Opening Up the Natural History Heritage for Europeana

D08 – DELIVERABLE 2.4.2

Project Acronym: OpenUp!

Grant Agreement No: 270890

Project Title: Opening up the Natural History Heritage for Europeana

OAI-Provider Interface production version

Part 3: Step by step example

D08 – Deliverable 2.4.2

Revision: Final

Authors:

Astrid Höller AIT Forschungsgesellschaft mbH

Odo Benda AIT Forschungsgesellschaft mbH

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

0 REVISION AND DISTRIBUTION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
Draft	2012-02-15	A. Höller	AIT	First Version (Draft)
1	2012-02-17	A. Höller, G. Koch, W. Koch	AIT	Revision
1b	2012-02-24	Coordination Team	BGBM	Minor editing
NB : The Software and user interface was developed in collaboration and with input from the TMG over the past 6 months				

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Distribution

Recipient	Date	Version	Accepted YES/NO
TMG	2012-02-17	1	
Work Package Leader WP2 (G. Malarky, NHM)	2012-02-22	1	YES
Project Coordinator	2012-02-24	1b	YES

TABLE OF CONTENTS

1 DESCRIPTION OF WORK

This document illustrates the complete procedure of harvesting, transforming and uploading data during the OpenUp! project. This includes harvesting datasources from the data provider BioCASe with the GBIF Harvesting and Indexing Toolkit (HIT), transforming the harvested ABCD records with Pentaho Kettle and finally uploading the created ESE records on the OAI-Provider-platform with the Zebra information management system. Figure 1 gives an overview of the whole process. As you can see the data has to pass six steps before it is finally delivered to Europeana.

1. Message from Data Provider that a new datasource is available
2. Harvesting the datasource with the GBIF-HIT Harvester
3. Transforming the ABCD files into ESE records with Pentaho Kettle (Data Transformation)
4. Informing the OpenUp! Meta Data Management that the data is transformed
5. (Informing us if the data was correct)
6. Uploading the records on the OAI-Provider-platform
7. Deliver the data to Europeana

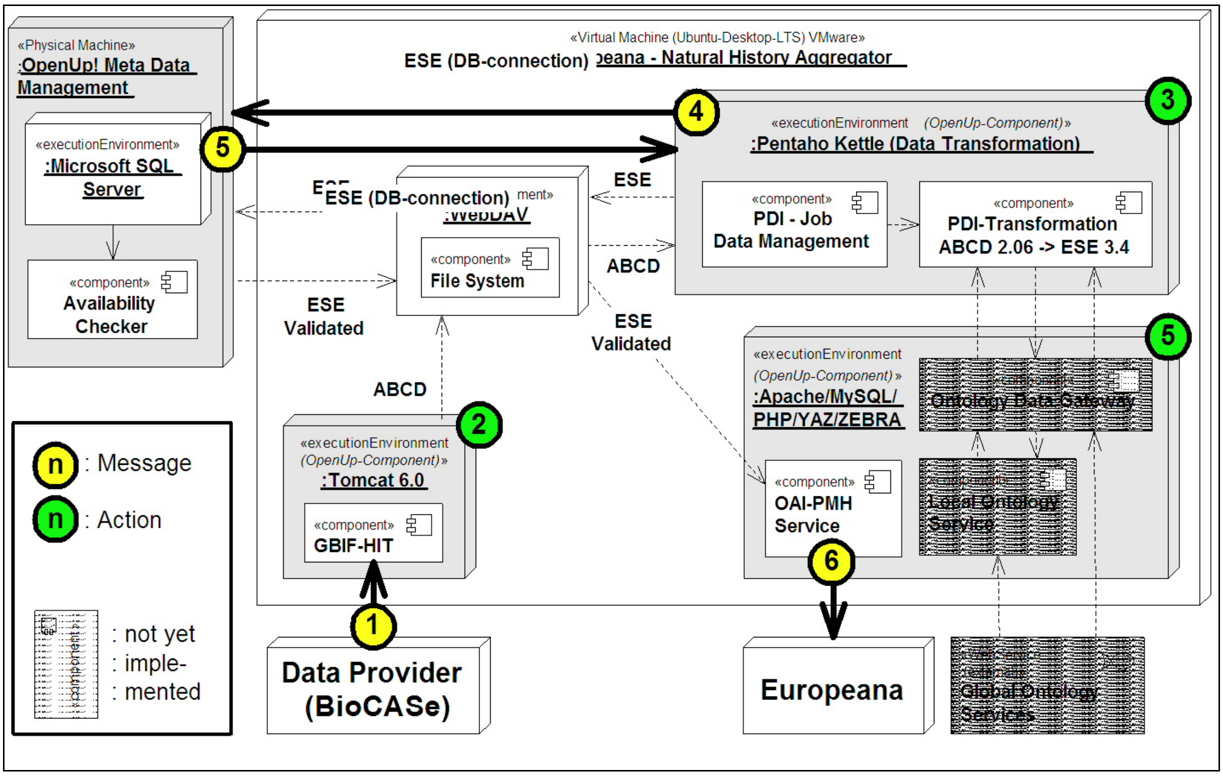

«Physical Machine»
:OpenUp! Meta Data Management
«executionEnvironment»
:Microsoft SQL Server
«component»
Availability Checker
«Virtual Machine (Ubuntu-Desktop-LTS) VMware»
ESE (DB-connection) peana - Natural History Aggregator
«executionEnvironment (OpenUp-Component)»
:Pentaho Kettle (Data Transformation)
«component»
PDI - Job Data Management
«component»
PDI-Transformation ABCD 2.06 -> ESE 3.4
ESE
ESE (DB-connection)
:WebDAV
«component»
File System
ABCD
ESE Validated
ESE Validated
ABCD
«executionEnvironment (OpenUp-Component)»
:Tomcat 6.0
«component»
GBIF-HIT
«executionEnvironment (OpenUp-Component)»
:Apache/MySQL/ PHP/YAZ/ZEBRA
Ontology Data Gateway
«component»
OAI-PMH Service
Local Ontology Service
Data Provider (BioCASe)
Europeana
Global Ontology Services
1
2
3
4
5
5
6
n : Message
n : Action
: not yet : imple- : mented

Figure 1 Diagram showing the main steps of the OpenUp! – Europeana ingest procedure

In the next chapters we will process an example datasource step by step. In this document we are concentrating on the three Action steps (compare Figure 1): **The HIT Harvester (step 2), Pentaho Kettle (step 3) and the OAI-PMH-Service (step 5).**

2 THE GBIF HARVESTING AND INDEXING TOOLKIT (HIT)

The GBIF HIT is a simple to use, simple to extend open source framework that allows you to easily manage data harvesting and quickly build specific indexes of harvested data.[1]

We are starting with installing the latest version of this software.

2.1 Installation guide[2]

Before you can start you have to make sure you meet the technical requirements. You need:

- A Java Runtime Environment Version 6 or higher[3]
- A web server with a servlet container (in our example it is Apache Tomcat)[4]
- MySQL version 5.1 or higher[5]

2.1.1 Creating the harvesting database (hit)[6]

When you have installed MySQL you can create the harvesting database. In this document it is called “hit”. To create the database type the following command:

mysql>create database hit DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;

Then you have to download the harvesting database's schema file (see footnote 6) and use the database with the command:

mysql>use hit;

Finally you can load the schema with the command

mysql>source ${download_location}/hit.sql

2.1.2 Setting up the indexing database (portal)[7]

To create the database “portal” use the command

mysql>create database portal DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;

[1] http://code.google.com/p/gbif-indexingtoolkit/ 17 Feb. 2012.

[2] http://code.google.com/p/gbif-indexingtoolkit/wiki/Installation 17 Feb. 2012.

[3] http://www.oracle.com/technetwork/java/index.html 17 Feb. 2012.

[4] http://tomcat.apache.org/) 17 Feb. 2012.

[5] http://www.mysql.com/ 17 Feb. 2012.

[6] http://code.google.com/p/gbif-indexingtoolkit/wiki/Installation#Set_up_the_harvesting_%28%27hit%27%29_database 17 Feb. 2012.

[7] http://code.google.com/p/gbif-indexingtoolkit/wiki/Installation#Set_up_the_indexing_%28%27portal%27%29_database 17 Feb. 2012.

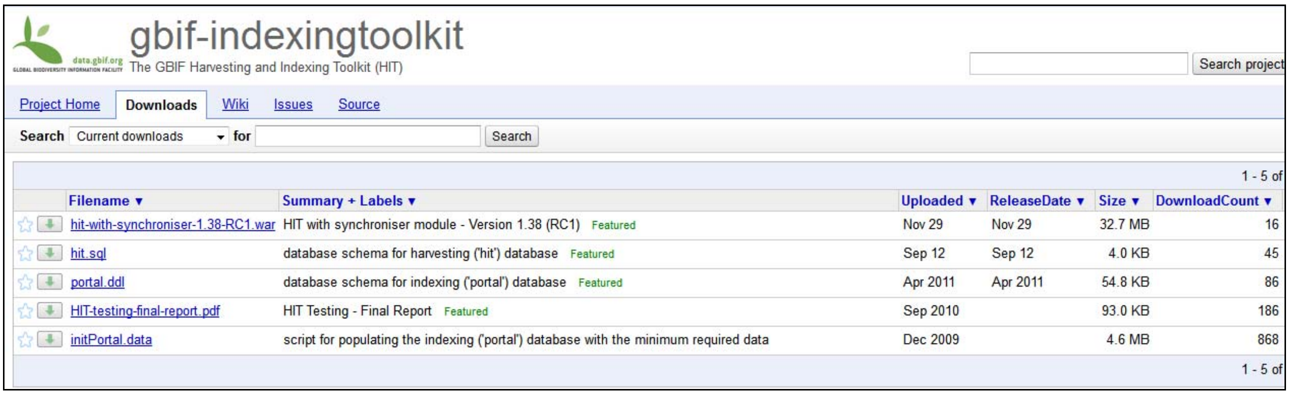
Then download the script responsible for populating the portal database (see footnote 7) and type

mysql>source ${download_location}/initPortal.data

2.1.3 Installing the application[8]

The instructions now given are specific for Tomcat.

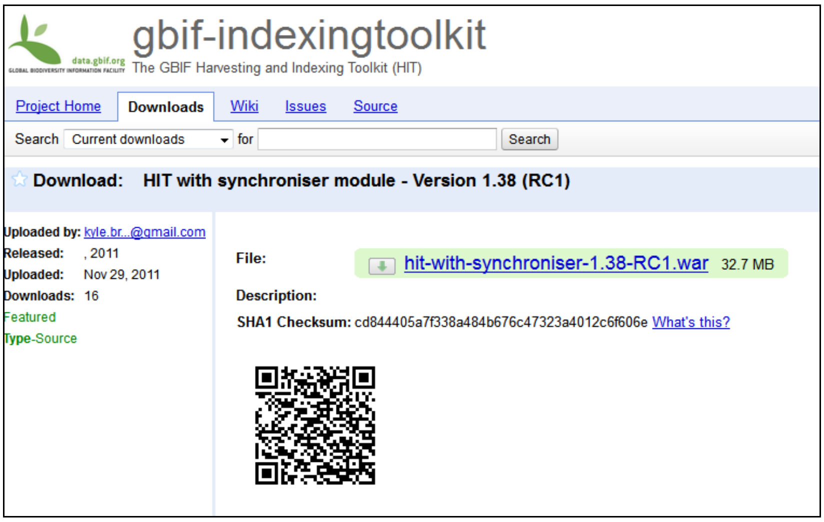
First you have to download the web archive file (with the suffix .war) from the project's download page[9]. In Figure 2 you can see this page with different files. The first one is the war-file we need: hit-with-synchroniser-1.38-RC1.war .



Filename ▾	Summary + Labels ▾	Uploaded ▾	ReleaseDate ▾	Size ▾	DownloadCount ▾
hit-with-synchroniser-1.38-RC1.war	HIT with synchroniser module - Version 1.38 (RC1) Featured	Nov 29	Nov 29	32.7 MB	16
hit.sql	database schema for harvesting ('hit') database Featured	Sep 12	Sep 12	4.0 KB	45
portal.ddl	database schema for indexing ('portal') database Featured	Apr 2011	Apr 2011	54.8 KB	86
HIT-testing-final-report.pdf	HIT Testing - Final Report Featured	Sep 2010		93.0 KB	186
initPortal.data	script for populating the indexing ('portal') database with the minimum required data	Dec 2009		4.6 MB	868

Figure 2 The download page of the gbif-indexingtoolkit

When clicking on this first file you will be forwarded to the page you can see in Figure 3. After clicking once again at the filename a download window will pop up and you can decide where to save the file.



Download: HIT with synchroniser module - Version 1.38 (RC1)

Uploaded by: kyle.br...@gmail.com
Released: , 2011
Uploaded: Nov 29, 2011
Downloads: 16
Featured
Type-Source

File: hit-with-synchroniser-1.38-RC1.war 32.7 MB

Description:

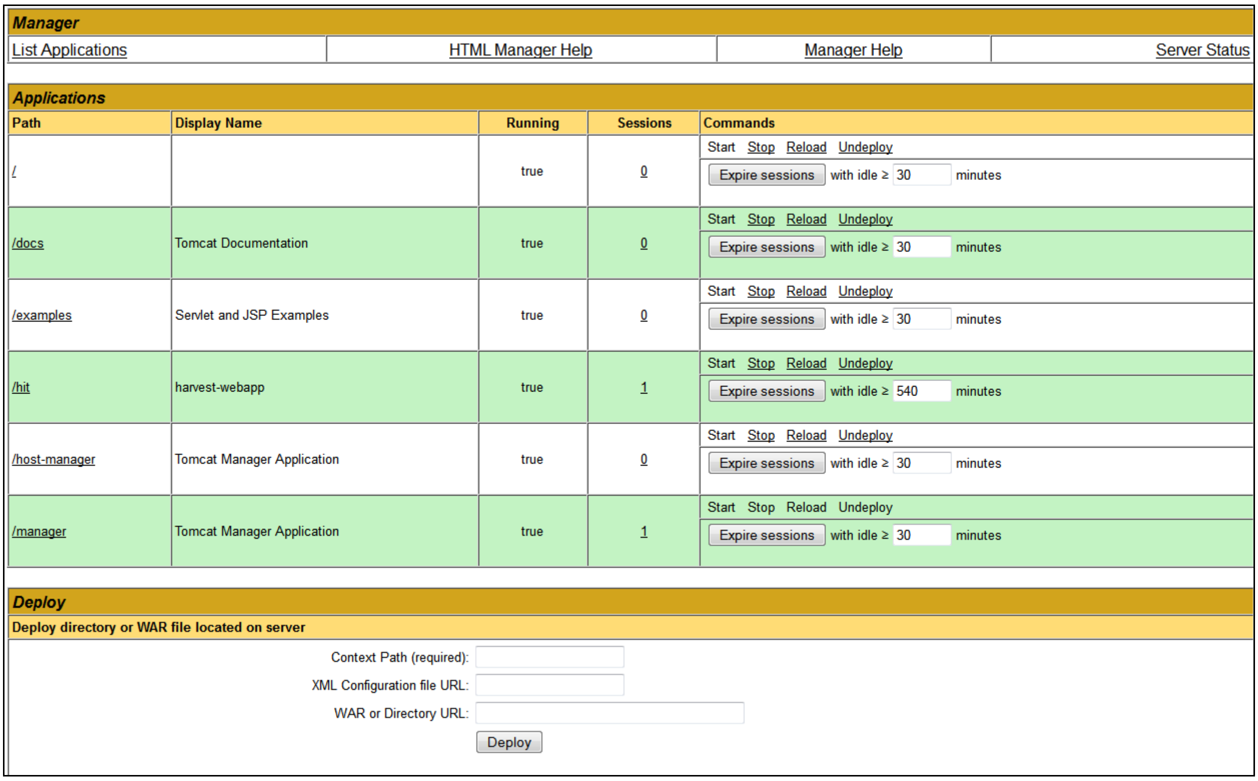
SHA1 Checksum: cd844405a7f338a484b676c47323a4012c6f606e What's this?

