish, Chinese, Arabic, Bulgarian, French, German, Hindi, Italian, Cebuano, Romanian, Russian.
- Accepted Input **Formats** TXT, HTML, XML, Doc, PDF and Java Serial, PostgreSQL, Lucene, Oracle Databases
- **GATE Developer** which is a GATE graphical user interface, like **Eclipse** for Java programmers, provides a graphical environment for research and development of language processing software.



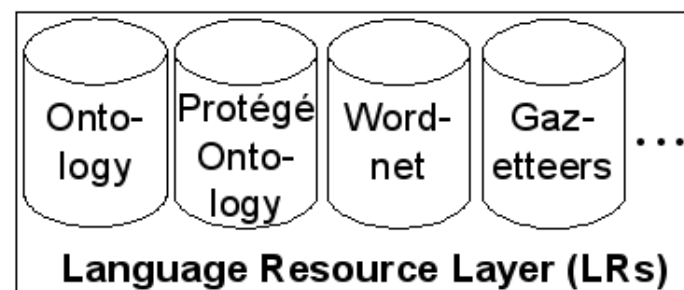
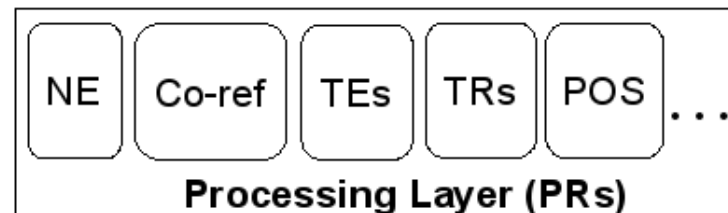
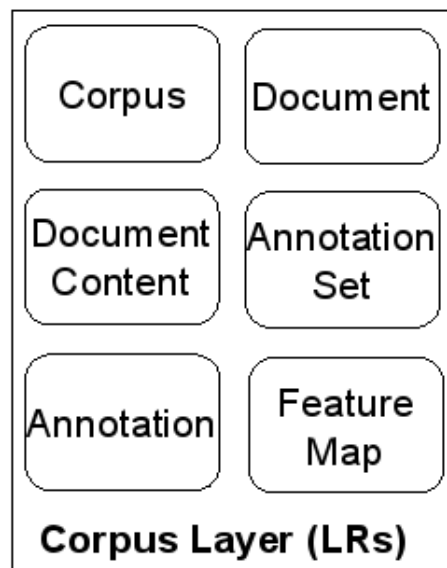
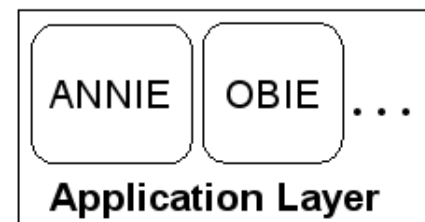
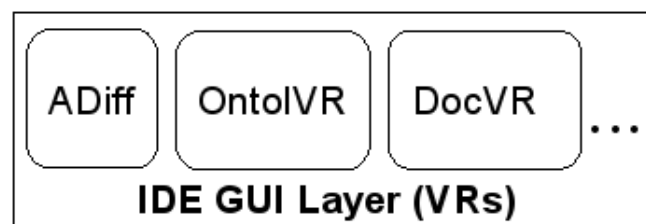
Gate Components and APIs

APIs (GATE Embedded)



