

---

# GATE Mimir and cloud services

---

Multi-paradigm indexing and search tool

Pay-as-you-go large-scale annotation

---



# GATE Mimir

- GATE Mimir is an indexing system for GATE documents.
- Mimir 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
- Mimir queries allow for any combination of these



# Why Mimir?

---

- Standard search covers the text.
- GATE documents also have annotations, which give access to the document's:
  - Structure (sections, titles, etc.)
  - Linguistic features (nouns, verbs, etc.)
  - Semantics
  - Etc.
- Examples: <http://demos.gate.ac.uk/mimir>
  - BBC News demo



# Mimir query language

---

- Simplest queries are free text
  - *text*
  - “quoted string”
- Searching against the document text



# Plain text queries

Harriet Harman

Search

Documents 1 to 20 of 81:

## [UK childcare needs to be more affordable - CentreForum \(cached\)](#)

quality' MP **Harriet Harman** was the architect

## [Birth weight among social mobility checks - Nick Clegg \(cached\)](#)

's deputy leader **Harriet Harman** said Mr Clegg

## [Ed Miliband's shadow cabinet and ministerial teams \(cached\)](#)

Miliband Opposition leader **Harriet Harman** Deputy Leader & in 2011. **HARRIET HARMAN** - DEPUTY LEADER

## [PM's response to Skinner Commons question 'shameful' \(cached\)](#)

. Deputy leader **Harriet Harman** wrote on Twitter

## [Daily Politics and Sunday Politics highlights of 2012 \(cached\)](#)

Sunday April 29 **Harriet Harman** on Hunt, Cameron and Clegg **Harriet Harman** struggles with bank on health by **Harriet Harman**  
PMQs: Harriet ...

## [Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance' \(cached\)](#)

Shadow Culture Secretary **Harriet Harman** says Jeremy Hunt culture secretary, **Harriet Harman**, told the

## [No Rupert Murdoch deal, says Alastair Campbell \(cached\)](#)



# Plain text queries

Harriet Harman says

Search

Documents 1 to 20 of 29:

**Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance'** (cached)

Shadow Culture Secretary **Harriet Harman says** Jeremy Hunt was

**Ed Miliband defends Iraq war condemnation** (cached)

, says Harman **Harriet Harman says** Labour will be

**Ed Miliband tells Labour: We're the optimists now** (cached)

, says Harman **Harriet Harman says** Labour will be

**Labour must have credible deficit plan, says Darling** (cached)

, says Harman **Harriet Harman says** Labour will be

**David Miliband says he won't join brother Ed's team** (cached)

, says Harman **Harriet Harman says** Labour will be

**Balls: Labour must fight cuts 'every inch of the way'** (cached)

, says Harman **Harriet Harman says** Labour will be

**Rob Ainsworth clashes with ex-aide over Trident** (cached)



# Token features

---

- Free text queries are actually searching the `string` features of the GATE Token annotations
- Can also search other token features, e.g. `root` (morphology) or `category` (POS)



# Morphology

Harriet Harman root:say

Search

Documents 21 to 18 of 38:

[David Cameron criticised for 'calm down dear' jibe \(cached\)](#)

former equality minister **Harriet Harman said** Mr Cameron's

[Queen's Speech: Biggest change to voter registration \(cached\)](#)

, Labour's **Harriet Harman said** the government was

[Harriet Harman struggles with bank bonus and job figures \(cached\)](#)

in Coventry, **Harriet Harman said**: "I

[PMQs: Harriet Harman and Nick Clegg on unemployment \(cached\)](#)

Labour, but **Harriet Harman said** unemployment was falling

[Leveson Inquiry: Jeremy Hunt fair on BSkyB, says top civil servant \(cached\)](#)

Shadow culture secretary **Harriet Harman said**: "David

[Jeremy Hunt: I followed due process over BSkyB \(cached\)](#)

But Labour's **Harriet Harman said** Mr Hunt had

[Ed Miliband 'will marry' but politics 'got in the way' \(cached\)](#)

, says Harman **Harriet Harman says** Labour will be



# Gaps

---

- Default combinator is *sequence* – terms must be adjacent
  - Different from typical search engines
- Can allow gaps with  $[n..m]$
- Arbitrary gap with AND
  - $x \text{ AND } y$  finds the shortest span that covers both



# Gaps

Harriet Harman [0..5] root:say

Search

Documents 1 to 20 of 43:

[Birth weight among social mobility checks - Nick Clegg](#) (cached)

's deputy leader **Harriet Harman said** Mr Clegg would

[Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance'](#) (cached)

Shadow Culture Secretary **Harriet Harman says** Jeremy Hunt was

[Jeremy Hunt quit call after Leveson BSkyB evidence](#) (cached)

Labour deputy leader **Harriet Harman also said** Mr Hunt should

[000375\\_http://www.bbc.co.uk/news/uk-14481315](http://www.bbc.co.uk/news/uk-14481315) (cached)

Miliband and deputy **Harriet Harman are in effect saying** Prime Minister David



# Annotations

---

- So far nothing a standard search engine couldn't do...
- But Mimir also indexes annotations
- Syntax:
  - `{AnnotationType feat1=val1 feat2=val2}`
  - Feature comparisons can be =, <, <=, >=, >
  - UI offers pop-up with available types/features
- Let's generalise – *any* person, not just Harriet

# Annotations

```
{Person} [0..5] root:say
```

Documents 1 to 20 of 5495:

## [Apple's Sir Jonathan Ive reaffirms desire to stay at company \(cached\)](#)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

## [Diamond Jubilee Tube train was faulty \(cached\)](#)

be happening' **Ms Siggs said**: "It

## [Warning over deep-ocean stowaways \(cached\)](#)

using the famous **Alvin sub say** the vehicle picked embarrassment. But **Dr Voight says** the experience is it," **Dr Voight said**. "We

## [School building system not fit for purpose, review says \(cached\)](#)

government-commissioned review by **Sebastian James of Dixons Group said** value for money by Education Secretary **Michael Gove, Mr James said**. Schools with shadow education secretary **Andy Burnham said** Mr Gove had ...

# Other operators

- Containment (use parentheses to group)
  - Query1 IN Query2
  - Query1 OVER Query2

```
( {Person} [0..5] root:say ) IN {Content}
```

Documents 1 to 20 of 4924:

[Apple's Sir Jonathan Ive reaffirms desire to stay at company \(cached\)](#)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

[Diamond Jubilee Tube train was faulty \(cached\)](#)

be happening' **Ms Siggs said**: "It

# Other operators

- Alternative
  - Query1 OR Query2
  - Query1 | Query2

```
{(Person} | {Organization}) [0..2] root:say
```

Documents 1 to 20 of 8611:

[Apple's Sir Jonathan Ive reaffirms desire to stay at company](#) (cached)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

[Diamond Jubilee Tube train was faulty](#) (cached)

for London (TfL) **has said**. Passengers were be happening' **Ms Siggs said**: "It broke down. **London Ambulance Service said** two ambulance crews ...



# Other operators

---

- Set difference
  - `Query1 MINUS Query2`
  - Returns all spans that match Query1 but are not also matches of Query2
  - E.g. sentences that *don't* mention a location

```
{Sentence} MINUS  
( {Sentence} OVER {Location} )
```



# Try it!

- <http://demos.gate.ac.uk/mimir>
  - BBC News demo
- Find:
  - Document titles
  - Date expressions
  - Amounts of money **being paid**
  - Amounts of money **being received**
- Hint: if you get too much noise, try
  - restricting to matches within a sentence
  - using `IN {Content}` to ignore boilerplate



# Semantics

---

- Annotations may be linked to a knowledge base, e.g. DBpedia (<http://dbpedia.org>)
- Annotation refers to an instance
  - [http://dbpedia.org/resource/Harriet\\_Harman](http://dbpedia.org/resource/Harriet_Harman)
- KB knows that this instance belongs to the class of politicians (and people, ...)
  - <http://dbpedia.org/ontology/Politician>
- SPARQL query language can retrieve instances that match constraints

# Semantics

- In news demo, Person, Location and Organization have `class` and `inst` features

```
{Person inst="http://dbpedia.org/resource/Harriet_Harman"}
```

Documents 1 to 20 of 81:

[UK childcare needs to be more affordable - CentreForum](#) (cached)

quality' MP **Harriet Harman** was the architect

[Birth weight among social mobility checks - Nick Clegg](#) (cached)

's deputy leader **Harriet Harman** said Mr Clegg



# Semantic search

---

- Can use SPARQL to query the KB at search time
- E.g. to find *all* politicians

```
{Person sparql="
SELECT DISTINCT ?inst WHERE {
?inst a :Politician }"}
```





# Custom UI

---

- Not the sort of query you want to construct by hand...
- Mimir provides an XML-over-HTTP query API to allow programmatic querying
- Can build custom UIs that hide the query language from users
  - <http://demos.gate.ac.uk/pin>
  - Try out the query builder, look at the underlying query

# Building an index - you need:

---

- Some annotated GATE documents
- A description of which annotations and features you want to index
  - “Index template”
  - Only the features you specify in the template will be available for searching
- A running instance of the Mimir webapp
  - You can download the WAR or build your own
- A way to push the documents to the server



# Pushing documents

---

- GATE PR
  - `uk.ac.gate.mimir:mimir-indexing-plugin:6.1`
- GCP – the “GATE Cloud Parallelizer”
  - Tool to deploy a saved GATE application multi-threaded on your own machine
  - Includes various “output handlers” to save annotations to disk, or push them into Mimir
  - <http://gate.ac.uk/gcp>
- or let us do it for you on GATE Cloud