Figure 3 Downloading the hit-with-synchroniser-1.38-RC1.war file

[8] http://code.google.com/p/gbif-indexingtoolkit/wiki/Installation#Install_the_application 17 Feb. 2012.

[9] http://code.google.com/p/gbif-indexingtoolkit/downloads/list 17 Feb. 2012.

When this is done go to your Tomcat Manager (see Figure 4).



Manager

List Applications	HTML Manager Help	Manager Help	Server Status

Applications

Path	Display Name	Running	Sessions	Commands
/		true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/docs	Tomcat Documentation	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/examples	Servlet and JSP Examples	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/hit	harvest-webapp	true	1	Start Stop Reload Undeploy Expire sessions with idle ≥ 540 minutes
/host-manager	Tomcat Manager Application	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/manager	Tomcat Manager Application	true	1	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes

Deploy

Deploy directory or WAR file located on server

Context Path (required):

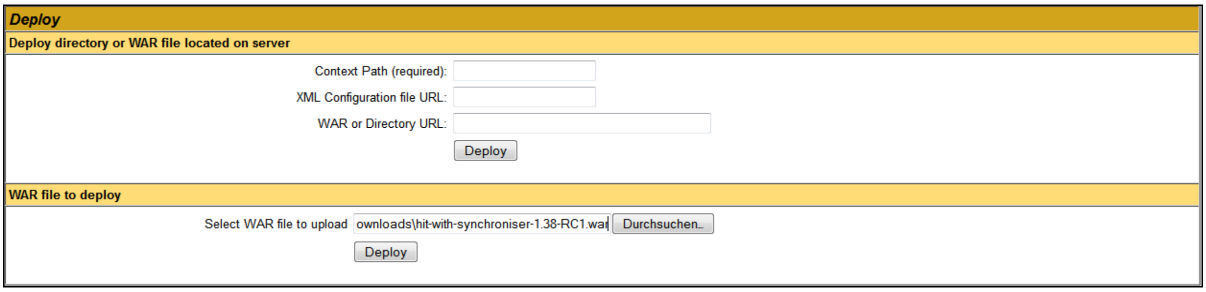
XML Configuration file URL:

WAR or Directory URL:

Deploy

Figure 4 The Tomcat Manager

At the bottom of the page you can see two sections: Deploy directory or WAR file located on server and WAR file to deploy. For deploying your before- downloaded war file you have to click on "Browse" and select the war file where you have saved it before. When the right file has been chosen click on "Deploy" (see Figure 5).



Deploy

Deploy directory or WAR file located on server

Context Path (required):

XML Configuration file URL:

WAR or Directory URL:

Deploy

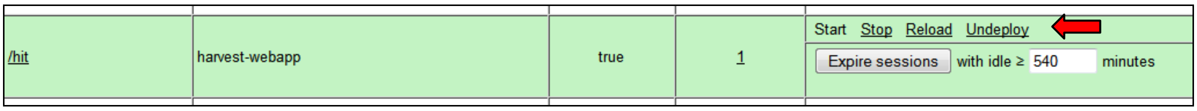
WAR file to deploy

Select WAR file to upload ownloads\hit-with-synchroniser-1.38-RC1.war Durchsuchen...

Deploy

Figure 5 Deploying the file hit-with-synchroniser-1.38-RC1.war

If there is already an old hit.war file you have to undeploy this file first (see Figure 6).



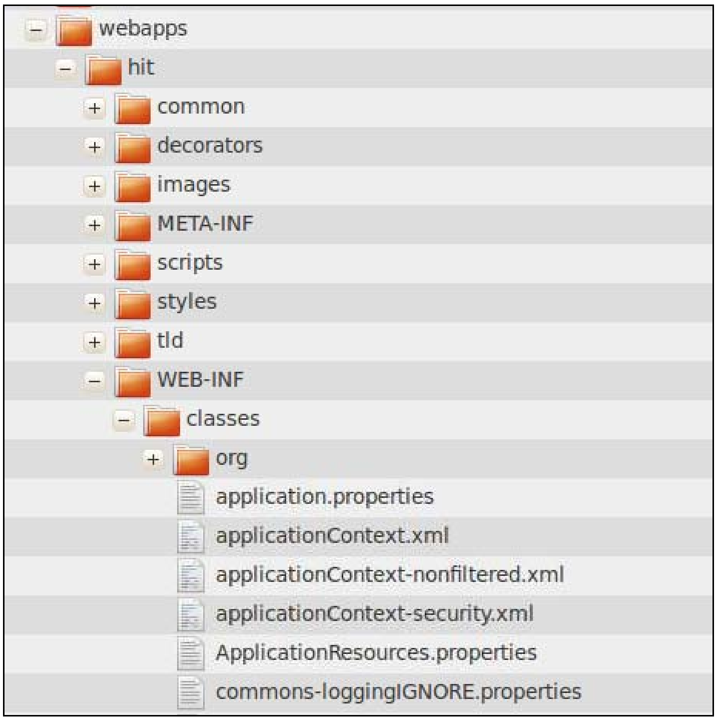
/hit	harvest-webapp	true	1	Start Stop Reload Undeploy Expire sessions with idle ≥ 540 minutes

Figure 6 Undeploying an old hit.war file

When the correct .war file has been deployed you have to adapt some parameters.

2.1.4 Configuring the application[10]

The first configuration takes place in var/lib/tomcat6/webapps/hit/WEB-INF/classes/application.properties (see Figure 7).



```
– webapps
  – hit
    + common
    + decorators
    + images
    + META-INF
    + scripts
    + styles
    + tld
    – WEB-INF
      – classes
        + org
          application.properties
          applicationContext.xml
          applicationContext-nonfiltered.xml
          applicationContext-security.xml
          ApplicationResources.properties
          commons-loggingIGNORE.properties
```

Figure 7 Going to application.properties

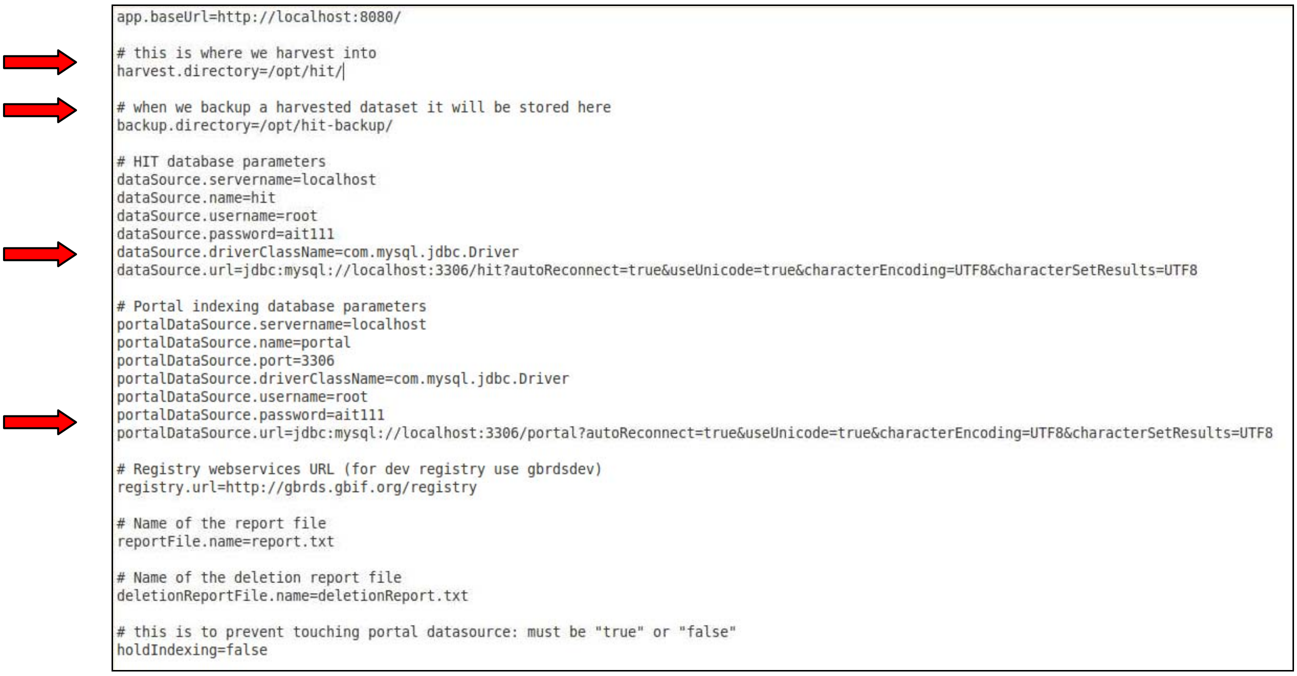
When you have opened application.properties you can modify the harvest and the backup directory (see Figure 8). As you can see our harvest directory is opt/hit and our backup directory opt/hit-backup.

Furthermore you can change your database parameters including the URL (see Figure 8).

Finally go to var/lib/tomcat6/webapps/hit/WEB-INF/classes/applicationContext-security.xml (compare Figure 7) to configure the user management (see Figure 9). You can change the default username and password. The password can only be replaced by a password that has been md5 encoded. You can use an online md5 encoder to create a new password.[11]

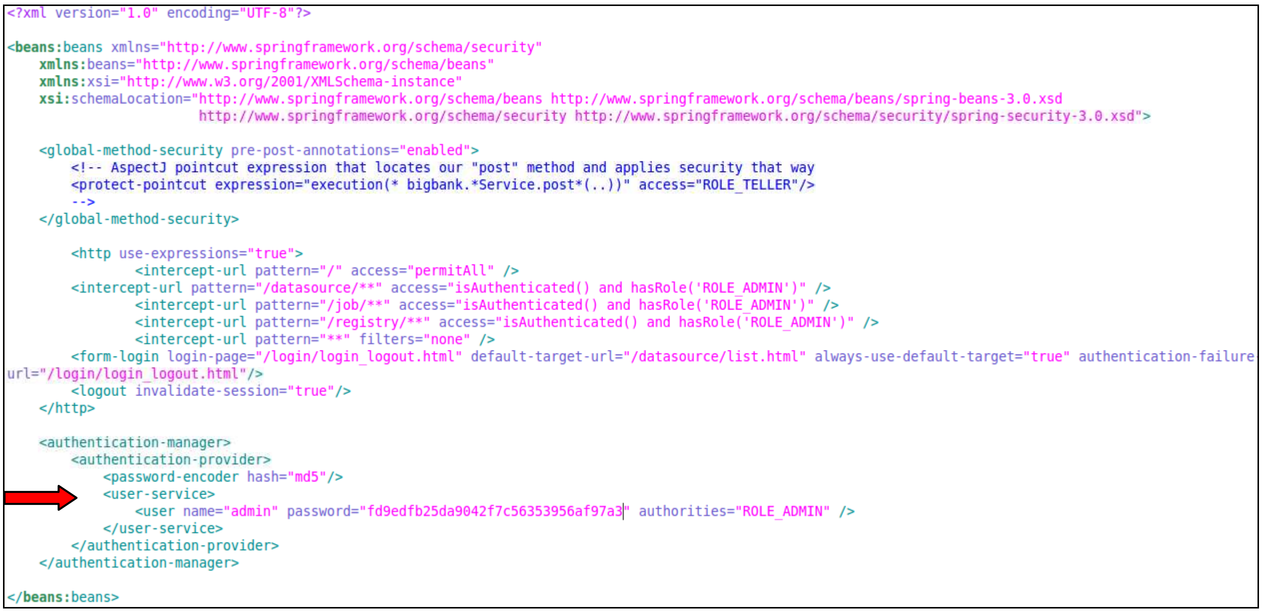
[10] http://code.google.com/p/gbif-indexingtoolkit/wiki/Installation#Configure_the_application 17 Feb. 2012.

[11] http://7thspace.com/webmaster_tools/online_md5_encoder.html 17 Feb. 2012.



```
app.baseUrl=http://localhost:8080/

# this is where we harvest into
harvest.directory=/opt/hit/

# when we backup a harvested dataset it will be stored here
backup.directory=/opt/hit-backup/

# HIT database parameters
dataSource.servername=localhost
dataSource.name=hit
dataSource.username=root
dataSource.password=ait111
dataSource.driverClassName=com.mysql.jdbc.Driver
dataSource.url=jdbc:mysql://localhost:3306/hit?autoReconnect=true&useUnicode=true&characterEncoding=UTF8&characterSetResults=UTF8

# Portal indexing database parameters
portalDataSource.servername=localhost
portalDataSource.name=portal
portalDataSource.port=3306
portalDataSource.driverClassName=com.mysql.jdbc.Driver
portalDataSource.username=root
portalDataSource.password=ait111
portalDataSource.url=jdbc:mysql://localhost:3306/portal?autoReconnect=true&useUnicode=true&characterEncoding=UTF8&characterSetResults=UTF8

# Registry webservices URL (for dev registry use gbrdsdev)
registry.url=http://gbrds.gbif.org/registry

# Name of the report file
reportFile.name=report.txt

# Name of the deletion report file
deletionReportFile.name=deletionReport.txt

# this is to prevent touching portal datasource: must be "true" or "false"
holdIndexing=false
```

Figure 8 Modifying the harvest and the backup directory and the database properties

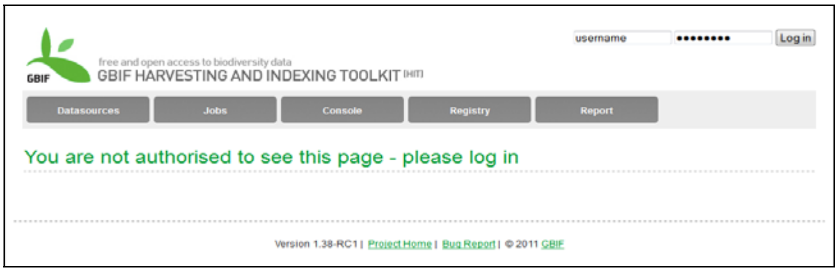


```
<?xml version="1.0" encoding="UTF-8"?>

<beans:beans xmlns="http://www.springframework.org/schema/security"
    xmlns:beans="http://www.springframework.org/schema/beans"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.springframework.org/schema/beans http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
                        http://www.springframework.org/schema/security http://www.springframework.org/schema/security/spring-security-3.0.xsd">

    <global-method-security pre-post-annotations="enabled">
        <!-- AspectJ pointcut expression that locates our "post" method and applies security that way
        <protect-pointcut expression="execution(* bigbank.*Service.post*(..))" access="ROLE_TELLER"/>
        -->
    </global-method-security>

        <http use-expressions="true">
                <intercept-url pattern="/" access="permitAll" />
        <intercept-url pattern="/datasource/**" access="isAuthenticated() and hasRole('ROLE_ADMIN')" />
                <intercept-url pattern="/job/**" access="isAuthenticated() and hasRole('ROLE_ADMIN')" />
                <intercept-url pattern="/registry/**" access="isAuthenticated() and hasRole('ROLE_ADMIN')" />
                <intercept-url pattern="**" filters="none" />
        <form-login login-page="/login/login_logout.html" default-target-url="/datasource/list.html" always-use-default-target="true" authentication-failure-
url="/login/login_logout.html"/>
        <logout invalidate-session="true"/>
    </http>

    <authentication-manager>
        <authentication-provider>
            <password-encoder hash="md5"/>
            <user-service>
                <user name="admin" password="fd9edfb25da9042f7c56353956af97a3" authorities="ROLE_ADMIN" />
            </user-service>
        </authentication-provider>
    </authentication-manager>

</beans:beans>
```

