

MÍMIR USER GUIDE

Valentin Tablan

Ian Roberts

Contents

1	Introduction	5
1.1	Core Concepts	6
1.2	Mimir Lifecycle	7
2	Quick Start	9
2.1	Set Up Your Environment	9
2.2	Build and Run a Mimir Web Application	10
2.3	Create, Populate, and Search an Index	10
3	Installing and Managing Mimir	12
3.1	Mimir Architecture	12
3.2	Building and Running a Mimir Web Application	13
3.2.1	The <code>mimir-cloud</code> Web Application	13
3.2.2	Binary Distribution	13
3.2.3	Prerequisites	14
3.2.4	Building	14
3.2.5	Configuring	14
3.2.6	Running	16
3.3	Upgrading from earlier versions of Mimir	16
3.4	Indexes in Mimir	17
3.4.1	Types of Index	17
3.4.2	Creating a Local Index	17
3.4.3	Working with Remote and Federated Indexes	19

3.4.4	Deleting Indexes	21
3.5	“Deleting” Documents from a Mimir Index	22
4	Indexing Documents with Mimir	24
4.1	Configuring the Indexer	24
4.2	Adding Documents to an Index	31
4.3	The Default Representation Scheme	32
5	Searching Mimir Indexes	35
5.1	The Mimir Query Language	35
5.1.1	String Queries	36
5.1.2	AND Operator: “&”	37
5.1.3	OR Operator: “ ”	37
5.1.4	IN and OVER Operators	37
5.1.5	MINUS Operator	37
5.1.6	Repeats Operator: “+”	38
5.1.7	Sequence Queries and Gaps	38
5.1.8	Escaping Reserved Words	38
5.2	Search Interfaces – How to Submit Queries to Mimir	39
5.2.1	Mimir Search Web Service	39
5.2.2	The Web UI Example User Interface	46
5.2.3	Embedding Mimir in a Grails Application	46
5.3	Ranking of Results	49
5.3.1	Configuring Ranking	49
6	Standard Mimir Plugins	50
6.1	The db-h2 Plugin	50
6.2	The measurements Plugin	51
6.2.1	Configuring the Measurements SAH	51
6.3	The sparql Plugin	53
6.3.1	Creating a SPARQL Helper	53

6.3.2	Format of SPARQL Queries	56
7	Extending and Customising Mimir	57
7.1	Creating New Semantic Annotation Helpers	57
7.1.1	The SemanticAnnotationHelper Interface	58
7.1.2	Abstract Base Classes	59
7.1.3	Packaging New Helper Types for Use with Mimir	60
7.2	Registering new Scorers	61
7.3	Embedding Mimir in a Custom Grails Application	61
7.3.1	Installing the Mimir Plugin	61
7.3.2	URL Mappings	62
7.3.3	Security Considerations	63
8	Additional Tools	64
8.1	Recovering a failed index	64
A	Change Log	66
A.1	Version 6.0 (May 2018)	66
A.2	Version 5.4 (February 2017)	66
A.3	Version 5.3 (January 2017)	67
A.4	Version 5.2 (June 2016)	67
A.5	Version 5.1 (June 2015)	67
A.6	Version 5.0.1 (October 2014)	67
A.7	Version 5.0 (February 2014)	68
A.8	Version 4.1.3 (September 2012)	68
A.9	Version 4.1.2 (August 2012)	69
A.10	Version 4.1.1 (May 2012)	69
A.11	Version 4.1 (May 2012)	69
A.12	Version 4.0 (February 2012)	69
A.13	Version 3.4.0 (November 2011)	70
A.14	Version 3.3.0 (October 2011)	70

A.15 Version 3.2.0 (May 2011) 70

Chapter 1

Introduction

Mímir is a multi-paradigm information management index and repository which can be used to index and search over text, annotations, semantic schemas (ontologies), and semantic meta-data (instance data). It allows queries that arbitrarily mix full-text, structural, linguistic and semantic queries and that can scale to gigabytes of text.

A typical semantic annotation project deals with large quantities of data of different kinds. Mímir provides a framework for implementing indexing and search functionality across all these data types, listed below in the order of increasing information density:

Text

All documents have a textual content. Support for full text search represents the most basic indexing functionality and it is required in most (if not all) cases. Even when semantic annotation is used to abstract away from the actual textual data, the original content still needs to be accessible so that it can be used to provide textual query fragments in the case of more complex conceptual queries.

Mímir uses inverted indexes¹ for indexing the document content (including additional linguistic information, such as part-of-speech or morphological roots), and for associating instance of annotations with the position in the input text where they occur. The inverted index implementation used by Mímir is based on MG4J².

Annotations

The first step in abstracting away from the plain text content is the production of *annotations*. Annotations are meta-data associated to text snippets in the documents. Mímir's view of annotations is based on that of GATE, with each annotation described by

- the document it belongs to;
- the start and end offset of the referred text snippet;
- the annotation type;
- an arbitrary set of <feature,value> pairs.

¹*Inverted Indexes* are data structures traditionally used in Information Retrieval to support indexing of text.

²<http://mg4j.dsi.unimi.it/>

An annotation index supports a more generic search paradigm. Depending on the type of annotations available, the user can search across different dimensions. If, for example, the documents are annotated with occurrences of **Person**, **Location**, **Organization** entities, then searches like {**Person**}, CEO of {**Organization**}, based in {**Location**} become possible. Storage of annotation data in Mimir indexes is handled by plugins, Mimir ships with two storage plugins by default, one storing annotation data in a relational database and the other in a Knowledge Base to support richer semantic querying.

ANNIC (ANNotations In Context)³ is a tool predating Mimir that supports the indexing of annotations, and that has been used to inform the design of Mimir.

Knowledge Base Data

Knowledge Base (KB) Data consists of an ontology populated with instances. The ontology represents the data schema and comprises a hierarchy of class types and a hierarchy of properties that are applicable between instances of classes. The instance data represents facts that are known to the systems and is typically at least partially derived from semantic annotation over documents. KB data is used to reach a higher level of abstraction over the information in the documents which enables conceptual queries such as “find all mentions of **Persons** who are employed by any organisation based in Yorkshire”.

A KB that is pre-populated with appropriate world knowledge can perform other generalisations that are natural to humans users, such as being able to identify Vienna as a valid answer to queries relating to Austria, Europe or the Western Hemisphere.

As mentioned above, Mimir can make use of a Knowledge Base to store information relating to annotations. The links between annotations, the textual data, and the knowledge base information are created by the inclusion into the text indexes of a set specially-created URIs that are associated with annotation data. Furthermore, URIs of entities from the Knowledge Base can be stored as annotation features.

Knowledge bases are typically represented as a collection of triples that are kept in highly-specialised and optimised triple stores, using standards such as RDF or one the versions of OWL⁴. The implementation used by Mimir is based on ORDI and OWLIM⁵.

1.1 Core Concepts

Mimir provides indexing infrastructure for annotated GATE⁶ documents. Users can start a Mimir server, submit documents to it for indexing, and execute queries against the set of indexed documents.

A Mimir index is a composite of multiple sub-indexes, which are defined in the *index template* that needs to be provided by the user when a new Mimir index is created (see Section 4.1 for details).

Token Indexes are sub-indexes that store the information associated with {**Token**} annotations. These provide a way to index the document content. Mimir does not directly index the document text. Instead it uses the sequence of {**Token**} annotation to construct

³See <http://gate.ac.uk/userguide/chap:annic>.

⁴See <http://www.w3.org/RDF/> and <http://www.w3.org/TR/owl-features/>.

⁵See <http://www.ontotext.com/ordi/> and <http://www.ontotext.com/owlim/>.

⁶<http://gate.ac.uk>

a representation of the document text. This provides more flexibility: if the user chooses to index the `string` feature of the tokens, that is equivalent to indexing the document text. Alternatively, the user could choose to pre-process their document with the GATE Morphological Analyser, and instead index the morphological roots of each token. This normalises the representation of words (by eliminating inflections) and allows different forms of the same word to be matched (e.g. *house* and *houses*). This is similar to stemming/lemmatising, a process traditionally employed in Information Retrieval, but it is more advanced and linguistically sophisticated, and allows matching e.g. *be*, *was*, *are* with each-other, which stemming would not be capable to.

Beside allowing the user to choose which token feature should be indexed, Mimir also allows multiple token features to be indexed in parallel sub-indexes. The user can actually choose to index **both** the token string and morphological root. In that case, the feature mentioned first in the *index template* becomes the default token feature. To search on any of the other token features, queries need to specify which feature they want to target (see Section 5.1.1 for details).

Annotation Indexes are the other type of Mimir sub-index. They are used to index information about annotations on the document. Which annotations should be indexed is described in the *index template*.

Both token and annotation indexes can be configured to also use **direct indexes**. Direct can be used to perform searches for terms starting from documents, for example finding the most frequently occurring word (or annotation) in a set of documents. This functionality is only available from the Java API and cannot be directly accessed by the system users via the web interface. More details can be found in Section 4.1.

1.2 Mimir Lifecycle

In versions prior to 5.0, a Mimir index would start its existence in *indexing* mode, when it would accept new documents for indexing. When all the documents had been indexed, the index would need to be *closed*, which would switch its operation mode to *searching*, and the index would then be able to answer queries. Once closed, an index could not accept any further documents for indexing. Starting with version 5.0, a Mimir index is continually accepting documents to be indexed and can answer queries that address the currently indexed document set. From being sent to Mimir for indexing to becoming available for search, documents go through several stages, which we describe next.

Documents submitted for indexing are initially accumulated in RAM, during which time they are not available for being searched. A *sync-to-disk* operation writes all the documents currently in RAM to disk, in the form of an *index batch*, after which the documents can be searched. Sync-to-disk operations happen automatically when too much document data has been accumulated in RAM, or after a given time interval has passed since the last sync. Alternatively, the user can also trigger a sync operation from index admin web interface.

Every sync-to-disk operation causes a new index *batch* to be created. All the batches are merged into an index cluster which is then used to serve queries. If the number of clusters gets too large, it can harm efficiency or the system can run into problems due to too large a number of files being open. To avoid this, the index batches can be compacted into a single batch. Mimir indexes will automatically do that once the number of batches exceeds

a certain threshold (which can be modified via API calls).

In order to keep its consistency, a Mimir index **must** be closed in an orderly fashion before the mimir server process is shut down. Shutting down the Mimir server (e.g. the `mimir-cloud` web application) will automatically close all currently open indexes. Users should never forcefully destroy the Mimir server process, as that would not allow the close operations to be performed, which can lead to data loss, or it can corrupt existing indexes.

Chapter 2

Quick Start

This chapter is aimed at the impatient reader who wants a working system as quickly as possible. The technical detail is deliberately kept at a minimum so, while you will hopefully end up with something that works, you will not necessarily understand how it all fits together. For that, please read the remainder of this guide.

2.1 Set Up Your Environment

We suggest you try this on a 64 bit operating system, as that is better suited for running Mímir. A 32 bit system would also work, but the maximum sizes for the indexes would be limited.

In order to build and run a Mímir server you will need the following pieces of software installed on your system:

Java Development Kit version 8 or newer. If you don't have one, you can download one from Oracle¹. Make sure you get the JDK and not the Java Runtime Environment (JRE), as that would not be suitable. Once installed, make sure your `JAVA_HOME` environment variable points to the location where the JDK was installed. Make sure that the `$JAVA_HOME/bin` location is on your `PATH`.

Apache Maven version 3.3 or later. You can download it from <http://maven.apache.org/>.

Working Internet Connection The next step, described below, is the building of the Mímir library. This starts by automatically downloading all the required dependencies, so it requires a working Internet connection. Once the software is built, it can work without an remote connection.

GATE Developer Mímir is an indexer for GATE Documents. The simplest way of generating some GATE documents to be indexed is by using the GATE Developer tool².

¹<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

²GATE Developer is available at <http://gate.ac.uk/download/>. Usage of GATE Developer is beyond the scope of this document, so we assume you have a basic understanding of how to use it. If not, a good place to start is the tutorials page at <http://gate.ac.uk/demos/developer-videos/>.

2.2 Build and Run a Mimir Web Application

After all the prerequisites are installed, we can move to building a Mimir application. For the purposes of this demo, we will build the `mimir-cloud` application, which is included in the source tree.

The following steps will help you build the `mimir-cloud` application. Commands that you have to execute are formatted in a distinctive font **like this**.

1. **Download the Mimir sources**, if you do not already have a copy. You can get either an archive of the entire source tree, or check it out directly from our GitHub repository. Instructions for doing so are available on Mimir's web page at: <http://gate.ac.uk/mimir/index.html>. Unpacking a source archive (or checking out the source code with subversion) will create a new directory called `mimir-(version)` (zip) or `mimir` (Git) containing all the source files.
2. **Build Mimir**: change to the top level directory where you unpacked the downloaded Mimir sources. If you can see the `mimir-core`, `mimir-client`, *etc.* directories, then you are in the correct directory. Execute the command `mvn install`. This will download all the required dependencies and compile all the Mimir libraries.
3. **Run the mimir-cloud application**: change to the `webapp/mimir-cloud` directory and execute the command `./grailsw run-app` on Linux/Mac or `grailsw.bat run-app` on Windows. This will start the application and will notify you which URL you should use in your browser to access it (normally `http://localhost:8080/`).

2.3 Create, Populate, and Search an Index

4. **Set-up your new Mimir application**: navigate to the administration page. You will be prompted to configure your Mimir instance. After clicking the link, enter the path to a local writable directory where new indexes will be created, and click the *Update* button.
5. **Create a new index**: navigate back to the administration page (by clicking the link at the top of the page). Under the *Local Indexes* section, click the *create a new local index* link. Give it a name (e.g. 'test'), and click the *create* button. Back on the administration page, click the name of the newly created index. This will take you to the index details page, where you can find the *Index URL* attribute. Make a note of its value, as you will need it later.
6. **Populate the new index**:
 - (a) Start GATE Developer, load the ANNIE application (Main Menu → File → Load ANNIE System → with Defaults).
 - (b) Open the CREOLE Plugin Manager ((Main Menu → File → Manage CREOLE Plugins), and load the Maven plugin with group ID `uk.ac.gate.mimir`, artifact ID `mimir-indexing-plugin` and version `6.0-SNAPSHOT`. Make sure the new plugin is loaded by checking the appropriate check-box.
 - (c) Load a new instance of *Mimir Indexing PR* (Main Menu → File → New Processing Resource → Mimir Indexing PR), and add it to the end of the ANNIE application.

- (d) Make sure that the `mimirIndexUrl` parameter for the new PR is set to the *Index URL* value obtained at Step 5.
 - (e) Load some test documents (e.g. some web pages from news web sites), create a GATE Corpus, add all the documents to the corpus, and set the newly corpus as the target for the ANNIE application.
 - (f) Run the ANNIE application. This will annotate the documents created during the previous step. The Mimir Indexing PR instance will make sure the annotated documents are sent for indexing to your new Local Index.
7. **Search the new index:** as soon as the index has started indexing document, you can use it to search by clicking the *search* link next to the name of your new index. There is a time delay between documents being submitted for indexing and them being available for searching. YOU can speed this process up by manually performing a *sync-to-disk* operation or by reducing the time interval between batches. Both of these actions are available on the index administration page.

