FAQ
Frequently Asked Questions
Featured

Updated May 5, 2011 by kleg...@gmail.com

A list of frequently asked questions

Questions are organised into 3 different categories:

GENERAL questions

What is the HIT?

It is a simple to use, extensible open source framework that allows you to easily manage data harvesting and quickly build specific indexes of harvested data.

Who are its end-users?

Data aggregators, or anybody who wants to build a central index from a network of distributed resources.

What could it be used for?

A Node or thematic community would use it to populate their own central index on top of which a portal could reside. GBIF will continue to use it to manage its internal indexing operations.

How does it work?

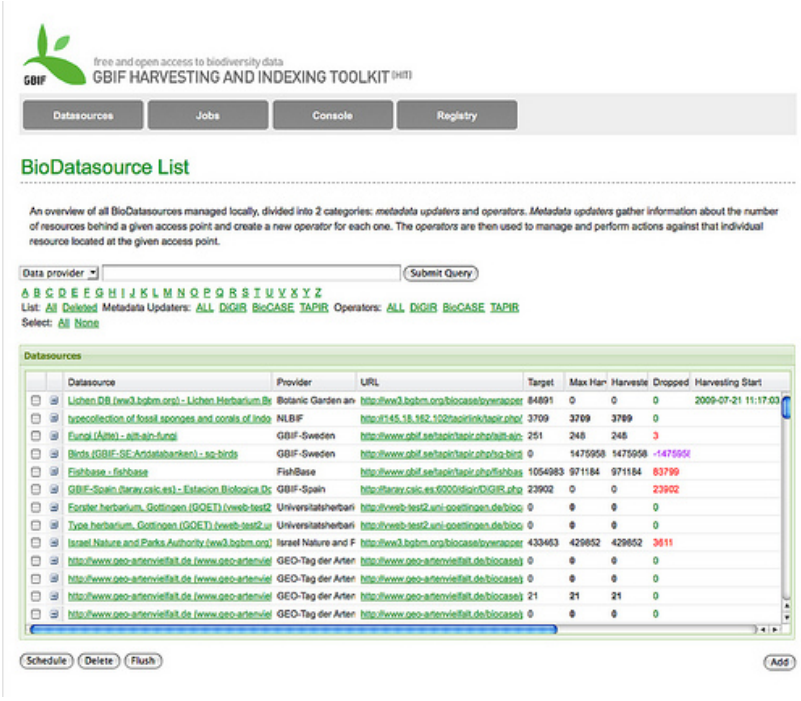
1. The HIT communicates with a central registry, gathering information about data publishers and their available access-points.
2. The HIT's operator then decides which datasets to schedule for harvesting.
3. Harvested records get saved to intermediary text files.
4. These records are then synchronised with the central index.

What protocols does it support?

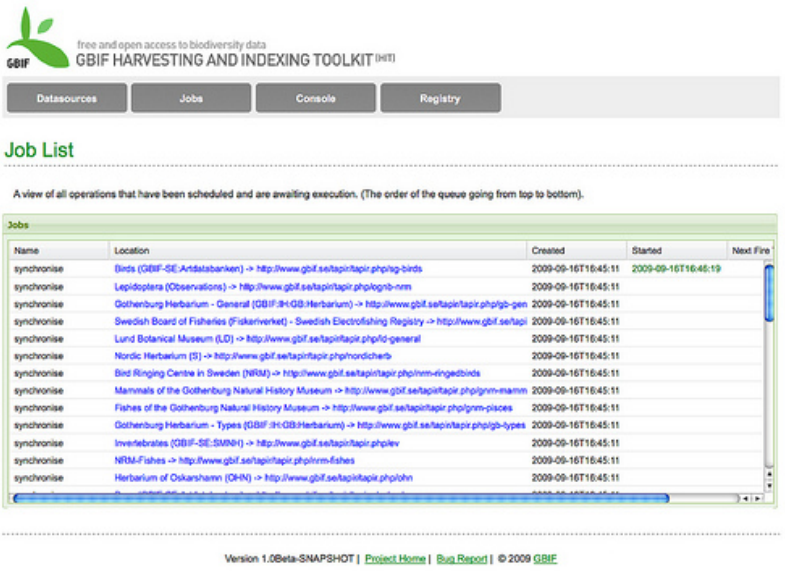
Currently, DiGIR, TAPIR, BioCASe, and Darwin Core Archive are all supported. The HIT is entirely customisable, however, and could be extended to include other protocols.

What are some of its main features?