Figure 9 Configuring the user management in applicationContext-security.xml

2.2 User Interface of HIT[12]

When going to *[your domain[13]]* you can see the following window (see Figure 10).



username ••••••••• Log in

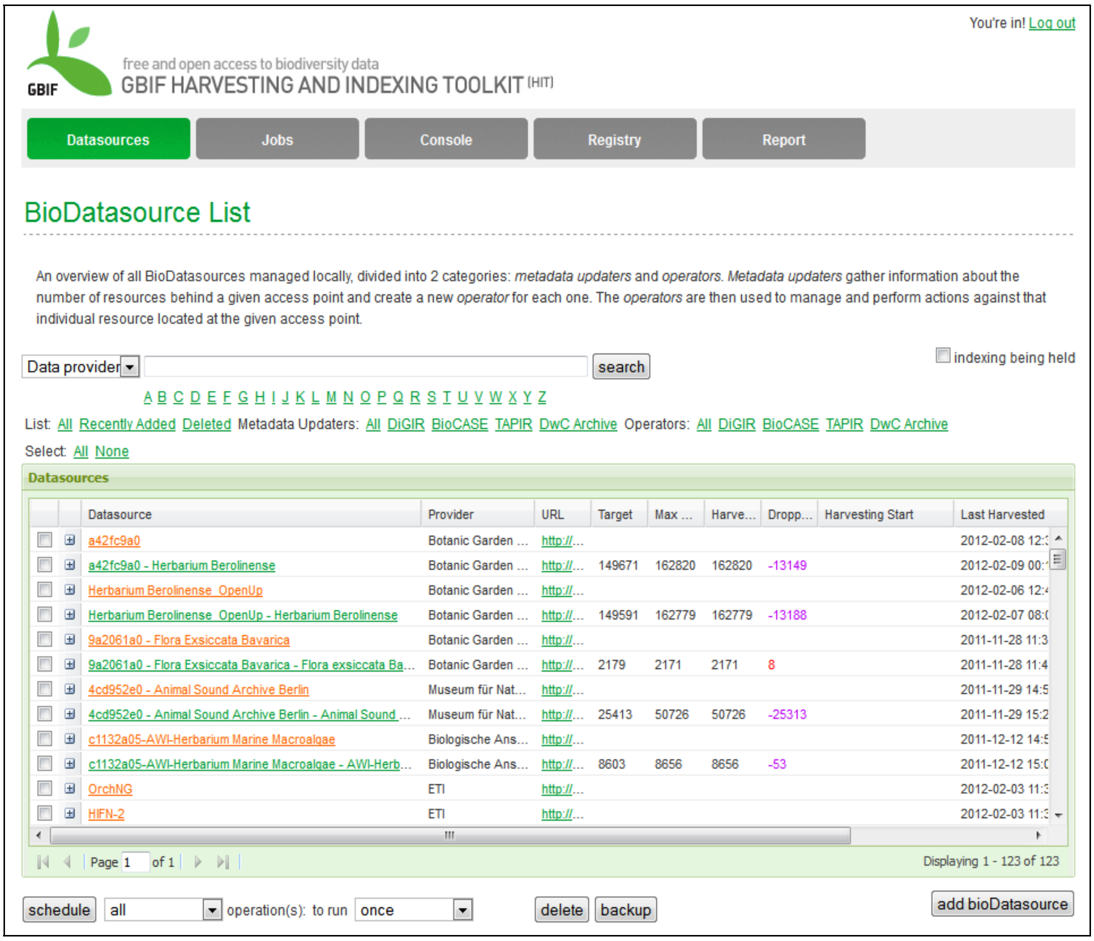
free and open access to biodiversity data
GBIF HARVESTING AND INDEXING TOOLKIT (HIT)

You are not authorised to see this page - please log in

Version 1.38-RC1 | Project Home | Bug Report | © 2011 GBIF

Figure 10 Logging in the GBIF Harvesting and Indexing Toolkit (HIT)

After logging in (in the upper right corner) you can see the interface of the HIT Harvester (see Figure 11).



You're in! Log out

free and open access to biodiversity data
GBIF HARVESTING AND INDEXING TOOLKIT (HIT)

BioDatasource List

An overview of all BioDatasources managed locally, divided into 2 categories: *metadata updaters* and *operators*. *Metadata updaters* gather information about the number of resources behind a given access point and create a new *operator* for each one. The *operators* are then used to manage and perform actions against that individual resource located at the given access point.

Data provider ▾ search □ indexing being held

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

List: All Recently Added Deleted Metadata Updaters: All DiGIR BioCASE TAPIR DwC Archive Operators: All DiGIR BioCASE TAPIR DwC Archive

Select: All None

Datasources

		Datasource	Provider	URL	Target	Max ...	Harve...	Dropp...	Harvesting Start	Last Harvested
□	⊞	a42fc9a0	Botanic Garden ...	http://...						2012-02-08 12:
□	⊞	a42fc9a0 - Herbarium Berolinense	Botanic Garden ...	http://...	149671	162820	162820	-13149		2012-02-09 00:
□	⊞	Herbarium Berolinense_OpenUp	Botanic Garden ...	http://...						2012-02-06 12:
□	⊞	Herbarium Berolinense_OpenUp - Herbarium Berolinense	Botanic Garden ...	http://...	149591	162779	162779	-13188		2012-02-07 08:
□	⊞	9a2061a0 - Flora Exsiccata Bavarica	Botanic Garden ...	http://...						2011-11-28 11:3
□	⊞	9a2061a0 - Flora Exsiccata Bavarica - Flora exsiccata Ba...	Botanic Garden ...	http://...	2179	2171	2171	8		2011-11-28 11:4
□	⊞	4cd952e0 - Animal Sound Archive Berlin	Museum für Nat...	http://...						2011-11-29 14:5
□	⊞	4cd952e0 - Animal Sound Archive Berlin - Animal Sound ...	Museum für Nat...	http://...	25413	50726	50726	-25313		2011-11-29 15:2
□	⊞	c1132a05-AWI-Herbarium Marine Macroalgae	Biologische Ans...	http://...						2011-12-12 14:
□	⊞	c1132a05-AWI-Herbarium Marine Macroalgae - AWI-Herb...	Biologische Ans...	http://...	8603	8656	8656	-53		2011-12-12 15:
□	⊞	OrchNG	ETI	http://...						2012-02-03 11:
□	⊞	HIFN-2	ETI	http://...						2012-02-03 11:

Page 1 of 1 Displaying 1 - 123 of 123

schedule all ▾ operation(s): to run once ▾ delete backup add bioDatasource

Figure 11 The HIT user interface

[12] http://code.google.com/p/gbif-indexingtoolkit/wiki/UserManual 17 Feb. 2012.

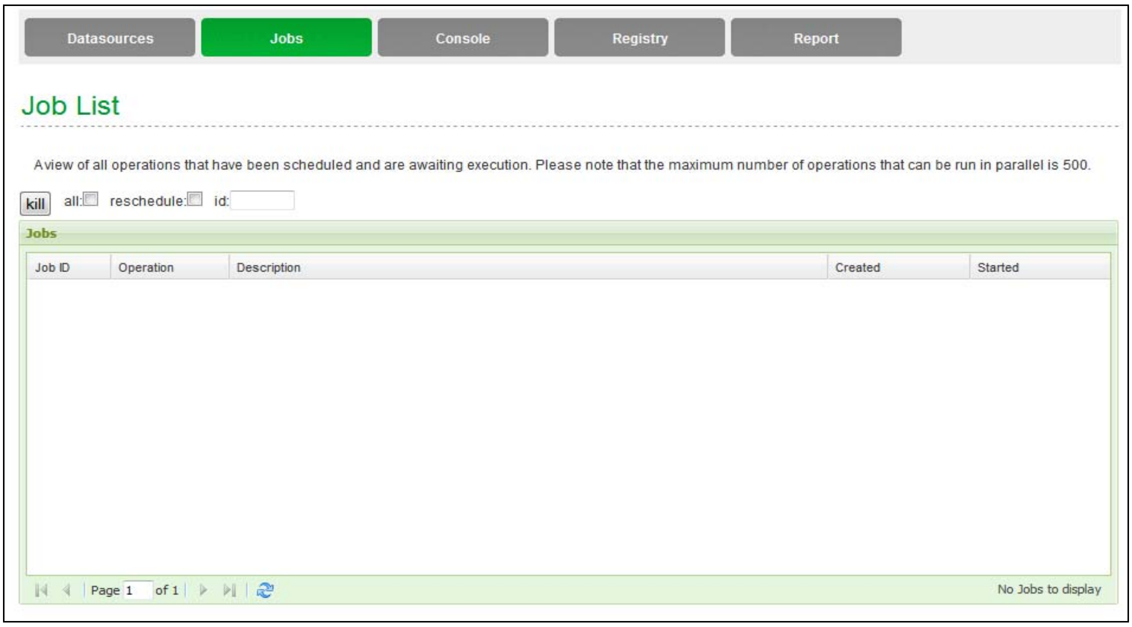
[13] for example http://localhost:8080/hit

There are five main sections: Datasources, Jobs, Console, Registry and Report. The tab used at the moment is always green (like Datasources in Figure 11), the others are grey.

In **Datasources** you can see all datasources that are available. The orange datasources are metadata updaters, the green ones operators. For both you can choose the protocol: DIGIR, BioCASE, TAPIR or DwC Archive. We are only working with the BioCASE protocol.

An operator is only created when a datasource has been created and the metadata updater has been successfully harvested. This case will be described in the next chapter.

When you click on the **Job**s tab you can see all jobs that have been started or jobs that are waiting for execution. The jobs are listed with their ID, name, description, their creation and their starting date (see Figure 12). If you decide to stop one or more jobs you can do that by filling in one id or checking “all” and click on the “kill” button. You can also reschedule a job.



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.

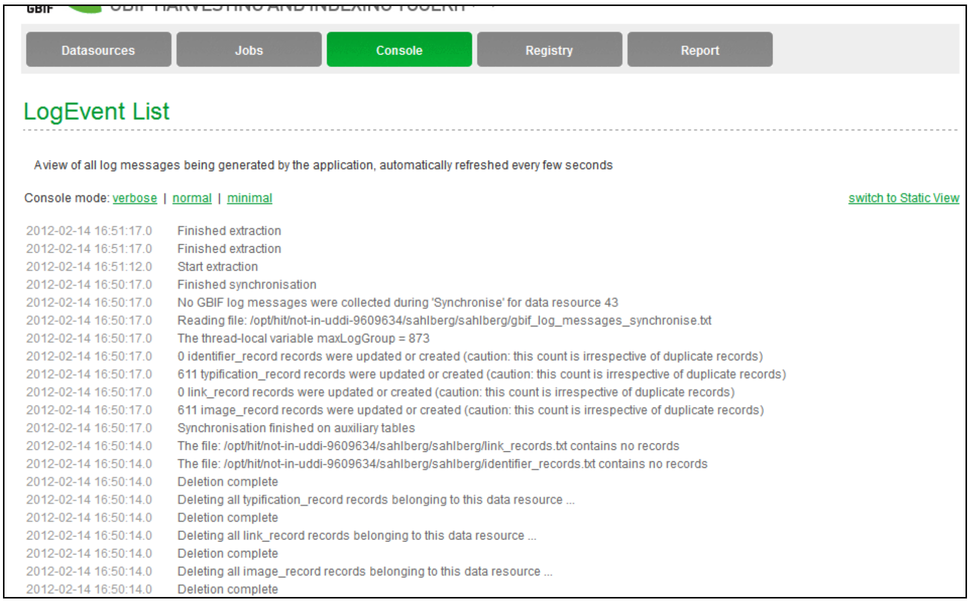
Jobs

Job ID	Operation	Description	Created	Started

No Jobs to display

Figure 12 The Jobs section

When a Job has been started its progress can be watched in the **Console** section. Every few seconds the log messages of the application are being refreshed with date and time (see Figure 13).



LogEvent List

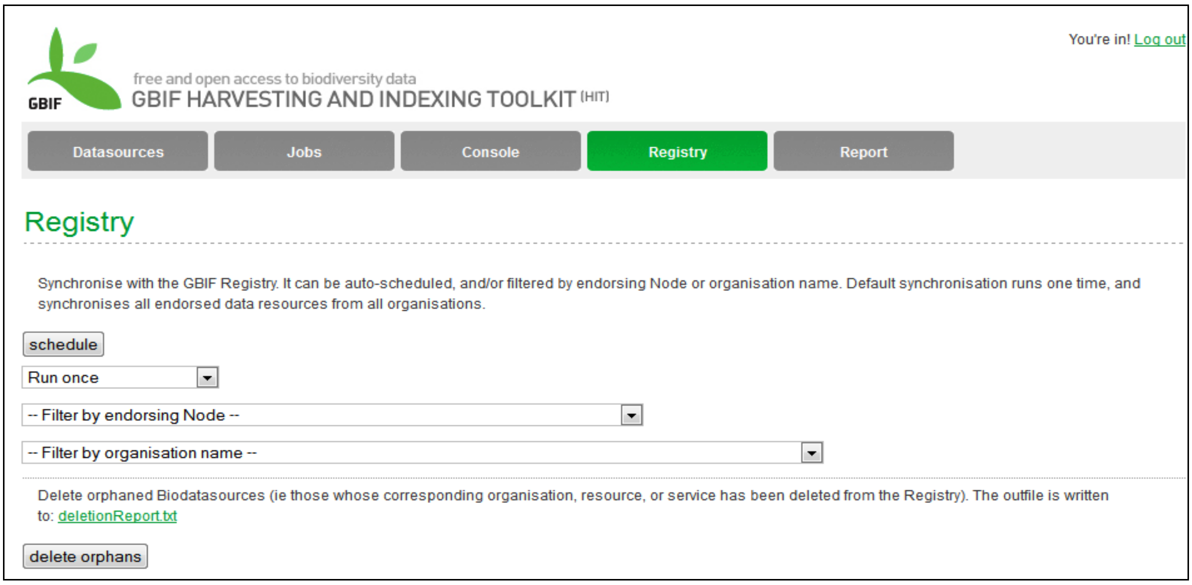
A view of all log messages being generated by the application, automatically refreshed every few seconds

Console mode: verbose | normal | minimal   switch to Static View

```
2012-02-14 16:51:17.0    Finished extraction
2012-02-14 16:51:17.0    Finished extraction
2012-02-14 16:51:12.0    Start extraction
2012-02-14 16:50:17.0    Finished synchronisation
2012-02-14 16:50:17.0    No GBIF log messages were collected during 'Synchronise' for data resource 43
2012-02-14 16:50:17.0    Reading file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/gbif_log_messages_synchronise.txt
2012-02-14 16:50:17.0    The thread-local variable maxLogGroup = 873
2012-02-14 16:50:17.0    0 identifier_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-14 16:50:17.0    611 typification_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-14 16:50:17.0    0 link_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-14 16:50:17.0    611 image_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-14 16:50:17.0    Synchronisation finished on auxiliary tables
2012-02-14 16:50:14.0    The file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/link_records.txt contains no records
2012-02-14 16:50:14.0    The file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/identifier_records.txt contains no records
2012-02-14 16:50:14.0    Deletion complete
2012-02-14 16:50:14.0    Deleting all typification_record records belonging to this data resource ...
2012-02-14 16:50:14.0    Deletion complete
2012-02-14 16:50:14.0    Deleting all link_record records belonging to this data resource ...
2012-02-14 16:50:14.0    Deletion complete
2012-02-14 16:50:14.0    Deleting all image_record records belonging to this data resource ...
2012-02-14 16:50:14.0    Deletion complete
```

Figure 13 The Console section with the Log Event List

The **Registry** tab is used to synchronise with the GBIF Registry. Before clicking on "schedule" you have the possibility to filter the datasources by endorsing Node or organisation name (see Figure 14).