To shut down the running web application, focus the command prompt window where you started the application and interrupt it by pressing the `Ctrl-C` key combination. On Windows you may be asked “Interrupt batch job (Y/N)?” – answer “N”.

To deploy the Mimir web application to an application server (such as Apache Tomcat) run the `gradlew assemble` command in the `mimir-cloud` directory. A `mimir-cloud-{version}.war` file will be created for you in the `build/libs` sub-directory. Alternatively, ready-built WAR files are available for release versions from <https://github.com/GateNLP/mimir/releases> and for nightly snapshots from the GATE Jenkins continuous integration server at <https://jenkins.gate.ac.uk/job/Mimir/>.

Chapter 3

Installing and Managing Mimir

3.1 Mimir Architecture

Mimir is divided into a number of related modules.

mimir-core The core Java library to create a Mimir index on disk, add GATE documents to the index, and then query the index once it has been built. Also provides some abstract helper classes for the annotation storage layer, but not the actual storage implementations (which are provided by separate plugins, leveraging the CREOLE plugin framework of GATE Embedded).

plugins/db-h2 The default annotation storage implementation. This stores annotation data using H2¹, an in-process embedded SQL database.

plugins/sparql A helper that can be layered on top of any other storage implementation to provide semantic querying against a separate knowledge base, accessible at a SPARQL endpoint.

plugins/measurements A special-purpose helper for Measurement annotations created by the GATE Tagger_Measurements plugin. Queries are normalised into SI units so can retrieve annotations that express the same measurement in different terms (e.g. an annotation for “90 seconds” would match a query for “1 to 2 minutes”).

mimir-connector Java library to allow indexing of GATE documents into a Mimir index.

mimir-client The client side of the Mimir remote protocol, to support distributed querying.

webapp/mimir-web A Grails² plugin providing both the user interface to create and query indexes over the web, and also the server side of the remote protocol to expose several distributed Mimir indexes as a single *federated* index for clients. This is provided as a plugin rather than an application to make it more easily customisable.

webapp/mimir-cloud An example Grails application, that uses the `mimir-web` plugin and also includes security support. This is the exact implementation used for Mimir

¹<http://h2database.com>

²<http://grails.org>

servers supplied on the GATE Cloud platform³. This application should be suitable without any modifications for most users.

3.2 Building and Running a Mimir Web Application

The `mimir-core` Java library provides support for indexes that are represented as an on-disk directory – named *Local Indexes* in the rest of this document (see the discussion about index types in Section 3.4.1). To get the full functionality of Mimir (including support for *Remote* and *Federated* indexes, as well as user interfaces for system administration and searching indexes) you will need to build and run a web application. All the web elements of Mimir are implemented as the `mimir-web` Grails plugin, which can easily be included in any Grails-based web application. The standard Mimir distribution provides such a web application, named `mimir-cloud`.

3.2.1 The `mimir-cloud` Web Application

The `mimir-cloud` web application is the actual version of the Mimir software that is used on the GATECloud platform. As such, it is configured for that particular usage scenario, where indexes have two different URLs (depending on whether they are accessed from within the same cloud region or not), and where the local configuration page is not made available to the user. However, some of this behaviour is switched off when the application detects that it is not running on the cloud, to allow it to be used as a general purpose Mimir-enabled web application.

In addition to the `mimir-web` plugin, it also includes some basic support for user authentication (using the Spring Security Grails plugin), and support for packaging and downloading local indexes. This is probably a suitable choice if you just need a stand-alone web application with Mimir functionality, and you do not intend to develop your own security solution.

If you are an experienced Grails developer and you intend to add your own security solution, then you should use the `mimir-web` Grails plugin directly in your own application, as described in Section 7.3.

While we include this application as an example of a fully-fledged Grails application using the `mimir-web` plugin, you may need to modify it slightly to make it mode suitable to your actual usage scenario.

3.2.2 Binary Distribution

Starting with version 6.0, a pre-built WAR of `mimir-cloud` is available for download from <https://github.com/GateNLP/mimir/releases>⁴. This can be run using any standard Java servlet container, such as Tomcat or Jetty. WAR files for nightly snapshot builds can be downloaded from <http://jenkins.gate.ac.uk/job/Mimir/>.

³<https://cloud.gate.ac.uk>: a platform for running GATE-based processes on the cloud.

⁴Earlier versions back to 5.2 are available from <https://sourceforge.net/projects/gate/files/>

3.2.3 Prerequisites

To build your own Mimir web application you will need:

- A Java 8 or later JDK. Mimir has been tested with the Sun/Oracle and OpenJDK JVMs on Linux and Mac OS X.
- Apache Maven 3.3 or later

While not strictly a pre-requisite, Mimir performs much better on 64-bit systems than on 32-bit ones, partly due to simply being able to assign more memory to the process, but also because the larger address space allows MG4J to memory-map many of the files that make up the index.

To run a local instance of Mimir you can use the standard `./grails run-app` command, but to deploy a production instance you will need a separate servlet container such as Tomcat.

3.2.4 Building

Building Mimir is a two step process. The core Java library, remote connector and plugins are built using Maven, the web components (which depend on the core) are built using Grails 3.3 via the included wrapper script – you do not need a separate installation of Gradle or Grails.

There is a top-level Maven pom file to build all the core components, you can run this with the standard `mvn install` command. Once the core components are built you can run the `mimir-cloud` web application using `./grails run-app` in the `mimir-cloud` directory.

The next step is to configure the `mimir-cloud` web application, and is described in the following section.

3.2.5 Configuring

When the Mimir Grails plugin is installed into a Grails application, the `./grails mimir-config` command adds a base set of configuration options to the application's `grails-app/conf/application.groovy`. This file contains a number of settings that affect the running of the Mimir components. In many cases the default options will be sufficient, but you should nevertheless check the configuration and make sure it is appropriate for your needs.

You can modify the `application.groovy` file directly, or you can use the “external configuration” mechanism in Grails to override these settings at runtime in a separate `application.groovy` in the working directory.

```
1 gate {
2     mimir {
3         plugins {
4             h2 {
5                 group = "uk.ac.gate.mimir"
```



```
6     artifact = "mimir-plugin-dbh2"
7     }
8     myCustomPlugin = "file:/data/mimir/plugins/myCustomPlugin"
9     }
10  }
11 }
```

Mimir uses the GATE plugin mechanism to allow different annotation helpers to be added dynamically at runtime. This section of the configuration specifies which plugins should be loaded when the web application starts up, and determines the kinds of annotation helpers you will be able to use in your indexes. You generally need at least the standard `db-h2` plugin to be able to do anything useful with Mimir, and you may want the `measurements` plugin as well if you will be searching on Measurement annotations and/or the `sparql` plugin if you have an external knowledge base. Section 4.1 has more information about the standard annotation helpers, and section 7.1 discusses how to implement your own custom ones.

Plugins can be specified in one of three ways:

- as a set of Maven co-ordinates “group”, “artifact” and an optional “version”, which will load the plugin from Maven repositories in the normal way. If “version” is omitted it defaults to the version of the `mimir-web` plugin, so the plugin versions that match this build of Mimir. The standard plugins are listed (commented out) in the sample configuration created by `mimir-config`, uncomment the ones you need.
- as an absolute URL string (the `myCustomPlugin` example above), which will load a directory-based GATE plugin from the given location.
- as a relative path string, which will be resolved against the web application root to form a URL which is then loaded as a directory-based plugin.

Maven-style plugins are typically resolved at runtime by downloading from repositories over the network, but `mimir-web` provides a utility to cache the plugins that are referenced by a configuration, along with their transitive dependencies, so they can be loaded from within the web application rather than downloaded from the internet (this is useful if, for example, your Mimir is running inside a locked-down firewall and cannot make outgoing web connections). To do this, once you have configured your plugins but before you build your WAR file, run `./grailsw run-command cache-mimir-plugins` in your `mimir-cloud` directory. This will create a cache in `src/main/webapp` which will be bundled into your WAR file when you build it.

```
1 // ...
2 queryTokeniserGapp =
3     "file:/path/to/query-tokeniser.xgapp"
```

Whereas GATE’s usual data model deals with annotations in terms of their *character* offsets from the start of the document, Mimir deals in terms of *tokens*. Queries for plain text strings in Mimir must be tokenised before they can be matched against the index, and the tokenisation applied to the queries must match that applied to the documents that have been indexed. The Mimir Grails plugin uses a saved GATE application state (gapp file) to

perform query tokenisation, the location of which is specified here. By default, Mimir uses the standard ANNIE English tokeniser with its default parameters.

If your tokenisation requirements are more complex, you can provide your own saved application that will produce annotations of type `Token` in the default annotation set. To bundle your tokeniser gapp inside the WAR file, put it under `src/main/resources` and use a “classpath:” URL, i.e. if your saved application is `src/main/resources/custom-tokeniser.xgapp` then set the `queryTokeniserGapp` parameter to `classpath:custom-tokeniser.xgapp`

3.2.6 Running

The easiest way to run the Mimir cloud web app is to use the normal Grails commands `./grailsw run-app`. However, for performance, it is preferable to create a WAR file using `./gradlew assemble` and run from that (using a tool such as `webapp-runner`) For anything more than the smallest toy index it is advisable to increase the memory available to Mimir by using the `-Xmx` option to your `webapp-runner` command.

For production deployments, a better option is to deploy the WAR file to a standalone servlet container such as Apache Tomcat. If you are using Ubuntu or Debian GNU/Linux, it is better to download the standard Tomcat ZIP package from Apache and use that rather than installing the Tomcat available through `apt-get` as the latter is configured by default with a security manager that interferes with Mimir.

When deployed to a servlet container the web application reads configuration at runtime in the standard Grails way, looking for files named `application.yml` or `application.groovy` in the working directory of the servlet container’s Java process. Any options specified in these files override the defaults from the `application.groovy` inside the WAR.

3.3 Upgrading from earlier versions of Mimir

Where possible Mimir tries to remain compatible with indexes and databases created by earlier versions of the software. Where this is not possible (e.g. due to changes in the on-disk format between versions, or revisions of the underlying libraries that cause binary incompatibility), a facility is provided to “upgrade” existing indexes to the latest format. If you attempt to use an old-format index in a new version of Mimir the index will show in the web interface as “failed”, but if you click the “details” button on the index information page you will see a link offering to upgrade the index to the latest version. The upgrade process is reasonably robust but **it is very strongly recommended to back up all the index files on disk before attempting to upgrade.**

Depending on the size of the index the upgrade process may take some time, when it completes the index will revert to the “ready” state and will then be available for searching or further indexing as normal.

3.4 Indexes in Mimir

3.4.1 Types of Index

A single instance of Mimir can host several indexes. Mimir supports *local* indexes, stored in the file system of the Mimir server, and *remote* indexes, which are a view of an index hosted in another Mimir instance (possibly on a different machine). Several indexes (of any type) can be combined into a *federated* index, which presents the group of indexes as a single virtual index. All the indexing and searching functionality of Mimir applies equally to all three index types.

Each Mimir index has a *state*, and the operations that can be performed on the index depend on which state it is currently in. Indexes spend most of their time in the *ready* state, when they can index new documents and answer queries. During various operations they may temporarily be in a different state, such as *closing* while the index is being shut down, typically because the Mimir server is itself being shut down. Sometimes a local index is *failed*, indicating a problem with the index. Typically a failed index will need to be deleted by the administrator, though it may be possible to recover most of the data using the index repair tool (see section 8.1).

Remote indexes inherit their state from the remote server, and federated indexes inherit their state by combining the states of their component indexes. A federated index may occasionally appear in the *working* state if its component indexes are not all in the same state, but the working state will usually resolve to a normal state once the component indexes have synchronised.

A typical setup for a large-scale indexing task would be to have a number of identical “slave” servers running Mimir, each with a single local index. A single “master” Mimir instance could then have one remote index definition pointing to each of the slaves, and a single federated index combining the remote indexes. This federated index would be the point of entry into the system and would share out indexing jobs (round-robin among the slaves) or search requests (to all the slaves in parallel) as appropriate.

3.4.2 Creating a Local Index

Indexes in Mimir are managed through the web interface. The front page of a newly-installed Mimir is shown in Figure 3.1. The *index templates* mentioned at the bottom, are used to define the properties of new indexes, and are described in more detail in Chapter 4. The Mimir Grails plugin provides a single example template based on ANNIE annotation types.

To create an empty local index ready to receive documents for indexing, select the *create a new local index* link. This will present a form (Figure 3.2) asking for the name of the new index and the template from which it should be created. The “Document URIs are external links” option affects the way documents are presented in the search interface. Every document in Mimir is identified by a URI, and if you intend to use document URIs that are actually resolvable URLs (for example if your documents came from a web crawl) then you should select this option to add a link to the original document to the search results. If the document URIs will not be resolvable URLs then leave the option un-selected. The index will be assigned a unique identifier and a new directory will be created under the



Figure 3.1: The default front page of a new Mimir

`indexBaseDirectory` you configured earlier to hold the index data. The newly-created index will start in the *ready* state (see Figure 3.3), ready to receive documents for indexing. For details of how to submit documents to the index, see Chapter 4.

This *index information* page can be accessed at any time by clicking the link for the relevant index name from the Mimir front page (Figure 3.4). At any time, the index can then be searched using the tools described in Chapter 5. Recently added documents only become available for searching after a *sync-to-disk* has taken place. Sync operations happen automatically at regular intervals, or can be triggered by the user by pressing the *Sync to Disk* button seen at the bottom of the index information page.

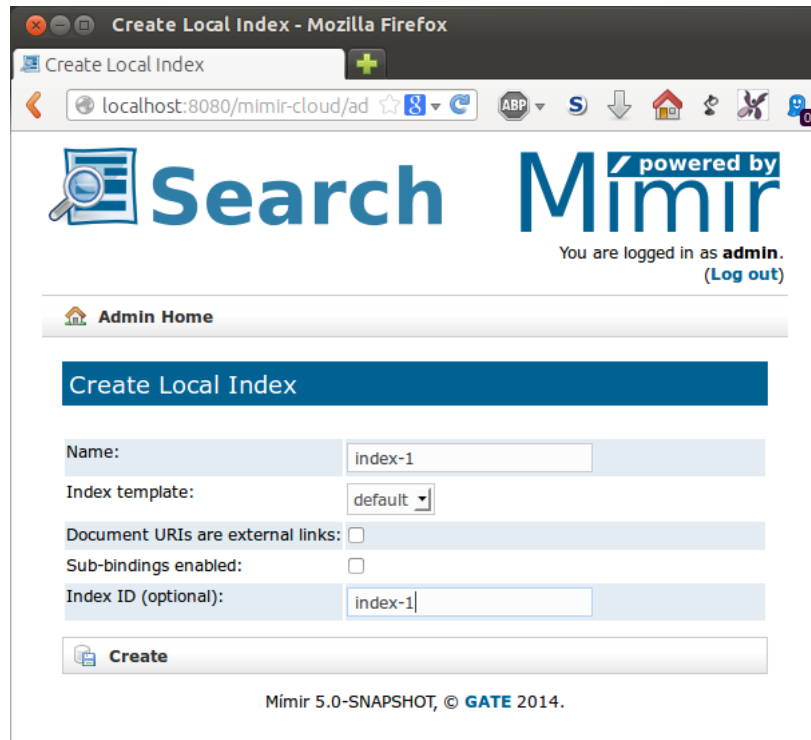


Figure 3.2: Creating a new local index

3.4.3 Working with Remote and Federated Indexes

The architecture of Mimir is designed to make working with remote and federated indexes as transparent as possible. The setup process will obviously vary for the different index types, but once created the process of submitting documents for indexing or of performing queries is exactly the same for all indexes.