# GATE Cloud

---

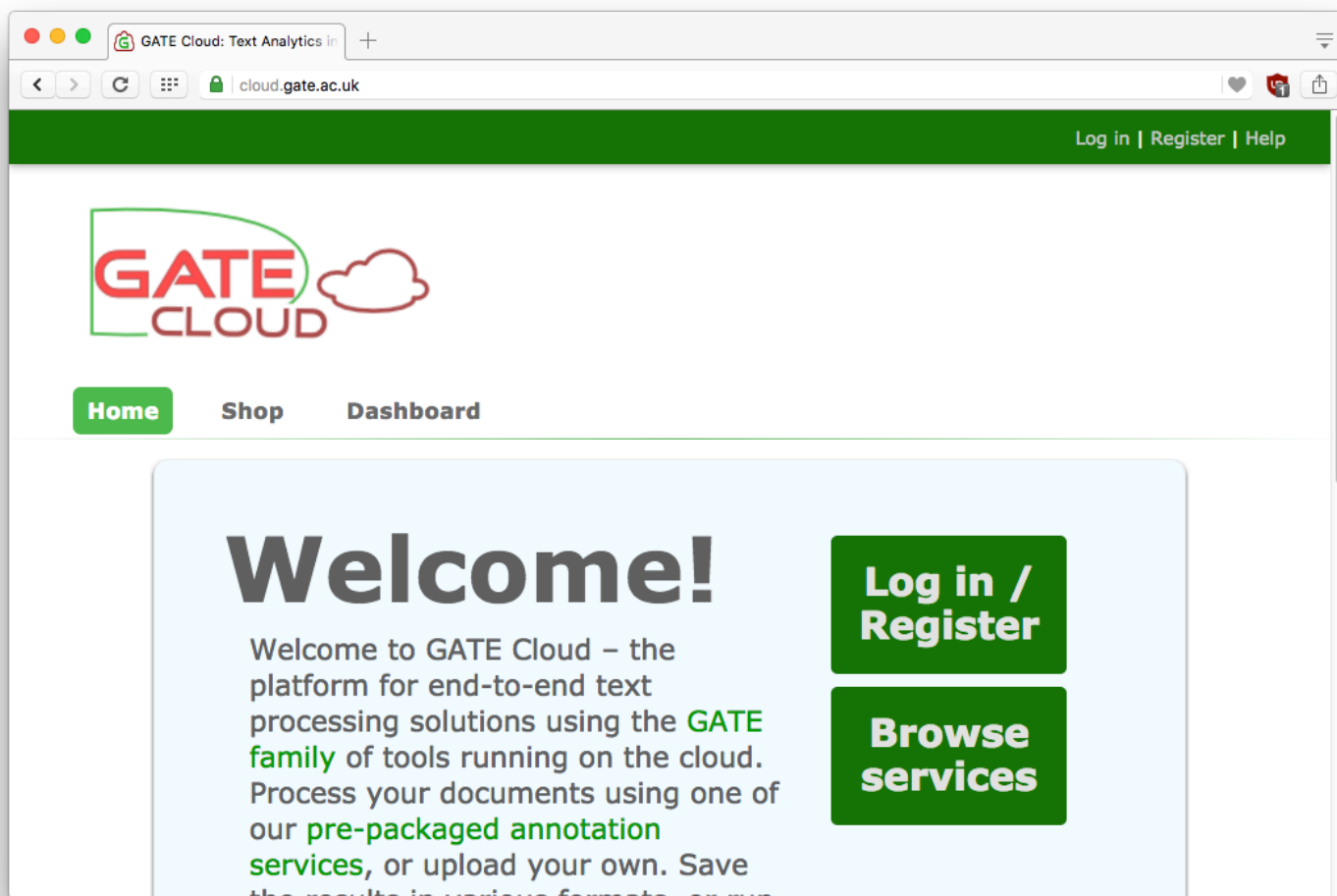
- A cloud based service from the GATE team
- Usual cloud benefits:
  - Pay-as-you-go, no upfront hardware costs
  - No sysadmin work
  - Web-based management tools
  - Always latest version, maintained by us
- Not-so-usual cloud benefits
  - Based on open-source software
  - Bring your own pipeline





# GATE Cloud

- <https://cloud.gate.ac.uk>





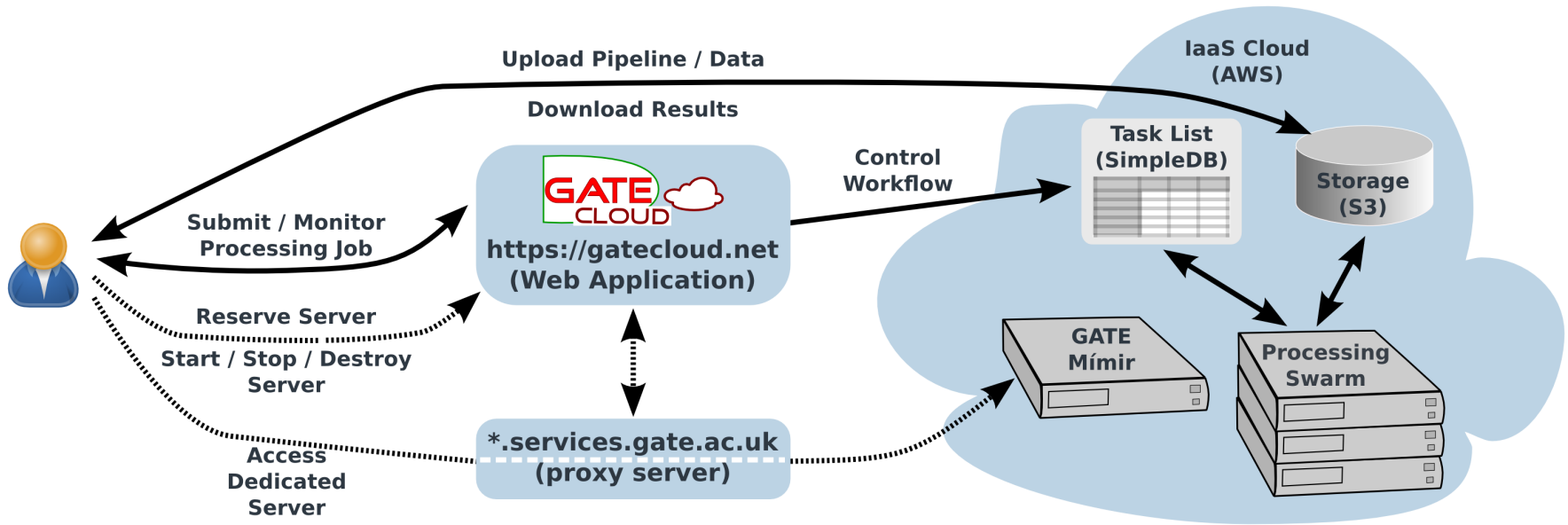
# GATE Cloud Features

---

- Pre-packaged pipelines
  - REST API for processing single documents
- On demand document processing (a.k.a. *Annotation Jobs*)
  - Parallel processing on Amazon EC2
  - Using one of our pipelines or your own
  - On-line job definition tool
  - Many output formats, including Mimir
- On demand servers, including Mimir
- Top up your account with vouchers from the University online shop



# Architecture





# Dedicated servers

---

- You saw these yesterday with Twitter collector
- Rent a dedicated Mimir server for your private use
- Start and stop it as required
- Pay only for the hours it is running
- Data (i.e. indexes) persistent across reboots
- Backup and restore facility available



# The shop

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/shopfront#tagged=Server`. The page header includes the user name "Ian Roberts's account" and links for "Log out" and "Help". The main content area features the "GATE CLOUD" logo and a navigation menu with "Home", "Services" (highlighted), and "Dashboard". Below the navigation is a filter section titled "Show only items tagged:" with various language and task tags. The "Server" tag is selected. Two service cards are visible: "Twitter Collector" and "Mimir Server (6.0)".

**Services**

Home **Services** Dashboard

Show only items tagged: Basque Biomed Bulgarian Catalan Chunker Croatian Custom Czech Danish Dutch English Environment Estonian Finnish French German Greek Indonesian Latvian Measurements Morphology Named Entity OpenNLP Opinion Mining Part-of-Speech Polish Politics Portugese Romanian Russian **Server** ✓ Slovak Slovenian SoBigData Spanish Summarization Swedish Term Recognition **Twitter (1)** Welsh

**Twitter Collector**

Collect tweets, view tweet statistics, and store results in your dashboard for further analysis.

£0.05 / CPU hour

**Mimir Server (6.0)**

Multiparadigm indexing server - index the results of your annotation jobs for rapid retrieval.

£0.50 / CPU hour



# Reserving a server

---

- The usual e-commerce experience
  - Sign up for an account
  - Buy a top-up voucher
  - Find the server you want in the shop
  - Press “reserve this machine” and follow the instructions
- Server appears in your *dashboard*
- Behind the scenes, creates a persistent data *volume* for your data

# Dashboard

## Ian Roberts: Your Dashboard

### Annotation Jobs

[Filter view...](#)

Name	Created At	State
------	------------	-------

### Data Bundles

You have 1 data bundle totalling 19.0 MB. The approximate monthly cost of this data is £0.0005691097 ([help](#))

Bundle ID	Name	Created At	Price per month
D-000011	News corpus	07 June 2016 21:47:47 BST	£0.0005691097

[Upload your own data](#)

### Cloud Machines

Recently purchased machines may take a few minutes to appear in this list.