You're in! Log out
GBIF
free and open access to biodiversity data
GBIF HARVESTING AND INDEXING TOOLKIT (HIT)
Datasources
Jobs
Console
Registry
Report

Registry

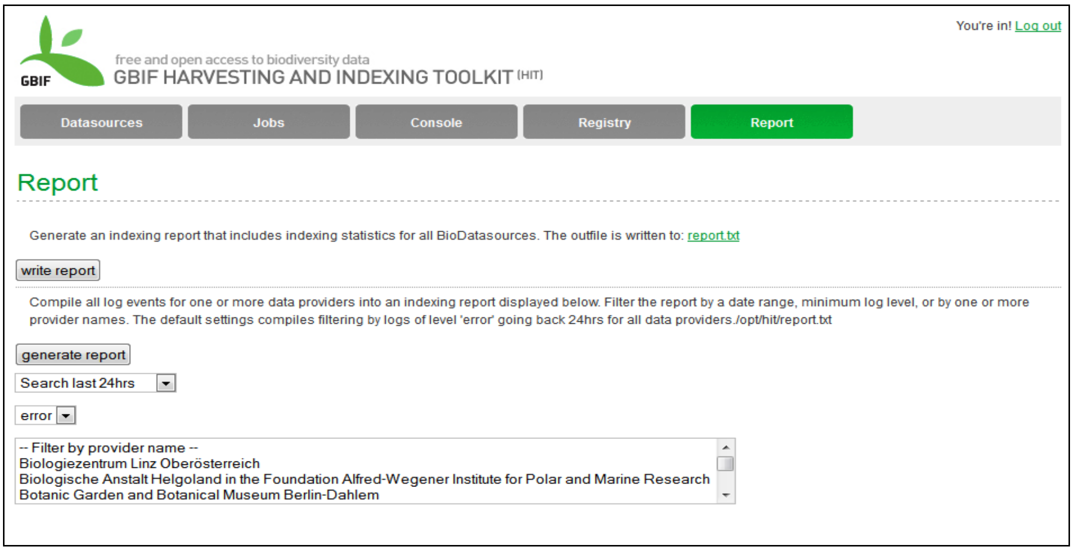
Synchronise with the GBIF Registry. It can be auto-scheduled, and/or filtered by endorsing Node or organisation name. Default synchronisation runs one time, and synchronises all endorsed data resources from all organisations.

schedule
Run once
-- Filter by endorsing Node --
-- Filter by organisation name --

Delete orphaned Biodatasources (ie those whose corresponding organisation, resource, or service has been deleted from the Registry). The outfile is written to: deletionReport.txt

Figure 14 The Registry tab

Finally you can write or generate a report in the **Report** section (see Figure 15). Again you have different options for filtering the result.



You're in! Log out

free and open access to biodiversity data
GBIF HARVESTING AND INDEXING TOOLKIT (HIT)

Report

Generate an indexing report that includes indexing statistics for all BioDatasources. The outfile is written to: report.txt

write report

Compile all log events for one or more data providers into an indexing report displayed below. Filter the report by a date range, minimum log level, or by one or more provider names. The default settings compiles filtering by logs of level 'error' going back 24hrs for all data providers./opt/hit/report.txt

generate report

Search last 24hrs

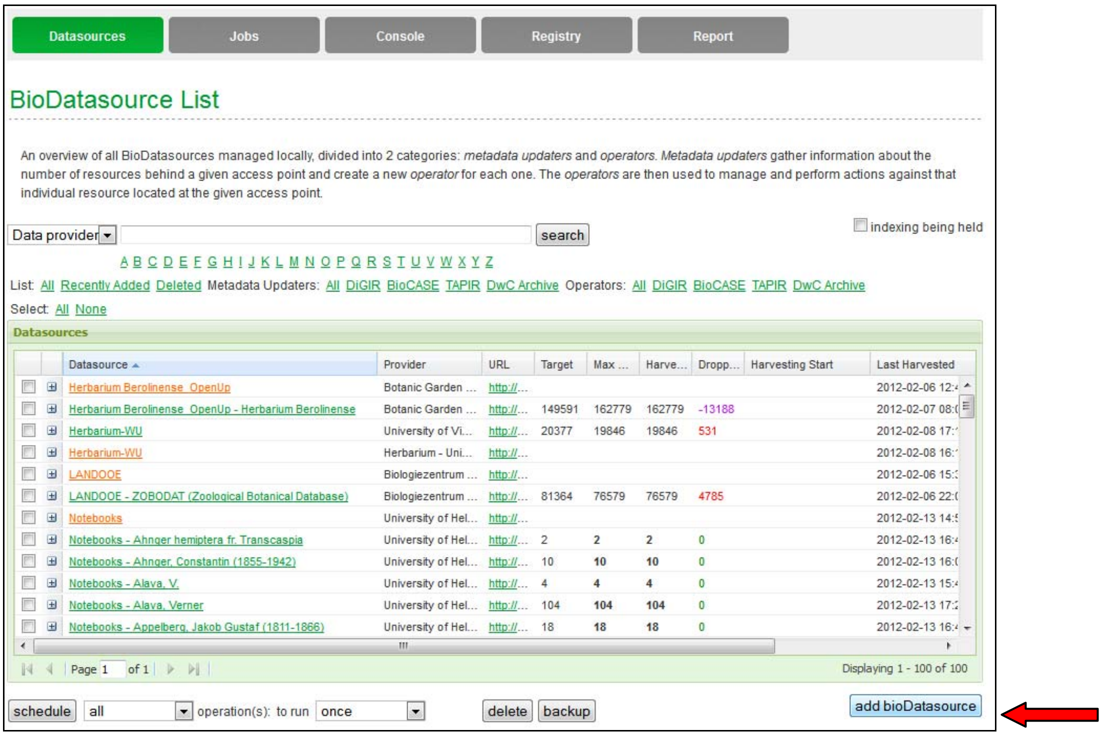
error

-- Filter by provider name --
Biologiezentrum Linz Oberösterreich
Biologische Anstalt Helgoland in the Foundation Alfred-Wegener Institute for Polar and Marine Research
Botanic Garden and Botanical Museum Berlin-Dahlem

Figure 15 Writing or generating a report

2.3 Adding a new bioDatasource and harvesting it

First of all click on “add bioDatasource” in the lower right corner (see Figure 16).



BioDatasource List

An overview of all BioDatasources managed locally, divided into 2 categories: *metadata updaters* and *operators*. *Metadata updaters* gather information about the number of resources behind a given access point and create a new *operator* for each one. The *operators* are then used to manage and perform actions against that individual resource located at the given access point.

Data provider | search | □ indexing being held

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

List: All Recently Added Deleted Metadata Updaters: All DiGIR BioCASE TAPIR DwC Archive Operators: All DiGIR BioCASE TAPIR DwC Archive

Select: All None

Datasources

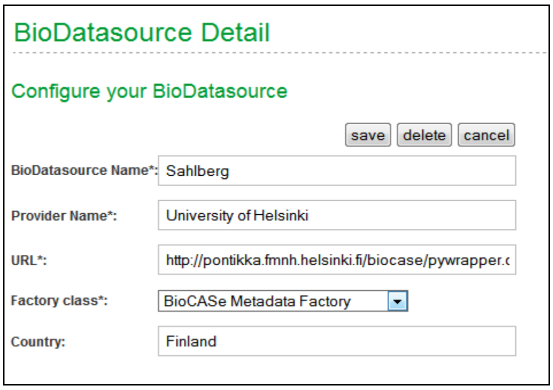
		Datasource ▲	Provider	URL	Target	Max ...	Harve...	Dropp...	Harvesting Start	Last Harvested
□	⊞	Herbarium Berolinense_OpenUp	Botanic Garden ...	http://...						2012-02-06 12:4
□	⊞	Herbarium Berolinense_OpenUp - Herbarium Berolinense	Botanic Garden ...	http://...	149591	162779	162779	-13188		2012-02-07 08:(
□	⊞	Herbarium-WU	University of Vi...	http://...	20377	19846	19846	531		2012-02-08 17:1
□	⊞	Herbarium-WU	Herbarium - Uni...	http://...						2012-02-08 16:1
□	⊞	LANDOOE	Biologiezentrum ...	http://...						2012-02-06 15:3
□	⊞	LANDOOE - ZOBODAT (Zoological Botanical Database)	Biologiezentrum ...	http://...	81364	76579	76579	4785		2012-02-06 22:(
□	⊞	Notebooks	University of Hel...	http://...						2012-02-13 14:5
□	⊞	Notebooks - Ahnger hemiptera fr. Transcaspia	University of Hel...	http://...	2	2	2	0		2012-02-13 16:4
□	⊞	Notebooks - Ahnger, Constantin (1855-1942)	University of Hel...	http://...	10	10	10	0		2012-02-13 16:(
□	⊞	Notebooks - Alava, V.	University of Hel...	http://...	4	4	4	0		2012-02-13 15:4
□	⊞	Notebooks - Alava, Verner	University of Hel...	http://...	104	104	104	0		2012-02-13 17:2
□	⊞	Notebooks - Appelberg, Jakob Gustaf (1811-1866)	University of Hel...	http://...	18	18	18	0		2012-02-13 16:4

Page 1 of 1 | Displaying 1 - 100 of 100

schedule | all | operation(s): to run | once | delete | backup | add bioDatasource

Figure 16 Clicking on “add bioDatasource”

After doing this you have to configure your datasource (see Figure 17). You have to fill in the name of the bioDatasource, the name of the provider, the URL and the factory class. It is very important to choose BioCASe in the drop-down-menu. If you want you can type in the name of the country.
When everything has been filled in correctly you can click on "save" and the datasource should now appear in orange in the datasource list (see Figure 18).

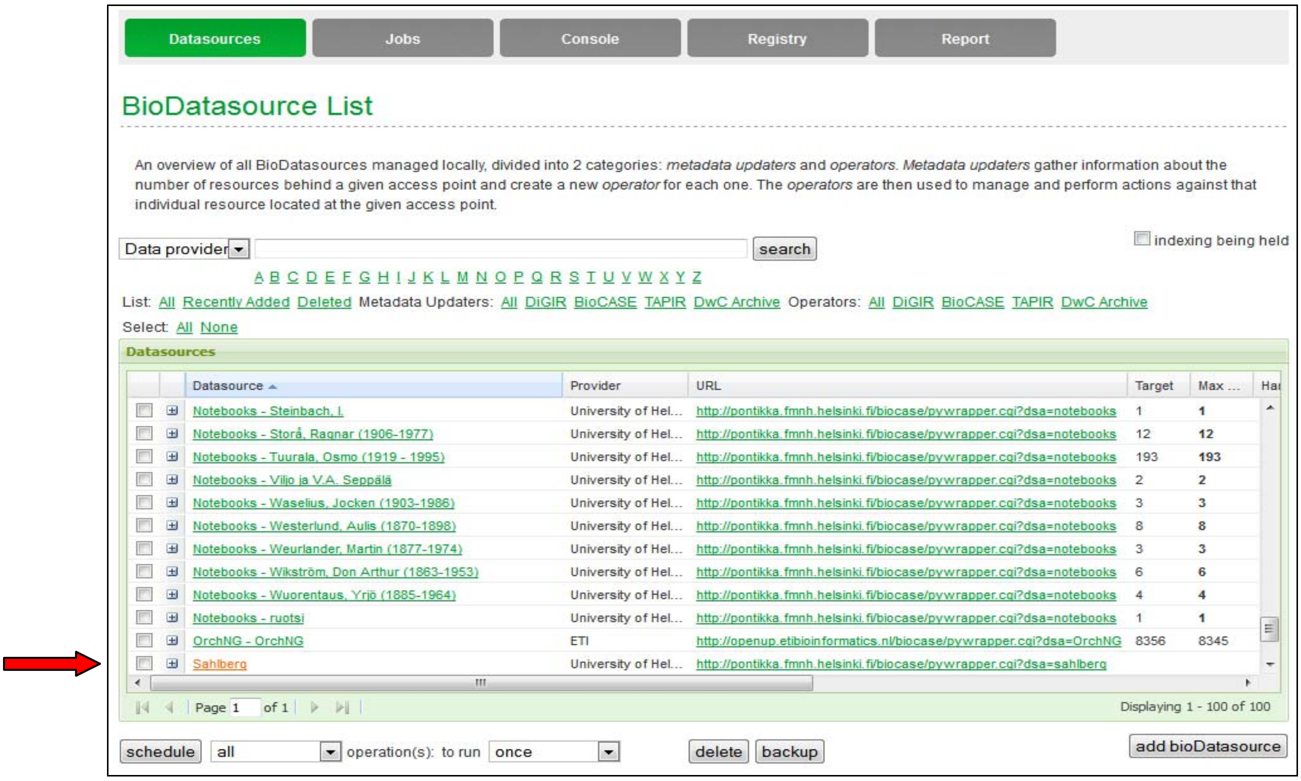


BioDatasource Detail

Configure your BioDatasource

BioDatasource Name*:	Sahlberg
Provider Name*:	University of Helsinki
URL*:	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.c
Factory class*:	BioCASe Metadata Factory
Country:	Finland

Figure 17 Adding a new datasource



BioDatasource List

An overview of all BioDatasources managed locally, divided into 2 categories: *metadata updaters* and *operators*. *Metadata updaters* gather information about the number of resources behind a given access point and create a new *operator* for each one. The *operators* are then used to manage and perform actions against that individual resource located at the given access point.