NOTES

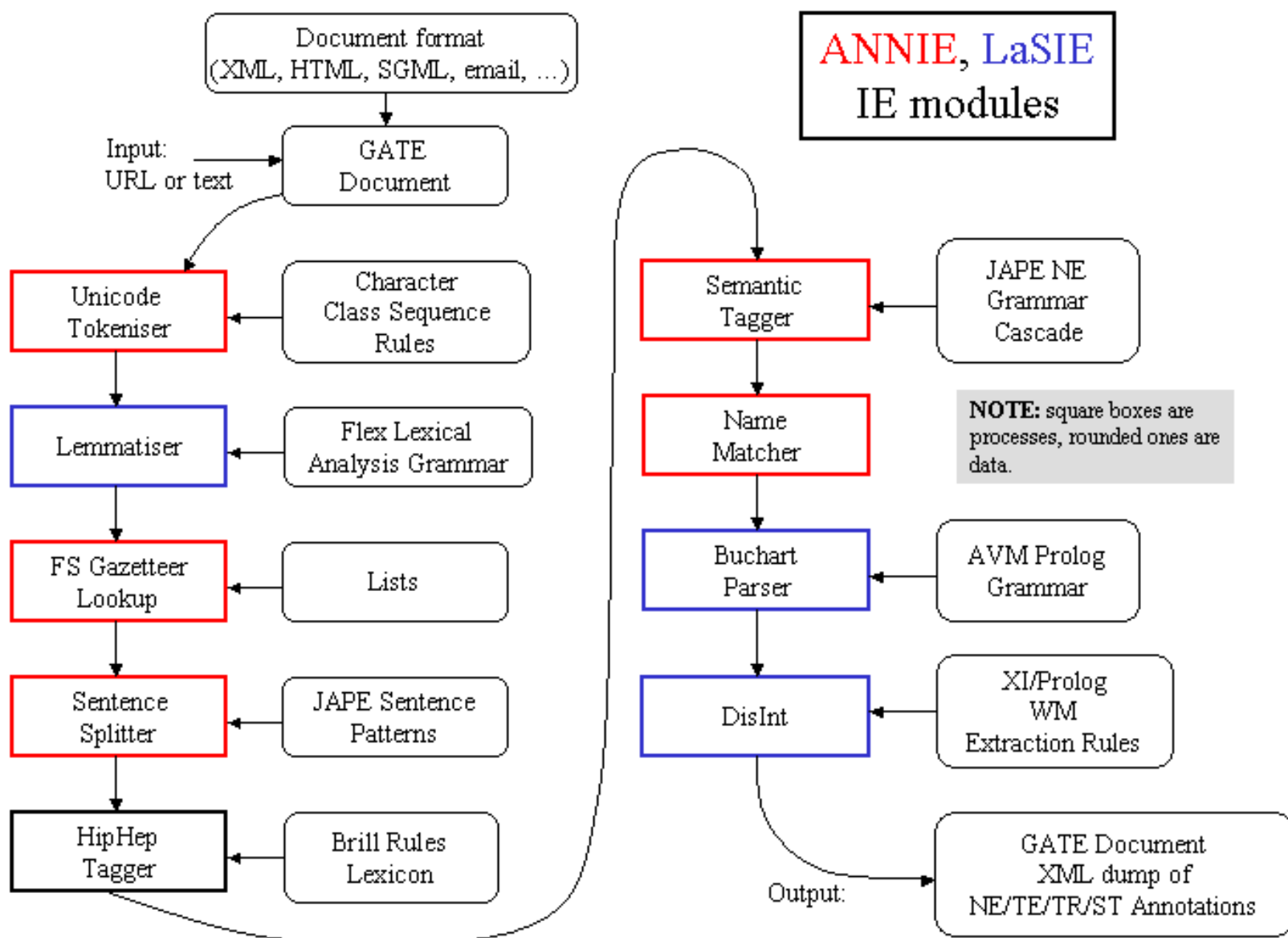
- everything is a replaceable bean
- all communication via fixed APIs
- low coupling, high modularity, high extensibility



ANNIE GATE Application

- A Nearly-New Information Extraction System
- Example Application for **English** Language Engineering
- A set of **modules**:
 - Tokenizer
 - Gazetteer
 - Sentence splitter
 - Part-of-speech tagger
 - Named entities transducer
 - Co-reference tagger.

ANNIE Architecture



- ANNIE Gazetteer: A list lookup component. The list files are located in \$GATE_HOME/plugins/ANNIE/resources/gazetteer
- JAPE Transducer: JAPE is a Java Annotation Patterns Engine. JAPE provides finite state transduction over annotations based on regular expressions. Example files are located in \$GATE_HOME/plugins/ANNIE/resources/NE
- ANNIE NE Transducer: (ANNIE named entity grammar) a semantic tagger based on the JAPE language.

- Provides indexing and searching the linguistic and semantic information generated by GATE

Demo

Installing Mimir

Install Grails

```
sudo add-apt-repository ppa:groovy-dev/grails  
sudo apt-get update  
sudo apt-get install grails-1.3.7
```

Get mimir

```
svn co https://gate.svn.sourceforge.net/svnroot/gate/mimir/trunk mimir
```

