

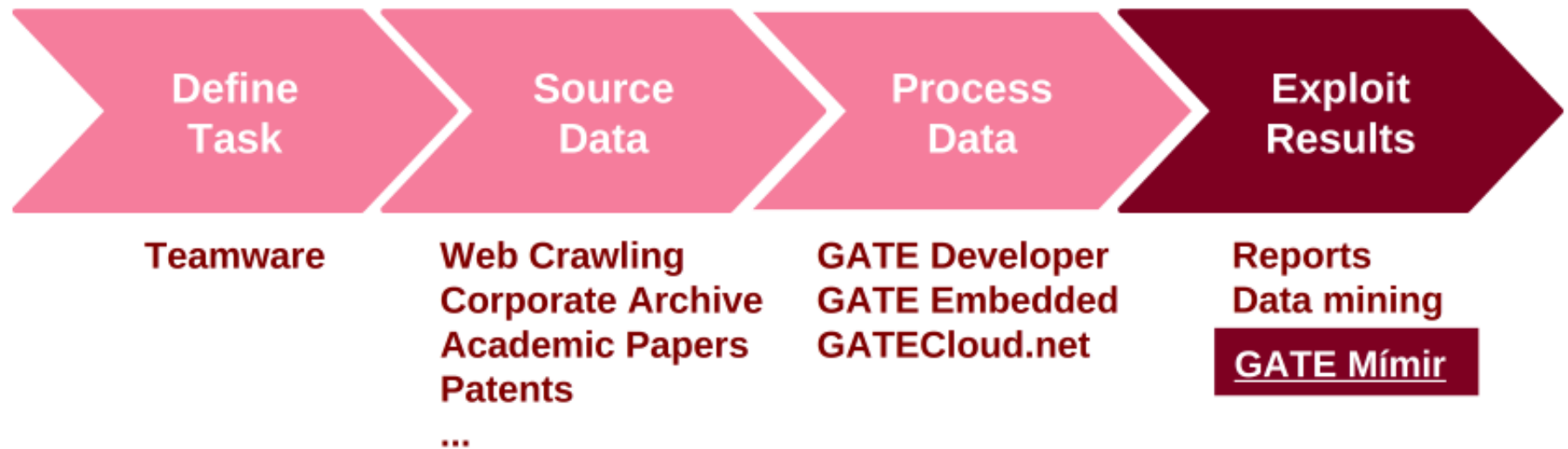


Module 4: Taking GATE to the Cloud

Part III: GATE Mímir Multi-paradigm Indexing

Text Mining Life Cycle

GATE

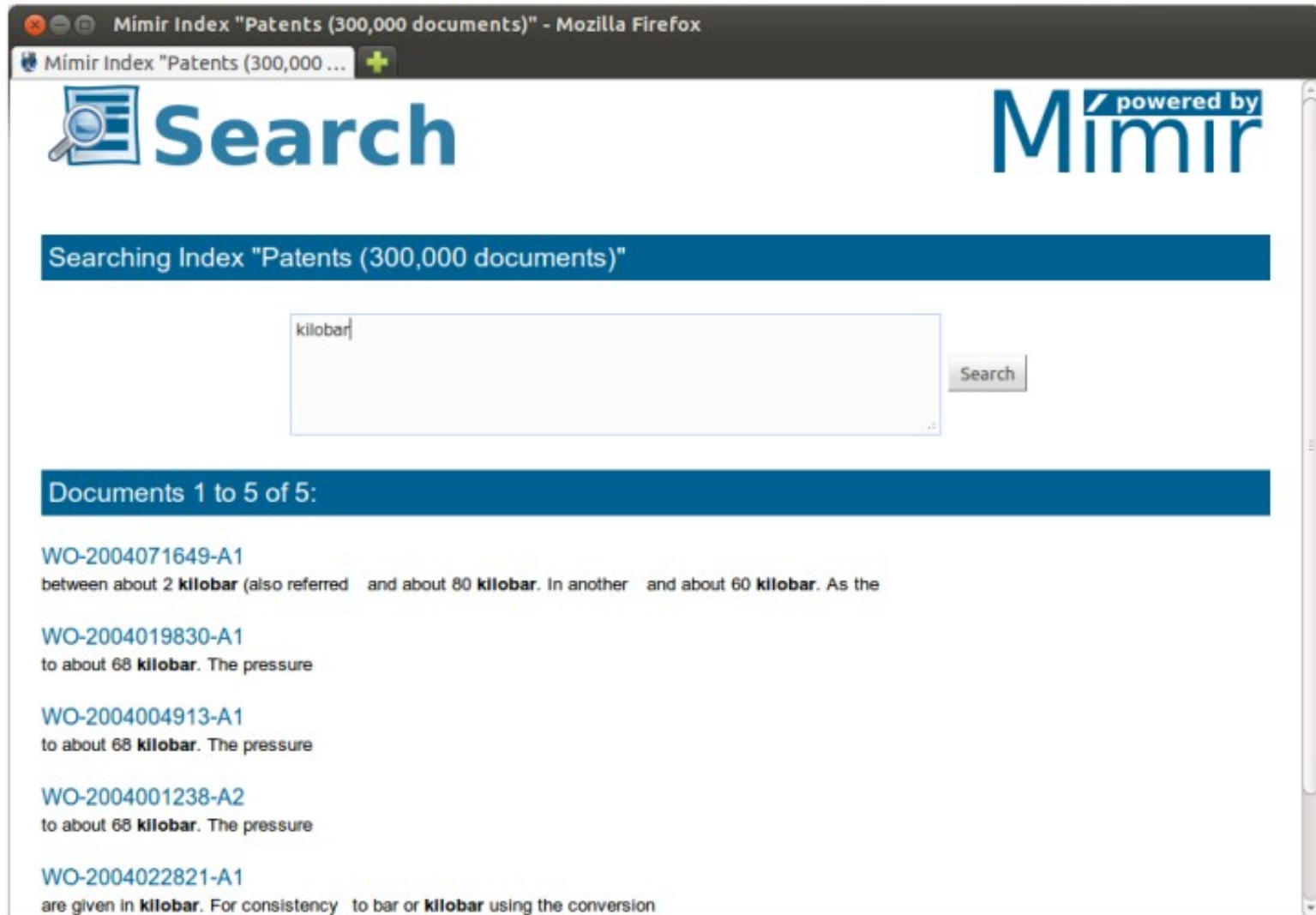


- GATE Mímir is an indexing system for GATE documents.
- Mímir can index:
 - Text: the original document content is indexed (based on Token annotations)
 - Annotations: annotations and features
 - Semantics: annotations can be linked to external ontologies which can be used at search time
- Mímir queries allow for any combination of these

- Standard search covers the text.
- GATE documents also have annotations, which give access to the document's:
 - Structure (sections, titles, etc.)
 - Semantics
 - Linguistic features (nouns, verbs, etc.)
 - Etc.
- Annotations are data
 - Computers like data, people not so much



Document text

GATE



Mimir Index "Patents (300,000 documents)" - Mozilla Firefox

Mimir Index "Patents (300,000 ..."

 **Search**  powered by Mimir

Searching Index "Patents (300,000 documents)"

Documents 1 to 5 of 5:

[WO-2004071649-A1](#)
between about 2 **kilobar** (also referred and about 80 **kilobar**. In another and about 60 **kilobar**. As the