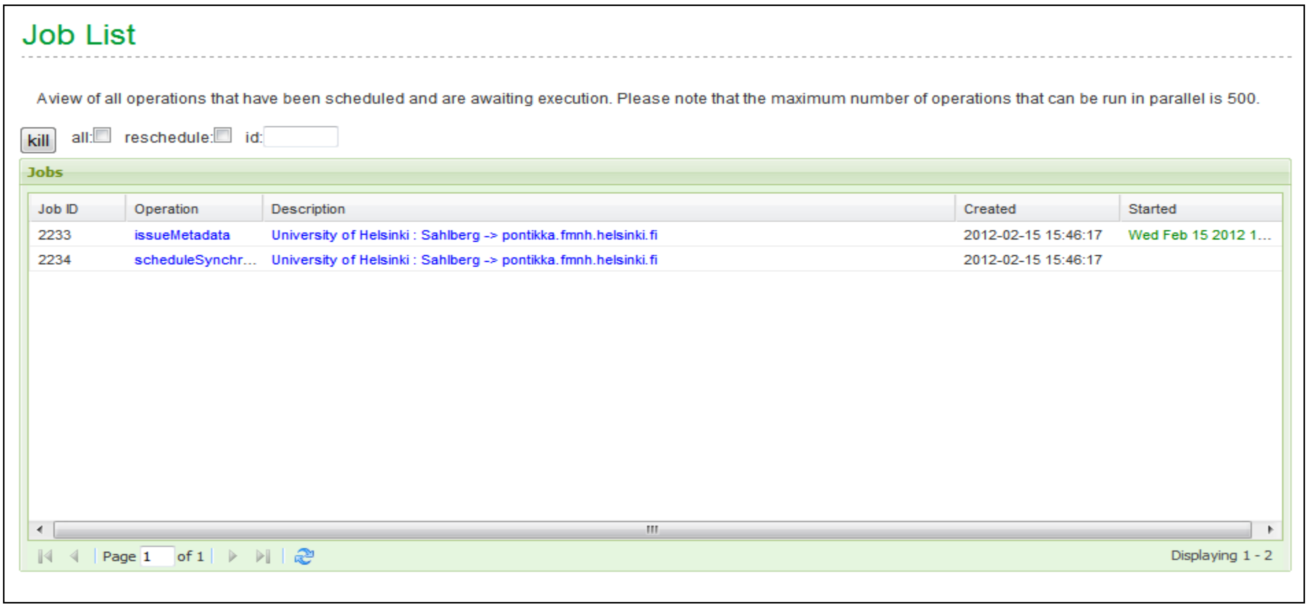
Data provider
search
□ indexing being held
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
List: All Recently Added Deleted Metadata Updaters: All DiGIR BioCASE TAPIR DwC Archive Operators: All DiGIR BioCASE TAPIR DwC Archive
Select: All None

Datasources

		Datasource	Provider	URL	Target	Max ...	Ha
□	⊞	Notebooks - Steinbach, I.	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	1	1	
□	⊞	Notebooks - Storå, Ragnar (1906-1977)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	12	12	
□	⊞	Notebooks - Tuurala, Osmo (1919 - 1995)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	193	193	
□	⊞	Notebooks - Viljo ja V.A. Seppälä	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	2	2	
□	⊞	Notebooks - Waselius, Jocken (1903-1986)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	3	3	
□	⊞	Notebooks - Westerlund, Aulis (1870-1898)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	8	8	
□	⊞	Notebooks - Weurlander, Martin (1877-1974)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	3	3	
□	⊞	Notebooks - Wikström, Don Arthur (1863-1953)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	6	6	
□	⊞	Notebooks - Wuorentaus, Yrjö (1885-1964)	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	4	4	
□	⊞	Notebooks - ruotsi	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=notebooks	1	1	
□	⊞	OrchNG - OrchNG	ETI	http://openup.etibioinformatics.nl/biocase/pywrapper.cgi?dsa=OrchNG	8356	8345	
□	⊞	Sahlberg	University of Hel...	http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=sahlberg			

Figure 18 The newly added datasource "Sahlberg"

Now tick the box in front of the datasource "Sahlberg" to select this metadata updater. Then click on "schedule". When you switch to the Jobs tab you can see the two Jobs are waiting to be executed: "issueMetadate" and "scheduleSynchronisation" (see Figure 19).



Job List

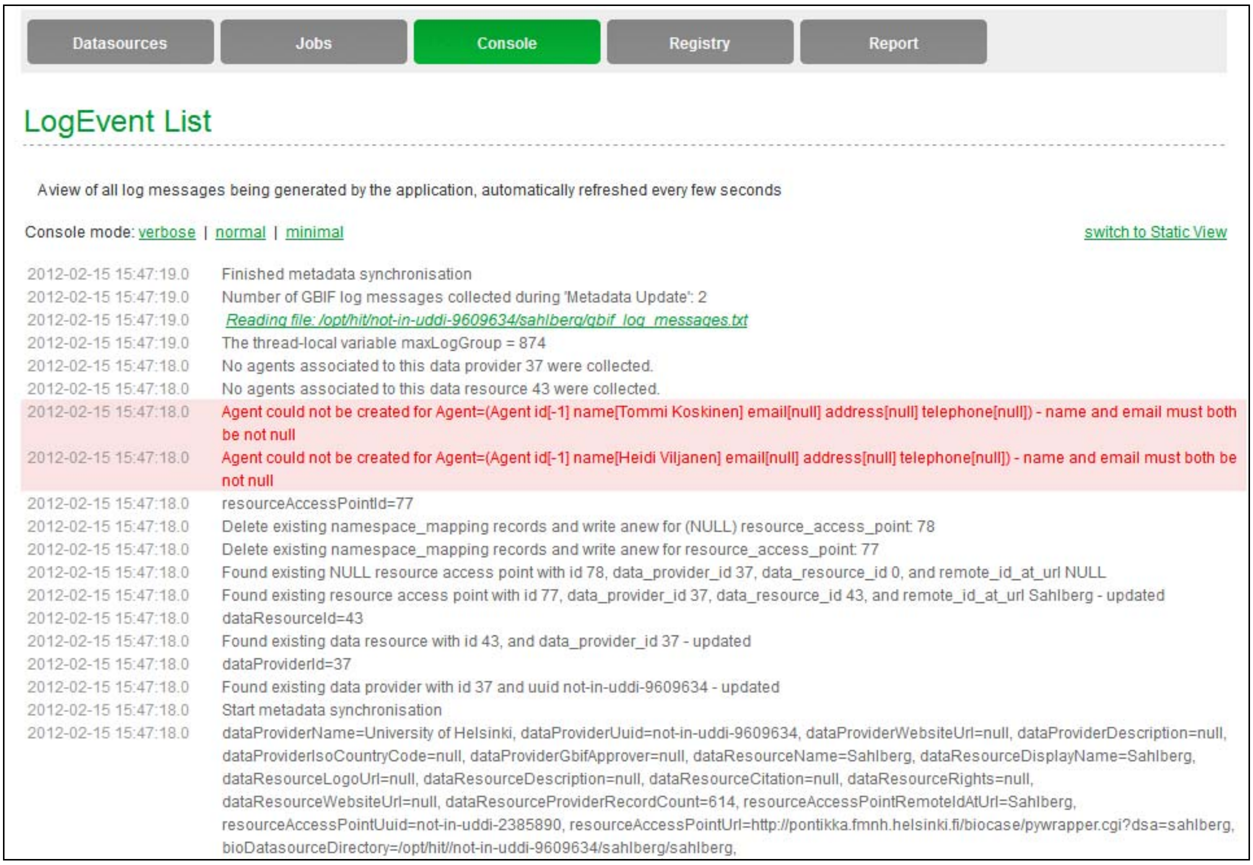
A view of all operations that have been scheduled and are awaiting execution. Please note that the maximum number of operations that can be run in parallel is 500.

Job ID	Operation	Description	Created	Started
2233	issueMetadata	University of Helsinki : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:46:17	Wed Feb 15 2012 1…
2234	scheduleSynchr…	University of Helsinki : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:46:17	

Figure 19 Job list after scheduling the metadata updater for "Sahlberg"

Now switch to the Console tab. As you can see in Figure 20 you can not only see the progress of the Jobs but also error messages if something is missing (marked red).

When the Jobs have been finished they must not appear anymore in the Job list. When going to Datasources again you can see that a "Sahlberg" operator has been created (see Figure 21).



LogEvent List

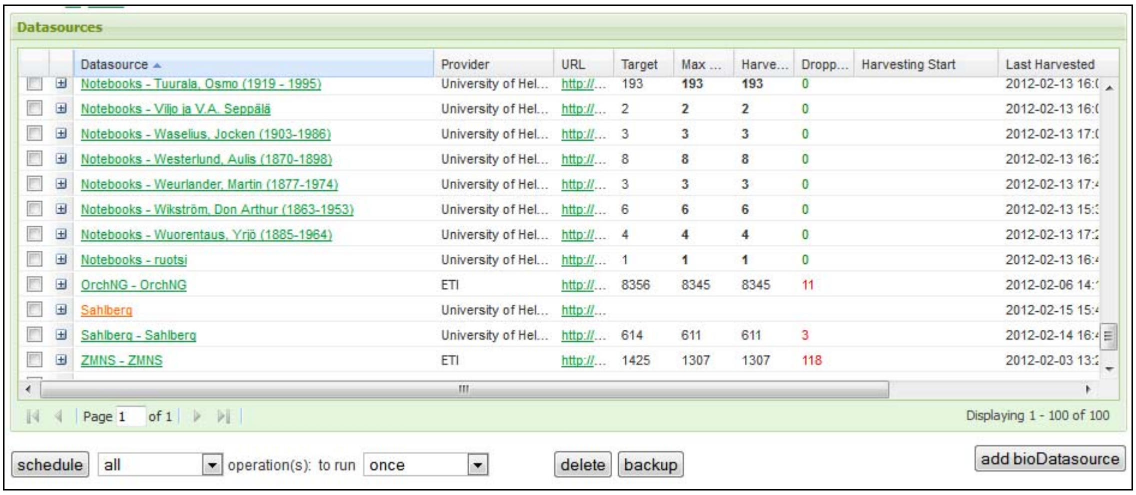
A view of all log messages being generated by the application, automatically refreshed every few seconds

Console mode: verbose | normal | minimal

switch to Static View

```
2012-02-15 15:47:19.0   Finished metadata synchronisation
2012-02-15 15:47:19.0   Number of GBIF log messages collected during 'Metadata Update': 2
2012-02-15 15:47:19.0   Reading file: /opt/hit/not-in-uddi-9609634/sahlberg/gbif_log_messages.txt
2012-02-15 15:47:19.0   The thread-local variable maxLogGroup = 874
2012-02-15 15:47:18.0   No agents associated to this data provider 37 were collected.
2012-02-15 15:47:18.0   No agents associated to this data resource 43 were collected.
2012-02-15 15:47:18.0   Agent could not be created for Agent=(Agent id[-1] name[Tommi Koskinen] email[null] address[null] telephone[null]) - name and email must both
                        be not null
2012-02-15 15:47:18.0   Agent could not be created for Agent=(Agent id[-1] name[Heidi Viljanen] email[null] address[null] telephone[null]) - name and email must both be
                        not null
2012-02-15 15:47:18.0   resourceAccessPointId=77
2012-02-15 15:47:18.0   Delete existing namespace_mapping records and write anew for (NULL) resource_access_point: 78
2012-02-15 15:47:18.0   Delete existing namespace_mapping records and write anew for resource_access_point: 77
2012-02-15 15:47:18.0   Found existing NULL resource access point with id 78, data_provider_id 37, data_resource_id 0, and remote_id_at_url NULL
2012-02-15 15:47:18.0   Found existing resource access point with id 77, data_provider_id 37, data_resource_id 43, and remote_id_at_url Sahlberg - updated
2012-02-15 15:47:18.0   dataResourceId=43
2012-02-15 15:47:18.0   Found existing data resource with id 43, and data_provider_id 37 - updated
2012-02-15 15:47:18.0   dataProviderId=37
2012-02-15 15:47:18.0   Found existing data provider with id 37 and uuid not-in-uddi-9609634 - updated
2012-02-15 15:47:18.0   Start metadata synchronisation
2012-02-15 15:47:18.0   dataProviderName=University of Helsinki, dataProviderUuid=not-in-uddi-9609634, dataProviderWebsiteUrl=null, dataProviderDescription=null,
                        dataProviderIsoCountryCode=null, dataProviderGbifApprover=null, dataResourceName=Sahlberg, dataResourceDisplayName=Sahlberg,
                        dataResourceLogoUrl=null, dataResourceDescription=null, dataResourceCitation=null, dataResourceRights=null,
                        dataResourceWebsiteUrl=null, dataResourceProviderRecordCount=614, resourceAccessPointRemoteIdAtUrl=Sahlberg,
                        resourceAccessPointUuid=not-in-uddi-2385890, resourceAccessPointUrl=http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=sahlberg,
                        bioDatasourceDirectory=/opt/hit//not-in-uddi-9609634/sahlberg/sahlberg,
```

Figure 20 The Log Event List after scheduling the metadata updater "Sahlberg"


Datasources

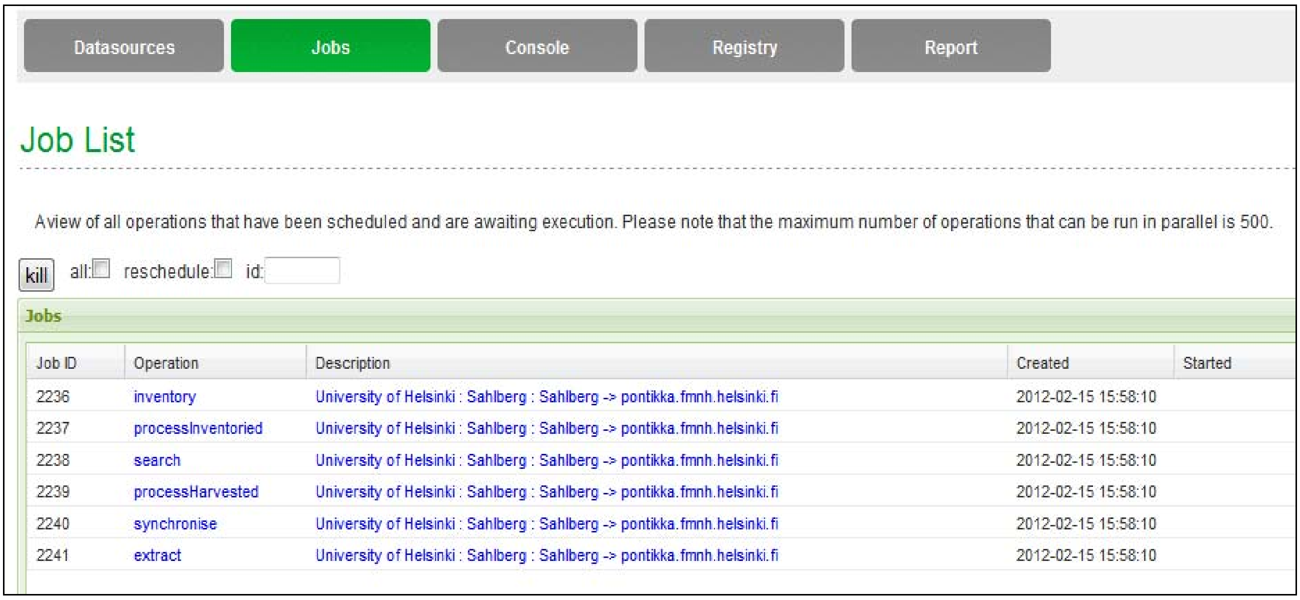
		Datasource ▲	Provider	URL	Target	Max ...	Harve...	Dropp...	Harvesting Start	Last Harvested
□	⊞	Notebooks - Tuurala, Osmo (1919 - 1995)	University of Hel...	http://...	193	193	193	0		2012-02-13 16:(
□	⊞	Notebooks - Viljo ja V.A. Seppälä	University of Hel...	http://...	2	2	2	0		2012-02-13 16:(
□	⊞	Notebooks - Waselius, Jocken (1903-1986)	University of Hel...	http://...	3	3	3	0		2012-02-13 17:(
□	⊞	Notebooks - Westerlund, Aulis (1870-1898)	University of Hel...	http://...	8	8	8	0		2012-02-13 16:2
□	⊞	Notebooks - Weurlander, Martin (1877-1974)	University of Hel...	http://...	3	3	3	0		2012-02-13 17:4
□	⊞	Notebooks - Wikström, Don Arthur (1863-1953)	University of Hel...	http://...	6	6	6	0		2012-02-13 15:3
□	⊞	Notebooks - Wuorentaus, Yrjö (1885-1964)	University of Hel...	http://...	4	4	4	0		2012-02-13 17:2
□	⊞	Notebooks - ruotsi	University of Hel...	http://...	1	1	1	0		2012-02-13 16:4
□	⊞	OrchNG - OrchNG	ETI	http://...	8356	8345	8345	11		2012-02-06 14:1
□	⊞	Sahlberg	University of Hel...	http://...						2012-02-15 15:4
□	⊞	Sahlberg - Sahlberg	University of Hel...	http://...	614	611	611	3		2012-02-14 16:4
□	⊞	ZMNS - ZMNS	ETI	http://...	1425	1307	1307	118		2012-02-03 13:2

Page 1 of 1

Displaying 1 - 100 of 100

Figure 21 The newly created operator "Sahlberg – Sahlberg"

Now it is time to gather records from the data provider. To achieve this you have to select the (green) operator "Sahlberg – Sahlberg" (tick the box) and click on "schedule". Right after this there should be six Jobs in the list: Inventory, processInventoried, search, processHarvested, synchronise and extract (see Figure 22). The order of these operations is essential for a correct harvesting process.