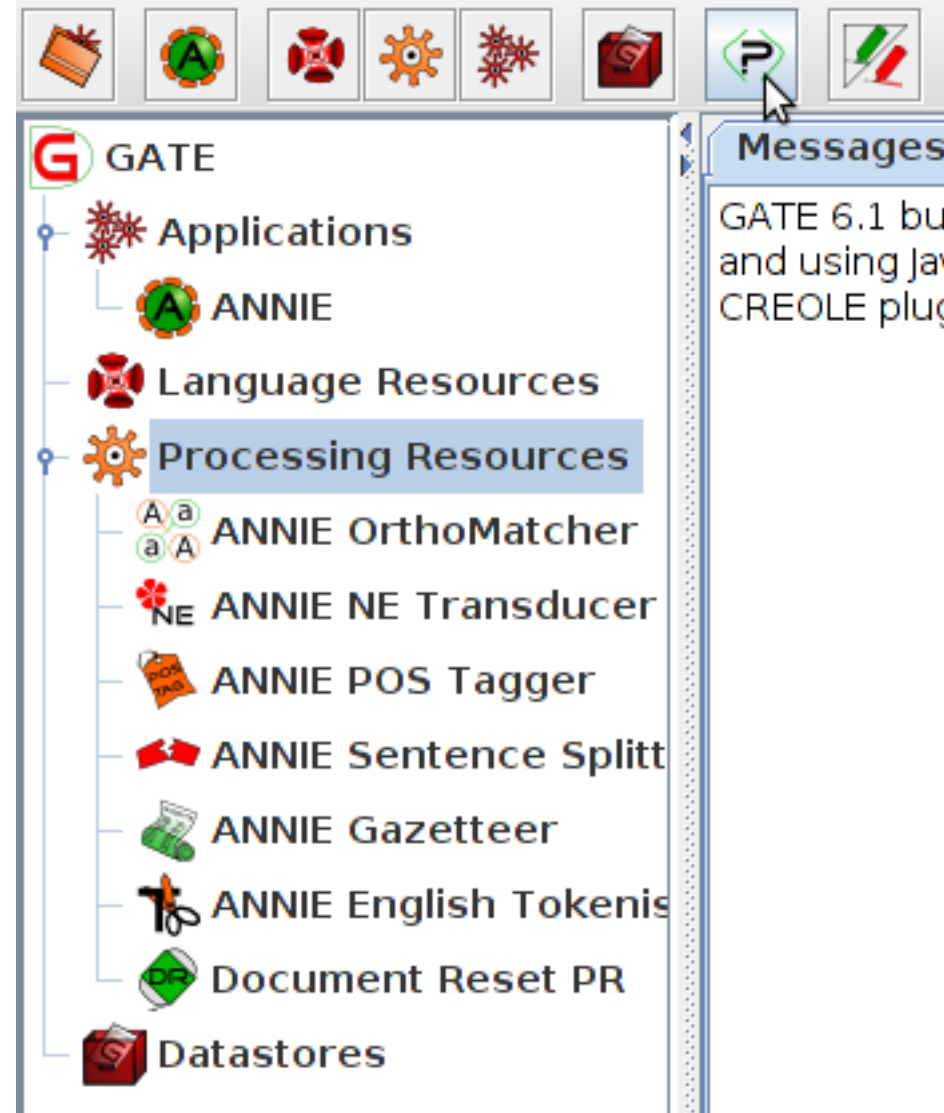
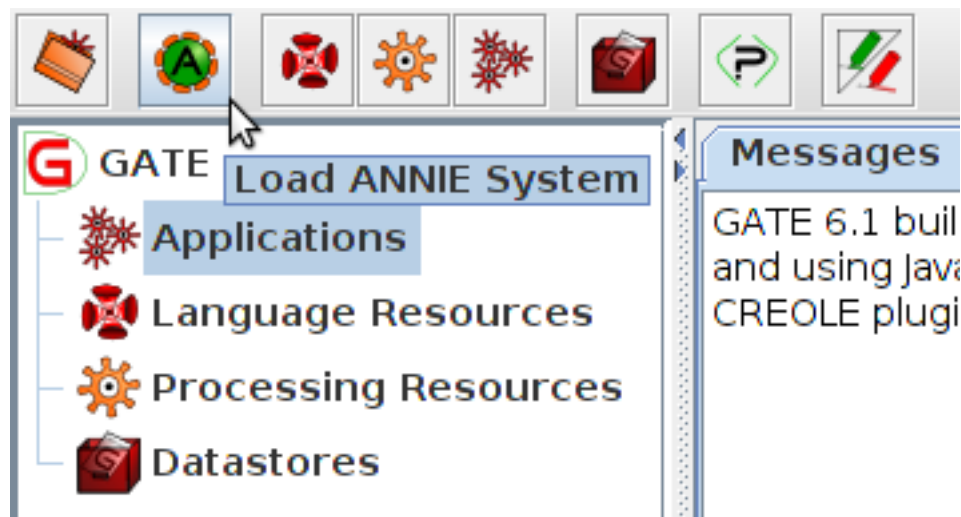
Build mimir

```
cd mimir  
ant
```

Build and start the mimir demo app

```
ant mimir-demo  
cd mimir-demo  
grails prod run-app
```

- Open GATE and Load ANNIE Systems with Defaults
- Then click the Manage CREOLE Plug-ins



Add Mimir Client Path

- Add Mimir as a Plugin and set mimir-client directory

The screenshot shows the 'Plugin Management Console' window. It contains a table of 'Known CREOLE directories' with columns for Name, URL, Load now, Load always, and Delete. A dialog box titled 'Enter an URL to the directory containing the' is open, showing a text input field with the path 'ne/magdy/Desktop/00/mimir-client/' and a 'Select a directory' button. The dialog also has 'OK' and 'Cancel' buttons. The background table lists various plugins like Alignment, ANNIE, Annotation_Merging, etc.

Name	URL	Load now	Load always	Delete
Alignment	file:/home/magdy/GATE-6.1/plugins/Alignment/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
ANNIE	file:/home/magdy/GATE-6.1/plugins/ANNIE/	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Annotation_Merging	file:/home/magdy/GATE-6.1/plugins/Annotation_Merging/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Copy_Annots_Between_Docs	file:/home/magdy/GATE-6.1/plugins/Copy_Annots_Between_Docs/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Gazetteer_LKB	file:/home/magdy/GATE-6.1/plugins/Gazetteer_LKB/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Gazetteer_Ontology_Based	file:/home/magdy/GATE-6.1/plugins/Gazetteer_Ontology_Based/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Groovy	file:/home/magdy/GATE-6.1/plugins/Groovy/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Information_Retrieval	file:/home/magdy/GATE-6.1/plugins/Information_Retrieval/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Inter_Annotator_Agreement	file:/home/magdy/GATE-6.1/plugins/Inter_Annotator_Agreement/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Jape_Compiler	file:/home/magdy/GATE-6.1/plugins/Jape_Compiler/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Keyphrase_Extraction_Algorithm	file:/home/magdy/GATE-6.1/plugins/Keyphrase_Extraction_Algorithm/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Arabic	file:/home/magdy/GATE-6.1/plugins/Lang_Arabic/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Cebuano	file:/home/magdy/GATE-6.1/plugins/Lang_Cebuano/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Chinese	file:/home/magdy/GATE-6.1/plugins/Lang_Chinese/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Hindi	file:/home/magdy/GATE-6.1/plugins/Lang_Hindi/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Romanian	file:/home/magdy/GATE-6.1/plugins/Lang_Romanian/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Language_Identification	file:/home/magdy/GATE-6.1/plugins/Language_Identification/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Learning	file:/home/magdy/GATE-6.1/plugins/Learning/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
LingPipe	file:/home/magdy/GATE-6.1/plugins/LingPipe/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