[WO-2004019830-A1](#)
to about 68 **kilobar**. The pressure

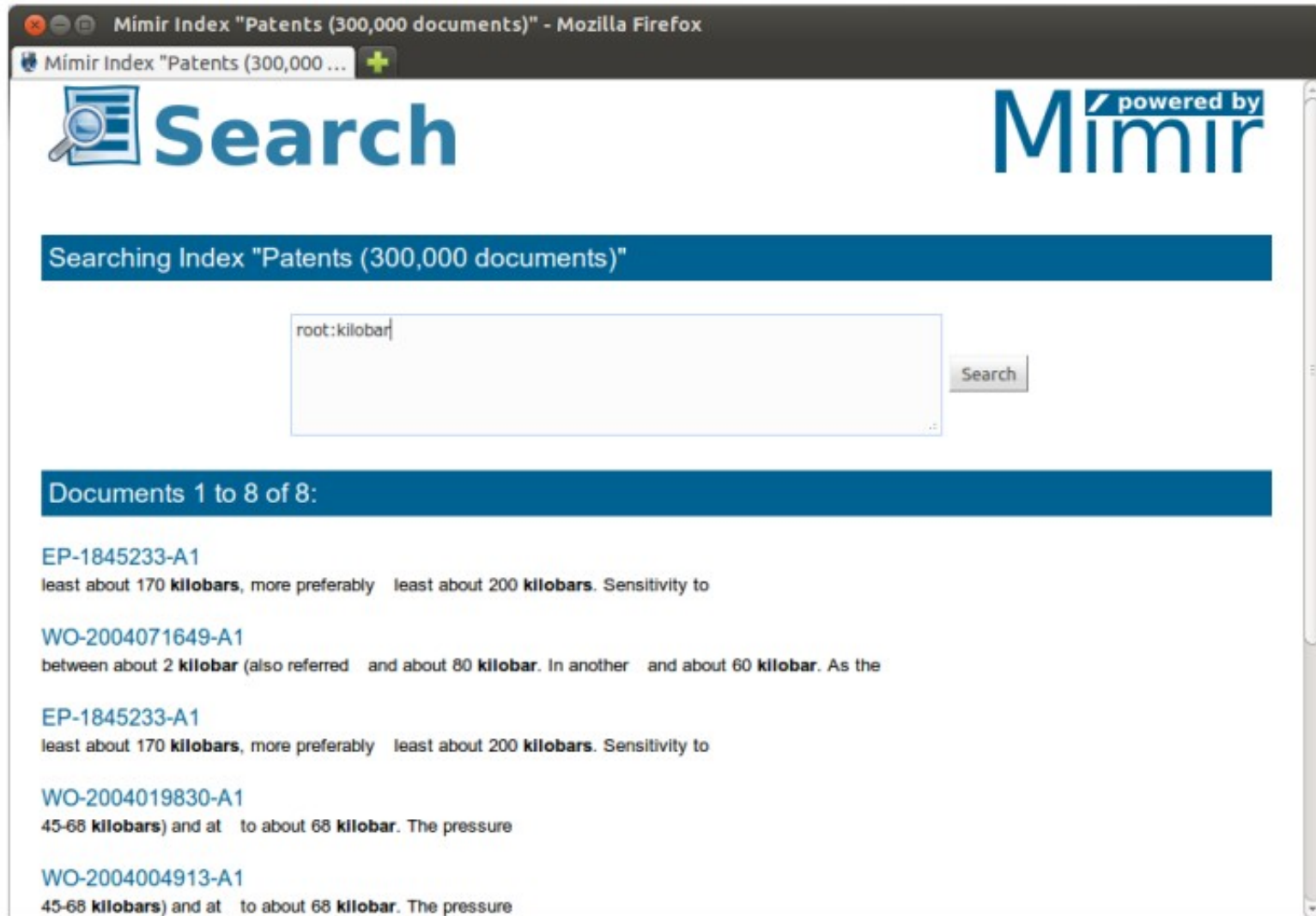
[WO-2004004913-A1](#)
to about 68 **kilobar**. The pressure

[WO-2004001238-A2](#)
to about 68 **kilobar**. The pressure

[WO-2004022821-A1](#)
are given in **kilobar**. For consistency to bar or **kilobar** using the conversion

Morphology: Root Form

GATE



The screenshot shows a web browser window titled "Mimir Index 'Patents (300,000 documents)' - Mozilla Firefox". The address bar shows "Mimir Index 'Patents (300,000 ...)". The page features a search bar with the text "root:kilobar" and a "Search" button. Below the search bar, a blue banner indicates "Searching Index 'Patents (300,000 documents)'". The results section is titled "Documents 1 to 8 of 8:" and lists five patent entries:



- [EP-1845233-A1](#)
least about 170 **kilobars**, more preferably least about 200 **kilobars**. Sensitivity to
- [WO-2004071649-A1](#)
between about 2 **kilobar** (also referred and about 80 **kilobar**. In another and about 60 **kilobar**. As the
- [EP-1845233-A1](#)
least about 170 **kilobars**, more preferably least about 200 **kilobars**. Sensitivity to
- [WO-2004019830-A1](#)
45-68 **kilobars**) and at to about 68 **kilobar**. The pressure
- [WO-2004004913-A1](#)
45-68 **kilobars**) and at to about 68 **kilobar**. The pressure

Morphology: Part of Speech

GATE

Mimir Index "Patents (300,000 documents)" - Mozilla Firefox

Mimir Index "Patents (300,000 ..."

