

BiNHum Harvesting and Indexing Toolkit

Contents

Pre-requisites.....	1
Basics	1
Web-services	1
Configuration	2
Edit the application.properties file	2
Edit applicationContext-security.xml.....	2
Before starting the app.....	3
Starting B-HIT.....	3
Overview.....	3
Add a datasource.....	4
Retrieve datasets	4
Perform inventory	5
Harvest data.....	5
Process harvested records.....	5

Pre-requisites

Basics

Tested with Java 8

Can be deployed with Tomcat (tested with Apache Tomcat/6.0.39 JRE 1.8.0_05-b13) or Jetty (tested with Jetty 8.1.14), or can be run inside Eclipse (tested with Eclipse Luna and Kepler)

MySQL Database v5.5

Web-services

Local gisgraphy installation (coordinates check) or maybe online version, see <http://www.gisgraphy.com/free-access.htm>

CoordinatesKdTree webservice (based on <https://github.com/AReallyGoodName/OfflineReverseGeocode>)

Configuration

Edit the application.properties file

Quality tests	
<ul style="list-style-type: none">Quality tests: activated / deactivated. The tab won't be displayed if turned OFF	qualityOnOff= OFF / ON
<ul style="list-style-type: none">Where the CoordinatesKDTree Web service is running	coordinatesWebservice= http://localhost:8080/CoordinatesWS
A local folder for temporary file manipulation	temporaryFolder= /home/user/test
Database properties	
<ul style="list-style-type: none">DB credentials	dataSource.username= USER dataSource.password= PASS
<ul style="list-style-type: none">DB name	dataSource.name= DATABASE_NAME
<ul style="list-style-type: none">MySQL-Server IP + DB name	dataSource.url=jdbc:mysql:// DATABASE_IP:3306/DATABASE_NAME?autoReconnect=true&useUnicode=true&characterEncoding=UTF8&characterSetResults=UTF8&MVCC=true
Application properties	
<ul style="list-style-type: none">Application server IP (ie. Localhost)	dataSource.servername= APPLICATIONSERVER_IP
<ul style="list-style-type: none">default URL with port (ie. http://localhost:8040/) The port must match your Tomcat/Jetty configuration !	baseUrl= http://APPLICATIONSERVER:PORT/
<ul style="list-style-type: none">directory (with all permissions for Tomcat/Jetty) where to store the harvested files	harvest.directory= /home/USER/PROJECT/

Edit applicationContext-security.xml

Edit the admin password, md5 encoded. The default password corresponds to « banana! »

```
<user name="admin" password="bb7a307e32b93a931da89d0a214dd47f" authorities="ROLE_ADMIN" />
```

Before starting the app

Run the database creation script (schemaOnly.sql).

Starting B-HIT

Based on your configuration, and the name chosen (eclipse config / war file name), open your favorite web-browser to <http://localhost:8040/Bibhum/datasource/list.html> .

Overview

The screenshot displays the B-HIT web application interface. At the top, there is a navigation menu with buttons for 'Datasources', 'Associated Data', 'Extra units', 'Jobs', 'Console', 'Report', 'Datasource Management', 'Data quality summary', and 'Data viewer'. The 'Datasources' button is highlighted. Below the menu, the page title is 'BioDatasource List'. A 'login' form is visible in the top right corner. The main content area contains a table with columns: 'Available Methods', 'Provider Name', 'Datasource', 'URL', 'Target', 'Harvested', 'Started', 'Last Inventory', 'Last Inv. processed', 'Last harvesting', 'Last harv. processed', and 'Country'. The table currently shows 'No biodatasource yet'. A 'schedule' button is located at the bottom left, and an 'add bioDatasource' button is at the bottom right. The footer indicates 'Based on Version 1.48, from GBIF | © 2012- BINHum'.