CREOLE resources in directory

- Compound Document
- Compound Document From Xml
- Compound Document Editor
- GATE Composite document
- Switch Member PR
- Delete Member PR
- Combine Members PR
- Segment Processing PR
- ExportAlignmentPR

+ Add a CREOLE repository

OK Cancel Help

- Make sure Mimir Plugin is loaded now and every time you open GATE

Known CREOLE directories Filter:

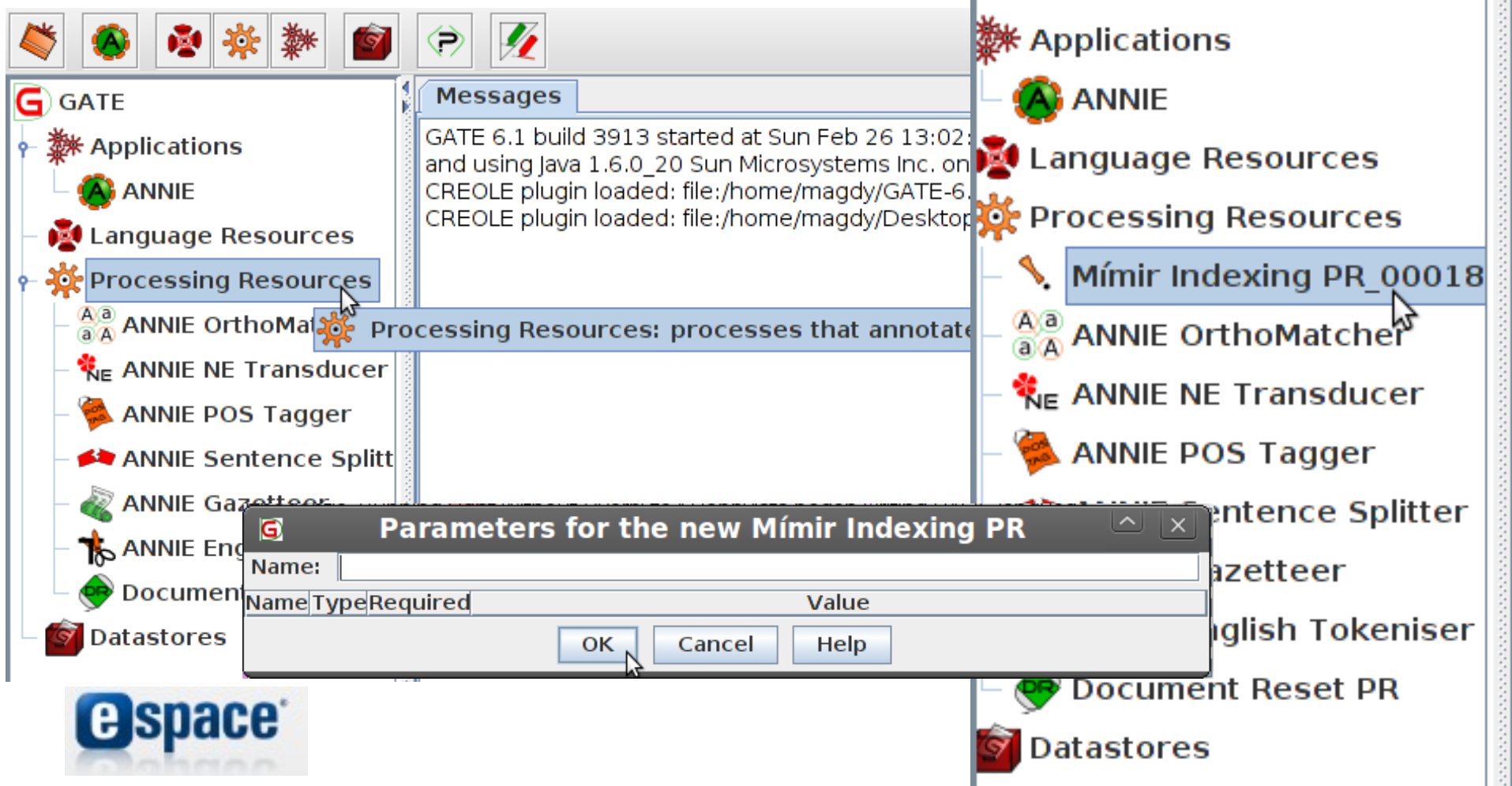
Name	URL	Load now	Load always	Delete
Annotation_Merging	file:/home/magdy/GATE-6.1/plugins/Annotation_Merging/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Copy_Annots_Between_Docs	file:/home/magdy/GATE-6.1/plugins/Copy_Annots_Between_Docs/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Gazetteer_LKB	file:/home/magdy/GATE-6.1/plugins/Gazetteer_LKB/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Gazetteer_Ontology_Based	file:/home/magdy/GATE-6.1/plugins/Gazetteer_Ontology_Based/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Groovy	file:/home/magdy/GATE-6.1/plugins/Groovy/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Information_Retrieval	file:/home/magdy/GATE-6.1/plugins/Information_Retrieval/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Inter_Annotator_Agreement	file:/home/magdy/GATE-6.1/plugins/Inter_Annotator_Agreement/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Jape_Compiler	file:/home/magdy/GATE-6.1/plugins/Jape_Compiler/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Keyphrase_Extraction_Algorithm	file:/home/magdy/GATE-6.1/plugins/Keyphrase_Extraction_Algorithm/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Arabic	file:/home/magdy/GATE-6.1/plugins/Lang_Arabic/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Cebuano	file:/home/magdy/GATE-6.1/plugins/Lang_Cebuano/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Chinese	file:/home/magdy/GATE-6.1/plugins/Lang_Chinese/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Hindi	file:/home/magdy/GATE-6.1/plugins/Lang_Hindi/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Lang_Romanian	file:/home/magdy/GATE-6.1/plugins/Lang_Romanian/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Language_Identification	file:/home/magdy/GATE-6.1/plugins/Language_Identification/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Learning	file:/home/magdy/GATE-6.1/plugins/Learning/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
LingPipe	file:/home/magdy/GATE-6.1/plugins/LingPipe/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Machine_Learning	file:/home/magdy/GATE-6.1/plugins/Machine_Learning/	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
mimir-client	file:/home/magdy/Desktop/4.0/mimir-client/	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