[Filter view...](#)

Reservation ID	Name	State	Order Number	Order Date
R-000144	Mimir Server (6.0)	inactive	O-000262	07 June 2018

# Reservation control panel

The screenshot shows a web interface for managing machine reservations. At the top, there is a navigation menu with 'Home', 'Services', and 'Dashboard' (highlighted in green). Below the navigation is the title 'Machine Reservation R-000144'. A table displays reservation details, including ID, Name, Machine type, Hourly price, State, and Instance ready status. There are buttons for 'Destroy Reservation', 'Rename', and 'Start Instance'. Below the table is a 'Backups' section with two slots, each containing a '<empty>' text and a 'Create new backup' button. At the bottom, the heading 'Reservation Details:' is visible.

ID	R-000144	<a href="#">Destroy Reservation</a>
Name	Mimir Server (6.0)	<a href="#">Rename</a>
Machine type	Mimir Server (6.0)	
Hourly price	£0.50	
State	inactive	<a href="#">Start Instance</a>
Instance ready	no	

### Backups

Slot 1	<empty>	<a href="#">Create new backup</a>
Slot 2	<empty>	<a href="#">Create new backup</a>

## Reservation Details:





# Controlling the server

---

- Start and stop instance
  - Startup/shutdown takes a few minutes – system will email you when server is ready
  - You pay the hourly price whenever the instance is running
- Backup and restore
  - Save the state of your data volume so you can roll back later
- Destroy reservation
  - If you no longer need the server, destroy it to discard the data volume and all backups
  - *This cannot be undone*



# Document processing

---


- Free-of-charge REST API for our standard pipelines
- POST your document, get annotations back
- Quota controlled and rate limited
  - Ask us if you need more quota
- Documentation on GATE Cloud site
- Web form on the shop page to test each pipeline using this API



# “Test this pipeline”


English Named Entity Recognizer

**1,200 free requests / day**  
Larger batches **£0.80 / CPU hour**

 **ANIE** is a named entity recognition pipeline that identifies basic entity types, such as *Person*, *Location*, *Organization*, *Money* amounts, *Time* and *Date* expressions.

It is the prototypical information extraction pipeline distributed with the **GATE framework** and forms the base of many more complex GATE-based IE applications.

[Annotation details](#)



The screenshot shows the GATE framework interface. On the left, a tree view displays the pipeline configuration: Applications, ANIE, Language Resources, Processing Resources, ANIE OrthoMatcher, and ANIE Transducer. On the right, a list of entity types is shown with checkboxes: Date, FirstPerson, JobTitle, Location, Lookup, Money, Organisation, Person, Sentence, and Unknown. Below this, a text snippet is displayed with various words highlighted in colored boxes corresponding to the entity types. For example, 'Deutsche Telekom' is highlighted in red (Person), 'wireless business' in green (Location), 'Breszone' in blue (Organization), 'last year' in yellow (Time), and 'autumn' in orange (Time).

## Test this pipeline

Type the content to annotate:

Or select a text file:  No file chosen

Output type:

Document format:

[Customize annotations](#)

[Test Pipeline](#)



# Pipeline results

English Named Entity Recognizer Mimir Index "example"

cloud.gate.ac.uk/shopfront/displayItem/annie-named-entity-recognizer

**Test this pipeline**

Type the content to annotate:

Or select a text file:  No file chosen

Output type:

Document format:

Annotation types:  Location  Person

**Annotations at this location**

**Location**

kind	locName
locType	city
rule	InLoc1
ruleFinal	LocFinal



# Calling the API

---

- You can call the API directly via HTTPS
- Generate an API key from your account page
  - The key consists of an identifier (username) and secret (password), used for basic authentication
- POST to endpoint, with appropriate parameters
  - <https://cloud.gate.ac.uk/info/help/online-api.html>



# Examples

---

- Curl

```
curl -H "Accept: application/json" -H  
"Content-Type: text/plain" --data-binary  
"Ian lives in Sheffield" -u  
"apiKey:password" https://cloud-  
api.gate.ac.uk/process-document/annie-  
named-entity-recognizer
```

- Java client library available from GitHub & Maven central
  - <https://github.com/GateNLP/cloud-client>
- Plugin to call the API from GATE Developer
- Other languages have similar tools
  - e.g. Python “requests” module



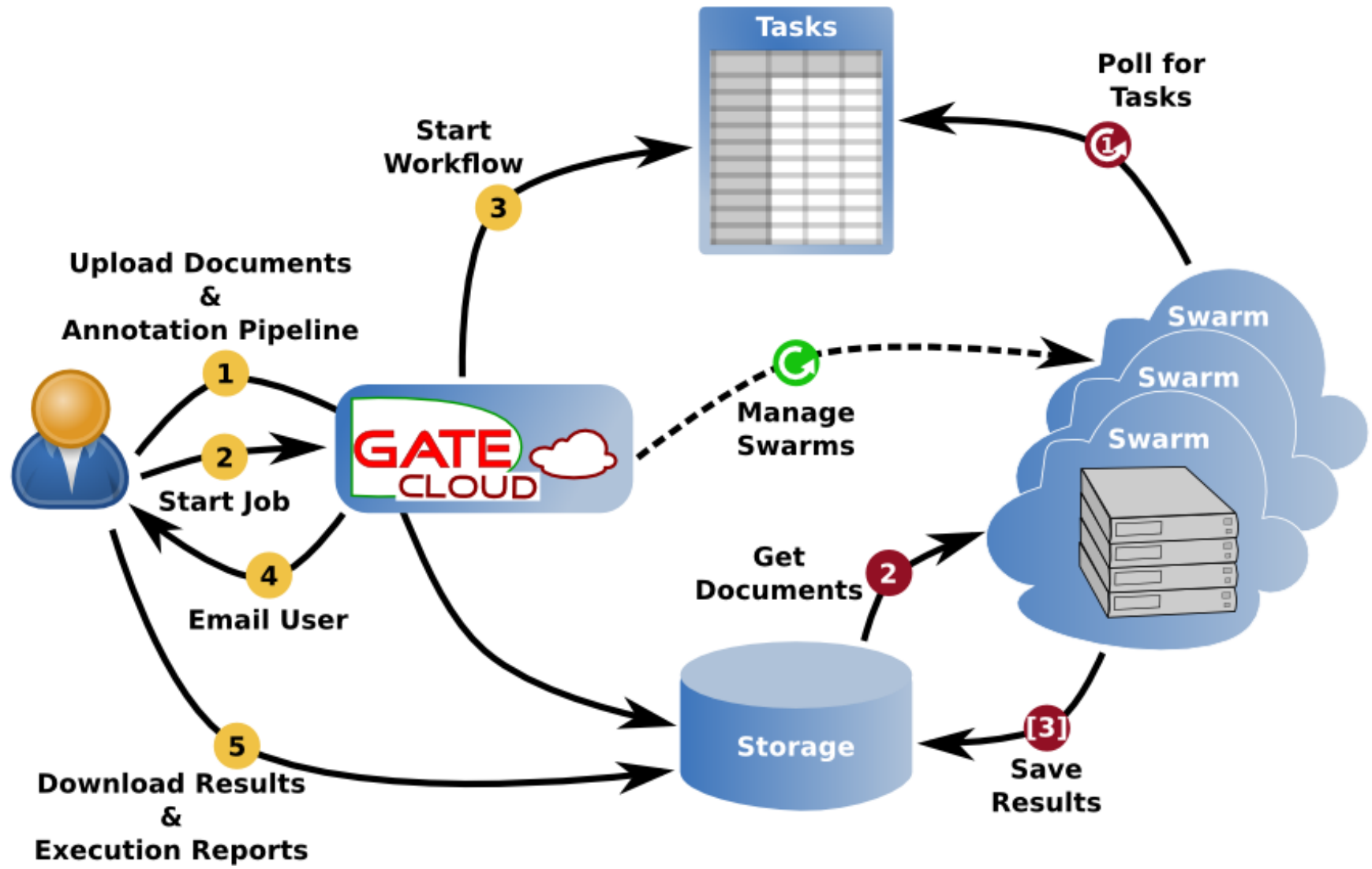
# Annotation jobs

---

- For larger amounts of data
- Parallel and distributed annotation of documents with a GATE application
  - Choose from the standard pipelines
  - Or upload your own pipeline
    - “Export for GATE Cloud”
- Upload your own documents
  - zip, tar, arc/warc archives, Twitter JSON
  - Or collect data directly from Twitter (yesterday)
- Output annotations in various formats, or send documents directly to Mimir



# Job lifecycle







# Execution environment

---

- Amazon EC2
- Ubuntu LTS, 64-bit
- Oracle Java 8
- ~2GB/thread RAM on average
- GCP 3.1, based on GATE Embedded 8.6
  - (8.5 and 8.4.1 also available for older apps)