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.

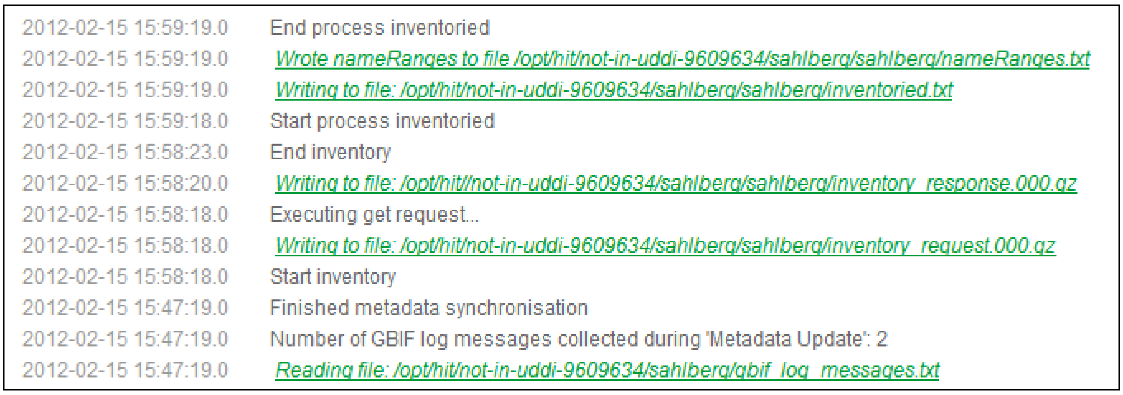
kill all:□ reschedule:□ id:

Jobs

Job ID	Operation	Description	Created	Started
2236	inventory	University of Helsinki : Sahlberg : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:58:10	
2237	processInventoried	University of Helsinki : Sahlberg : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:58:10	
2238	search	University of Helsinki : Sahlberg : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:58:10	
2239	processHarvested	University of Helsinki : Sahlberg : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:58:10	
2240	synchronise	University of Helsinki : Sahlberg : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:58:10	
2241	extract	University of Helsinki : Sahlberg : Sahlberg -> pontikka.fmnh.helsinki.fi	2012-02-15 15:58:10	

Figure 22 The Job list after scheduling the operator "Sahlberg – Sahlberg"

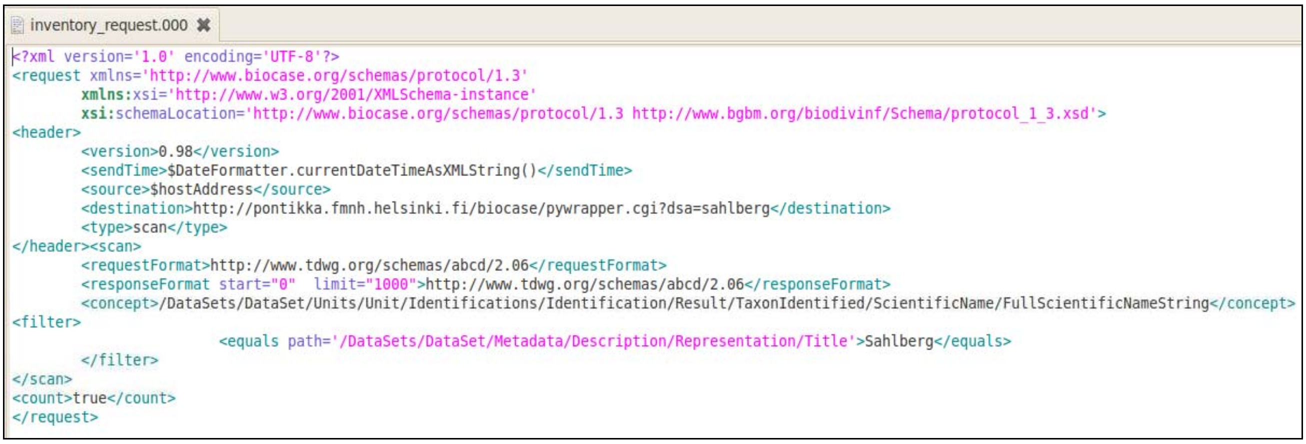
During the **Inventory** operation a list of all scientific names occurring in the datasource is generated. You can follow this process in the Console section (see Figure 23). In contrary to the other protocols for BioCASe no count information is ever collected.



```
2012-02-15 15:59:19.0   End process inventoried
2012-02-15 15:59:19.0   Wrote nameRanges to file /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/nameRanges.txt
2012-02-15 15:59:19.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/inventoried.txt
2012-02-15 15:59:18.0   Start process inventoried
2012-02-15 15:58:23.0   End inventory
2012-02-15 15:58:20.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/inventory_response.000.gz
2012-02-15 15:58:18.0   Executing get request...
2012-02-15 15:58:18.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/inventory_request.000.gz
2012-02-15 15:58:18.0   Start inventory
2012-02-15 15:47:19.0   Finished metadata synchronisation
2012-02-15 15:47:19.0   Number of GBIF log messages collected during 'Metadata Update': 2
2012-02-15 15:47:19.0   Reading file: /opt/hit/not-in-uddi-9609634/sahlberg/gbif_log_messages.txt
```

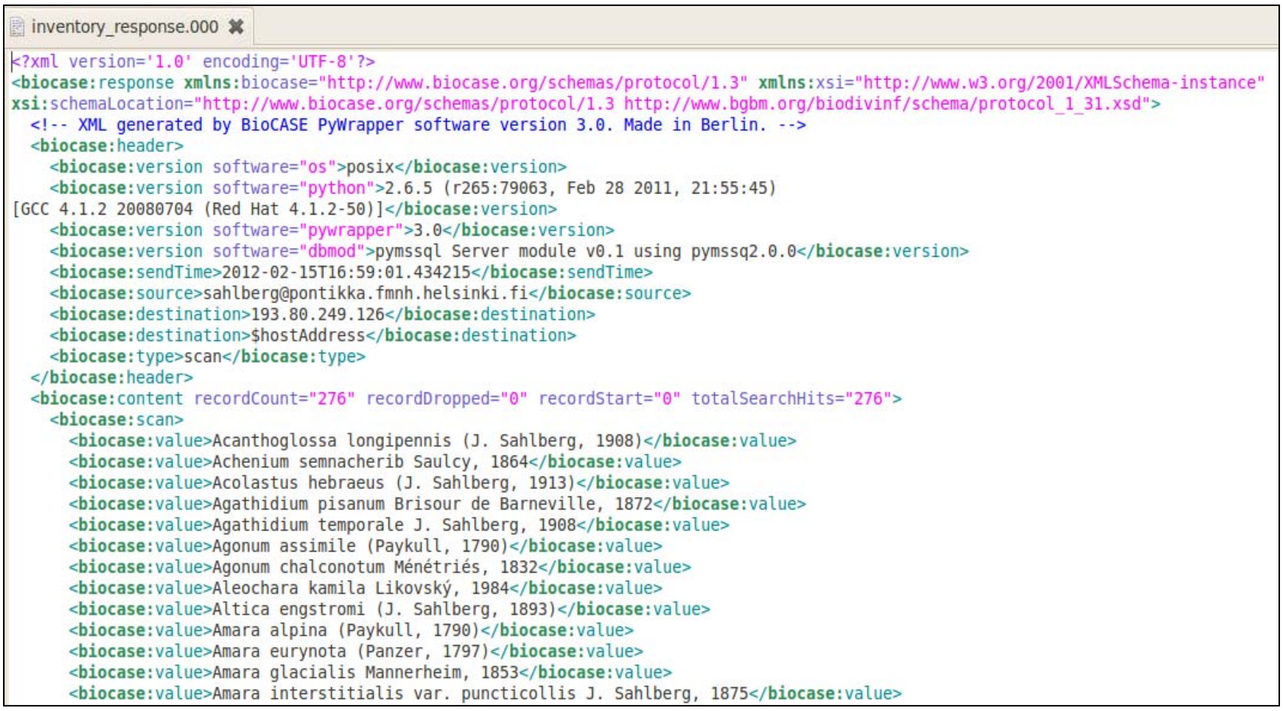
Figure 23 Console section during the Inventory operation

As you can see there is an inventory_request (see Figure 24) and an inventory_response (see Figure 25). Both are saved in /opt/hit (…) – the harvest directory we have determined during the HIT installation.



```
<?xml version='1.0' encoding='UTF-8'?>
<request xmlns='http://www.biocase.org/schemas/protocol/1.3'
        xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
        xsi:schemaLocation='http://www.biocase.org/schemas/protocol/1.3 http://www.bgbm.org/biodivinf/Schema/protocol_1_3.xsd'>
<header>
        <version>0.98</version>
        <sendTime>$DateFormatter.currentDateTimeAsXMLString()</sendTime>
        <source>$hostAddress</source>
        <destination>http://pontikka.fmnh.helsinki.fi/biocase/pywrapper.cgi?dsa=sahlberg</destination>
        <type>scan</type>
</header><scan>
        <requestFormat>http://www.tdwg.org/schemas/abcd/2.06</requestFormat>
        <responseFormat start="0"  limit="1000">http://www.tdwg.org/schemas/abcd/2.06</responseFormat>
        <concept>/DataSets/DataSet/Units/Unit/Identifications/Identification/Result/TaxonIdentified/ScientificName/FullScientificNameString</concept>
<filter>
                        <equals path='/DataSets/DataSet/Metadata/Description/Representation/Title'>Sahlberg</equals>
        </filter>
</scan>
<count>true</count>
</request>
```

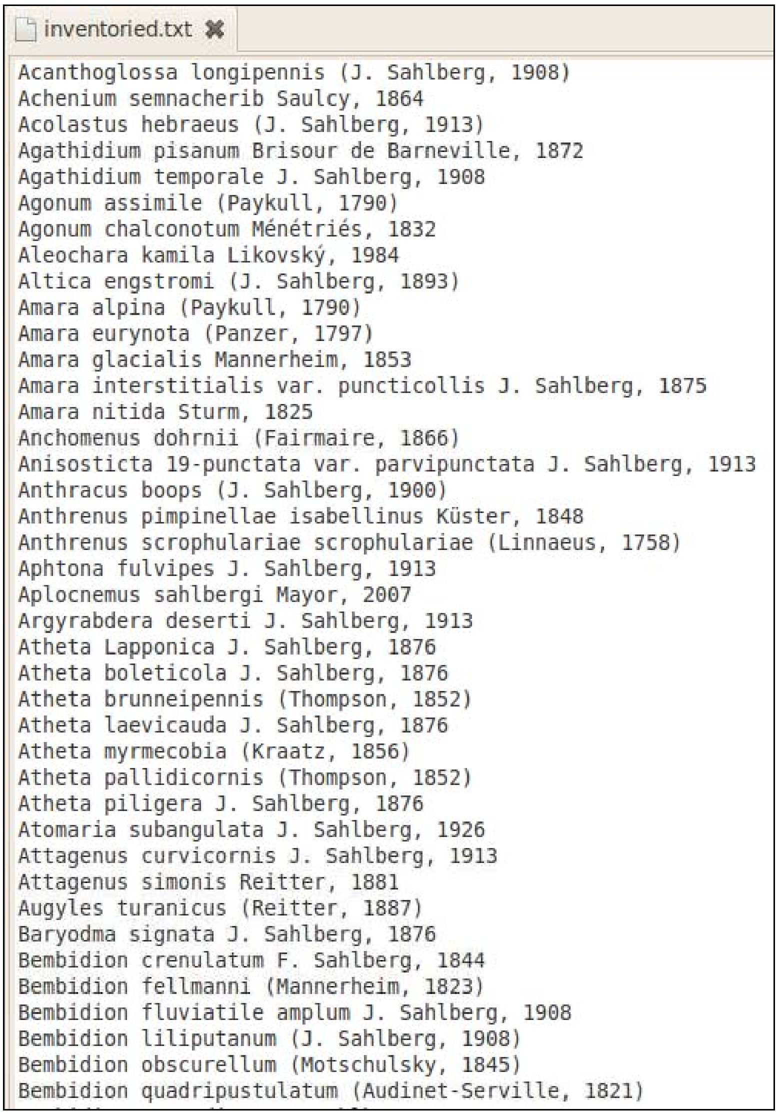
Figure 24 The inventory_request of the Inventory operator



```
<?xml version='1.0' encoding='UTF-8'?>
<biocase:response xmlns:biocase="http://www.biocase.org/schemas/protocol/1.3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.biocase.org/schemas/protocol/1.3 http://www.bgbm.org/biodivinf/schema/protocol_1_31.xsd">
  <!-- XML generated by BioCASE PyWrapper software version 3.0. Made in Berlin. -->
  <biocase:header>
    <biocase:version software="os">posix</biocase:version>
    <biocase:version software="python">2.6.5 (r265:79063, Feb 28 2011, 21:55:45)
[GCC 4.1.2 20080704 (Red Hat 4.1.2-50)]</biocase:version>
    <biocase:version software="pywrapper">3.0</biocase:version>
    <biocase:version software="dbmod">pymssql Server module v0.1 using pymssq2.0.0</biocase:version>
    <biocase:sendTime>2012-02-15T16:59:01.434215</biocase:sendTime>
    <biocase:source>sahlberg@pontikka.fmnh.helsinki.fi</biocase:source>
    <biocase:destination>193.80.249.126</biocase:destination>
    <biocase:destination>$hostAddress</biocase:destination>
    <biocase:type>scan</biocase:type>
  </biocase:header>
  <biocase:content recordCount="276" recordDropped="0" recordStart="0" totalSearchHits="276">
    <biocase:scan>
      <biocase:value>Acanthoglossa longipennis (J. Sahlberg, 1908)</biocase:value>
      <biocase:value>Achenium semnacherib Saulcy, 1864</biocase:value>
      <biocase:value>Acolastus hebraeus (J. Sahlberg, 1913)</biocase:value>
      <biocase:value>Agathidium pisanum Brisour de Barneville, 1872</biocase:value>
      <biocase:value>Agathidium temporale J. Sahlberg, 1908</biocase:value>
      <biocase:value>Agonum assimile (Paykull, 1790)</biocase:value>
      <biocase:value>Agonum chalconotum Ménétriés, 1832</biocase:value>
      <biocase:value>Aleochara kamila Likovský, 1984</biocase:value>
      <biocase:value>Altica engstromi (J. Sahlberg, 1893)</biocase:value>
      <biocase:value>Amara alpina (Paykull, 1790)</biocase:value>
      <biocase:value>Amara eurynota (Panzer, 1797)</biocase:value>
      <biocase:value>Amara glacialis Mannerheim, 1853</biocase:value>
      <biocase:value>Amara interstitialis var. puncticollis J. Sahlberg, 1875</biocase:value>
```