Remote indexes

A *remote* index is a mechanism whereby one Mimir instance can transparently index documents in, or send queries to, an index that is located in a different Mimir instance, typically running on separate hardware. To connect one *master* Mimir instance to an index running in another *slave* instance, first visit the index information page for the relevant index on the slave and make a note of its *remote URL* (typically a URL of the form `http://server:port/mimir/remote/{UUID}`). Now on the front page of the master instance, select the *connect to a new remote index* link. This will present a form (Figure 3.5) asking for a name for the remote index (which need not be the same as the name of the index on the slave), and a *remote URL* which is the one you made a note of from the slave above. You should never create a remote index pointing to another index in the same Mimir instance. Such a configuration is not supported and will lead to errors!

The remote index defined on the master server will synchronise its state with that of the

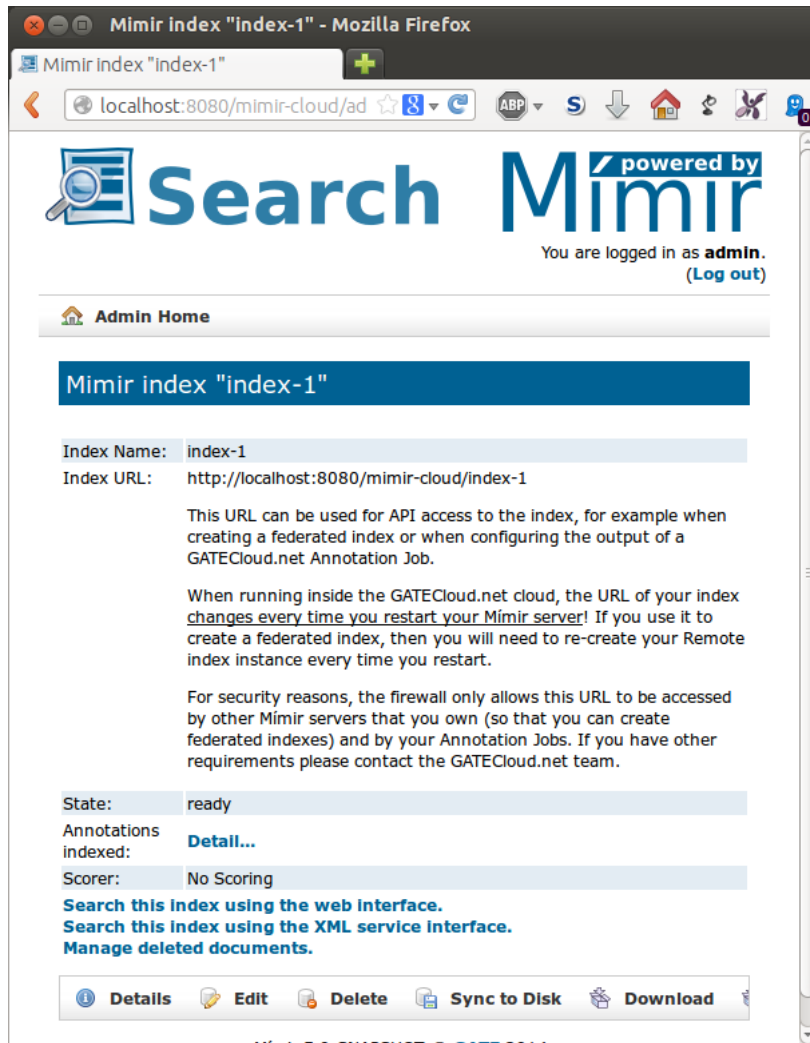


Figure 3.3: Results of creating a new local index

Local Indexes (?)

The following local indexes are configured:

1. [index-1 \(search\)](#)

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

Figure 3.4: List of local indexes on the Mimir front page

underlying index on the slave, and once created will be usable exactly like a local index. However remote indexes are rarely used directly, as in most cases it is more efficient to operate on the slave instance itself. The main benefit of remote indexes comes when they are used as part of a *federated index*.

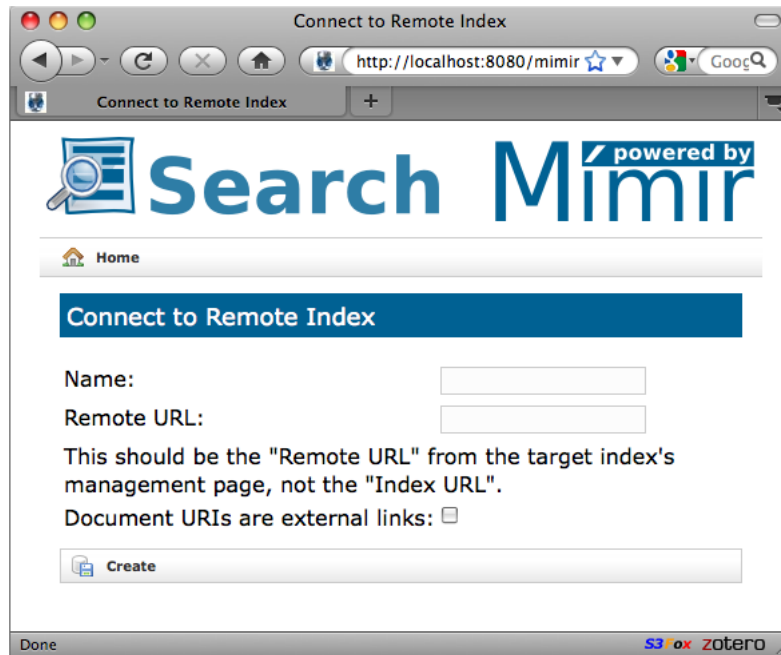


Figure 3.5: Connecting to a remote index

Federated indexes

A *federated index* is a device to bundle several indexes (which can themselves be local, remote or federated) together so they can be used as a single index. Documents for indexing are shared out between the component *sub-indexes*, and searches are performed by all sub-indexes in parallel. Thus a federation of five indexes each containing 200,000 documents will typically run queries faster than a single index containing 1 million documents. To create a federated index, go to the Mimir front page and select the *create a new federated index* link. This will present a form (Figure 3.6) asking for a name for the federated index. The form also includes a multiple-selection list to specify the sub-indexes to be included in the federated index. Select the appropriate entries from this list using the usual multiple list selection mechanism (ctrl-click on Windows or Linux, cmd-click on Mac OS X) and press the *Create* button to create the index. Once created the federated index will be usable exactly like a local or remote index.

3.4.4 Deleting Indexes

If an index registered with Mimir is no longer required it can be deleted by selecting the *Delete* button from the index information page (accessible by clicking on the name of the relevant index on the Mimir front page). For remote and federated indexes this simply deletes the “registration” of the index with Mimir, which can be easily re-created as above. For local indexes it also offers the option to delete the underlying index files from disk. If a local index is deleted without deleting the disk files then the index can be re-created later using the *import an existing index for searching* option from the Mimir front page.



Figure 3.6: Creating a new federated index

Mimir will not allow the deletion of an index which is currently part of a federated index in the same Mimir instance. To delete such an index, it must first be removed from the federated index. This guarantee only applies to indexes within a *single* Mimir instance — Mimir does not prevent the deletion of an index on a slave instance which is being used as a remote index by a master instance (it prevents the deletion of the remote index definition in the master but not the slave index it points to). However to do so would put the remote index on the master (and hence any federated index that it is part of) into the *failed* state, preventing further use until the problem is resolved.

3.5 “Deleting” Documents from a Mimir Index

While Mimir indexes are not directly modifiable once they have been created, there are situations in which it is necessary to remove documents that should not have been indexed in the first place, or documents that may be considered libellous, etc. To support this, Mimir provides a mechanism to mark individual documents in the index as “deleted”, and any documents so marked will be excluded from future queries. It is not possible to completely delete the data from the index files on disk, short of completely re-building the index from scratch, but documents marked as deleted are not accessible through any of the public Mimir APIs or user interfaces.

To mark a document as deleted (or to remove an existing deletion marker, making the document available for queries again), use the “Manage deleted documents” link from the index’s administration page. This will present the screen shown in figure 3.7, with a text box into which you can type one or more (space-separated) document IDs, and choose

whether to mark them as deleted or as “not deleted” (i.e. to remove any existing deletion markers for those document IDs).

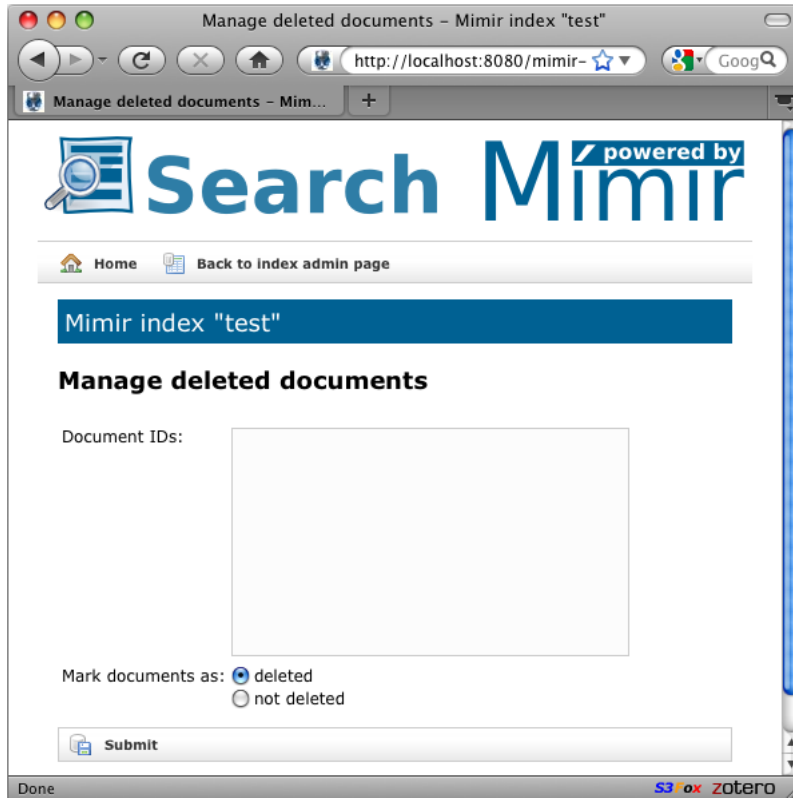


Figure 3.7: Managing deleted documents

Note that the IDs required here are not the URIs that were provided with the documents when they were indexed, but the internal Mimir document IDs which are numbers starting from 0, as returned in the hit lists and “getDocumentId” by the Mimir query APIs (see section 5.2.1).

Chapter 4

Indexing Documents with Mimir

Mimir is designed to index semantically annotated documents. It accepts as input GATE documents¹ and produces a set of indexes as a result. The way the text and annotations of the input documents are converted into indexes is controlled through configuration options.

4.1 Configuring the Indexer

In the Mimir web interface, the configuration of a new index is represented by an *index template*. This specifies:

- which annotation types and features to index
- which annotation sets contain these annotations
- (optionally) which document features should be indexed
- how to handle the document format and metadata

Index templates can be managed using the “Click here to manage the index templates” link at the bottom of the Mimir front page. An index template is specified in a structured “domain specific language” using Groovy — Listing 4.1 shows an example of the default template provided by the Mimir Grails plugin.

¹<http://gate.ac.uk/userguide/chap:corpora>

The various sections of the template are as follows:

Imports

```
1  import gate.creole.ANNIEConstants
2  import gate.mimir.SemanticAnnotationHelper.Mode
3  import gate.mimir.index.OriginalMarkupMetadataHelper
4  import gate.mimir.db.DBSemanticAnnotationHelper as DefaultHelper
5  import gate.mimir.util.DocumentFeaturesMetadataHelper
```

The template can optionally start with import statements to import any Java classes that are used further on in the template.

Token Definitions

```
7  tokenASName = ""
8  tokenAnnotationType = ANNIEConstants.TOKEN_ANNOTATION_TYPE
9  tokenFeatures = {
10     string()
11     category()
12     root()
13 }
```

The next section of the template deals with the *tokens* that Mimir will base its index on. Mimir sees every document as a stream of tokens rather than a stream of characters, and all the annotations indexed by Mimir are stored in terms of their starting and ending *tokens*, not character offsets. Thus for Mimir to work correctly it needs to know how to split up the document into tokens and what information to store about each token. For this purpose it uses GATE annotations, and indexes the values of features on the annotations.

The following options can be configured:

tokenASName The name of the annotation set in which the token annotations can be found (for example `tokenASName = "mimir"`). To use the default annotation set, which has no name, this may be left unspecified, or explicitly set to the empty string: `""`, or to `null` (without quotes).

tokenAnnotationType The annotation type that should be used as tokens. This entry is required, and can generally be simply set to the default `ANNIEConstants.TOKEN_ANNOTATION_TYPE` (with a suitable `import` at the top of the template).

tokenFeatures A block of code giving the features from each token annotation that should be indexed.

The *tokenFeatures* block should list the features to be indexed as shown in the example, each feature name followed by parentheses. For advanced users an `MG4J TermProcessor` instance may be provided inside these parentheses. By default, if no term processors are specified, the *first* feature is converted to lowercase and the subsequent features are not modified. Since terms in a query are processed using the same processor as those

in the index, this has the effect of making searches on the first feature case-insensitive, and searches on the other features case-sensitive. To stop any processing being done, you should supply a `it.unimi.dsi.mg4j.index.NullTermProcessor` value, by specifying e.g. `string(NullTermProcessor.getInstance())`, after including the relevant `import` statement at the top.

Semantic Annotations

```

15 semanticASName = ""
16 semanticAnnotations = {
17     index {
18         annotation helper:new DefaultHelper(annType: 'Sentence')
19     }

```

The next section defines the *semantic annotations* that Mimir will include in the index. Each `index` block defines a set of semantic annotation types that will be indexed and stored together in one index. The choice of how to group annotation types together into indexes can affect the indexing speed, as the annotations within one index are processed sequentially by a single thread, whereas types in separate indexes are indexed in parallel.

Each annotation type to be indexed is introduced by “`annotation`”. This is a method call in the Groovy DSL which takes the following named arguments:

helper The *semantic annotation helper* Java object that should be used to index this annotation type.

type The annotation type that is to be indexed. When using the default semantic annotation helper types this can be omitted.²

Semantic Annotation Helpers

Semantic annotations are stored in special indexes that associate URIs with document positions. During indexing, the role of the helper implementations is to store the necessary information about each annotation to be indexed in a persistent form and return one or more URIs that identify it.

One could make a distinction between *generic* semantic annotation helper types, which can be configured to handle any annotation type and features, and *special-purpose* helpers that are designed to handle specific annotation types. Mimir supplies a generic helper implementations in the `db-h2` plugin that store annotation information in a relational database. For the most standard cases, this default helper implementation should be sufficient. One sample special-purpose helper for `Measurement` annotations (as generated by the GATE `Tagger_Measurements` plugin) is also provided, in the `measurements` plugin. This is intended both to be useful in its own right and to serve as a template for how to implement your own helpers for other complex annotation types. The `sparql` plugin provides a helper that can wrap any other helper and add the ability to query for URI-valued features by making a query to a SPARQL endpoint.