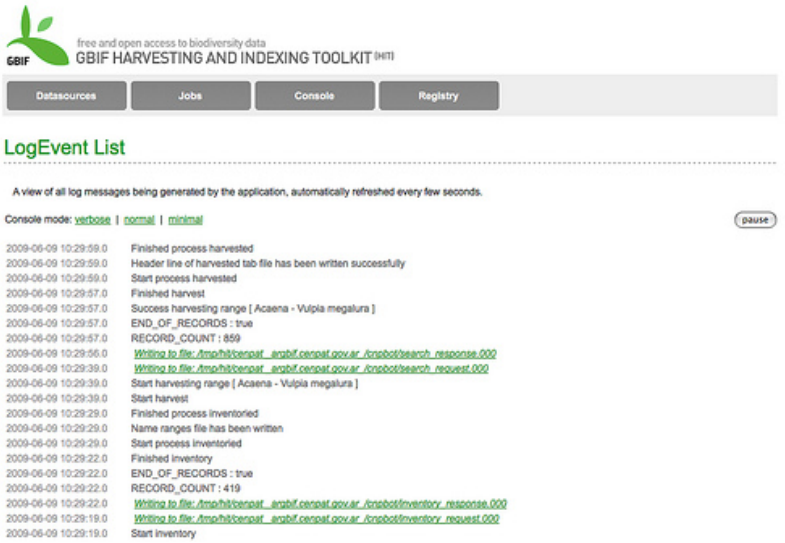
- A user-friendly dashboard from which to manage all harvesting activities.



- Multi-threaded execution, allowing several harvests to be carried out concurrently.



- Harvesting and indexing operations entirely decoupled from each other, allowing for increased performance and flexibility.
- Controlled workflows.
- Comprehensive logging, permitting the user to quickly identify errors.



- Data validation.

What are its current limitations?

- It can only index into a database mirroring the GBIF data structure. Adopters could add extensions allowing for indexing into other structures.
- It does not contain time-based job scheduling (Due in 1.1 release candidate)

INSTALLATION questions

What should I do when getting a java.security exception?

On platforms such as Ubuntu, the Java security manager can be quite restrictive. To make them less restrictive, and hence to allow the application to be able to run, it's necessary to modify the catalina.policy file, which is how the security policies implemented by the Java SecurityManager are configured. This file is located in the tomcat conf directory. Open this file, and add the following snippet of code (replacing the ${tomcat.directory} with the actual path where your tomcat installation is located).

```
// HIT webapp permission
grant codebase "file:${tomcat.directory}/webapps/hit/-" {
        permission java.io.FilePermission "${tomcat.directory}/webapps/hit/-",
"read,write,delete";
        permission java.util.PropertyPermission "*", "read,write";
        permission java.util.logging.LoggingPermission  "control";
        permission java.lang.RuntimePermission "getClassLoader";
        permission java.lang.RuntimePermission "preferences";
        permission java.lang.RuntimePermission "shutdownHooks";
        permission java.lang.reflect.ReflectPermission "suppressAccessChecks";
};
```

Some additional properties other users have added to get things running include:

```
permission java.net.SocketPermission "*:3306", "connect, resolve";
permission java.lang.RuntimePermission "getProtectionDomain";
permission java.util.PropertyPermission "cglib.debugLocation", "read";
```

Then restart Tomcat, and the exception should disappear. Otherwise, create a new issue or post a question to the mailing list.

What should I do when getting an outOfMemoryError permGen space?

The problem here is that the PermGen space is too small. To fix this, it is necessary to increase its size by playing with the flags -XX:PermSize=128m and -XX:MaxPermSize=256m (the default value is 64M).

Please see the section entitled Tomcat configuration for help with setting this JVM environment variable.

Hopefully this fixes the problem, otherwise, try incrementing it in increments of 128MB until the error goes away (assuming you actually have sufficient memory to allocate)

What should I do when getting a SEVERE: Exception starting filter struts java.lang.NullPointerException at com.opensymphony.xwork2.util.FileManager$FileRevision.needsReloading

This error first appeared on Windows XP. It seems like this happens when tomcat is installed in a directory that has spaces, such as C:\Program Files\apache-tomcat-6.0.20\ Moving the installation to C:\apache-tomcat-6.0.20 fixed the problem, and allowed the application to run.

OPERATION/USAGE questions

Why do I get a connection timed out error?

- Answer: It could be because your connection to the internet has been lost.
- Answer: It could be because a port, or ip has been blocked on your machine. It is a good idea to try pinging the publisher you are trying to connect to from the same machine where the HIT is installed. It is also a good idea to check your firewall settings.

Why do I harvest LESS than 100% of the target records?

- Answer: It could be because there was a problem constructing the name ranges file. To ensure that the proper name ranges were constructed:
  - examine the name ranges file for any peculiarities - relates to this FAQ
- Answer: It could be because there was a problem parsing some of the XML responses. To double check, you could do the following:
  - check the logs to see that there were no parsing errors
  - check the actual search response(s), to see that they in fact contain records and that these correspond to the appropriate name range.

Why is there a problem with the ranges in my name ranges file?

- Answer: The most probable answer, is that there was a problem with the inventory response in that it did not contain proper scientific names. To double check: it would be advisable to:
  - check that the inventory response(s) was actually returned
  - check that there were no XML parsing errors

- check the logs to see that no illegal characters had to be replaced from any of the names
- check the logs to see that there were no invalid scientific names as these may have been skipped

Why do I harvest MORE than 100% of the target records?

- Answer: One source of inflated record count in Darwin-Core archives can be illegal line terminating characters. A record containing such a character would break in two and appear to the parser as two lines with an insufficient number of columns. Consequently these two lines would be replaced by blank lines but still appear in the record count turning a single line into two.
  - Search for lines containing line terminating characters *inside* the records and remove these