Main menu:

- Datasources : main entry point to add datasources/datasets, launch metadata operations, inventory+harvesting+processing data
- Associated datasources: entry point for associated data (relationships), to harvest and process associated data only
- Extra units: entry point for single units retrieval, based on a list of unit IDs
- Jobs: overview of waiting and running jobs
- Console: overview of log events
- Report: generation of reports (statistics)
- Datasource management: to hide or delete datasets
- Data quality : to launch quality tests, exports test results and display results
- Data viewer: to display data stored in the database, either the raw data or the improved data from the quality tests

Add a datasource

Click on the button « add bioDatasource »

BioDatasource Detail

Configure your BioDatasource

BioDatasource Name:

Provider Name (abbrev):

=Accesspoint URL:

Provider Full Name:

Provider Website URL:

Provider Address:

Factory class:

Country:

Enter a name for the datasource, the provider name abbreviated (ie. BGBM), the provider fullname (ie. Botanischer Garten und Botanisches Museum Berlin-Dahlem), the accesspoint (ie. http://ww3.bgbm.org/biocase/py-wrapper.cgi?dsa=test_ABCD21), the provider website (ie. <http://www.bgbm.org>), the provider postal address and country, and the type of datasource (ie. Biocase/digir/Tapir/darwincore archive). Then save. The new datasource is now listed on the <http://localhost:8040/Bibhum/datasource/list.html> page.

To see the available operations for this datasource, click on the « available methods » checkbox on the left.

Available Methods	Provider Name	Datasource	URL	Target	Harvested	Started	Last Inventory	Last Inv. processed	Last harvesting	Last harv. processed	Country
<input checked="" type="checkbox"/> Metadata update	BGBM	Test ABCD	http://ww3.bgbm.org/biocase/py-wrapper.cgi?dsa=test_ABCD21								Germany

Retrieve datasets

In order to trigger this « Metadata update » operation, check the box and click on « Schedule » at the bottom of the page.

A new Job is created and can be followed under the « Job » tab (</Bibhum/job/list.html>).

Job List

A view of all operations that have been scheduled and are awaiting execution. Please note that the maximum number of operations that can be run in parallel is 500.

kill: alt: reschedule: id:

Show entries

Search:

ID	Name	Description	Created	Started
1	IssueMetadata	BGBM : Test ABCD -> ww3.bgbm.org	2015-03-04 16:05:16	

Showing 1 to 1 of 1 entries

Previous Next

Refresh the page to update the job status :

ID	Name	Description	Created	Started
1	IssueMetadata	BGBM : Test ABCD -> ww3.bgbm.org	2015-03-04 16:05:16	2015-03-04 16:06:10

Going back to the main page, a new line appeared in the list of datasource for the discovered dataset, with its corresponding available methods. These methods depend on the type of provider. ABCD-Archive will enable the methods XXXXXXXXXXXXXXXXXXXX, Dwc-A will enable YYYYYYYYYY.

Available Methods	Provider Name	Datasource	URL	Target	Harvested	Started	Last inventory	Last inv. processed	Last harvesting	Last harv. processed	Country
<input checked="" type="checkbox"/> Inventory <input checked="" type="checkbox"/> Harvest <input checked="" type="checkbox"/> Process harvested records <input checked="" type="checkbox"/> Harvest sibling units <input checked="" type="checkbox"/> Process sibling units	BGBM	Test ABCD - Herbarium Berlinense	http://ww3.bgbm.org/blocase/pywrapper.cgi?dsa=test_ABCD21	192551	0						Germany
<input type="checkbox"/>	BGBM	Test ABCD	http://ww3.bgbm.org/blocase/pywrapper.cgi?dsa=test_ABCD21						04-03-2015		Germany

DEBUGGING HINT: The most common reason that a metadata update fails, is that the accesspoint is wrong.

The target column contains the theoretical number of units, according to the providers metadata.

Perform inventory

Select the first operation and click on « schedule ». The queries are stored on the disc, in compressed files. The location path is defined in the configuration file, and the directory for each datasource can be found in the MySQL table `bio_datasource`, column « basedirectory ».

Harvest data

Data will be harvested if and only if a list of units has been retrieved with the Inventory. The outputs from this operation are:

1. One or more search requests (with enumerated extensions corresponding to the order in which they were dispatched, i.e. `search_request.000`)
2. One or more search responses (with enumerated extensions corresponding to the order in which they were dispatched, i.e. `search_response.000`). Often, there will only be a single response per request, but sometimes there can be multiple responses for a single request!

The default range is 200 units per request.

Process harvested records

In this operation, an Operator (BioDatasource) will collect all the search responses, parse them, and write the parsed values to the database. In case of re-harvesting a dataset, the system will check if the data changed since the last harvesting. It will calculate the new checksum of each downloaded file, and will modify data in the DB if and only if the checksum changed. Each checksum is stored in the DB table « `sha1responses` ».