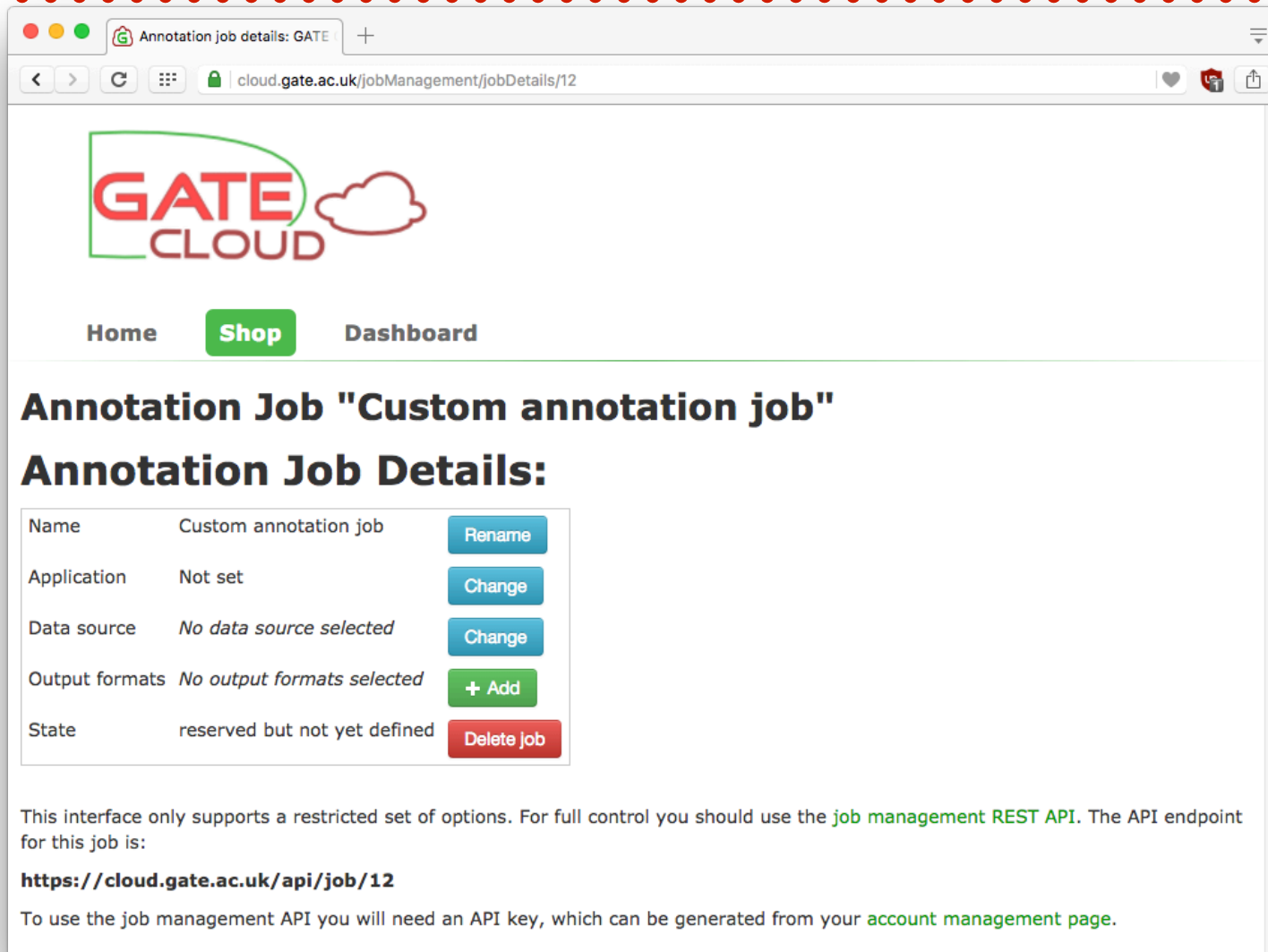


# Reserving a job

---

- Use the “reserve an annotation job” link from the pipeline page
- Or select “custom annotation job” to provide your own pipeline
  - This is what we will do here, since we need the morphological analyser as well as the basic ANNIE
  - 8.6, 8.5 or 8.4.1 options available – you want 8.6
- Once reserved, jobs appear in your dashboard

# Managing a job



The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The page features the GATE Cloud logo and navigation links for Home, Shop, and Dashboard. The main content area displays the title "Annotation Job 'Custom annotation job'" and "Annotation Job Details:". Below this is a table of job properties with corresponding action buttons.

Name	Custom annotation job	<a href="#">Rename</a>
Application	Not set	<a href="#">Change</a>
Data source	<i>No data source selected</i>	<a href="#">Change</a>
Output formats	<i>No output formats selected</i>	<a href="#">+ Add</a>
State	reserved but not yet defined	<a href="#">Delete job</a>

This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).



# Managing a job - application

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The page title is "Annotation job details: GATE". A modal dialog box titled "Upload application" is open, showing the text "Application zip file:" followed by a "Choose File" button and the filename "annie-with-morph.zip". Below the dialog, a table lists job details:

Name	Custom an	
Application	annie-with-morph.zip	<a href="#">Change</a>
Data source	No data source selected	<a href="#">Change</a>
Output formats	No output formats selected	<a href="#">+ Add</a>
State	reserved but not yet defined	<a href="#">Delete job</a>

At the bottom of the page, there is a note: "This interface only supports a restricted set of options. For full control you should use the job management REST API. The API endpoint"



# Managing a job - input

---

- Annotation job input and output files are modelled as “data bundles”
  - Persistent sets of data
  - You pay per GB per month for storage
  - Delete the bundle when you no longer require it
- Bundle can be created by uploading files
- Output of a job becomes another bundle
  - So you can feed the output of one job to the input of another



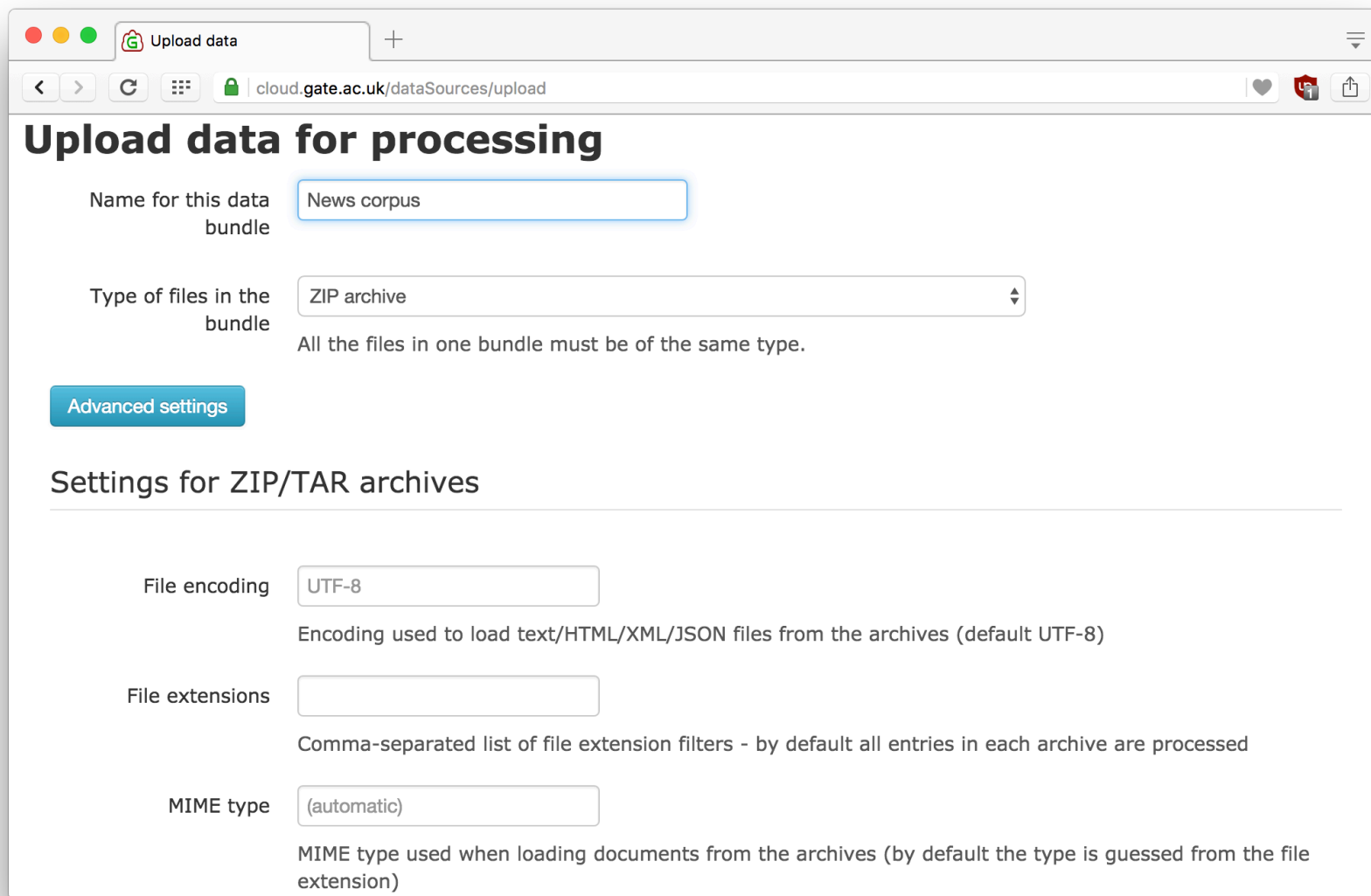
# Managing a job - input

A screenshot of a web browser window showing the GATE interface. The browser tab is titled "Annotation job details: GATE" and the address bar shows "cloud.gate.ac.uk/jobManagement/jobDetails/12". The main content area displays "Annotation Job 'Custom annotation job'" with a table of job details. A modal dialog box titled "Select data source" is overlaid on the page, containing the text: "No suitable data sources found. You can also [upload your own data](#), or use the API if you have an existing data set on Amazon S3". A "Cancel" button is located at the bottom right of the dialog. The table in the background has the following visible rows:

Name	Custom an...
Application	annie-with-
Data source	No data so...
Output formats	No output formats selected
State	reserved but not yet defined

At the bottom of the page, there is a link: <https://cloud.gate.ac.uk/dataSources/upload> and a snippet of text: "options. For full control you should use the job management REST API. The API endpoint".