CREOLE resources in directory
Mimir Indexing PR

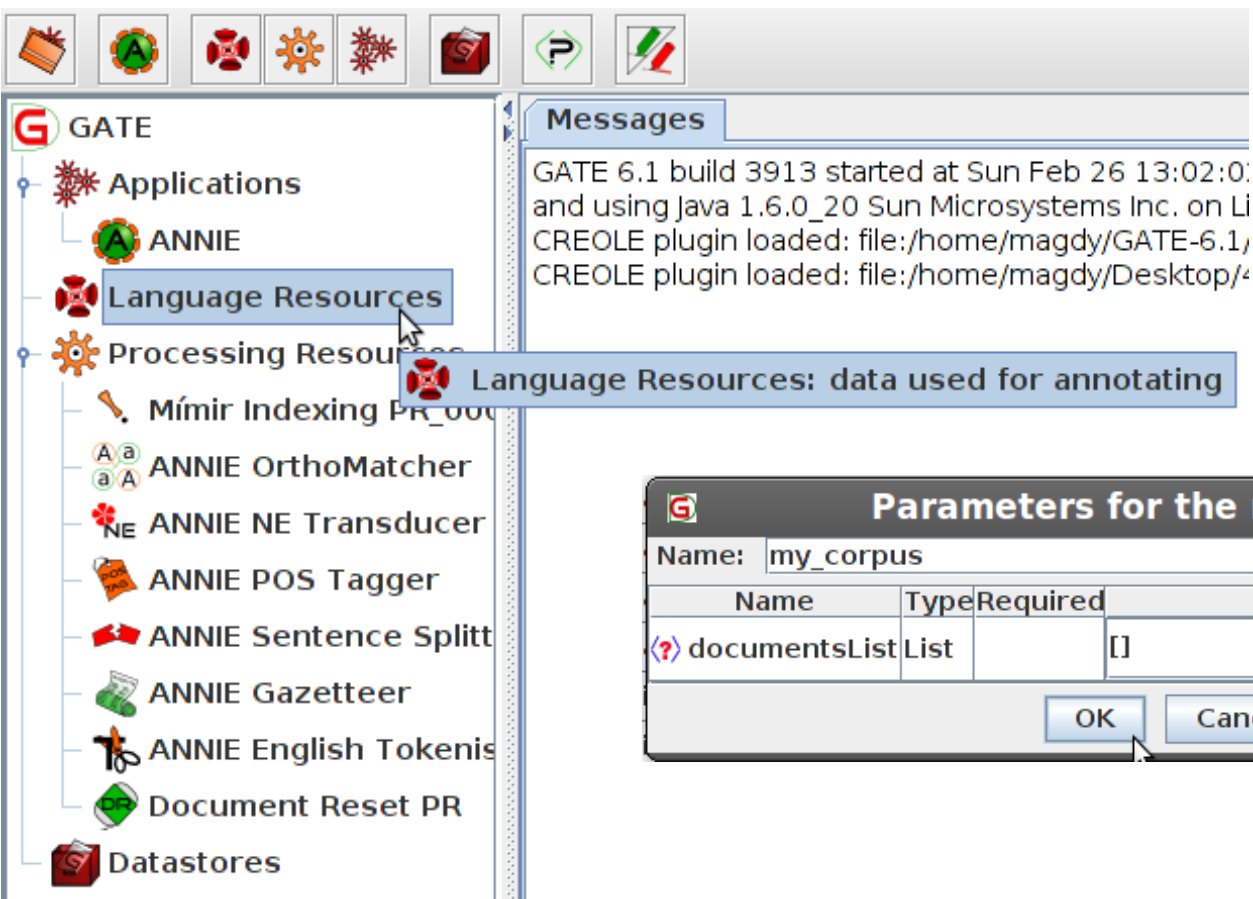
+ Add a CREOLE repository

OK Cancel Help

- Add Mimir Indexing PR to Processing Resources



- Create a New Corpus from Language Resource



GATE

- Applications
 - ANNIE
 - Language Resources**
 - Processing Resources
 - Mimir Indexing Process
 - ANNIE OrthoMatcher
 - ANNIE NE Transducer
 - ANNIE POS Tagger
 - ANNIE Sentence Splitter
 - ANNIE Gazetteer
 - ANNIE English Tokeniser
 - Document Reset PR
- Datastores