 **Search**  powered by Mimir

Searching Index "Patents (300,000 documents)"

category:JJ

Documents 1 to 20 of 299967:

[EP-1793116-A2](#)
inlet 12 , **connectable** to a source a source of **pressurised** fuel, such pressurised fuel, **such** as a fuel ...

[EP-1793158-A2](#)
YU A replacement **light** bulb assembly (in an incandescent **light** fixture utilizes a fixture utilizes a **light** emitting diode (...

[EP-1492320-A9](#)
product represents a **comprehensive** real-time solution for represents a comprehensive **real-time** solution for rating solution for rating **incoming** and outgoing MMS ...


[EP-1793109-A1](#)
method and a **corresponding** apparatus for controlling measured data, **such** a data drift calibration of air **low** sensor. The ...

Document Structure

GATE

Mimir Index "Patents (300,000 documents)" - Mozilla Firefox

Mimir Index "Patents (300,000 ..."

 **Search** powered by **Mimir**

Searching Index "Patents (300,000 documents)"

{Abstract}

Documents 1 to 20 of 99994:

[EP-1826232-A1](#)
SK TR YU **A plastic component for household appliances, in p ... as well as chemical, wear and thermal resistance.** The present invention

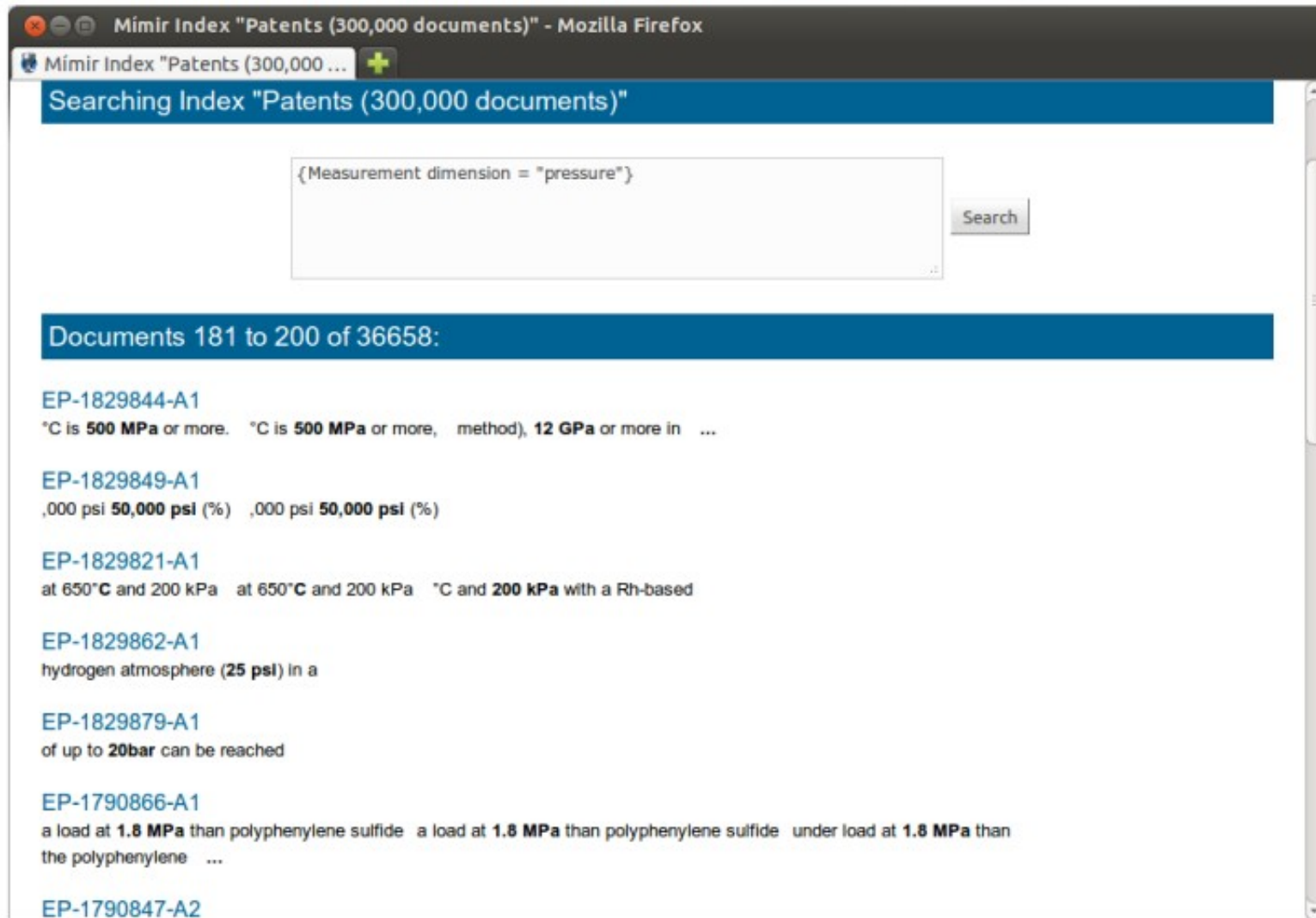
[EP-1826275-A1](#)
SK TR YU **The method of fast microbiological analysis of a s ... some ATP when the unwanted ATP has been consumed.** At present,