# Managing a job - input

A screenshot of a web browser window showing the 'Upload data for processing' page. The browser's address bar shows 'cloud.gate.ac.uk/dataSources/upload'. The page title is 'Upload data for processing'. There are three input fields: 'Name for this data bundle' with the value 'News corpus', 'Type of files in the bundle' with a dropdown menu set to 'ZIP archive', and 'Advanced settings' which is expanded to show 'Settings for ZIP/TAR archives'. This section includes three more input fields: 'File encoding' set to 'UTF-8', 'File extensions' (empty), and 'MIME type' set to '(automatic)'. Each input field has a descriptive text below it.

Upload data

cloud.gate.ac.uk/dataSources/upload

## Upload data for processing

Name for this data bundle

Type of files in the bundle

All the files in one bundle must be of the same type.

[Advanced settings](#)

### Settings for ZIP/TAR archives

File encoding   
Encoding used to load text/HTML/XML/JSON files from the archives (default UTF-8)

File extensions   
Comma-separated list of file extension filters - by default all entries in each archive are processed

MIME type   
MIME type used when loading documents from the archives (by default the type is guessed from the file extension)



# Managing a job - input

The screenshot shows a web browser window with the following content:

- Browser Tab:** Data bundle D-000011
- Address Bar:** cloud.gate.ac.uk/dataManagement/dataBundleDetails/11
- Section Header:** Data bundle D-000011
- Metadata Table:**

ID	D-000011	
Name	News corpus	<a href="#">Rename</a>
Date created	June 7, 2016 9:47:47 PM BST	
Monthly price	GBP0.03	
Type	ZIP archive	
Input mime type	text/html	
Actions	<a href="#">Delete this bundle</a>	

This will also delete the underlying files
- Section Header:** Upload data
- Text:** You can add more data to this bundle - when you have finished adding files, click the "Finished" button
- Text:** Upload a ZIP archive:
- Form:**  news-corpus-large.zip
- Form:**
- Text:** 0 files added so far.



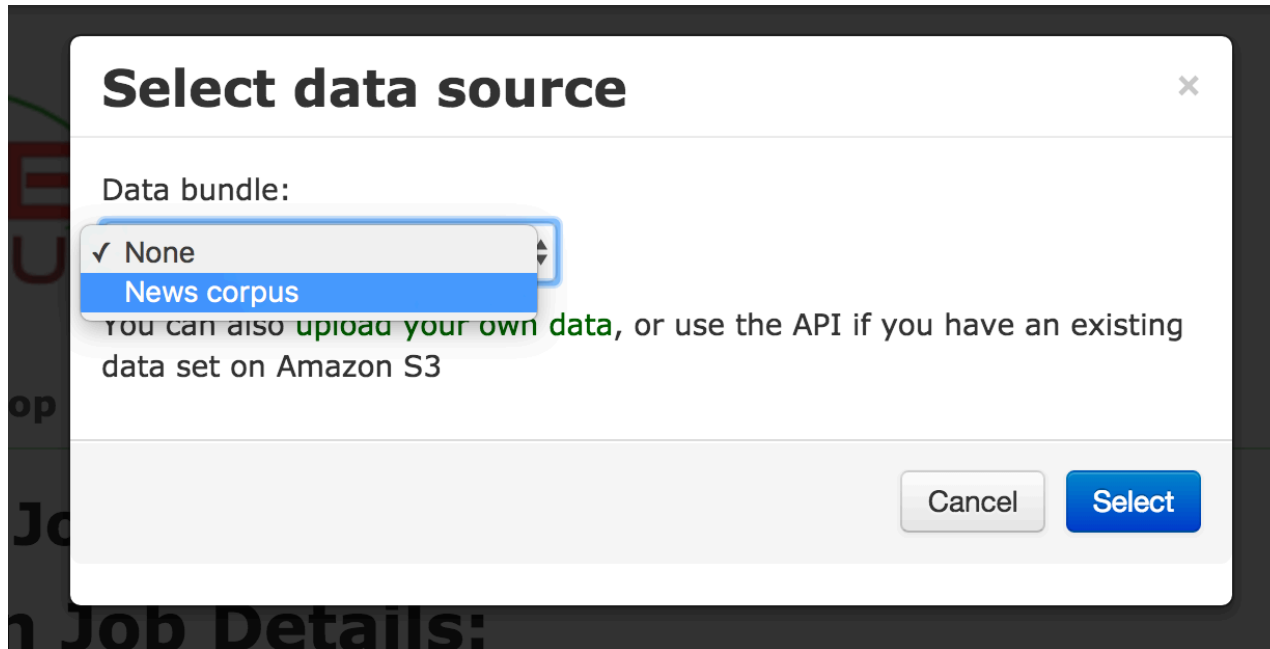


# Managing a job - input

A screenshot of a web browser window. The address bar shows the URL 'cloud.gate.ac.uk/dataManagement/dataBundleDetails/11'. The page title is 'Data bundle D-000011'. A green notification bar at the top of the content area says 'Data bundle D-000011 closed successfully'. Below this is a table of metadata for the data bundle. The table includes fields for ID, Name, Date created, Total size, Monthly price, Type, and Input mime type. There are two buttons: a blue 'Rename' button next to the Name field and a red 'Delete this bundle' button under the Actions field. Below the table, it says 'This will also delete the underlying files'. Underneath, it states 'This bundle contains 1 data files.' and shows a table with one file: 'news-corpus-large.zip'. At the bottom of the page, there are links for 'About' and 'Contact'.

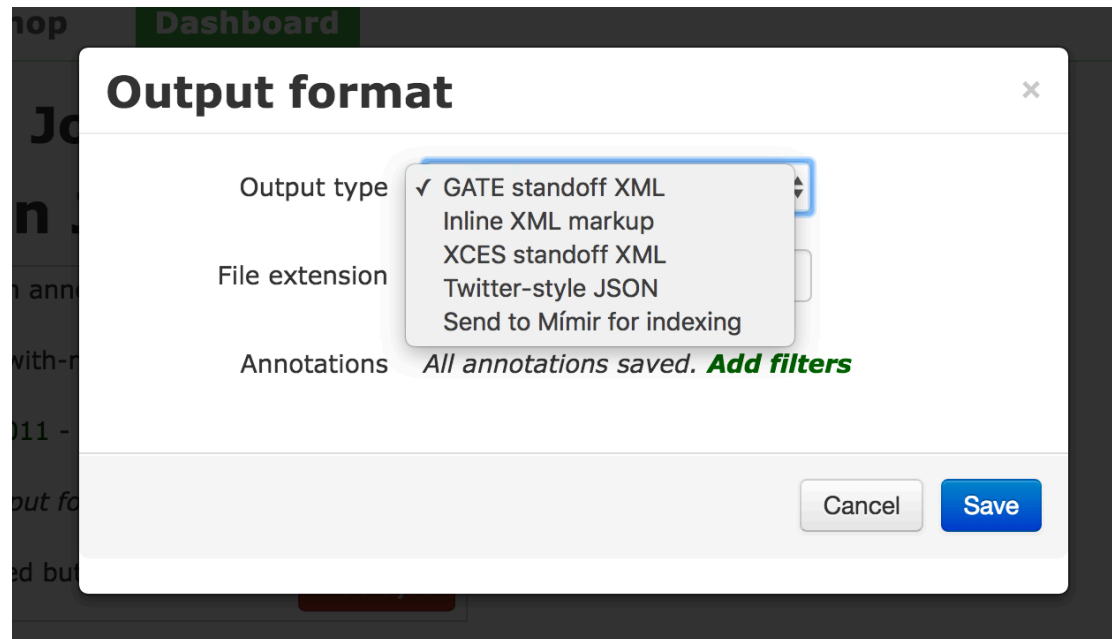
# Managing a job - input

- Back at the job editor



# Managing a job - output

- Various output formats supported
  - Save annotated documents as GATE XML
  - Twitter-style JSON
  - Index in Mimir





# Outputting to Mimir

---

- Start up the Mimir server we reserved earlier
- Create an appropriate index template
- Create an index

# Create a template

Machine Reservation R-000 Create IndexTemplate

mimir-2z17eea4hs.services.gate.ac.uk/admin/indexTemplate/create

## Create IndexTemplate

Name:

Comment:

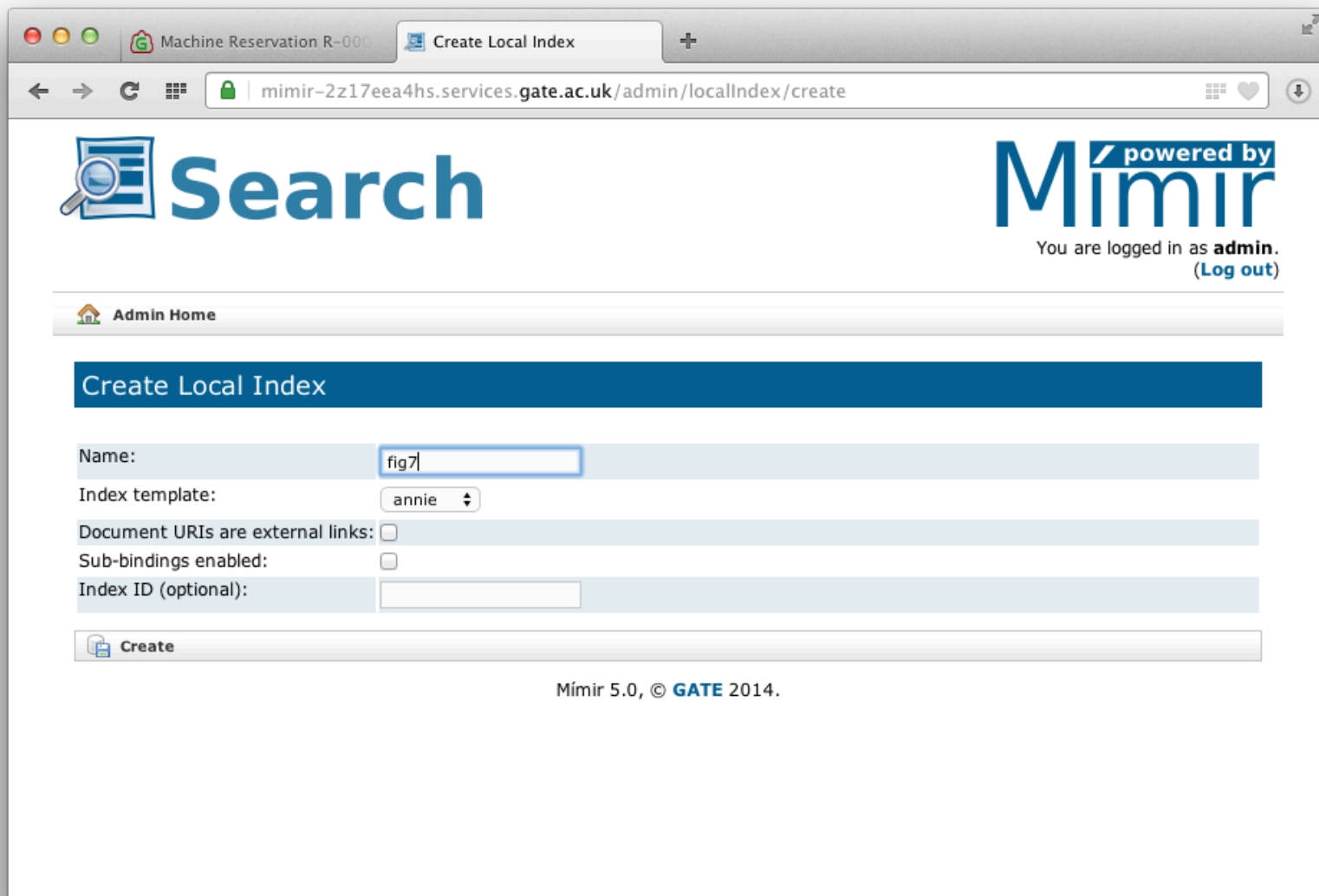
Configuration:

```
semanticAnnotations = {
  index {
    annotation helper:new DefaultHelper(annType:'Sentence')
    annotation helper:new DefaultHelper(annType:'Document',
mode:Mode.DOCUMENT)
  }
  index {
    annotation helper:new DefaultHelper(annType:'Person',
nominalFeatures:['gender'])
    annotation helper:new DefaultHelper(annType:'Location',
nominalFeatures:['locType'])
    annotation helper:new DefaultHelper(annType:'Organization')
  }
  index {
    annotation helper:new DefaultHelper(annType:'Percent')
    annotation helper:new DefaultHelper(annType:'Money')
    annotation helper:new DefaultHelper(annType:'Date',
nominalFeatures:['kind'])
    annotation helper:new DefaultHelper(annType:'Address',
nominalFeatures:['kind'])
  }
}
documentRenderer = new OriginalMarkupMetadataHelper()
documentMetadataHelpers = [documentRenderer]
```

Create


Mimir 5.0, © GATE 2014.

# Create an index


A screenshot of a web browser window showing the 'Create Local Index' page in the Mimir administration interface. The browser's address bar shows the URL 'mimir-2z17eea4hs.services.gate.ac.uk/admin/localIndex/create'. The page features a 'Search' header with a magnifying glass icon and the text 'powered by Mimir'. A user is logged in as 'admin' with a '(Log out)' link. Below the header is an 'Admin Home' link. The main content area is titled 'Create Local Index' and contains a form with the following fields: 'Name:' with the value 'fig7', 'Index template:' with a dropdown menu showing 'annie', 'Document URIs are external links:' with an unchecked checkbox, 'Sub-bindings enabled:' with an unchecked checkbox, and 'Index ID (optional):' with an empty text input field. At the bottom of the form is a 'Create' button. The footer of the page reads 'Mimir 5.0, © GATE 2014.'

Machine Reservation R-000 Create Local Index

mimir-2z17eea4hs.services.gate.ac.uk/admin/localIndex/create

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

You are logged in as **admin**.  
([Log out](#))

 [Admin Home](#)

## Create Local Index


Name:

Index template:

Document URIs are external links:

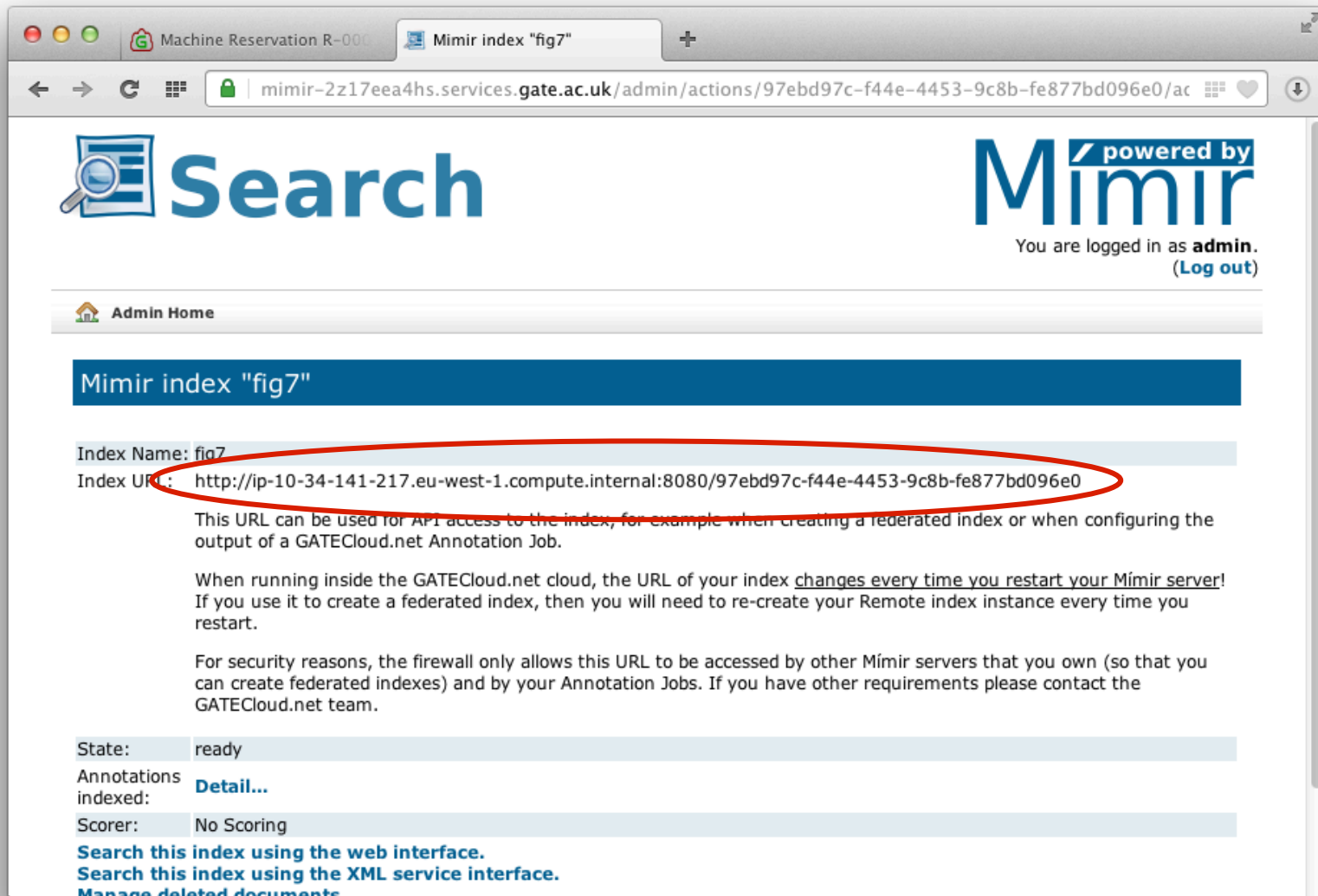
Sub-bindings enabled:

Index ID (optional):

 **Create**

Mimir 5.0, © GATE 2014.

# Index details – note URL



The screenshot shows a web browser window with the URL `mimir-2z17eea4hs.services.gate.ac.uk/admin/actions/97ebd97c-f44e-4453-9c8b-fe877bd096e0/ac`. The page header includes a 'Search' logo and 'powered by Mimir'. The user is logged in as 'admin' with a '(Log out)' link. Below the header, there is an 'Admin Home' link and a blue bar with the text 'Mimir index "fig7"'. The main content area displays the following details:

Index Name: fig7  
Index URL: <http://ip-10-34-141-217.eu-west-1.compute.internal:8080/97ebd97c-f44e-4453-9c8b-fe877bd096e0>

This URL can be used for API access to the index, for example when creating a federated index or when configuring the output of a GATECloud.net Annotation Job.