Messages

GATE 6.1 build 3913 started at Sun Feb 26 13:02:00 and using Java 1.6.0_20 Sun Microsystems Inc. on Linux
 CREOLE plugin loaded: file:/home/magdy/GATE-6.1/
 CREOLE plugin loaded: file:/home/magdy/Desktop/

Language Resources: data used for annotating

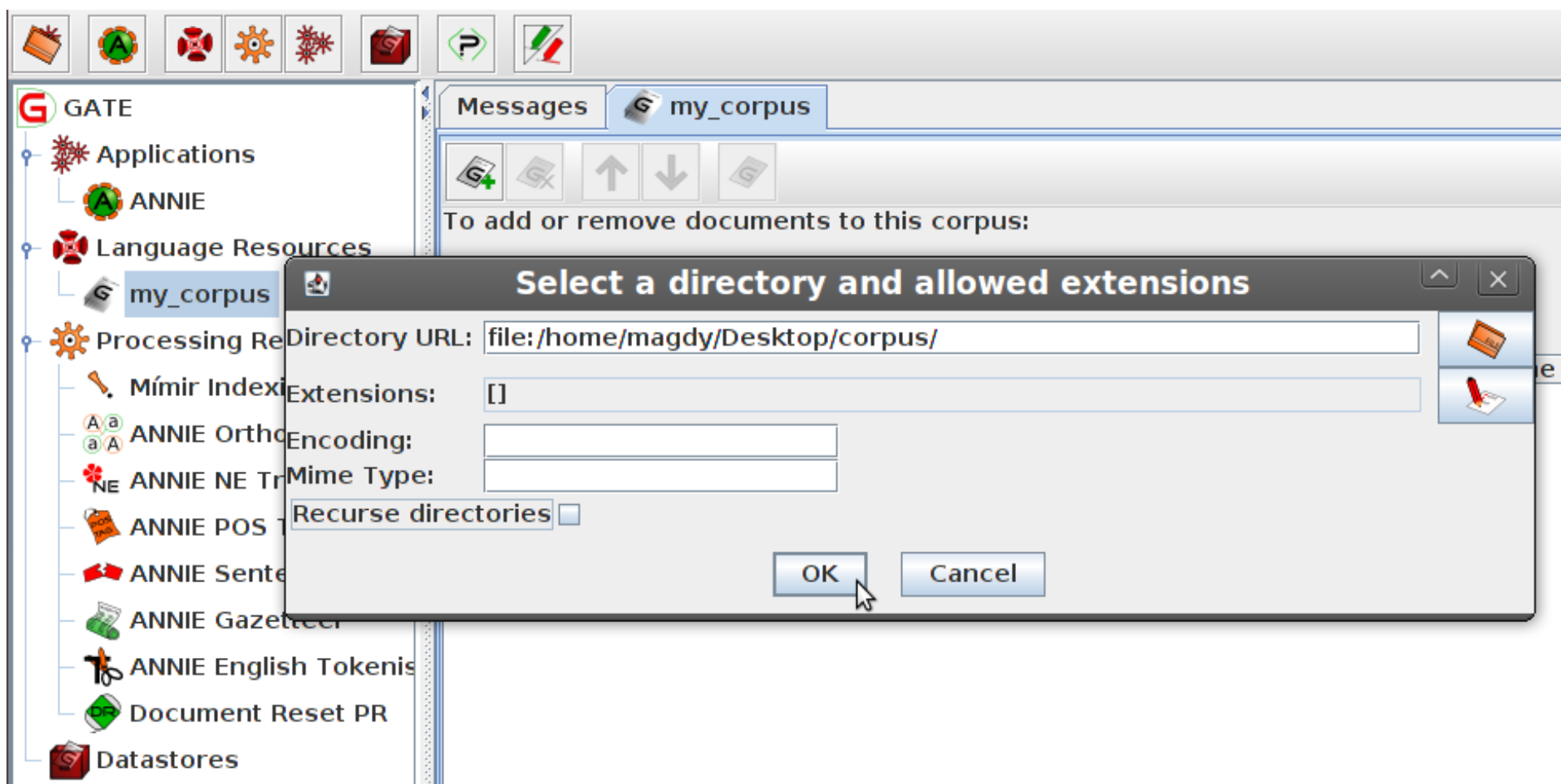
Parameters for the new GATE Corpus

Name:

Name	Type	Required	Value
documentsList	List	<input type="checkbox"/>	<input type="text" value="[]"/>

OK Cancel Help

- Right Click the Corpus and populate it with Documents



Edit the Default Index Template

- Open <http://localhost:8080/mimir-demo> in your browser and go to the configuration page
- Then go to the Index Templates section and manage them
- Then Click on the default Index Template to edit it.