²In particular, if the specified helper has a method “`getAnnotationType()`” then this will be called and the returned value used as the annotation type. All the standard helpers provided with Mimir extend `AbstractSemanticAnnotationHelper` which implements this method.

The plugins that include all the provided semantic annotation helpers are discussed in detail in Chapter 6. Detailed documentation for configuring each of the helpers is available there.

Note for users upgrading from Mimir 3.2.0 and earlier: the previous index template DSL style using the annotation type as the method name and the `nominalFeatures` etc. as parameters is still supported but should be considered deprecated. You should consider porting your index templates to the new style, as support for the old style may be removed in a future release.

Document Features

Starting with Mimir version 3.4.0, annotation helpers can also be used to index document features.

```

1  import gate.creole.ANNIEConstants
2  import gate.mimir.SemanticAnnotationHelper.Mode
3  import gate.mimir.index.OriginalMarkupMetadataHelper
4  import gate.mimir.db.DBSemanticAnnotationHelper as DefaultHelper
5  import gate.mimir.util.DocumentFeaturesMetadataHelper
6
7  tokenASName = ""
8  tokenAnnotationType = ANNIEConstants.TOKEN_ANNOTATION_TYPE
9  tokenFeatures = {
10     string()
11     category()
12     root()
13 }
14
15 semanticASName = ""
16 semanticAnnotations = {
17     index {
18         annotation helper:new DefaultHelper(annType: 'Sentence')
19     }
20     index {
21         annotation helper:new DefaultHelper(annType: 'Person',
22                                             nominalFeatures:["gender", "title"],
23                                             textFeatures:["name"])
24         annotation helper:new DefaultHelper(annType: 'Location',
25                                             nominalFeatures:["locType", "
26                                                         continent"])
27         annotation helper:new DefaultHelper(annType: 'Organization',
28                                             nominalFeatures:["orgType"])
29         annotation helper:new DefaultHelper(annType: 'Date',
30                                             integerFeatures:["normalized"])
31         annotation helper:new DefaultHelper(annType: 'Document',
32                                             integerFeatures:["date"], mode:Mode.
33                                             DOCUMENT)
34     }
35 }
36 documentRenderer = new OriginalMarkupMetadataHelper()
37 documentFeaturesHelper = new DocumentFeaturesMetadataHelper("date",
38                                                         "source", "id", "type")
39 documentMetadataHelpers = [documentRenderer, documentFeaturesHelper
40 ]
41
42 // miscellaneous options - these are the defaults
43 //timeBetweenBatches = 1.hour
44 //maximumBatches = 20

```

Listing 4.1: The default index template provided with Mimir

```
25      annotation helper: new DefaultHelper(annType: 'Document',
                                           integerFeatures: ["date"], mode: Mode.
                                           DOCUMENT)
```

The above declaration (note the `mode` parameter!) creates a new Semantic Annotation Helper that uses the document features instead of the features from any given annotation. The helper behaves as if a single annotation, of the declared type (in our case *Document*), existed that covers the whole extent of the document, and has the same features as the GATE Document being indexed.

Things to note:

- All helper implementations supplied with Mimir are capable of working in DOCUMENT mode, so you can use them for indexing document features.
- The default value for the `mode` parameter for all supplied helper implementations is ANNOTATION. Not specifying a mode value preserves the default functionality (from versions preceding 3.4.0), i.e. indexing **annotation** features.
- You can have as many helpers as you want working in DOCUMENT mode, in parallel.
- The specified value for the `annType` parameter is used by the helper for simulating the presence of an actual annotation spanning the whole document; Mimir then behaves as if such annotations actually existed. This implies that you cannot re-use the name of an annotation type that is already being indexed. For example, if you are already indexing actual annotations of type *Document*, then you will need to choose a different name for the virtual annotation type used when indexing document features.

Document Rendering and Metadata

```
28      documentRenderer = new OriginalMarkupMetadataHelper()
29      documentFeaturesHelper = new DocumentFeaturesMetadataHelper("date",
                                                                    "source", "id", "type")
30      documentMetadataHelpers = [documentRenderer, documentFeaturesHelper
                                  ]
```

The next part of the template concerns how document-level metadata is indexed, and how this can be combined with the document text to render the document content at search time, with matches of the query highlighted. These tasks are performed by objects that implement the interfaces `DocumentMetadataHelper` and `DocumentRenderer` respectively (both in the `gate.mimir` package). Mimir provides a single class `gate.mimir.index.OriginalMarkupMetadataHelper` which implements both interfaces, so in most cases the same object can be used for both jobs.

An index template must define one `documentRenderer` and may define any number of `documentMetadataHelpers` (in a square-bracketed list). If the renderer is an `OriginalMarkupMetadataHelper` (or a subclass) then the renderer object must be included in the list of metadata helpers in order to function correctly. Other metadata helpers may be added to the list if required.

In the listing above, we use one instance of `OriginalMarkupMetadataHelper` as a document renderer. To enable it to function, we also include the same object instance in the list of metadata helpers. Additionally, we also construct an instance of `DocumentFeaturesMetadataHelper`, which we name `documentFeaturesHelper`, and we then add to the list of metadata helpers. `DocumentFeaturesMetadataHelper` instances can be used to store additional metadata in the index being constructed. The metadata values to be stored must be provided in the form of GATE document features on the documents being indexed. When such values are present at indexing time, they are serialised and stored in the index. At search time, the stored metadata fields can be retrieved back from the index. Note that the values used must be serialisable to be usable (i.e. they must implement the `java.io.Serializable` interface).

Miscellaneous options

```
32 // miscellaneous options - these are the defaults
33 // timeBetweenBatches = 1.hour
34 // maximumBatches = 20
```

Finally, additional miscellaneous options can be specified at the end of the template. The supported options are:

timeBetweenBatches the maximum amount of time that the indexer should wait between writing batches to disk. Since only batches that have been dumped to disk are searchable, this specifies the maximum time a document should be held in RAM after having been submitted for indexing but before it becomes available to be searched. The value can either be a plain number (of milliseconds) or a Groovy `TimeCategory` duration expression such as `10.minutes`. If unspecified, the default is one hour (3600000 milliseconds). Note that it is always possible to force the system to dump the current batch to disk immediately via the index administration page.

maximumBatches the maximum number of constituent batches before a compaction operation is triggered. The default is 20, and it should rarely be necessary to modify this as index compaction is transparent – the index behaves exactly the same whether or not it has recently been compacted³.

Direct Indexes

Starting with version 5.0, Mimir can build direct indexes as well as inverted ones. By default only inverted indexes are created, which are used to associate terms to documents. Direct indexes encode the inverse relation from documents to terms, hence a direct index can be used to find out which terms occur in any given document.

To enable direct indexes for tokens, the configuration in the index template needs to be modified like in the following example:

```
1 tokenFeatures = {
2   string(directIndex:true)
3   category(directIndex:true)
4   root()
```

³The main difference is that a compacted index requires fewer open file handles to operate.


```
5 }
```

In this case, direct indexes will be built for the `string` and `category` features of `{Token}` annotations, but not for the `root` feature.

When building a direct index for a feature that is not present on every token, tokens that do not have the feature will be reported as “null” in terms query result sets. In many cases this is not desirable (for example if only a small minority of tokens have the feature, the nulls will swamp the real values when calculating statistics), and for these situations Mimir offers a custom MG4J term processor that will omit entirely any token where the feature is missing or empty. You can configure this using

```
1 import gate.mimir.util.IgnoreEmptiesTermProcessor
2
3 tokenFeatures = {
4     featureName(directIndex:true,
5         termProcessor:IgnoreEmptiesTermProcessor.INSTANCE)
6     // ...
7 }
```

In the case of semantic annotations, direct indexes are enabled in a similar fashion:

```
1 index(directIndex:true) {
2     annotation helper:new DefaultHelper(...)
3     ...
4 }
```

Note that direct indexes can only be enabled at the level of a `index` element in the template, and not for individual annotation types.

Direct indexes are stored in separate files from the default indirect indexes, so they will not affect the functionality that does not require direct indexes at all.

Direct indexes can currently only be searched via the Java API provided by the `gate.mimir.search.terms` package.

4.2 Adding Documents to an Index

Once an index has been created, the next stage is to add documents to the index. Mimir provides an HTTP API for this which accepts documents for indexing via HTTP POST requests that include the document in Java serialised format. The easiest way to make use of this API is via GCP (the GATE Cloud Paralleliser batch processing tool) using a `MimirOutputHandler`. This GCP output handler makes use of the `gate.mimir.index.MimirConnector` (in the `mimir-connector` module) to actually make the remote call, and you can use the same API in your own code. To add a GATE document to an open index simply call:

```
1 MimirConnector.defaultConnector().sendToMimir(document, uri,
2     indexUrl);
```

... with the following parameters:

document a `gate.Document` for indexing.

uri the URI that should be used to identify the document in the Mimir index. May be `null`, in which case Mimir will generate a URI, but in most cases there will be a more meaningful identifier that could be used.

indexUrl a `java.net.URL` pointing to the location of the Mimir index. This is the “Index URL” given on the index information page.

The document to be indexed must, of course, contain the token and semantic annotations that the index expects.

It is possible to create your own private instance of `MimirConnector` rather than simply using the default one, but this is not necessary in normal use.

4.3 The Default Representation Scheme

The default generic SAH implementations try to minimise the amount of data stored in their underlying database or semantic repository by creating representation templates that are shared between all occurrences of annotations with the same values for the features. There are two levels of templates, the first defined by the values of nominal features, and the second that uses the values of all the other features. This is intended to reflect the typical scenario where most annotations are defined by a small set of nominal features, with a few of them having features with arbitrary values. Most annotation types would then only make use of level-1 templates, with a few of them employing both level-1, and level-2 templates.

For each input annotation the following IDs are retrieved (or generated on first occurrence):

Level-1 template ID The annotation type and the values for all its nominal features form a tuple. The first time each tuple configuration is seen, it is allocated a level-1 ID. Subsequent annotations that match an already existing tuple will re-use the same level-1 ID. For example, in Figure 4.1 all annotations of type *Location* with feature *city* will use the level-1 ID ‘1’.

Level-2 template ID The level-1 template ID together with the values for all the remaining (i.e. non-nominal) features form a second tuple. Unique configurations of these tuples are allocated level-2 IDs. It should be noted that most NLP annotations tend to include only nominal features, so they would not require a level-2 ID. The `{Location}` annotations shown in Figure 4.1 have a non-nominal feature, so they each get a level-2 ID allocated to them. Note, however, that all further mentions of e.g. the *Thames* would re-use the same IDs, even when phrased differently in the text, e.g. “*the river Thames*”, or “*La Tamise*”.

Mention ID The level-1 ID and the annotation length (number of tokens) forms a tuple, which is associated with a mention ID – in figure 4.1 *Location* annotations with feature *city* covering one token will take the mention ID “Location:1”. If present, the level-2 ID and the annotation length also get a mention ID. For example, all mentions of “the River Thames” are associated with the mention ID “Location:5” (because they refer to the Thames, and are 3 tokens long).

Finally, the one or two mention IDs associated with each annotation are added to an *annotation index*, using the annotation start token as the position.

We index two separate mention IDs associated with either level-1 or level-2 IDs, in order to speed-up searches that only make use of nominal features. For annotation types that

have non-nominal features, the number of level-2 IDs will be orders of magnitude greater than that for level-1. If a search only relies on nominal constraints (a large proportion of searches tend to fall into this category), then the query can be answered much faster by only accessing the smaller number of posting lists for the matching level-1 IDs.

Document:	London	is	located	on	the	Thames	.
position:	0	1	2	3	4	5	6
string:	london	is	located	on	the	thames	.
root:	london	be	locate	on	the	thames	.
part-of-speech:	NNP	VBZ	VBN	IN	DT	NNP	.
Location:	type=city					type=river	

Token indexes

root index	
.	0(6)
be	0(1)
locate	0(2)
london	0(0)
on	0(3)
thames	0(5)
the	0(4)

Location templates

L1 ID	type
1	city
2	river

L2 ID	L1 ID	instURI
1	1	dbpedia.org/resource/London
2	2	dbpedia.org/resource/Thames_river

Mention ID	L1 ID	L2 ID	length
Location:1	1	-	1
Location:2	1	1	1
Location:3	2	-	1
Location:4	2	2	1
Location:5	2	2	3

PoS index	
.	0(6)
DT	0(4)
IN	0(3)
NNP	0(0, 5)
VBN	0(2)
VBZ	0(1)

Location index

{Location} index	
Location:1	0(0)
Location:2	0(0)
Location:3	0(5)
Location:4	0(5)

Figure 4.1: A very simple example document and the corresponding contents of a Mimir index. We assume that the only document ID is 0.

Different *views* of the document text are generated by different token features, which are stored in separate sub-indexes. The document string has been down-cased prior to indexing; we do not show the **string** index, as it is very similar to the one for the **root** feature. The values used for Part-of-Speech (PoS) are standard tags as produced by GATE's PoS Tagger: DT=determiner, IN=preposition, NNP=proper noun, VBN=verb - past participle, VBZ=verb - 3rd person singular present.

A single annotation type (**{Location}**) is being indexed, with two different occurrences, and we assume the only non-nominal feature to be the DBpedia instance URI. Note that "Location:5" (i.e. a mention of the Thames that is 3-tokens long) does not actually occur in the document text, so it is not present in the index. We have included it here as an example of an annotation of length greater than 1.

Chapter 5

Searching Mimir Indexes

From a user's point of view, Mimir is a tool for searching a collection of semantically annotated documents. It provides facilities for searching over different views of the document text, for example one can search the document words, the part-of-speech of those words, or their morphological roots. Beside searching the document text, Mimir also supports searches over the documents' semantic annotations, where queries are based on annotation types and restrictions over the values of annotation features. These different search paradigms can be combined freely into complex queries, with support for sequences, repetitions, and Boolean operators.

A search session entails the formulation of a query, running the query with the Mimir query engine, and consuming the query results. Mimir queries are expressed in a text-based query language which is described in section 5.1. The way these queries are submitted to Mimir depends on how it is deployed, the various interfaces are discussed in section 5.2.

5.1 The Mimir Query Language

A Mimir query is either a simple query (i.e. a `String` query, section 5.1.1, or an `Annotation` query, section 5.1.1), or a compound one, which comprises a set of sub-queries linked by operators. If no operator is placed between any two sub-queries, then the `Sequence` operator (see section 5.1.7) is implied. This means that several queries written one after another are interpreted as one sequence query. For example, a query like *'the brown dog'* is interpreted as a sequence query, having three sub-queries, each of them being a `String` query. This would match occurrences of the exact phrase *'the brown dog'* in the indexed documents. Note that this is different from the standard behaviour of search engines, which would simply match documents in which all three query terms occur, in whichever order. That type of search is also supported in Mimir, through the `AND` operator (see section 5.1.2). Parentheses can be used for grouping where the syntax would otherwise be ambiguous.

5.1.1 String Queries

The simplest form of query is a query term. This will match all occurrences of the query term in the indexed documents.