Figure 25 The inventory_response of the Inventory operator

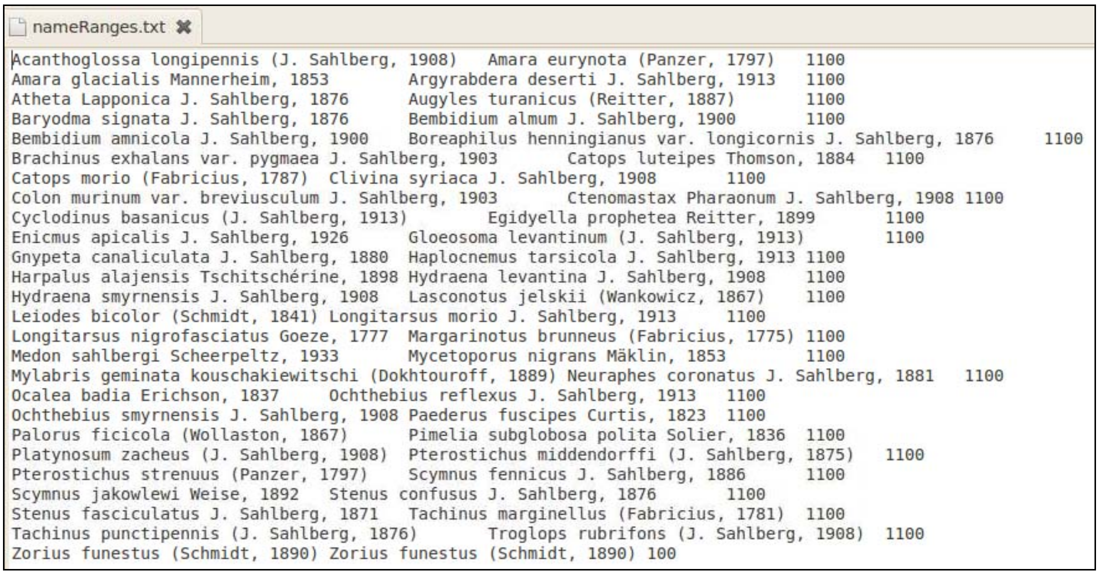
Figure 26 shows the result of the **processInventoried** operation: the text document *inventoried.txt* with an alphabetical list of all scientific names.



```
Acanthoglossa longipennis (J. Sahlberg, 1908)
Achenium semnacherib Saulcy, 1864
Acolastus hebraeus (J. Sahlberg, 1913)
Agathidium pisanum Brisour de Barneville, 1872
Agathidium temporale J. Sahlberg, 1908
Agonum assimile (Paykull, 1790)
Agonum chalconotum Ménétriés, 1832
Aleochara kamila Likovský, 1984
Altica engstromi (J. Sahlberg, 1893)
Amara alpina (Paykull, 1790)
Amara eurynota (Panzer, 1797)
Amara glacialis Mannerheim, 1853
Amara interstitialis var. puncticollis J. Sahlberg, 1875
Amara nitida Sturm, 1825
Anchomenus dohrnii (Fairmaire, 1866)
Anisosticta 19-punctata var. parvipunctata J. Sahlberg, 1913
Anthracus boops (J. Sahlberg, 1900)
Anthrenus pimpinellae isabellinus Küster, 1848
Anthrenus scrophulariae scrophulariae (Linnaeus, 1758)
Aphtona fulvipes J. Sahlberg, 1913
Aplocnemus sahlbergi Mayor, 2007
Argyrabdera deserti J. Sahlberg, 1913
Atheta Lapponica J. Sahlberg, 1876
Atheta boleticola J. Sahlberg, 1876
Atheta brunneipennis (Thompson, 1852)
Atheta laevicauda J. Sahlberg, 1876
Atheta myrmecobia (Kraatz, 1856)
Atheta pallidicornis (Thompson, 1852)
Atheta piligera J. Sahlberg, 1876
Atomaria subangulata J. Sahlberg, 1926
Attagenus curvicornis J. Sahlberg, 1913
Attagenus simonis Reitter, 1881
Augyles turanicus (Reitter, 1887)
Baryodma signata J. Sahlberg, 1876
Bembidion crenulatum F. Sahlberg, 1844
Bembidion fellmanni (Mannerheim, 1823)
Bembidion fluviatile amplum J. Sahlberg, 1908
Bembidion liliputanum (J. Sahlberg, 1908)
Bembidion obscurellum (Motschulsky, 1845)
Bembidion quadripustulatum (Audinet-Serville, 1821)
```

Figure 26 Alphabetical list of all scientific names

Another document containing all the name ranges that were constructed is created too: *nameRanges.txt* (see Figure 27).

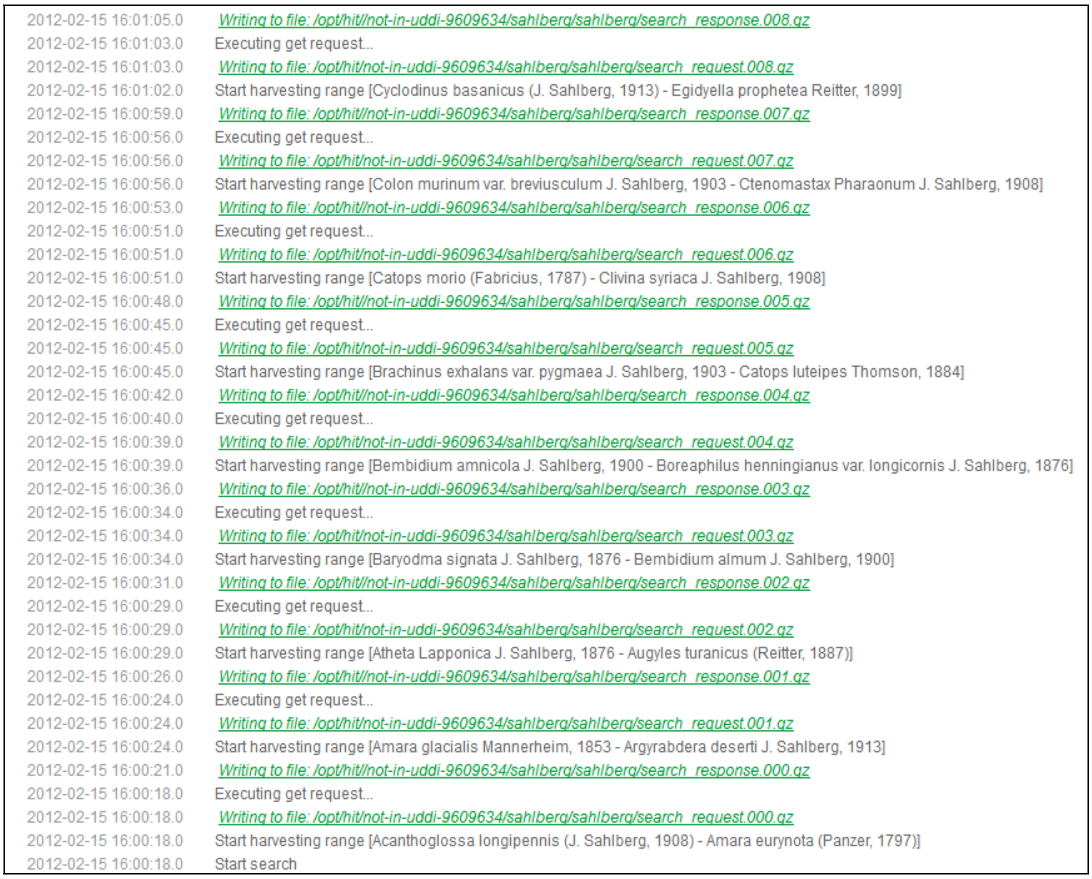


nameRanges.txt

```
Acanthoglossa longipennis (J. Sahlberg, 1908)   Amara eurynota (Panzer, 1797)   1100
Amara glacialis Mannerheim, 1853        Argyrabdera deserti J. Sahlberg, 1913   1100
Atheta Lapponica J. Sahlberg, 1876      Augyles turanicus (Reitter, 1887)       1100
Baryodma signata J. Sahlberg, 1876      Bembidium almum J. Sahlberg, 1900       1100
Bembidium amnicola J. Sahlberg, 1900    Boreaphilus henningianus var. longicornis J. Sahlberg, 1876     1100
Brachinus exhalans var. pygmaea J. Sahlberg, 1903       Catops luteipes Thomson, 1884   1100
Catops morio (Fabricius, 1787)  Clivina syriaca J. Sahlberg, 1908       1100
Colon murinum var. breviusculum J. Sahlberg, 1903       Ctenomastax Pharaonum J. Sahlberg, 1908 1100
Cyclodinus basanicus (J. Sahlberg, 1913)        Egidyella prophetea Reitter, 1899       1100
Enicmus apicalis J. Sahlberg, 1926      Gloeosoma levantinum (J. Sahlberg, 1913)        1100
Gnypeta canaliculata J. Sahlberg, 1880  Haplocnemus tarsicola J. Sahlberg, 1913 1100
Harpalus alajensis Tschitschérine, 1898 Hydraena levantina J. Sahlberg, 1908    1100
Hydraena smyrnensis J. Sahlberg, 1908   Lasconotus jelskii (Wankowicz, 1867)    1100
Leiodes bicolor (Schmidt, 1841) Longitarsus morio J. Sahlberg, 1913     1100
Longitarsus nigrofasciatus Goeze, 1777  Margarinotus brunneus (Fabricius, 1775) 1100
Medon sahlbergi Scheerpeltz, 1933       Mycetoporus nigrans Mäklin, 1853        1100
Mylabris geminata kouschakiewitschi (Dokhtouroff, 1889) Neuraphes coronatus J. Sahlberg, 1881   1100
Ocalea badia Erichson, 1837     Ochthebius reflexus J. Sahlberg, 1913   1100
Ochthebius smyrnensis J. Sahlberg, 1908 Paederus fuscipes Curtis, 1823  1100
Palorus ficicola (Wollaston, 1867)      Pimelia subglobosa polita Solier, 1836  1100
Platynosum zacheus (J. Sahlberg, 1908)  Pterostichus middendorffi (J. Sahlberg, 1875)   1100
Pterostichus strenuus (Panzer, 1797)    Scymnus fennicus J. Sahlberg, 1886      1100
Scymnus jakowlewi Weise, 1892   Stenus confusus J. Sahlberg, 1876       1100
Stenus fasciculatus J. Sahlberg, 1871   Tachinus marginellus (Fabricius, 1781)  1100
Tachinus punctipennis (J. Sahlberg, 1876)       Troglops rubrifons (J. Sahlberg, 1908)  1100
Zorius funestus (Schmidt, 1890) Zorius funestus (Schmidt, 1890) 100
```