[EP-1826264-A1](#)
SK TR YU **The present invention is directed to an animal mod ... al therapeutic or an anti-orthopox virus vaccine.** The present invention

[EP-1826218-A1](#)
SI SK TR **An antibody against a peptide having an amino acid ... ting effect on BAFF with the use of the antibody.**

Semantics: Pressure Measurements

GATE



Mimir Index "Patents (300,000 documents)" - Mozilla Firefox

Mimir Index "Patents (300,000 documents)"

Searching Index "Patents (300,000 documents)"

{Measurement dimension = "pressure"}

Search

Documents 181 to 200 of 36658:

[EP-1829844-A1](#)
°C is **500 MPa** or more. °C is **500 MPa** or more, method), **12 GPa** or more in ...

[EP-1829849-A1](#)
,000 psi **50,000 psi** (%) ,000 psi **50,000 psi** (%)

[EP-1829821-A1](#)
at 650°C and 200 kPa at 650°C and 200 kPa °C and **200 kPa** with a Rh-based

[EP-1829862-A1](#)
hydrogen atmosphere (**25 psi**) in a

[EP-1829879-A1](#)
of up to **20bar** can be reached

[EP-1790866-A1](#)
a load at **1.8 MPa** than polyphenylene sulfide a load at **1.8 MPa** than polyphenylene sulfide under load at **1.8 MPa** than the polyphenylene ...


[EP-1790847-A2](#)

Bring it all Together

GATE

Mimir Index "Patents (300,000 documents)" - Mozilla Firefox

Mimir Index "Patents (300,000 ..."

 **Search** powered by Mimir

Searching Index "Patents (300,000 documents)"

`k
category::J [0..3] {Measurement dimension = "pressure"}
) IN {Abstract}`

Documents 1 to 20 of 82:

[EP-1829912-A1](#)
maintain at an **arbitrary pressure from 0.1 to 100 MPa** maintain at an **arbitrary pressure from 0.1 to 100 MPa**
maintain at an **arbitrary pressure from 0.1 to 100 MPa** (absolute pressure ...

[EP-1826177-A2](#)
pressures of as **high as around 1,000 bar**), pipeline pressures of as **high as around 1,000 bar**), pipeline

[EP-1821289-A2](#)
stress of the **plated film is 400 MPa** or more.

[EP-1840990-A1](#)
(a detected **partial pressure of 1.0x10-12 Pa**) of temperature (a detected **partial pressure of 1.0x10-12 Pa**) of temperature

Document tokens

- Syntax:

- *Text*, *root:text*, *category:CAT*

- Example:

- *kilobar*

- *root:be*

- *category:JJ*

Document Annotation

- Syntax:

- {Annotation},

- {Annotation feature1=v1 f2=v2 ... }

- Example:

- {Abstract}

Sequence

- Syntax:

- (QUERY) (QUERY) (QUERY) ...