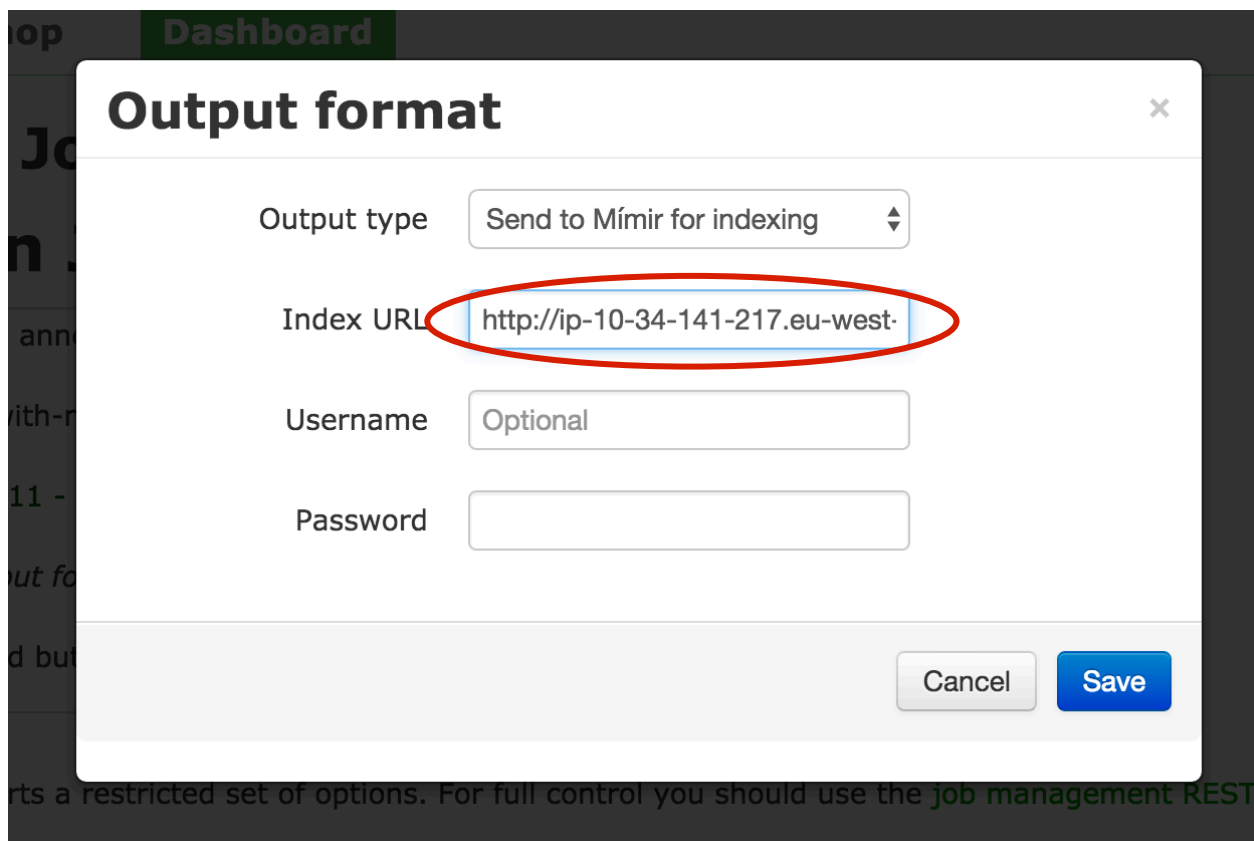
When running inside the GATECloud.net cloud, the URL of your index changes every time you restart your Mimir server! If you use it to create a federated index, then you will need to re-create your Remote index instance every time you restart.

For security reasons, the firewall only allows this URL to be accessed by other Mimir servers that you own (so that you can create federated indexes) and by your Annotation Jobs. If you have other requirements please contact the GATECloud.net team.

State: ready  
Annotations indexed: [Detail...](#)  
Scorer: No Scoring

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

# Managing a job - output

A screenshot of a web application interface showing a dialog box titled "Output format". The dialog box has a close button (x) in the top right corner. It contains four input fields: "Output type" with a dropdown menu showing "Send to Mimir for indexing"; "Index URL" with a text input field containing "http://ip-10-34-141-217.eu-west-" and a red oval highlighting the text; "Username" with a text input field containing "Optional"; and "Password" with an empty text input field. At the bottom right of the dialog box are two buttons: "Cancel" and "Save". The background of the application shows a "Dashboard" tab and some text fragments like "op", "Jo", "n.", "ann", "with-r", "11 -", "ut fo", "d bu", and "rts a restricted set of options. For full control you should use the job management REST".

**Output format** ×

Output type

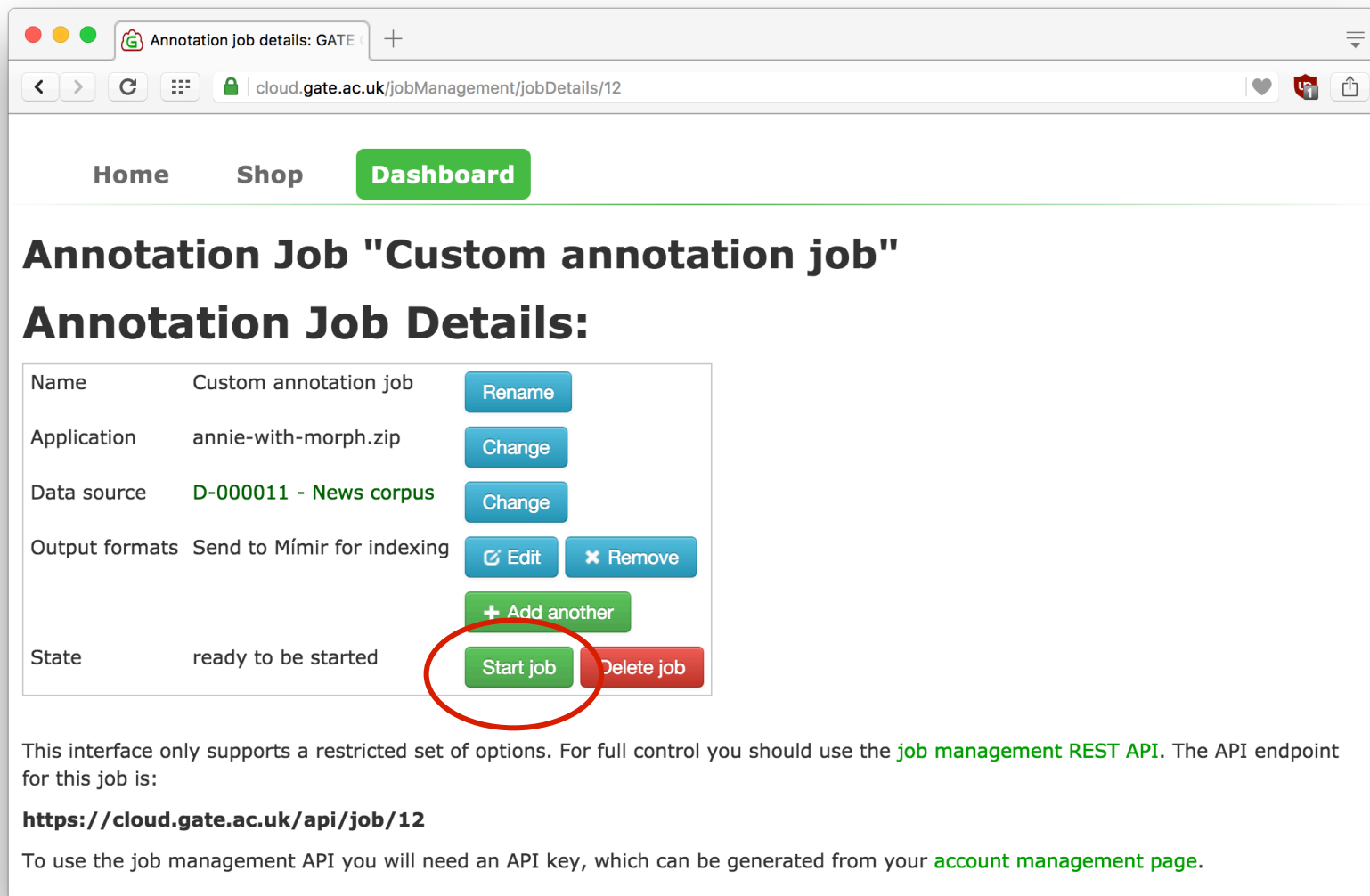
Index URL

Username

Password



# Start the job

A screenshot of a web browser showing the GATE job management interface. The browser's address bar shows the URL 'cloud.gate.ac.uk/jobManagement/jobDetails/12'. The page has a navigation menu with 'Home', 'Shop', and 'Dashboard' (the latter is highlighted in green). The main heading is 'Annotation Job "Custom annotation job"'. Below this is a section titled 'Annotation Job Details:' containing a table of job properties and their corresponding control buttons. The 'Start job' button is circled in red. Below the table, there is a paragraph of text and a URL.

Name	Custom annotation job	<a href="#">Rename</a>
Application	annie-with-morph.zip	<a href="#">Change</a>
Data source	D-000011 - News corpus	<a href="#">Change</a>
Output formats	Send to Mimir for indexing	<a href="#">Edit</a> <a href="#">Remove</a>
		<a href="#">+ Add another</a>
State	ready to be started	<a href="#">Start job</a> <a href="#">Delete job</a>

This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).