If the Mimir index being interrogated includes multiple token indexes, then the particular index to be searched can be specified by prefixing the query term with the index name and a colon, for example the query `'root:be'`¹ will match all morphological forms of the verb *to be*. If the name of the string index is omitted, then the first configured index is used. By convention (reflected in the default Mimir configuration) the first string index is used to store the terms text, so the default behaviour is to search over the document text, as expected. Double-quoted strings are treated as plain term queries against the first token index in a similar way.

In fact the above is a slight simplification, as bare terms (and double-quoted strings) are actually tokenised before being searched for. This is because Mimir views documents as streams of tokens, not characters, and the query must match the tokenisation that was used to index the documents. For example, the default GATE tokeniser treats “don’t” as two tokens, “do” and “n’t”, so a query for *don’t* as a single token would fail. To get around this, Mimir runs a GATE application over the string of the query to generate Token annotations, and then constructs a query for these tokens in sequence (see section 5.1.7). Named index queries (“root:be”) are not tokenised, so if you want to avoid tokenising a particular query for any reason (e.g. if you suspect there is a mis-tokenised document in your index) you can explicitly name the appropriate index (typically “string”, i.e. `string:don't`).

Annotation Queries

If annotation indexes were used during indexing, Mimir allows searching for annotation-based patterns. An annotation is a piece of metadata associated with a text segment, with a **type** and optionally **features**. An annotation query takes the following form: `{Type feature1=value1 feature2=value2 ...}`. The annotation type is required, the feature constraints are optional.

While the example above uses equality for the feature constraints, other operators are also available. Here is the full list:

equality: represented by the sign `=`, matches annotations which have the given value for the specified feature. The equality operator is applicable to features of any type.

comparison operators: represented by one of the following symbols: `<`, `<=`, `>`, `>=`, with the usual meaning. These operators can apply to features of type `nominal`, `numeric`, or `text`.

regular expressions: can be specified using the syntax `REGEX(pattern, flags)`, where the `pattern` represents the regular expression sought, and the `flags` are optional, and can be used to change the way matching is performed. See <http://www.w3.org/TR/xpath-functions/#regex-syntax> for a full specification of the regular expression support. The REGEX operator can only be used for `nominal`, and `text` features.

Examples:

¹This assumes that an index named `root` exists, and was used to store the morphological root of the words.

{Person gender = female} – searches for annotations of type `Person`, which have a feature named `gender`, with the value *female*.

{Measurement type = scalar normalisedValue > 0 normalisedValue < 10 normalisedUnit = m} – searches for scalar measurements, with a unit of *metre*, and a normalised value between 0 and 10². Note that the same feature name can be used in several constraints, in which case only annotations where the feature meets *all* the constraints will be matched by the query. For disjunctive queries, use the OR operator described below.

In order to be able to search on a particular feature, that feature must have been specified in the index template when the index was created – Mimir indexes only the features it has been told to index. There may be additional “synthetic” features available at query time depending on the semantic annotation helper that the index uses for the given annotation type, for example the SPARQL helper allows queries on the “feature” named “sparql”, the measurements helper allows queries for “spec” etc.

5.1.2 AND Operator: “&”

The ‘AND’ (also ‘&’) operator can be used to specify queries that should match document segments that include at least one hit from each of the sub-queries. The results returned will always be the shortest document segments that satisfy the query.

5.1.3 OR Operator: “|”

OR queries are used to search hits that match one of a set of alternative query expressions. This is indicated by using the ‘OR’ (also ‘|’) operator between the sub-queries. A query of the form `Query1 | Query2` will return hits that match either sub-query `Query1` or sub-query `Query2`.

5.1.4 IN and OVER Operators

The operators `IN` and `OVER` are used to search for hits of a query that contain, or are contained in the hits of another query. For example:

`Query1 IN Query2` will match all the hits of `Query1` that are contained in a hit of `Query2`.

`Query1 OVER Query2` will match all hits of `Query1` that contain (are overlapping) a hit of `Query2`.

5.1.5 MINUS Operator

The `MINUS` operator is a binary operator that takes a left and a right operand. It returns the hits from the left sub-query that are not also valid hits for right sub-query. For example:

²The extended support for Measurement annotations is discussed further in section 6.2.

Query1 MINUS Query2 will match all the hits of Query1 that are not also hits of Query2.

5.1.6 Repeats Operator: “+”

The + operator can be used to match text segments that comprise a sequence of hits from the same sub-query. The length of the sequence is specified through a number (representing the **maximum** number of repetitions) or through two numeric values (representing the **minimum** and **maximum** number of repetitions). For example:

`to+3` will match one, two, or three repeated occurrences of the word *to*. The returned hits will be of the form “*to*”, “*to to*”, or “*to to to*”).

`{Person}+2..5` will match sequences of 2, 3, 4, or 5 adjacent **Person** annotations.

`({Location locType = city} | {Location locType = country})+3` will match any sequence of up to three **Location** annotations where each one refers to either a city or a country.

Note that there is no support for a repetition count of zero (an optional match) – you will need to reformulate the query to cover the versions with and without the optional match separately and combine them with an OR, for example `(term1 term2+2 term3) | (term1 term3)`. Similarly there is no support for unbounded wildcards (*n* times or more).

5.1.7 Sequence Queries and Gaps

As sequence is the default operator in Mimir, there is no graphical sign for it: simply writing a set of queries one after another will cause a search for sequences of hits, one from each sub-query. For example, the query “*the energy level*” is actually a sequence query where the first sub-query searches for the word “*the*”, the second for “*energy*”, and the last for “*level*”.

It is sometimes useful to include gaps in a sequence query, that is to allow arbitrary text fragments (of specified length) to occur in-between the hits from some of the sub-queries. This can be done by using the gap markers “[*n*]”, or “[*m*..*n*]”. These will match a sequence of length *n*, or with a length of between *m* and *n* of arbitrary tokens.

For example the query “*the [2] root:time*” will match phrases like “*the best of times*” or “*the worst of times*”, whereas the query “*the [2..10] root:time*” would also match “*the best use of one’s time*” (where the gap consists of six tokens – five words and an apostrophe).

5.1.8 Escaping Reserved Words

Some words are part of the query language definition so they cannot be used directly as query terms. If that is desired, then these constructs must be escaped as shown in the following table:

Reserved input	Escaped form
{ }	\{ \}
()	\(\)
[]	\[\]
:	\:
+	\+
	\
&	\&
?	\?
\	\\
.	\.
”	\"
=	\=
IN	“IN”
OVER	“OVER”
OR	“OR”
AND	“AND”

Escaping reserved constructs in the Mimir query language

5.2 Search Interfaces – How to Submit Queries to Mimir

The Mimir Grails plugin supplies two search interfaces by default, with the infrastructure to implement other interfaces as required. An XML-based service interface allows other applications to submit queries to the indexes hosted by a Mimir web application by POSTing requests over HTTP (described in section 5.2.1). There is also an example user-facing search interface called *Web UI*, intended primarily for testing and demonstration purposes (described in section 5.2.2). Both of these interfaces interact with the underlying indexes through the `SearchService` Grails service provided by the plugin. When embedding the Mimir Grails plugin in another Grails application this service is the primary means for application code to interact with Mimir, and is described in section 5.2.3.

5.2.1 Mimir Search Web Service

The Mimir web application exposes the search functionality as a web service that can be accessed through a simple HTTP interface. All requests are performed by calling an action with a set of parameters; the results of a call are encoded in XML and returned as the response to the request. All the example URLs in this section assume the `mimir-cloud` application with its default URL mappings, running on `localhost` port 8080.

The Mimir web service can be accessed at a URL like:

`http://localhost:8080/mimir-cloud/{index ID}/search/{action}`, where the `action` value is the name of one of the supported actions, described below. The actual URL (with the correct index ID included) can be obtained from the *index information page* presented by the Mimir web application. Parameters may be supplied as query parameters with a GET request or in normal `application/x-www-form-urlencoded` form in a POST request. Alternatively, they may be supplied as XML (if the request content type is `text/xml` or `application/xml`) of the form:

```
<request xmlns="http://gate.ac.uk/ns/mimir">
  <firstParam>value</firstParam>
  <secondParam>value</secondParam>
</request>
```

The first request to the service will return a session cookie, which must be passed back with all subsequent requests.

When accessing the service URL with no value provided for `action`, a help page will be returned presenting the documentation associated with the XML web service.

The following actions are available:

help

Function	Obtain service documentation.
Parameters	none
Returns	A help message describing how to use the service.

postQuery

Function	Starts a new query. This call returns immediately, as the query will execute asynchronously in a background thread.
Parameters	queryString: the text of the query, using the Mimir query language.
Returns	An XML message with the ID of the new query, or an error message if there were any problems while parsing the query. Example request: <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/postQuery?queryString=%22the%22</pre> Example response: <pre><?xml version="1.0"?> <message xmlns='http://gate.ac.uk/ns/mimir'> <state>SUCCESS</state> <data> <queryId>a28656e2-18f4-4b58-b9d3-9a9378eb14d0</queryId> </data> </message></pre>

documentsCount

Function	Gets the number of result documents.
Parameters	queryId: the ID for the query, as returned by the <code>postQuery</code> action.

Returns	<p>An XML message encapsulating a numeric value, or an error message if there were any problems. The value returned is -1 if the search has not yet completed, or the total number of result documents otherwise.</p> <p>Example request:</p> <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentsCount?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0</pre> <p>Example response:</p> <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <value>8209</value> </data> </message></pre> <p>Example error response:</p> <pre><?xml version="1.0"?> <message xmlns='http://gate.ac.uk/ns/mimir'> <state>ERROR</state> <error>Query ID a28656e2-18f4-4b58-b9d3-9a9378eb14d1 not known! </error> </message></pre>
---------	--

documentsCurrentCount

Function	Gets the number of result documents found so far.
Parameters	queryId : the ID for the query, as returned by the postQuery action.
Returns	<p>An XML message encapsulating a numeric value, or an error message if there were any problems. After the search completes, the value returned is identical to that returned by calling documentsCount.</p> <p>Example request:</p> <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentsCurrentCount?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0</pre> <p>Example response:</p> <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <value>142</value> </data> </message></pre> <p>Example error response:</p> <pre><?xml version="1.0"?> <message xmlns='http://gate.ac.uk/ns/mimir'> <state>ERROR</state> <error>Query ID a28656e2-18f4-4b58-b9d3-9a9378eb14d1 not known! </error> </message></pre>

documentId

Function	Obtains the document ID for the document at a given rank (position in the results list).
Parameters	<p>queryId: the ID for the query, as returned by the postQuery action.</p> <p>rank: the rank (position on the results list) for the requested document.</p>

Returns	<p>An XML message encapsulating a numeric value, or an error message if there were any problems.</p> <p>Example request:</p> <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentId?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0&rank=3</pre> <p>Example response:</p> <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <value>11</value> </data> </message></pre>
---------	--

documentScore

Function	Obtains the score for the document at a given rank (position in the results list).
Parameters	<p>queryId: the ID for the query, as returned by the postQuery action.</p> <p>rank: the rank (position on the results list) for the requested document.</p>
Returns	<p>An XML message encapsulating a numeric value, or an error message if there were any problems.</p> <p>Example request:</p> <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentScore?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0&rank=3</pre> <p>Example response:</p> <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <value>12.330469310919446</value> </data> </message></pre>

documentHits

Function	Obtains a set of hits. Each hit is defined by a document ID, a position and a length, both of which are defined in terms of tokens, not characters (see Section 4.1 for details).
Parameters	queryId : the ID for the query, as returned by the postQuery action. rank : the rank (position on the results list) for the requested document.
Returns	An XML message encapsulating a set of <hit> elements, one for each individual hit. Example request: <code>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentHits?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0&rank=3</code> Example response: <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <hits> <hit documentId='11' position='257' length='1'/> <hit documentId='11' position='266' length='1'/> <hit documentId='11' position='290' length='1'/> <hit documentId='11' position='303' length='1'/> <hit documentId='11' position='309' length='1'/> <hit documentId='11' position='316' length='1'/> <hit documentId='11' position='320' length='1'/> <hit documentId='11' position='332' length='1'/> <hit documentId='11' position='335' length='1'/> <hit documentId='11' position='342' length='1'/> <hit documentId='11' position='348' length='1'/> </hits> </data> </message></pre>

documentText

Function	Action for obtaining (a segment of) the text of a document.
Parameters	queryId : the ID for the query that has returned the document ID being used, as returned by the postQuery action. rank : the rank (position on the results list) for the requested document. termPosition the position of the first returned token. This parameter is optional; defaults to 0 is not provided, which means the first token of the document. length the number of tokens to be returned. This parameter is optional, if omitted, all the document tokens will be returned.

Returns	<p>An XML message containing the text of all the individual tokens and, if available, the spaces between them.</p> <p>This action could be used, for example, to obtain text snippets around a query hit.</p> <p>Example request:</p> <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentText?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0&rank=1&termPosition=100&length=10</pre> <p>Example response:</p> <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <text position='100'>25</text> <text position='101'>C</text> <space> </space> <text position='102'>1</text> <text position='103'>/</text> <text position='104'>08</text> <space> </space> <text position='105'>C</text> <text position='106'>25</text> <text position='107'>C</text> <space> </space> <text position='108'>1</text> <text position='109'>/</text> </data> </message></pre>
---------	--

documentMetadata

Function	Returns the title and URI associated with a document. Optionally, other metadata fields can also be obtained. All these values were provided at indexing time.
Parameters	<p>queryId: the ID for the query that has returned the document ID being used, as returned by the postQuery action.</p> <p>rank: the rank (position on the results list) for the requested document.</p> <p>fieldNames: (optional) a comma-separated list of other field names to be returned.</p>
Returns	<p>An XML message encapsulating the several string values, or an error message if there were any problems.</p> <p>Example request:</p> <pre>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/documentMetadata?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0&documentId=1</pre> <p>Example response:</p> <pre><?xml version="1.0"?> <message xmlns="http://gate.ac.uk/ns/mimir"> <state>SUCCESS</state> <data> <documentTitle> Virtual job-hunting: Technology fills situations vacant </documentTitle> <documentURI>http://www.bbc.co.uk/news/business-12194581</documentURI> </data> </message></pre>

renderDocument

Function	Renders the document text and hits for a given document, in the context of a given query. The HTML of the document is rendered directly to the response stream of the connection.
Parameters	queryId: the ID for the query, as returned by the <code>postQuery</code> action. rank: the rank (position on the results list) for the requested document.
Returns	HTML content. The hits are rendered as <code>...</code> . Example request: <code>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/renderDocument?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0&rank=1</code>

renderDocument

Function	Alternative version of the previous action. Renders the document text and hits for a given document, outside of the context of any given query. The HTML of the document is rendered directly to the response stream of the connection.
Parameters	documentId: the ID for the document, as returned by the <code>documentId</code> action.
Returns	HTML content. No hits are highlighted, as there is no query context available. Example request: <code>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/renderDocument?documentId=11</code>

close

Function	Closes a query, releasing all resources allocated for supporting it. After a query is closed, no more actions can be performed for it. It is important to close queries, as each running query uses up memory on the server. Queries are also closed automatically after a period of inactivity (upon session expiration, the time for which is defined in the configuration of the web application server – this is why it is important to pass the session cookie you received from <code>postQuery</code> back to the server with subsequent calls).
Parameters	queryId: the ID for the query, as returned by the <code>postQuery</code> action.
Returns	An XML message with a success or failure value. Example request: <code>http://localhost:8080/mimir-cloud/a4300d00-2dd1-4797-8eaa-e65b0c7d879b/search/close?queryId=a28656e2-18f4-4b58-b9d3-9a9378eb14d0</code> Example response: <pre><?xml version="1.0"?> <message xmlns='http://gate.ac.uk/ns/mimir'> <state>SUCCESS</state> </message></pre> Example failure (using the ID for an already closed query): <pre><?xml version="1.0"?> <message xmlns='http://gate.ac.uk/ns/mimir'> <state>ERROR</state> <error>Query ID a28656e2-18f4-4b58-b9d3-9a9378eb14d0 not known! </error> </message></pre>



Figure 5.1: Front page of the Web UI user interface

5.2.2 The Web UI Example User Interface

The Web UI search tool is a browser-based search interface intended to serve as a platform for experimentation with a Mimir index, and as a demonstration of the capabilities of the Mimir framework and API. It is written using the Google Web Toolkit, and the source code is included in the Mimir Grails plugin.