Index Templates (?)

[Click here](#) to manage the index templates.

IndexTemplate List

Name	Comment
default	The default index configuration



Add some annotations to the Default Index Template

Name:	<input type="text" value="default"/>
Comment:	<input type="text" value="The default index configuratic"/>
Configuration:	<pre>import gate.creole.ANNIEConstants import gate.mimir.SemanticAnnotationHelper.Mode import gate.mimir.index.OriginalMarkupMetadataHelper import gate.mimir.db.DBSemanticAnnotationHelper as DefaultHelper tokenASName = "" tokenAnnotationType = ANNIEConstants.TOKEN_ANNOTATION_TYPE tokenFeatures = { string() category() root() } semanticASName = "" semanticAnnotations = { index { annotation helper:new DefaultHelper(annType:'Sentence') } index { annotation helper:new DefaultHelper(annType:'Person', nominalFeatures:["gender"]) annotation helper:new DefaultHelper(annType:'Location', nominalFeatures:["locType"]) annotation helper:new DefaultHelper(annType:'Organization', nominalFeatures:["orgType"]) annotation helper:new DefaultHelper(annType:'Date', integerFeatures:["normalized"]) annotation helper:new DefaultHelper(annType:'Document', integerFeatures:["date"], mode:Mode.DOCUMENT) annotation helper:new DefaultHelper(annType:'Lookup', nominalFeatures:["majorType"]) annotation helper:new DefaultHelper(annType:'JobTitle', nominalFeatures:[]) annotation helper:new DefaultHelper(annType:'skill', nominalFeatures:[]) annotation helper:new DefaultHelper(annType:'Address', nominalFeatures:["kind"]) } } documentRenderer = new OriginalMarkupMetadataHelper() documentMetadataHelpers = [documentRenderer]</pre>

Add a new Index

Local Indexes (?)

There are no local indexes configured in this Mimir instance.

You can [create a new local index](#), or [import an existing index for searching](#).

Create Local Index

Name:

Index template:

Document URIs are external links: ☐



Create

Edit the Index you created and set the Scorer Algorithm

(1) Local Indexes (?)

The following local indexes are configured:

1. [my_corpus_index](#)

You can [create a new local index](#), or [import an existing index for searching](#).

Mimir index "my_corpus_index"

Index Name: my_corpus_index
Index URL: http://localhost:8080/mimir-demo/90665bdd-9b17-4db4-89b3-2fde933a447e
State: indexing
Annotations indexed: [Detail...](#)

Annotation type Features

Sentence	<none>
Person	gender
Location	locType
Organization	orgType
Date	normalized
Document	date
Lookup	majorType
JobTitle	<none>
skill	<none>
Address	kind

Scorer: No Scoring

[Details](#) [Edit](#) [Delete](#) [Close](#)



(2)

Edit LocalIndex

Index ID: 90665bdd-9b17-4db4-89b3-2fde933a447e
Name: my_corpus_index
State: indexing
Index Directory: /home/magdy/Desktop/4.0/mimir-indexes
Scorer: BM25
Document URIs are external links: ☐

[Save](#) [Delete](#)

(3)

Copy the Index URL

Mimir index "my_corpus_index"

Index Name: my_corpus_index

Index URL: <http://localhost:8080/mimir-demo/90665bdd-9b17-4db4-89b3-2fde933a447e>

State: indexing

Annotations indexed: [Detail...](#)

Scorer: BM25

 Details  Edit  Delete  Close

Paste Index URL in Mimir and Run ANNIE on the Corpus

File Options Tools Help

GATE

- Applications
 - ANNIE
- Language Resources
- Processing Resources
 - Mimir Indexing PR_00018
 - ANNIE OrthoMatcher
 - ANNIE NE Transducer
 - ANNIE POS Tagger
 - ANNIE Sentence Splitter
 - ANNIE Gazetteer
 - ANNIE English Tokeniser
 - Document Reset PR
- Datastores

Messages | **my_corpus** | **ANNIE**

Loaded Processing resources

Name	Type
------	------

Selected Processing resources

!	Name	Type
	Document Reset PR	Document Reset PR
	ANNIE English Tokeniser	ANNIE English Tokeniser
	ANNIE Gazetteer	ANNIE Gazetteer
	ANNIE Sentence Splitter	ANNIE Sentence Splitter
	ANNIE POS Tagger	ANNIE POS Tagger
	ANNIE NE Transducer	ANNIE NE Transducer
	ANNIE OrthoMatcher	ANNIE OrthoMatcher
	Mimir Indexing PR_00018	Mimir Indexing PR

Corpus: **my_corpus**

Runtime Parameters for the "Mimir Indexing PR_00018" Mimir Indexing PR:

Name	Type	Required	Value
mimirIndexUrl	URL	✓	http://localhost:8080/mimir-demo/90665bdd-9b17-4db4-89b3-2fde933a447e
mimirPassword	String		
mimirUsername	String		

Run this Application

Serial Application Editor | Initialisation Parameters | **Run this application F3**