# Job running

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The navigation menu includes **Home**, **Shop**, and **Dashboard**. The main heading is **Annotation Job "Custom annotation job"**. A green notification bar states "Job successfully started". Below this, the **Annotation Job Details:** section contains a table with the following information:

Name	Custom annotation job	<a href="#">Rename</a>
Application	annie-with-morph.zip	
Data source	D-000011 - News corpus	
Output formats	Send to Mimir for indexing	
State	active - processing	<a href="#">Interrupt job</a>

Below the table is a progress bar showing 33.33333333299999995% completion.

This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).



# When job completes...

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/jobManagement/jobDetails/12`. The navigation menu includes **Home**, **Shop**, and **Dashboard**. The main heading is **Annotation Job "Custom annotation job"**. A green notification bar states "Job successfully started". Below this, the **Annotation Job Details:** section contains a table of job information and action buttons.

Name	Custom annotation job	<a href="#">Rename</a>
Application	annie-with-morph.zip	
Data source	D-000011 - News corpus	
Output formats	Send to Mimir for indexing	
State	completed	<a href="#">Reset job</a> <a href="#">Delete job</a>
Final report	<a href="#">Download</a>	
Debug logs	<a href="#">Download</a>	

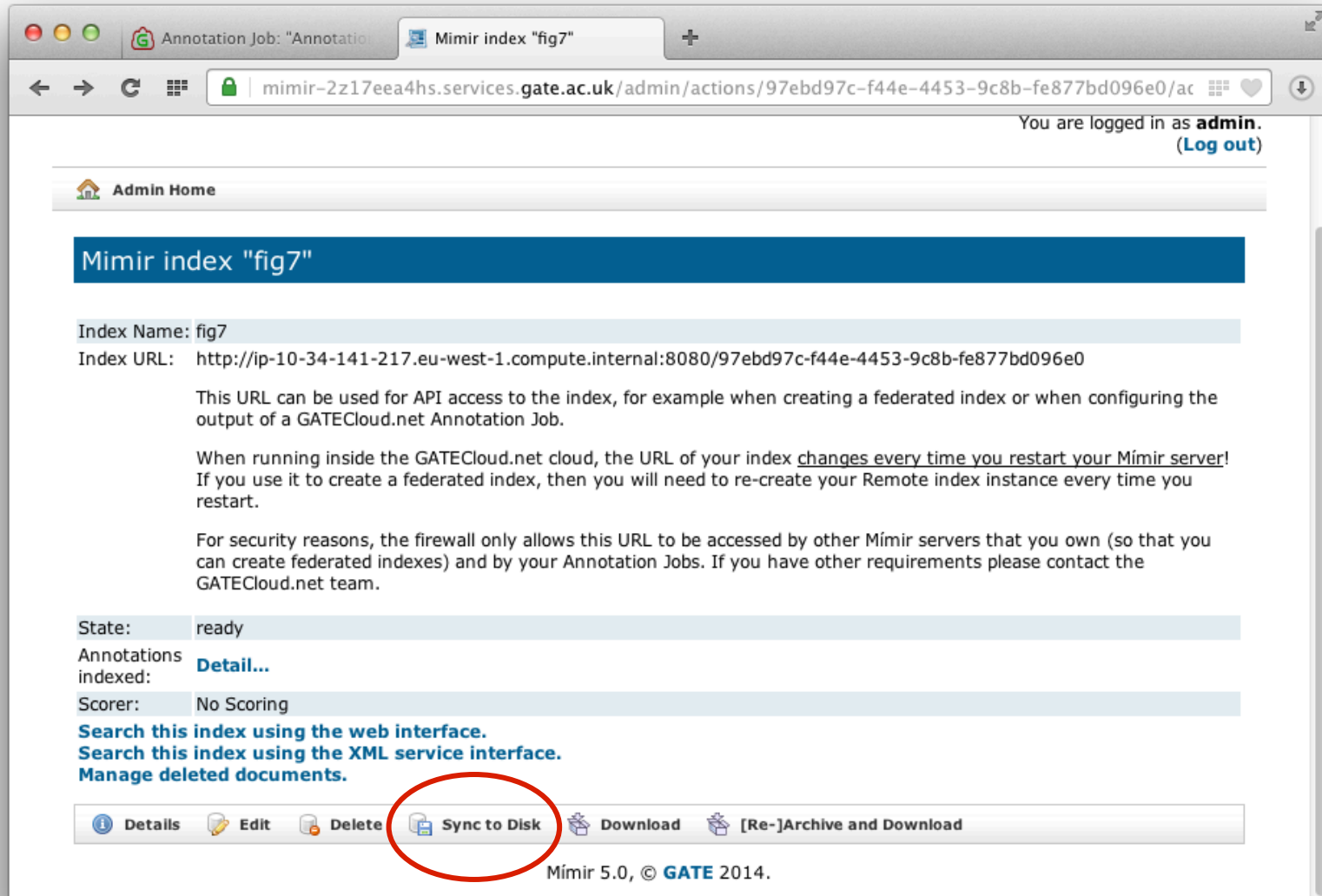
This interface only supports a restricted set of options. For full control you should use the [job management REST API](#). The API endpoint for this job is:

**<https://cloud.gate.ac.uk/api/job/12>**

To use the job management API you will need an API key, which can be generated from your [account management page](#).

Waiting for cloud.gate.ac.uk...

# “Sync” the index



Annotation Job: "Annotation" Mimir index "fig7"

← → ↻ ☰ [mimir-2z17eea4hs.services.gate.ac.uk/admin/actions/97ebd97c-f44e-4453-9c8b-fe877bd096e0/ac](https://mimir-2z17eea4hs.services.gate.ac.uk/admin/actions/97ebd97c-f44e-4453-9c8b-fe877bd096e0/ac) You are logged in as **admin.** ([Log out](#))

[Admin Home](#)

## Mimir index "fig7"

Index Name: fig7  
Index URL: <http://ip-10-34-141-217.eu-west-1.compute.internal:8080/97ebd97c-f44e-4453-9c8b-fe877bd096e0>

This URL can be used for API access to the index, for example when creating a federated index or when configuring the output of a GATECloud.net Annotation Job.

When running inside the GATECloud.net cloud, the URL of your index changes every time you restart your Mimir server! If you use it to create a federated index, then you will need to re-create your Remote index instance every time you restart.

For security reasons, the firewall only allows this URL to be accessed by other Mimir servers that you own (so that you can create federated indexes) and by your Annotation Jobs. If you have other requirements please contact the GATECloud.net team.

State: ready  
Annotations indexed: [Detail...](#)  
Scorer: No Scoring

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

[Details](#) [Edit](#) [Delete](#) [Sync to Disk](#) [Download](#) [\[Re-\]Archive and Download](#)

Mimir 5.0, © GATE 2014.



# “Sync” the index

---

- Mimir accumulates documents in RAM
- Documents saved to disk after (by default) one hour, or when memory threshold reached
- Documents become searchable once saved to disk
- “Sync” button forces an immediate save, if you know no more documents due
  - Can continue to send more documents, but only sync-ed ones available for search



# Search your new index

A screenshot of a web browser window displaying the Mimir search interface. The browser's address bar shows the URL: `mimir-2z17eea4hs.services.gate.ac.uk/97ebd97c-f44e-4453-9c8b-fe877bd096e0/search/index#qu`. The page features a search bar with the query `{Person gender = "female"}` and a "Search" button. A blue banner at the top indicates "Searching Index 'fig7'". The results section, titled "Documents 1 to 20 of 760:", lists search results including a Guardian article from 2009 and a book by Sarah Raven.

Annotation Job: "Annotation" Mimir Index "fig7"

mimir-2z17eea4hs.services.gate.ac.uk/97ebd97c-f44e-4453-9c8b-fe877bd096e0/search/index#qu

Search

powered by Mimir

You are logged in as admin. (Log out)

Searching Index "fig7"

{Person gender = "female"}

Search

Documents 1 to 20 of 760:

[www.guardian.co.uk/business/2009/dec/20/shoppers-defy-big-freeze?](http://www.guardian.co.uk/business/2009/dec/20/shoppers-defy-big-freeze?INTCMP=ILCNETTXT3487)  
INTCMP=ILCNETTXT3487

39 5. Sarah Raven's Wild Flowers Wild Flowers by Sarah Raven £29.

[www.guardian.co.uk/business/2008/dec/27/high-street-retailers-retail?](http://www.guardian.co.uk/business/2008/dec/27/high-street-retailers-retail?INTCMP=ILCNETTXT3487)  
INTCMP=ILCNETTXT3487



# Try it!

- Sign up for an account on <https://cloud.gate.ac.uk>
- Use your voucher code
- Reserve a Mimir 6.1 server
- Start it up, log in
- Create a new index template using the contents of `index-template.groovy`
- Create a new local index using this template
- Visit index admin page and note the URL



# Try it!

- Reserve a “custom annotation job (8.6)”
- Application zip file is annie-with-morph.zip
- Create a data bundle for the input, and upload news-corpus-large.zip
  - Mime type: text/html, Encoding: UTF-8





# Try it!

- Set one output to MIMIR, using the Index URL you noted above
  - Make sure to not include any spaces in the index URL
- Run the job, and when finished sync the index



# Try it!

- Try some searches on your new index
  - E.g. stock price movements  
{Organization} (up | down) ({Money} | {Percent})
- When finished, make sure you stop the Mimir server, destroy the reservation, and delete the data bundle you created



# Questions?

---

- More info
  - <https://cloud.gate.ac.uk>
  - <https://gate.ac.uk/mimir>