In the `mimir-cloud` web application with its default URL mappings, the Web UI interface for an index in searching mode is available at `http://localhost:8080/mimir-cloud/{index ID}/search/index`. The initial page, shown in figure 5.1, provides a text area into which you can type queries in the Mimir query language. It provides auto-completion for annotation types and features (by asking the index what types it was configured with when it was created). Clicking the Search button starts a search on the server.

Hits are shown below the search box, as shown in figure 5.2, with the hit text highlighted in bold and with three tokens of left and right context. The document title is a link, in this example to the original document as the index was created with the “Document URIs are external links” option. The “cached” link shows Mimir’s cached copy of the document, with all the hits from that document highlighted in red. For indexes where the document URIs are not external links the document title would link directly to the cached version and there would be no separate “cached” link.

At the bottom of the page is a row of pagination links (figure 5.3).

5.2.3 Embedding Mimir in a Grails Application

Both the XML web service and the Web UI interface ultimately use a Grails service provided by the Mimir plugin to search their indexes. If you install the Mimir plugin into your own Grails application this service will be your primary entry point to make use of Mimir



The screenshot shows the Mimir search interface. At the top left is a magnifying glass icon over a document, followed by the word "Search" in large blue font. To the right is the Mimir logo, which includes the word "Mimir" in a large blue font and "powered by" in a smaller font above it. Below the search bar is a dark blue header with the text "Searching Index 'BBC News'". Underneath is a search input field containing the placeholder text "{Person}" and a "Search" button. Below the search bar is another dark blue header with the text "Documents 1 to 20 of 8209:". The main content area displays three search results, each with a blue title and a list of links in black text. The first result is "Ed Miliband's shadow cabinet and ministerial teams (cached)" with links to "BBC News - Ed Miliband's shadow cabinet", "World UK England N. Ireland Scotland Wales Business", and "Twitter Email Print Ed Miliband's shadow cabinet ...". The second result is "NI Assembly election: full list of candidates (cached)" with links to "Help Accessibility Help N. Ireland Politics Home World World UK England N. Ireland Scotland Wales Business", ". East Antrim", and "Roy Beggs: UUP Stewart ...". The third result is "Public Sector pay: The numbers (cached)" with links to "World UK England N. Ireland Scotland Wales Business", "Cabinet Office Minister Francis Maude's own department", and ", headed by Eric Pickles, has nine ...".

Figure 5.2: Web UI search results page



The screenshot shows the Mimir search interface with pagination links. At the top left is a magnifying glass icon over a document, followed by the word "Search" in large blue font. To the right is the Mimir logo, which includes the word "Mimir" in a large blue font and "powered by" in a smaller font above it. Below the search bar is a dark blue header with the text "Searching Index 'BBC News'". Underneath is a search input field containing the placeholder text "{Person}" and a "Search" button. Below the search bar is another dark blue header with the text "Documents 1 to 20 of 8209:". The main content area displays three search results, each with a blue title and a list of links in black text. The first result is "The demise of the Elgar £20 note (cached)" with links to "World UK England N. Ireland Scotland Wales Business", "20 note By Kevin Peachey Personal finance reporter", and "image of composer Edward Elgar which will no ...". The second result is "Charities fight downturn with business know-how (cached)" with links to "World UK England N. Ireland Scotland Wales Business", "business know-how By Emily Blewett", and "BBC News Neil Blewett BBC News Neil Cain has made the ...". Below the search results is a dark blue pagination bar with white text showing the page numbers 1 through 20. The number 1 is highlighted in a darker blue, indicating the current page. Below the pagination bar is the text "Mimir 4.0-SNAPSHOT, © GATE 2011."

Figure 5.3: Web UI pagination links for a large search

functionality, so this section explains what you need to know to use it effectively.

The `searchService` is a normal Grails service which can be autowired into your own services, controllers, etc. The service itself is very simple, offering only the following methods:

postQuery(index, queryString) start running a query against the given index. The index can be specified either as a string containing the `indexId` (the last component of the index URL, typically a UUID) or as an `Index` domain object (the database object representing a local, remote or federated index). Returns a *query ID* string.

getQueryRunner(queryId) retrieves the `QueryRunner` for the given running query ID. `QueryRunner` is the interface through which you can interact with the running query.

closeQueryRunner(queryId) indicates that the given query runner is no longer required. It is important to call this method when you have finished with a query runner, as each runner owns resources such as background threads which need to be properly cleaned up.

Once a query has been started, its `QueryRunner` provides access to the statistics, the hits themselves, and the text in the matched documents. The most important methods are summarised below, but for full details you should look at the interface definition itself, in the `gate.mimir.search` package of `mimir-core`.

getDocumentsCount() Gets the number of result documents that have been found. While the query has not yet finished running, this method returns `-1`.

getDocumentsCurrentCount() Gets the number of distinct documents that have so far been found to contain hits. This number may increase at any time while the query is currently active. Once the query completes, this method returns the same value as `getDocumentsCount()`.

getDocumentHits(rank) Gets the details for the hits found inside a given document. The document is specified by its rank (the position in the list of result documents). The value supplied for the `rank` parameter must be between 0 (inclusive) and the value returned by `getDocumentsCount()` (exclusive). The return value from this method is a list of `Binding` objects, each representing one hit.

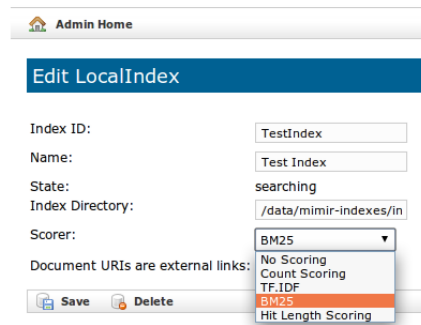
getDocumentID(rank) Gets the ID in the underlying index of the *n*th document that matched this query. This ID is needed when deleting documents from the index.

getDocumentTitle/URI(rank) Gets metadata about the document with the given rank.

getDocumentText(rank, start, length) Gets the text of the document with the given rank, starting at the *start*th token and extending for *length* tokens. The return value is a pair of parallel string arrays, one containing the text of the tokens and the other containing the text between each token and the following one.

renderDocument(rank, Appendable) Render the document content, with hits highlighted, using the document renderer configured for the index. The content is written to the specified `Appendable` (a `StringBuilder`, `Writer`, etc.).

The `getDocumentHits()` method returns a list of `Binding` objects, which provide several methods, the most important ones being `getTermPosition()` (the offset of the first token covered by the hit) and `getLength()` (the number of tokens it covers).



Admin Home

Edit LocalIndex

Index ID:

Name:

State:

Index Directory:

Scorer:

Document URIs are external links:

- BM25
- No Scoring
- Count Scoring
- TF.IDF
- BM25
- Hit Length Scoring

Figure 5.4: Configuring a Scorer

5.3 Ranking of Results

Starting with version 4.0, Mimir supports the ranking of results. This means that a scoring algorithm is applied to assess the relevance of each returned document and that documents are returned in decreasing order of their scores. To support this, Mimir will need to execute the query twice: once to calculate the scores, and a second time to collect the actual hits. Consequently, the query execution will be a bit slower when using ranking.

5.3.1 Configuring Ranking

Each searchable index can be configured to use a scorer. When initially created, all indexes have scoring disabled. This can be changed from the admin interface, by editing the index configuration, as shown in Figure 5.4.

If the provided scoring algorithms are not suitable for your needs, you can add new ones as discussed in Section 7.2.

Chapter 6

Standard Mimir Plugins

Mimir uses the GATE Embedded *CREOLE plugin* mechanism to load semantic annotation helper classes. A number of plugins are supplied by default with the Mimir distribution, and those plugins are described in this chapter. Information on how to create new plugins to provide user-defined helper classes can be found in section 7.1.

6.1 The db-h2 Plugin

The db-h2 plugin is a plugin that provides a *generic* semantic annotation helper implementation that can be configured for any annotation type with any features. The helper provided by db-h2 uses an embedded relational database engine (<http://www.h2database.com/>) to store the annotation data, and generally provides the best performance of the standard generic helpers.

`gate.mimir.db.DBSemanticAnnotationHelper` is the helper class provided by the db-h2 plugin. It has a constructor that takes a `Map` of configuration parameters, and Groovy provides special “named argument” support for `Map`-valued method and constructor parameters, allowing the following idiom in the index template DSL:

```
1 // note the “import X as Y”, which is another Groovy feature to create an
2 // alias for an imported class name
3 import gate.mimir.db.DBSemanticAnnotationHelper as DefaultHelper
4 // ...
5 semanticAnnotations = {
6     index {
7         annotation helper:new DefaultHelper(annType:'Person',
8             nominalFeatures:["gender", "title"], textFeatures:["name"]))
9     }
10 }
```

The supported constructor arguments are:

mode: whether to index actual GATE annotations and their features, or to index a single “virtual” annotation spanning the whole document from document-level features. The value must be either `Mode.ANNOTATION` (the default) or `Mode.DOCUMENT`,

where `Mode` is the enum type `gate.mimir.SemanticAnnotationHelper.Mode`

- annType:** the annotation type which the helper is to process, or for document-mode helpers, the “virtual” annotation type under which the specified features will be indexed.
- nominalFeatures:** the names of the features to be indexed that have nominal values. An annotation feature is said to be nominal if the range of possible values is clearly defined and limited in size. There is no hard rule regarding the size of the set of permitted values, but, for optimal results, this should not exceed a few tens of values.
- integerFeatures:** the names of the features to be indexed that have integer values (i.e. values that can be converted to a Java `long` value).
- floatFeatures:** the names of the features to be indexed that have floating-point numeric values (i.e. values that can be converted to a Java `double` value).
- textFeatures:** the names of the features to be indexed that have arbitrary text values (as opposed to the nominal case of a fixed list of possible values).
- uriFeatures:** the names of the features to be indexed that have URIs as values.
- indexNulls:** (boolean, defaults to `true`) if this is set to `false` then instances that have no values for *any* of the specified features will not be indexed at all. This is something that generally only makes sense for document-mode helpers, when you want to index a feature that is only present on some documents. Setting `indexNulls:false` allows you to use `OVER {AnnType}` queries to restrict searches to only the documents that have that feature.

The DB-based helper does not distinguish between text- and URI-valued features, indexing both types in the same way, but it accepts both kinds as arguments.

6.2 The measurements Plugin

The GATE `Tagger_Measurements` plugin, introduced in GATE 6.1, is able to recognise many different kinds of measurement expressions in text. It normalises the value and unit of each measurement into the SI system of measurements and stores these values as features of the `Measurement` annotation. For example, the text “45 cm” would be annotated with a normalised unit of metres and a normalised value of 0.45, the text “18 in” would also be normalised to metres, in this case with a normalised value of 0.4572.

The Mimir `measurements` plugin provides a SAH that implements the same normalisation on queries. It processes queries for a “synthetic” feature called “spec” which represents a measurement specification in a controlled language and converts constraints on this feature into the corresponding constraints on the real normalised value and unit features that have been indexed. For example, a search for `{Measurement spec="1 to 3 feet"}` would be treated as a query for measurements whose normalised unit is metres and whose normalised value is between 0.3048 and 0.9144, which would match both the “45 cm” and “18 in” examples above.

6.2.1 Configuring the Measurements SAH

To use the `measurements` helper you need to first ensure that the `measurements` plugin is loaded into your Mimir instance, then create an index template that specifies an instance of the helper:

```

1  import gate.mimir.measurements.MeasurementAnnotationHelper
2
3  // ...
4  semanticAnnotations = {
5      index {
6          // Measurement helper with default settings
7          annotation helper:new MeasurementAnnotationHelper(
8              delegateHelperType:DefaultHelper)
9      }
10 }

```

Note that the measurement helper does not need any “annType” or “xxxFeatures” parameters, as it is hard-coded to work only for annotations that are produced by the measurement tagger PR. However the constructor does take a Map with other named arguments:

```

6      // Example of how to configure a custom “units” file
7      annotation helper:new MeasurementAnnotationHelper(
8          delegateHelperType:DefaultHelper,
9          unitsFile:'resources/americanUnits.dat',
10         locale:'en_US')

```

The following parameters are supported:

delegateHelperType (required) a Class object representing the type of generic helper that the Measurements helper should delegate to. This class must provide a 6-argument constructor taking the annotation type (a String) and five String arrays for the nominal, integer, float, text and URI feature names respectively.

unitsFile the location of the units.dat file used to configure the measurements parser. If not specified, a default file provided with the measurements plugin is used. This value can be an absolute URL (file:/path/to/units.dat) or a relative path which will be resolved against the measurements plugin directory.

commonWords the location of the common words file used by the measurements parser. As with the unitsFile parameter, if omitted a default file bundled with the plugin is used.

locale the locale under which the measurements will be parsed. Defaults to “en_GB” if unspecified.

encoding the character encoding used to read the configuration files. Defaults to “UTF-8” if unspecified.

annType the annotation type, if something other than the default of “Measurement”

The measurements SAH is pre-configured with the feature names that the measurements tagger produces, and attempting to specify any feature name parameters such as nominalFeatures will cause an error.

Measurements helper implementation

The MeasurementAnnotationHelper extends the DelegatingSemanticAnnotationHelper base class described above. It does not add any behaviour at indexing time, simply passing all the annotations through directly to its delegate. However it overrides the getMentions search method to support the “spec” feature.

When a query including a spec feature constraint is received, the helper parses this spec using the measurements parser to obtain a normalised unit and value or values for the measurement sought. It then constructs a number of new constraint sets that match annotations compatible with the spec and then for each of these alternatives, runs these constraints in combination with the other non-spec constraints of the original query against the delegate helper. The final set of URIs returned is the union of the results obtained from the delegate for all the alternative reformulations of the spec constraint.

As well as being useful in its own right for Measurement annotations, the measurements helper serves as an example of how to implement your own special-purpose helper based on the delegating base class. Feel free to use it as a template for your own helper implementations.