- (QUERY) [n..m] (QUERY)...

- Example:

- {Person} [0..3] {Money}

Boolean operators

- Syntax:

- (QUERY) **AND** (QUERY)

- (QUERY) **OR** (QUERY)

- Example:

- {Person} **AND** {Money}

Containment

- Syntax:

- (QUERY) **IN** (QUERY)

- (QUERY) **OVER** (QUERY)

- Example:

{Organization} AND {Money} IN {Sentence}

Hands-on: Search the BBC News Demo Index

- Go to: <http://demos.gate.ac.uk/mimir/>
- Open the BBC News Demo
- Find:
 - Document titles
 - Date expressions
 - Volume measurements
 - Amounts of money being paid
 - Amounts of money being received

- Measurement feature 'dimension' can be time, length, volume, mass, etc.
- Use **root:word** to find all forms of a word (e.g. root:pay)
- Search for co-occurrence within a sentence to reduce noise.

Building an Index: Requirements

- Mimir server instance:
 - On GATECloud.net
 - Build your own from sources
- Index template
- A source of annotated GATE documents
- A method for pushing documents to the index:
 - Mimir Client PR, running inside Developer
 - Mimir output handler in GATE Cloud Paralleliser
 - Mimir output handler on GATECloud.net

Specifies which:

- Token features should be indexed
- (semantic) annotation types should be indexed
- features for each annotation type
- features of the document

... should be indexed.

Hands-on: Build a Mimir Index

- Rent a Mimir server on GATECloud.net
- Navigate to your Dashboard page →
Reservation
- Start the server, wait for start-up notification.
- Navigate to your new server
- Create a new index template
 - Give it a name
 - Use content of *index-template.groovy* for the configuration text.

Hands-on: Build a Mimir Index

-
- Create a new **local index**, using the **above template**
 - Go back to the admin page, and open the page for the new index.
 - Make a note of the **Index URL**

Hands-on: Build a Mimir Index

- Buy a new Custom Annotation Job
 - Application file: [annie-plus-morph-application.zip](#)
 - Input: [news-corpus-large.zip](#)
 - Mime type: **text/html**, Encoding: **UTF-8**
- Set one output to MIMIR, using the **Index URL** from above.
 - Make sure to not include any spaces in the IndexURL
- Run the job.
- **When finished**, close the index.

Hands-on: Explore the New Index

The logo for GATE, consisting of the word "GATE" in red, uppercase letters, enclosed within a green, rounded rectangular border.

- Find stock movements;
- Find people working for organizations;
- Find mergers;
- Find payments;
- Etc.

Simple stock movement example:

```
{Organization} (up | down) ({Money} | {Percent})
```

Backup index:

<http://demos.gate.ac.uk/mimir/fig/search/index>

Custom Mimir Interfaces

GATE

Mimir supports an XML based RESTful web service which can easily be used to provide custom interfaces which hide the query complexity from users.

<http://demos.gate.ac.uk/pin/>

People in the News - Chromium

People in the News

demos.gate.ac.uk/pin/?name=&bornIn=Sheffield&FamousAs=Politician

PEOPLE IN THE NEWS

Looking For...

Name:

Fuzzy Name Matching

Born In:

Famous As:

In Articles...

Published Between and

Classified As:

Ignore Boilerplate Text

Search

Results 1 to 1 of 1 [Show Underlying Mimir Query](#)

[Scottish election: Respect Coalition Against Cuts profile](http://www.bbc.co.uk/news/uk-scotland-13048761)
<http://www.bbc.co.uk/news/uk-scotland-13048761> - [Cached](#)

Contains 2 matches:

... Bow - whose sitting MP **Oona King** had voted for the war ...

... success came when Galloway overturned **Oona King's** 10,000- ...

Powered by GATE Mimir
© The University of Sheffield, 2011-2012

Hands-on: Clean Up

- Stop your Mimir server
- Destroy the reservation.

Questions?

The logo for GATE, featuring the word "GATE" in red, bold, uppercase letters inside a green rounded rectangular border.

<http://gate.ac.uk>

<http://gatecloud.net>