Double click any document and check Annotations yourself

The screenshot displays the eSpace application window. On the left is a 'Language Resources' sidebar with a list of HTML documents, including 'Cotonti.html_0007E' and 'Copleft.html_0007I'. The main window has a tab bar at the top with 'Messages', 'my_corpus', 'ANNIE', and 'Copleft.html_0...'. Below the tabs are buttons for 'Annotation Sets', 'Annotations List', 'Annotations Stack', 'Co-reference Editor', 'Text', and a search icon. The central text area shows a document with various annotations: 'effective. In particular, permissive free software licenses such as BSD allow re-distributors to remove some or all these rights, and do not require the distribution of source code. [edit] History The use of "Copyleft; All Wrongs Reserved" in 1976 An early example of copyleft was the Tiny BASIC project started in the newsletter of the People's Computer Company in 1975. Dennis Allison wrote a specification for a simple version of the BASIC programming language. [4] This design did not support text strings and only used integer arithmetic. The goal was for the program to fit in 2 to 3 kilobytes of memory. The Tiny BASIC contents of the newsletter soon became Dr. Dobb's Journal of Tiny BASIC with a subtitle of "Calisthenics & Orthodontia, Running Light Without Overbyte." Hobbyists began writing BASIC language interpreters for their microprocessor-based home computers and sending the source code to Dr. Dobb's Journal and other magazines to be published. By the middle of 1976, Tiny BASIC interpreters were available for the Intel 8080, the Motorola 6800 and MOS Technology 6502 processors. This was a free software project before the internet allowed easy transfer of files. Computer hobbyists would exchange paper tapes, cassettes or even retype the files from the printed listings. [5] Jim Warren, editor of Dr. Dobb's Journal, wrote in the July 1976 ACM Programming Language newsletter about the motivations and methods of this successful project. He started with this: "There is a viable alternative to the problems raised by Bill Gates in his irate letter to computer hobbyists concerning 'ripping off' software. When software is free, or so inexpensive that it's easier to pay for it than to duplicate it, then it won't be 'stolen'." The method was to have an experienced professional do the overall design and then outline an implementation strategy. Knowledgeable amateurs would implement the design for a variety of computer systems. Warren predicted this strategy would be continued and expanded. [5] The May 1976 issue of Dr. Dobbs Journal had Li-Chen Wang's Palo Alto Tiny BASIC for the Intel 8080 microprocessor. The listing began with the usual title, author's name and date but it also had "@COPYLEFT ALL WRONGS RESERVED". [6] A fellow Homebrew Computer Club member, Roger Rauskolb, modified and improved Li-Chen Wang's program and this was published in the December 1976 issue of Interface Age magazine. [7] Roger added his name and preserved the COPYLEFT Notice. A later instance of copyleft arose when Richard Stallman was working on a Lisp interpreter. Symbolics asked to use the Lisp interpreter, and Stallman agreed to supply them with a public domain version of his work. Symbolics extended and improved the Lisp interpreter, but when Stallman wanted access to the improvements that Symbolics had made to his interpreter, Symbolics refused. Stallman then, in 1984, proceeded to work towards eradicating this emerging behavior and culture of proprietary software, which

On the right side of the main window is a vertical list of annotation categories with checkboxes: Address, DEFAULT_TOKEN, Date, FirstPerson, Identifier, JobTitle, Location, Lookup, Money, Organization, Person, Sentence, SpaceToken, Split, Temp, TempDate, TempYear, Title, Token, Unknown, UriPre, skill, and Original markups. Below this list is a table with columns for 'MatchesAnnots' and 'MimeType'. The 'MatchesAnnots' column contains '{null=[[' and the 'MimeType' column contains 'text/html'. Below the table is a 'gate.SourceURL' field with the value 'file:/home'.

Close and Search the Index

Mimir index "my_corpus_index"

(1)

Index Name: my_corpus_index
Index URL: http://localhost:8080/mimir-demo/90665bdd-9b17-4db4-89b3-2fde933a447e
State: indexing
Annotations indexed: [Detail...](#)
Scorer: BM25

 Details  Edit  Delete  Close

(2)

Mimir index "my_corpus_index"

Index Name: my_corpus_index
Index URL: http://localhost:8080/mimir-demo/90665bdd-9b17-4db4-89b3-2fde933a447e
State: searching
Annotations indexed: [Detail...](#)
Scorer: BM25

[Search this index using the web interface.](#)
[Search this index using the XML service interface.](#)
[Manage deleted documents.](#)

 Details  Edit  Delete

Example Query

Searching Index "my_corpus_index"

{Person}
of
{Location}

Search

Documents 1 to 2 of 2:

[Canada.html_0005B](#)

General the Viscount **Alexander of Tunis** (centre) commissioned by the **Lieutenant Governor of Quebec** , the Honourable Federal Publications (**Queen of Canada**) . http ...

[Canada.html_0005B](#)

General the Viscount **Alexander of Tunis** (centre) commissioned by the **Lieutenant Governor of Quebec** , the Honourable Federal Publications (**Queen of Canada**) . http ...



Thank you

References

GATE Website (it is huge)

[**http://gate.ac.uk**](http://gate.ac.uk)

Mother of all Knowledge

[**http://www.wikipedia.com**](http://www.wikipedia.com)