6.3 The sparql Plugin

The `sparql` plugin provides a semantic annotation helper that wraps another helper, adding flexible *semantic query* support. It is intended to be used with annotations that have one or more URI-valued features whose values refer to entities in an external knowledge base (accessible at a standard *SPARQL endpoint*). The SPARQL helper has no effect at indexing time, simply delegating all calls through to the underlying helper, but at search time it allows queries for the synthetic feature “`sparql`”. This feature value is taken to be a SPARQL “SELECT” query, which is posted to the configured SPARQL endpoint. The variables in the SELECT query must correspond to the names of features that have been indexed by the underlying helper, and each row in the result set becomes a standard Mimir query to the underlying helper. Any annotations that match any of these new queries will be treated as a match for the `sparql` constraint. This process is described in detail below.

For example, given a helper configured for the public DBpedia endpoint `http://dbpedia.org/sparql`, the following Mimir query:

```
{Person sparql = "SELECT DISTINCT ?inst WHERE {
  ?inst <http://dbpedia.org/ontology/birthPlace>
  <http://dbpedia.org/resource/Sheffield> }"}}
```

would search for all Person annotations that have an “`inst`” feature containing the URI of an entity in DBpedia that represents a person born in Sheffield.

6.3.1 Creating a SPARQL Helper

The SPARQL semantic annotation helper class is called `gate.mimir.sparql.SPARQLSemanticAnnotationHelper`. It has a `Map` constructor taking the following parameters:

- delegate (required)** the underlying semantic annotation helper that this SPARQL helper should wrap.
- sparqlEndpoint (required)** the address of the SPARQL endpoint that this helper should use when making SPARQL queries.
- queryPrefix** an optional prefix that will be prepended to the string specified in the `sparql` synthetic feature to form the actual SPARQL query that will be sent to the endpoint. Typically this would be used to define appropriate namespace prefixes.

querySuffix an optional suffix that will be appended to the end of the SPARQL queries. Thus the actual SPARQL query submitted to the endpoint is *queryPrefix* + `sparql` feature + *querySuffix*.

sparqlEndpointUser and sparqlEndpointPassword username and password used to authenticate to the SPARQL endpoint (only HTTP basic authentication is supported). May be omitted if your endpoint does not require authentication.

sparqlRequestMethod the mechanism by which the query will be passed to the endpoint. This value is an enum type `gate.mimir.sparql.RequestMethod` with three possible values:

GET (default) the query will be URL encoded and appended to the URL as a parameter `?query=...`

POST_ENCODED the query will be URL encoded as in the GET case and sent in a POST request with content type `application/x-www-form-urlencoded`

POST_PLAIN the query will be sent as-is as the body of a POST request with content type `application/sparql-query`

The helper also accepts the usual “annType” and “xxxFeatures” parameters but these are not normally required – if the delegate helper is a subclass of `AbstractSemanticAnnotationHelper` (which is the case for all the standard helpers) then the SPARQL helper will take its feature names from the delegate, so the only time the features need to be specified explicitly for the SPARQL helper is if the delegate is a custom helper type that does not extend `AbstractSemanticAnnotationHelper`.

For example, the following configuration would set up a helper for Person annotations operating against DBpedia, to support the example query above:

```

1  import gate.mimir.db.DBSemanticAnnotationHelper as DBH
2  import gate.mimir.sparql.SPARQLSemanticAnnotationHelper as
   SPARQLHelper
3  import gate.mimir.sparql.RequestMethod as RM
4  // ...
5  semanticAnnotations = {
6    index {
7      annotation helper:new SPARQLHelper(
8        sparqlEndpoint:'http://dbpedia.org/sparql',
9        sparqlRequestMethod:RM.POST_ENCODED,
10       delegate:new DBH(annType:"Person", uriFeatures:['inst']))
11    }
12  }

```

Alternatively, the helper could be configured with a `queryPrefix` to set up some useful namespace prefixes:

```

1  semanticAnnotations = {
2    index {
3      annotation helper:new SPARQLHelper(
4        sparqlEndpoint:'http://dbpedia.org/sparql',
5        sparqlRequestMethod:RM.POST_ENCODED,
6        queryPrefix:
7        'PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#> \
8        PREFIX xsd:<http://www.w3.org/2001/XMLSchema#> \
9        PREFIX owl:<http://www.w3.org/2002/07/owl#> \
10       PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> \
11       PREFIX dbo:<http://dbpedia.org/ontology/> \
12       PREFIX dbr:<http://dbpedia.org/resource/> ',

```



```

13     delegate:new DBH(annType:"Person", uriFeatures:['inst']))
14   }
15 }

```

Note the backslashes which are a Groovy feature to permit a String literal to be broken across several lines, and also the trailing space before the closing quotation mark – the helper simply concatenates the prefix, query and suffix without any additional space, so required spaces must be part of the prefix or suffix string itself. This would allow the example query above to be rewritten more succinctly as:

```
{Person sparql = "SELECT DISTINCT ?inst WHERE {
  ?inst dbo:birthPlace dbr:Sheffield }"}

```

For annotation types that have only one URI-valued feature it may be desirable to include the “SELECT DISTINCT ?inst WHERE { ” in the prefix and add a suffix of “ }”, which would reduce the query down to

```
{Person sparql = "?inst dbo:birthPlace dbr:Sheffield"}

```

If your index template includes several ontology-based annotation types sharing the same SPARQL endpoint and prefixes then listing these in full for each annotation type will result in a large and unwieldy template. However, since the index template is itself a Groovy script it is possible to declare methods to factor out the common code. Method declarations must be placed *outside* the semanticAnnotations block:

```

1  import gate.mimir.db.DBSemanticAnnotationHelper as DBH
2  import gate.mimir.sparql.SPARQLSemanticAnnotationHelper as
   SPARQLHelper
3  import gate.mimir.sparql.RequestMethod as RM
4
5  def standardHelper(type) {
6    return new SPARQLHelper(
7      sparqlEndpoint:'http://dbpedia.org/sparql',
8      sparqlRequestMethod:RM.POST_ENCODED,
9      queryPrefix:
10     'PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#> \
11     PREFIX xsd:<http://www.w3.org/2001/XMLSchema#> \
12     PREFIX owl:<http://www.w3.org/2002/07/owl#> \
13     PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> \
14     PREFIX dbo:<http://dbpedia.org/ontology/> \
15     PREFIX dbr:<http://dbpedia.org/resource/> ',
16     delegate:new DBH(annType:type, uriFeatures:['inst']))
17 }
18
19 // ...
20 semanticAnnotations = {
21   index {
22     annotation helper:standardHelper('Person')
23     annotation helper:standardHelper('Location')
24     annotation helper:standardHelper('Organization')
25   }
26 }

```

6.3.2 Format of SPARQL Queries

This section describes in more detail how the SPARQL queries relate to the annotations indexed by the underlying semantic annotation helper. As a simple example we consider the query for people born in Sheffield:

```
{Person sparql = "SELECT DISTINCT ?inst WHERE {
  ?inst dbo:birthPlace dbr:Sheffield }"}
```

The helper will submit the SPARQL query to its configured endpoint, and receive a response of the form:

inst
http://dbpedia.org/resource/Gordon_Banks
http://dbpedia.org/resource/Michael_Palin
http://dbpedia.org/resource/David_Blunkett
...

This will then generate a series of queries to the underlying helper of the form:

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

and any annotation that matches any of these queries will be returned as a match for the `sparql` constraint.

The SPARQL query can bind more than one variable, and the values of the variable bindings can be RDF literals as well as URIs, they convert to queries on the underlying helper in the same way.

Chapter 7

Extending and Customising Mimir

The standard semantic annotation helpers provided by Mimir are adequate for many use cases, but if your application needs more functionality that they cannot provide it is easy to add your own custom helper implementations using a plugin mechanism. This process is described in section 7.1. Similarly, the basic Mimir cloud application shows a simple way to use the Mimir Grails plugin, but it provides no other features. If you have different requirements, you can install the Mimir plugin into your own custom Grails application, as described in section 7.3.

7.1 Creating New Semantic Annotation Helpers

Semantic annotation helpers (SAHs) are the mechanism that Mimir uses to store information about annotations and allow this information to be queried at search time. A SAH is associated with a particular annotation type in the Mimir index configuration, and performs two functions:

During indexing for each annotation of the relevant type, store information about that annotation in some persistent form and return to Mimir one or more URIs that represent that annotation. These URIs are included in the main MG4J index and associated with the location in the document where the annotation was found.

During searching given a set of feature value constraints, use the persistent store created during indexing to determine the URIs associated with annotations that satisfy the constraints.

Conceptually, SAH implementations can be divided into two types. *Generic* helpers are those that can index any annotation types and features, and *special-purpose* helpers are those that are designed to work with specific types of annotation. There are two generic SAH implementations provided with Mimir by default. You would create a new generic SAH implementation if you wanted to store annotation data in a different underlying storage format. Mimir provides one example special-purpose SAH for Measurement anno-

tations, which can serve as a template for how to implement your own helpers for other annotation types.

7.1.1 The SemanticAnnotationHelper Interface

The `gate.mimir.SemanticAnnotationHelper` interface is the contract that all helpers must implement. It specifies three groups of methods that must be implemented:

Lifecycle Methods

The interface includes `init/close` lifecycle methods, taking an `MimirIndex` parameter. The `MimirIndex` object provides access to an `IndexConfig` object which defines the configuration of the index, including the location of the index files on disk, and provides a mutable “context” map that can be used to share objects among the various SAH objects. The `init` method is called by Mimir when the index is opened, before any other requests are passed to the helper, and the corresponding `close` method is called when the index is shut down.

Indexing Methods

When indexing annotations Mimir calls the following methods:

documentStart(document) Called when the indexer starts processing a particular document to allow the helper to perform any per-document setup tasks. This method is guaranteed to be called once per document, before any calls to `getMentionUris`.

getMentionUris(annotation, length, indexer) Called once for each semantic annotation of this helper’s type in the document. The helper is expected to use the annotation’s length (in tokens) and feature values to determine the relevant URI or URIs that represent this annotation, and return them.

documentEnd() Called after all the annotation for a particular document have been processed.

These methods are always called from a single thread, as long as the same helper object is not used for more than one annotation type.

Note that the annotation length passed to `getMentionUris` is measured in *tokens*, not characters. Because Mimir operates on streams of tokens, semantic annotations that partially overlap a token will be considered by Mimir to cover the whole token. I.e. given the hypothetical example:

```
... started on 10/05/1987 by John Smith ...
-----
```

where tokens are represented by `---`, an annotation that covers just the “87” would be indexed as if it covered the whole “1987” token.

Search Methods

The final method in the interface is `getMentions(annotationType, constraints, queryEngine)`. This method is called by Mimir when searching for annotations, and the helper must use its stored data to determine all the possible mention URIs that satisfy the provided constraints, and return them along with their lengths (in tokens) as provided to `getMentionUris` when the annotations were indexed.

There is a second overloading of this method specified in the interface which is a convenience for callers when all the constraints are simple feature value equality constraints, but generally implementers of new SAH types can ignore this as Mimir provides an abstract base class that converts the `Map` form of constraints into the more general `List<Constraint>` form and calls the other method.

The `getMentions()` methods may be called from multiple threads at the same time, so implementations should be thread-safe.

7.1.2 Abstract Base Classes

Mimir provides an abstract class `AbstractSemanticAnnotationHelper` which, as described above, implements the `Map` version of `getMentions` in terms of the `List<Constraint>` version, and also provides empty implementations of `documentStart` and `documentEnd`. As well as this, it provides accessor methods to access the list of feature names of each of the five types (nominal, integer, float, text and URI) that a particular helper object supports. `AbstractSemanticAnnotationHelper` enjoys special support in the Mimir Grails plugin, allowing clients to determine what feature names an index supports for each annotation type whose helper extends `AbstractSemanticAnnotationHelper`. All the standard helper implementations provided with Mimir extend this base class.

Special-purpose helpers for particular annotation types typically operate by mapping the features of their target annotations and/or the feature constraints in a query into a different set of features or constraints which can then be handled by a generic helper. The Measurement helper described in section 6.2 operates in this way. To support this pattern, Mimir provides an abstract `DelegatingSemanticAnnotationHelper` which implements all the SAH interface methods to simply delegate to another helper instance. Subclasses can then override the methods as appropriate to map their features or constraints into terms that the underlying helper can understand and then call the `super` method to pass these parameters on to the delegate.

`DelegatingSemanticAnnotationHelper` extends `AbstractSemanticAnnotationHelper` so it advertises the features it supports in the usual way. However it is important to note that the various `get*FeatureNames` methods of the delegating helper do *not* call their counterparts in the delegate, which allows a delegating helper to advertise different features from those supported by its delegate.

The SAH lifecycle

Semantic annotation helper objects go through a specific lifecycle in Mimir. When creating a new index for indexing, the helpers defined for each semantic annotation type are instan-

tiated by calling their constructors from the Groovy DSL (see the measurements example below). Once instantiated, the `init(Indexer)` methods of each helper in turn are called (one after the other, in a single thread, so if you are sharing objects among your helpers through the context you can be sure that you have exclusive access to the context map during the call to your `init` method).

The actual indexing process takes place in several threads in a pipelined manner. When a document arrives for indexing it is first processed by the token indexer (to index the token features), then the semantic indexers specified by the `index { ... }` blocks in the DSL in turn. Each indexer operates in its own thread, with documents passing from one to the next via queues. So each document is only processed by one thread at a time but under load you may have the token indexer dealing with document 3 at the same time as semantic indexers are dealing with documents 2 and 1.

When indexing is complete the helpers' `close(Indexer)` methods are called (again, in sequence in one thread). The index is now closed and the SAH objects can be garbage collected.

The index configuration, including all the SAH objects, is serialised to XML using XStream (<http://xstream.codehaus.org>). Therefore it is important to mark as `transient` any fields of your helper class that should not be serialised (e.g. temporary in-memory caches, etc.).

When an index is opened for searching the XML configuration is deserialised to re-create the helper objects, and their `init(QueryEngine)` methods are called. Note that as with Java object serialisation XStream does *not* call object constructors when de-serialising, so any initialisation must happen in the `init` method or in a `readResolve` method, and not in the constructor.

Annotation queries result in calls to the relevant helper's `getMentions` method, which has been discussed in detail above.

Finally, when the index is shut down the `close(QueryEngine)` methods of the helpers are called in sequence.

7.1.3 Packaging New Helper Types for Use with Mimir

To create a new helper type and make it available to Mimir you need to first write your Java class that implements `SemanticAnnotationHelper`, typically via one of the abstract base classes. To make your helper available to Mimir requires a couple of extra steps.

Helper implementations are loaded into Mimir using the standard GATE CREOLE plugin mechanism. The simplest approach is to create a standard GATE plugin with a Maven POM file that uses the `gate-plugin-base` parent, install the plugin to your local Maven cache or deploy it to a repository, then configure it as a dependency in your `application.groovy` with

```
1 gate {
2   mimir {
3     plugins {
4       myPlugin {
5         group = "com.example.mygroup"
6         artifact = "my-helper"
```

```
7     version = "0.1-SNAPSHOT"
8   }
9 }
10 }
11 }
```

as described in Section 3.2.5.

7.2 Registering new Scorers

The set of scoring algorithms made available in the admin interface is obtained by the `mimir-web` plugin by querying a service defined in the host application. For `mimir-cloud` this is `grails-app/services/gate/mimir/cloud/MimirScorerService.groovy`.

The default configuration includes the following scorers:

```
1 def scorers = [
2   'Count Scoring':      { -> new DSQE(new CountScorer()) },
3   'TF.IDF':            { -> new DSQE(new TfIdfScorer()) },
4   'BM25':              { -> new DSQE(new BM25Scorer()) },
5   'Hit Length Scoring': { -> new BindingScorer() },
6 ]
```

Each entry maps a Groovy closure to the name of a scorer. The closure must return an object implementing the `gate.mimir.search.score.MimirScorer` interface, which itself extends the MG4J interface `it.unimi.dsi.mg4j.search.score.DelegatingScorer`.

Mimir provides the `gate.mimir.search.score.DelegatingScoringQueryExecutor` class which can be used to wrap one of the scorer implementations provided by MG4J. If none of those are suitable, you can write your own custom implementations, either under `src/main/groovy` of your application or by putting them in their own library and adding a suitable dependency to `build.gradle`, and edit the `MimirScorerService` to expose them under appropriate names.

7.3 Embedding Mimir in a Custom Grails Application

The `mimir-cloud` Grails application provided with Mimir provides access to the functionality of the Mimir Grails plugin but is deliberately kept simple. For more complex needs, you can install the Mimir plugin into a custom Grails application.

7.3.1 Installing the Mimir Plugin

Both snapshot and release versions of the Mimir plugin are published to the standard GATE Maven repository, and can be installed into your own Grails application (3.3 or later) in the normal way by adding a dependency to `build.gradle`:

```
1 repositories {
2     // ...
3     maven { url "http://repo.gate.ac.uk/content/groups/public" }
4 }
5
6 dependencies {
7     // ...
8     compile "uk.ac.gate.mimir:mimir-web:6.0-SNAPSHOT"
9 }
```

After installing the plugin into your app you should run `./grailsw mimir-config`. This adds a number of files to the host application:

grails-app/conf/application.groovy default Mimir settings are appended to the end of this file (which is created if it did not previously exist). These can be customised as described in section 3.2.5.

grails-app/controllers/(default package)/MimirUrlMappings.groovy a default set of URL mappings for Mimir controllers. These are discussed in more detail below.

grails-app/services/(default package)/MimirScorerService.groovy a Grails service implementing `ScorerSource`, which defines the default set of available scorers.

7.3.2 URL Mappings

The standard URL mappings for Mimir adhere to the following conventions:

- All administrative actions have URLs that start with `/admin`. This is intended to simplify the job of providing security via an external mechanism, such as the default container-provided security defined by the servlet specification.
- The URLs for all actions referring to a given index have a common prefix (by default `/{indexId}/`). This is used to support remote access through the `mimir-client` library, which relies on a base URL to which it adds known suffixes to access given functionality (e.g. the search actions are all found under `/{indexId}/search/{actionName}`). The `mimir-client` library is used, for example, in the implementation of remote indexes.

These default mappings may be modified, but the protocols used by client code to push documents into an index, to search using the XML API, and to interact with remote indexes all make assumptions about certain relative paths. Therefore, if you need to allow remote access via the `mimir-client` library, then you must make sure that:

- there is a URL prefix that is common to all actions provided by the `IndexManagementController` and the `SearchController`. Let us call this common prefix the `{indexUrl}`;
- all actions provided by the `IndexManagementController` should be published under `{indexUrl}/manage/{actionName}`;
- all actions provided by the `SearchController` should be published under `{indexUrl}/search/{actionName}`;
- all actions of the search controller must have `parseRequest:true`;
- the names of the actions in the controllers above must not be changed.

More detail is provided in the comments in the default generated mapping file.

7.3.3 Security Considerations

There are a number of established and mature Grails plugins that implement user registration, authentication and security in various ways for Grails applications. By design, the Mímir plugin does not provide any security of its own, allowing it to integrate with any security solution. For production deployments you may wish to consider any or all of the following.

- Place your application behind an Apache HTTPD or other similar front-end server, and proxy only those URLs that need to be public (/assets, /{yourindex}/search, etc.)
- Install a security plugin such as `spring-security-core`, and restrict the Mímir pages to certain users. For example the /admin pages could be restricted to administrative users, the Web UI pages to registered searchers, etc. This is how the `mimir-cloud` application is set up.
- If your Mímir instance is a slave in a federated index, use firewall rules or a Grails filter to restrict access to the */search URLs to only the master Mímir server.

The default Mímir plugin can handle remote indexes that use HTTP *basic* authentication, but if you use another authentication mechanism on the XML search API or the remote protocol then you will need to modify the default client code to authenticate itself appropriately. The remote protocol uses the `WebUtils` class in `mimir-client` to make its HTTP calls, you would need to subclass this to implement your authentication method and wire it in by overriding the definition of the `webUtilsManager` Spring bean supplied by the Mímir plugin – see the plugin source code for full details.

Chapter 8

Additional Tools

This chapter documents additional tools that are provided with Mimir for special use cases, but which are not required for general day-to-day operation.

8.1 Recovering a failed index

Mimir indexes are ordinarily quite robust, but there are certain circumstances in which an index can become corrupted and marked as “failed” in the Mimir web UI. Typically this only happens when the index has not been shut down cleanly, for example if an out-of-memory condition occurs during indexing. The majority of these failures fall into two categories, either a crash during a “sync to disk” operation which leaves a corrupted batch on disk, or a crash after all batches have been saved but before the index has been completely closed, which leaves the document metadata zip files corrupted. In both of these cases the vast majority of the indexed data can usually be recovered using the index repair tool. The last documents to be indexed will likely be lost – exactly how many are lost depends on a number of factors including exactly when the failure occurred, the length of the `timeBetweenBatches` setting in the index template, etc. – but the tool attempts to minimise the number of lost documents as far as possible.

The repair tool is a command line application which operates directly on the index files on disk. In order to run the tool the index must not be “open” and in use by a running Mimir, so you must either shut down your running Mimir application or delete the local index from the web UI (*without* deleting the underlying index files from disk!).

Before attempting the repair process it is **very strongly recommended** to make a backup copy of the index files. If the repair process itself fails (e.g. with an out of memory error) it can leave the index in a completely unrecoverable state, and you will have to restore from your backup, correct the problem (e.g. allocate more memory) and try again.

The simplest way to run the repair tool is via the `truncate-index.sh` bash script at `WEB-INF/utis` inside the compiled `mimir-cloud` WAR file.

```
bash truncate-index.sh /path/to/the/index-12345.mimir
```

The final parameter is the full path to the top-level directory of the index you want to repair. If the repair is successful you should then be able to re-start your Mimir application and/or re-import the fixed index using the “import an existing index” option.

See the comment block at the top of the script for full details of the available parameters.

The recovery process in detail

The repair process consists of a number of phases.

1. Ensure the document metadata zip files are all complete, repairing the last one if necessary
2. Examine all the index batches and determine the latest point at which all the sub-indexes successfully dumped a batch in sync. This is referred to as the “last good batch”. Delete any batches beyond this point.
3. If the (repaired) zip files contain at least as many document as the good batches, then simply truncate the zip collection to match the last good batch and the repair process is complete.
4. Otherwise, the zip files are the limiting factor, as the zip collection ends in the middle of a “good” batch. Determine which batch this is, delete all the subsequent batches, then truncate what is now the last batch to match the length of the zip collection.

The final step can require a lot of memory if the last batch is large (e.g. a recently compacted head batch), it may be necessary to allocate more memory to the repair process by editing the shell script.

In the best case (all batches successfully synced to disk, just the zip collection failed to close) this will recover all but the last one or two documents. The worst case is when the index failed during the very first sync to disk, in which case nothing will be recoverable, but in this case there should be no more than one hour or so of documents that need to be re-indexed. Most cases will fall somewhere between these extremes, and the number of documents lost depends on the `timeBetweenBatches` configured in the index template. A shorter time between batches means less potential for data loss but more work for the indexer.

Appendix A

Change Log

This appendix details the main changes in each Mimir release.

A.1 Version 6.0 (May 2018)

- Now depends on GATE Embedded 8.5.
- Build process completely revamped to build the core components and plugins with Maven, and the web components with Gradle/Grails 3.3.
- New versions of many dependencies including MG4J and dsiutils. Note that this breaks compatibility with existing indexes from Mimir versions before 6.0, they will need to be upgraded using the upgrade option on the index management web page in order to be usable with Mimir 6.0 (see section 3.3)

A.2 Version 5.4 (February 2017)

- Now depends on GATE Embedded 8.4.
- A few critical bug fixes:
 - Fixed a bug that prevented fetching of results when querying a remote index through the XML search API
 - Set the proper query tokeniser and executor when creating a brand new index from an index template – previously these were only being set for existing indexes opened from disk.
- Robustness improvements to make it more likely that an index which was not closed properly (e.g. if Mimir was killed due to low memory conditions) will be recoverable. There were previously a couple of critical points where a crash would leave the index in a state where it could not be recovered even with the use of the repair tool.

A.3 Version 5.3 (January 2017)

- Now depends on GATE Embedded 8.3.
- Added a tool to repair failed indexes, see section 8.1 for details.
- Added the ability to completely ignore annotations that have none of the features configured in the index template. This is more useful for document-mode helpers rather than normal annotation-mode ones, see section 6.1 for an explanation.

A.4 Version 5.2 (June 2016)

- Now depends on GATE Embedded 8.2.
- The Grails plugin and `mimir-cloud` application have been upgraded to Grails 2.5.4 for better compatibility with Java 8. Note that the plugin and app will *not* work with Grails 3.
- The query language can now handle non-ASCII characters in unquoted strings.
- The Mimir indexing PR is now much more efficient, able to send more than one document in each HTTP call.

A.5 Version 5.1 (June 2015)

- Mimir 5.1 depends on GATE Embedded 8.1.
- Bug fixes in various corner cases, in particular for very sparse semantic annotations (where annotations of a particular type are found in relatively few documents).
- Robustness improvements in `mimir-client` indexing code.
- The SPARQL semantic annotation helper can now send queries to the SPARQL endpoint using POST requests instead of GET, and now works correctly with endpoints that require HTTP basic authentication.

A.6 Version 5.0.1 (October 2014)

Two critical fixes:

- Deletion of documents now works correctly, it had been broken in version 5.0
- Fixed clustering logic for multi-batch indexes.

A.7 Version 5.0 (February 2014)

- Mímir indexes are now updateable: new documents can be submitted for indexing at any time.
- Mímir indexes are now live: they can index new documents and serve queries at the same time. Manually *closing* indexes before they become searchable is no longer required.
- The *mimir-demo* example web application has been removed.
- The *mimir-cloud* has been modified to make it more suitable as a generic example web application.
- The sesame Mímir plugin has been removed. For standard annotation indexing we recommend using the db-h2 plugin. For handling formal semantics, we recommend using the SPARQL plugin.
- New query operator: **MINUS** (also ‘-’) performs the set minus operation on result sets (see Section 5.1.5).
- Mímir now supports the construction of direct indexes (see Section 4.1). Direct indexes are used to support a new family of queries, that use document ID as query terms, and which return terms as results. Currently these are only available as a Java API, and can be found in the `gate.mimir.search.terms` package.
- Semantic annotation helpers are now capable of ‘describing’ a matched mention. The S-A-H implementations included in the main distribution provide default implementations for this functionality, which can be replaced by pluggin-in alternative versions.
- The on-disk format for Mímir indexes has changed. This was required in order to support live indexing and searching.
- Mímir has been upgraded to use MG4J version 5.2.1. Newly created indexes will now be semi-succint, which is the highest performance implementation.
- Mímir now uses Grails 2.2.3 and GWT 2.6.0 to build the *mimir-cloud* web application.
- Bugfix: you can now use a string on the right hand side of a `<`, `>`, `<=` and `>=` in annotation queries. This was always documented, but did not work before.
- Many other bugfixes.

A.8 Version 4.1.3 (September 2012)

- Bug fix in ranking query runner (used to search local indexes): a document ID was used instead of a document rank when requesting metadata fields.

A.9 Version 4.1.2 (August 2012)

- Bug fix to void null pointer exceptions when the API is used to access query results in a federated index without first checking the number of available documents. Calling methods with an invalid rank parameter will now cause an index out of bounds exception.

A.10 Version 4.1.1 (May 2012)

- It is now possible to specify an index ID for a newly created/imported local, remote or federated index, rather than having to create the index with a random UUID and then change the ID later.
- Bugfix: stopped the web search UI from showing 'null' for context tokens outside of the document, when a hit result occurs close to the end of the document.
- Bugfix: the annotation type needed to be specified twice in the index template when using the SPQARQL plugin.
- Bugfix: the web search UI was not updating correctly when a query completed without matching any results.

A.11 Version 4.1 (May 2012)

- A bugfix was applied to avoid leaking threads and memory in the new ranking query runner implementation (the class `gate.mimir.search.RankingQueryRunnerImpl`).
- Mimir now uses the mg4j-big variant of the MG4J library. This uses 64 bit integers (Java longs) for document identifiers, and allows for larger indexes to be created.
- The dependency to MG4J and related libraries is now managed through the maven-central repository.

A.12 Version 4.0 (February 2012)

- Changed the results presentation to be document-centric, as opposed to hit-centric.
- Overhauled the query API (in all modalities: Java local, Java remote, and XML remote) to work in document centric mode and to remove the main pain points identified.
- Simplified all the query APIs by making them almost completely synchronous.
- Added support for ranking the results (see Sections 5.3 and 7.2).
- New implementations for all the query runners (used when searching local, remote and federated indexes).
- Replaced the old GWT based UI with a new implementation (see Section 5.2.2).
- Added the mimir-cloud web application to the source tree (see Section 3.2.1).

A.13 Version 3.4.0 (November 2011)

- Added support for indexing document metadata, i.e. features (see Section 4.1).
- Mimir Grails Plugin: moved some configuration options from the external file to a database field, so that it can now be changed using the admin web UI.
- API: simplified the construction of all default Semantic Annotation Helpers. They all get a single no-argument constructor, and set of setter method for editing the various properties (Java Bean style). The Groovy interface does not change, as Groovy will automatically convert a constructor call that takes a Map to a call for the no-argument constructor, followed by all the required setPropertyXYZ calls.
- Completely removed the (previously deprecated) `ordi` plugin, as it relies on software that is no longer supported by the original authors.
- Removed the `mimir-demo` example application from the source tree. It can now be automatically generated using an Ant call (see Section 3.2).
- Licence changed to LGPL.

A.14 Version 3.3.0 (October 2011)

- Added support for marking documents as “deleted” (see section 3.5).
- Major changes to the format of the Index Template Groovy DSL (see section 4.1). The old format provided by Mimir 3.2.0 is still supported for existing semantic annotation helper types, but new helper types in future may not be supported in the old style DSL.
- Added the *SPARQL* semantic annotation helper (see section 6.3).
- Updated versions of a number of libraries (H2 database to 1.3.160, OWLIM to 3.5, MG4J to 4.0, fastutil to 6.4, dsiutils to 2.0).
- The `ordi` semantic annotation helper plugin is now deprecated. Use the `sesame` plugin instead, which supports the same on-disk format for its annotation storage but uses a different library to access it.
- Fixed various bugs and memory leaks (see subversion logs for full details).

A.15 Version 3.2.0 (May 2011)

First public release of Mimir, under an AGPL licence.