Figure 27 The nameRanges.txt document

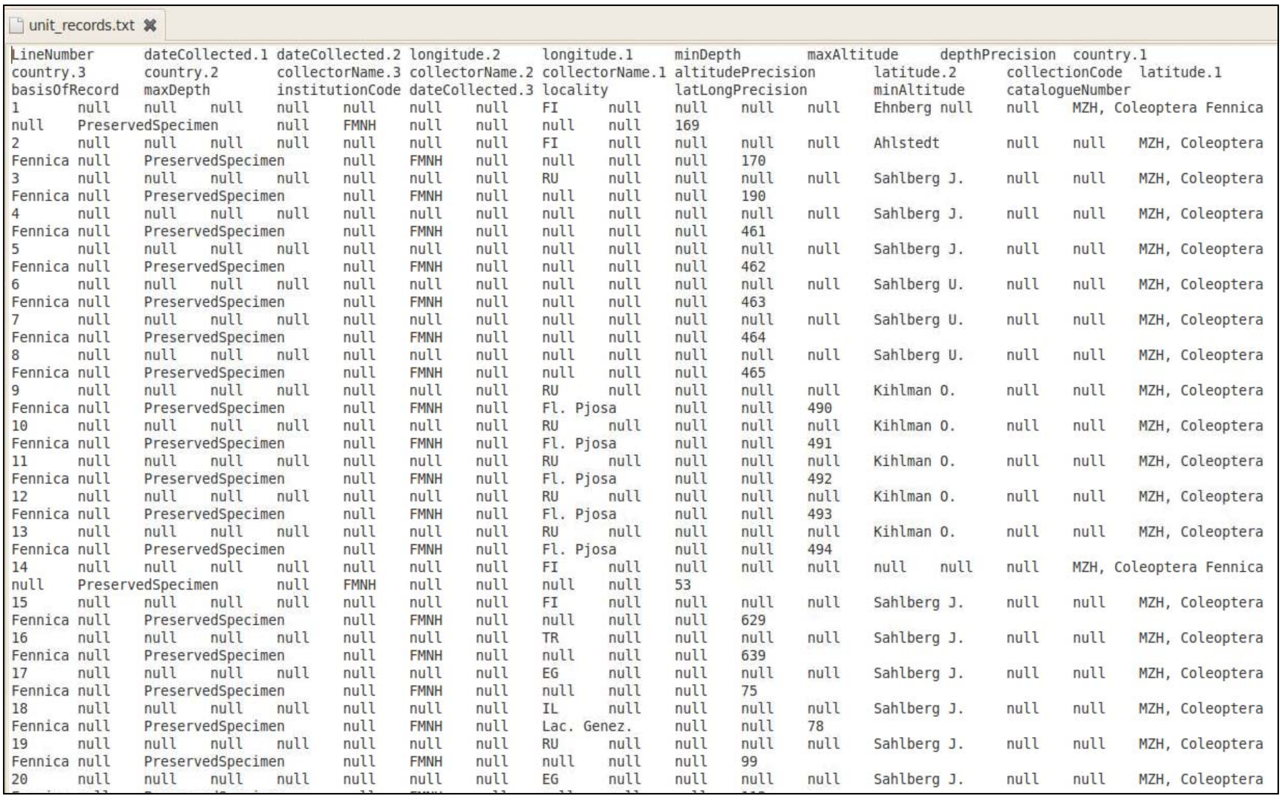
After this it is time for the **search** operation. In this phase the later with Pentaho Kettle transformed abcd records are created. There is always a search_request and a search_response (see Figure 28).



```
2012-02-15 16:01:05.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.008.gz
2012-02-15 16:01:03.0   Executing get request...
2012-02-15 16:01:03.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.008.gz
2012-02-15 16:01:02.0   Start harvesting range [Cyclodinus basanicus (J. Sahlberg, 1913) - Egidyella prophetea Reitter, 1899]
2012-02-15 16:00:59.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.007.gz
2012-02-15 16:00:56.0   Executing get request...
2012-02-15 16:00:56.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.007.gz
2012-02-15 16:00:56.0   Start harvesting range [Colon murinum var. breviusculum J. Sahlberg, 1903 - Ctenomastax Pharaonum J. Sahlberg, 1908]
2012-02-15 16:00:53.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.006.gz
2012-02-15 16:00:51.0   Executing get request...
2012-02-15 16:00:51.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.006.gz
2012-02-15 16:00:51.0   Start harvesting range [Catops morio (Fabricius, 1787) - Clivina syriaca J. Sahlberg, 1908]
2012-02-15 16:00:48.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.005.gz
2012-02-15 16:00:45.0   Executing get request...
2012-02-15 16:00:45.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.005.gz
2012-02-15 16:00:45.0   Start harvesting range [Brachinus exhalans var. pygmaea J. Sahlberg, 1903 - Catops luteipes Thomson, 1884]
2012-02-15 16:00:42.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.004.gz
2012-02-15 16:00:40.0   Executing get request...
2012-02-15 16:00:39.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.004.gz
2012-02-15 16:00:39.0   Start harvesting range [Bembidium amnicola J. Sahlberg, 1900 - Boreaphilus henningianus var. longicornis J. Sahlberg, 1876]
2012-02-15 16:00:36.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.003.gz
2012-02-15 16:00:34.0   Executing get request...
2012-02-15 16:00:34.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.003.gz
2012-02-15 16:00:34.0   Start harvesting range [Baryodma signata J. Sahlberg, 1876 - Bembidium almum J. Sahlberg, 1900]
2012-02-15 16:00:31.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.002.gz
2012-02-15 16:00:29.0   Executing get request...
2012-02-15 16:00:29.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.002.gz
2012-02-15 16:00:29.0   Start harvesting range [Atheta Lapponica J. Sahlberg, 1876 - Augyles turanicus (Reitter, 1887)]
2012-02-15 16:00:26.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.001.gz
2012-02-15 16:00:24.0   Executing get request...
2012-02-15 16:00:24.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.001.gz
2012-02-15 16:00:24.0   Start harvesting range [Amara glacialis Mannerheim, 1853 - Argyrabdera deserti J. Sahlberg, 1913]
2012-02-15 16:00:21.0   Writing to file: /opt/hit//not-in-uddi-9609634/sahlberg/sahlberg/search_response.000.gz
2012-02-15 16:00:18.0   Executing get request...
2012-02-15 16:00:18.0   Writing to file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/search_request.000.gz
2012-02-15 16:00:18.0   Start harvesting range [Acanthoglossa longipennis (J. Sahlberg, 1908) - Amara eurynota (Panzer, 1797)]
2012-02-15 16:00:18.0   Start search
```

Figure 28 The search operation creates search_requests and search_responses

If the response was encoded using ABCD, there is one core file after the **processHarvested** operation: *unit_records.txt* (see Figure 29). It contains a header line with column names, with each line representing a single Unit (record) element.



unit_records.txt

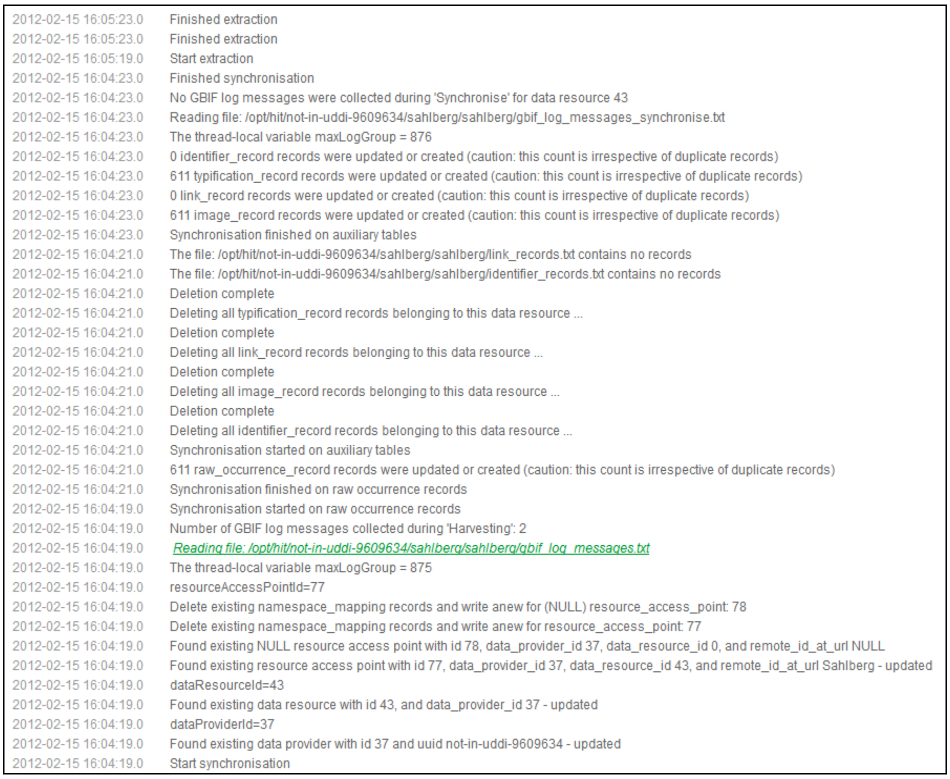
```
LineNumber      dateCollected.1 dateCollected.2 longitude.2     longitude.1     minDepth        maxAltitude     depthPrecision  country.1
country.3       country.2       collectorName.3 collectorName.2 collectorName.1 altitudePrecision       latitude.2      collectionCode  latitude.1
basisOfRecord   maxDepth        institutionCode dateCollected.3 locality        latLongPrecision        minAltitude     catalogueNumber
1       null    null    null    null    null    null    null    FI      null    null    null    null    Ehnberg null    null    MZH, Coleoptera Fennica
null    PreservedSpecimen       null    FMNH    null    null    null    null    169
2       null    null    null    null    null    null    null    FI      null    null    null    null    Ahlstedt        null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    170
3       null    null    null    null    null    null    null    RU      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    190
4       null    null    null    null    null    null    null    null    null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    461
5       null    null    null    null    null    null    null    null    null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    462
6       null    null    null    null    null    null    null    null    null    null    null    null    Sahlberg U.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    463
7       null    null    null    null    null    null    null    null    null    null    null    null    Sahlberg U.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    464
8       null    null    null    null    null    null    null    null    null    null    null    null    Sahlberg U.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    465
9       null    null    null    null    null    null    null    RU      null    null    null    null    Kihlman O.      null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    Fl. Pjosa       null    null    490
10      null    null    null    null    null    null    null    RU      null    null    null    null    Kihlman O.      null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    Fl. Pjosa       null    null    491
11      null    null    null    null    null    null    null    RU      null    null    null    null    Kihlman O.      null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    Fl. Pjosa       null    null    492
12      null    null    null    null    null    null    null    RU      null    null    null    null    Kihlman O.      null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    Fl. Pjosa       null    null    493
13      null    null    null    null    null    null    null    RU      null    null    null    null    Kihlman O.      null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    Fl. Pjosa       null    null    494
14      null    null    null    null    null    null    null    FI      null    null    null    null    null    null    null    MZH, Coleoptera Fennica
null    PreservedSpecimen       null    FMNH    null    null    null    null    53
15      null    null    null    null    null    null    null    FI      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    629
16      null    null    null    null    null    null    null    TR      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    639
17      null    null    null    null    null    null    null    EG      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    75
18      null    null    null    null    null    null    null    IL      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    Lac. Genez.     null    null    78
19      null    null    null    null    null    null    null    RU      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
Fennica null    PreservedSpecimen       null    FMNH    null    null    null    null    99
20      null    null    null    null    null    null    null    EG      null    null    null    null    Sahlberg J.     null    null    MZH, Coleoptera
```

Figure 29 The unit_records.txt file

In addition six file all relating back to the core file are created during this process:

- *image_records.txt* - a text file containing a header line with column names, with each line representing a *multimedia record* relating to a given Unit (record) element.
- *identifier_records.txt* - a text file containing a header line with column names, with each line representing an *identifier record* (i.e. GUID) relating to a given Unit (record).
- *identification_records.txt* - a text file containing a header line with column names, with each line representing an *Identification element* relating to a given Unit (record) element.
- *higher_taxon_records.txt* - a text file containing a header line with column names, with each line representing higher taxon elements relating to some Unit (record) element.
- *link_records.txt* - a text file containing a header line with column names, with each line representing a *link record* (i.e. URL) relating to a given Unit (record) element.
- *typification_records.txt* - a text file containing a header line with column names, with each line representing a *typification record* (i.e. type status) relating to a given Unit (record) element.

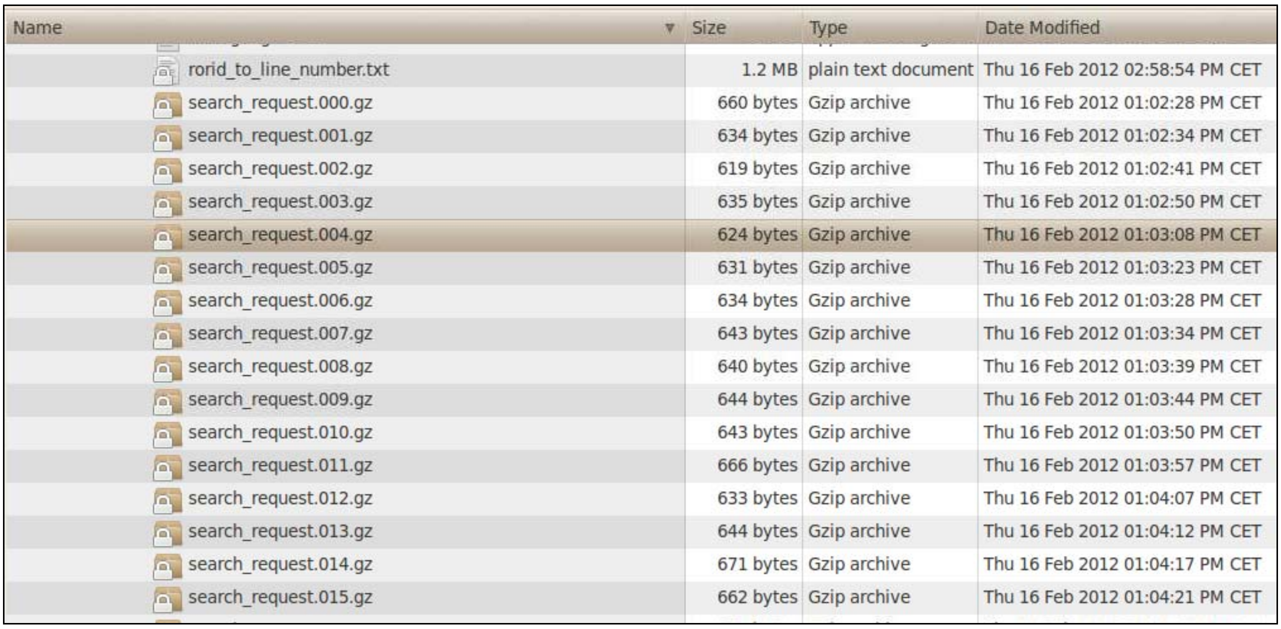
Finally there are the **synchronisation** and the **extraction** operations. During the synchronisation the data is updated and old data is deleted (see Figure 30).



```
2012-02-15 16:05:23.0    Finished extraction
2012-02-15 16:05:23.0    Finished extraction
2012-02-15 16:05:19.0    Start extraction
2012-02-15 16:04:23.0    Finished synchronisation
2012-02-15 16:04:23.0    No GBIF log messages were collected during 'Synchronise' for data resource 43
2012-02-15 16:04:23.0    Reading file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/gbif_log_messages_synchronise.txt
2012-02-15 16:04:23.0    The thread-local variable maxLogGroup = 876
2012-02-15 16:04:23.0    0 identifier_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-15 16:04:23.0    611 typification_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-15 16:04:23.0    0 link_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-15 16:04:23.0    611 image_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-15 16:04:23.0    Synchronisation finished on auxiliary tables
2012-02-15 16:04:21.0    The file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/link_records.txt contains no records
2012-02-15 16:04:21.0    The file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/identifier_records.txt contains no records
2012-02-15 16:04:21.0    Deletion complete
2012-02-15 16:04:21.0    Deleting all typification_record records belonging to this data resource ...
2012-02-15 16:04:21.0    Deletion complete
2012-02-15 16:04:21.0    Deleting all link_record records belonging to this data resource ...
2012-02-15 16:04:21.0    Deletion complete
2012-02-15 16:04:21.0    Deleting all image_record records belonging to this data resource ...
2012-02-15 16:04:21.0    Deletion complete
2012-02-15 16:04:21.0    Deleting all identifier_record records belonging to this data resource ...
2012-02-15 16:04:21.0    Synchronisation started on auxiliary tables
2012-02-15 16:04:21.0    611 raw_occurrence_record records were updated or created (caution: this count is irrespective of duplicate records)
2012-02-15 16:04:21.0    Synchronisation finished on raw occurrence records
2012-02-15 16:04:19.0    Synchronisation started on raw occurrence records
2012-02-15 16:04:19.0    Number of GBIF log messages collected during 'Harvesting': 2
2012-02-15 16:04:19.0    Reading file: /opt/hit/not-in-uddi-9609634/sahlberg/sahlberg/gbif_log_messages.txt
2012-02-15 16:04:19.0    The thread-local variable maxLogGroup = 875
2012-02-15 16:04:19.0    resourceAccessPointId=77
2012-02-15 16:04:19.0    Delete existing namespace_mapping records and write anew for (NULL) resource_access_point: 78
2012-02-15 16:04:19.0    Delete existing namespace_mapping records and write anew for resource_access_point: 77
2012-02-15 16:04:19.0    Found existing NULL resource access point with id 78, data_provider_id 37, data_resource_id 0, and remote_id_at_url NULL
2012-02-15 16:04:19.0    Found existing resource access point with id 77, data_provider_id 37, data_resource_id 43, and remote_id_at_url Sahlberg - updated
2012-02-15 16:04:19.0    dataResourceId=43
2012-02-15 16:04:19.0    Found existing data resource with id 43, and data_provider_id 37 - updated
2012-02-15 16:04:19.0    dataProviderId=37
2012-02-15 16:04:19.0    Found existing data provider with id 37 and uuid not-in-uddi-9609634 - updated
2012-02-15 16:04:19.0    Start synchronisation
```

Figure 30 The synchronisation and the extractions operations in the Console section

The extraction operation creates the ABCD records as search_responses with continuing numbers in .gz format (see Figure 31).



Name	Size	Type	Date Modified
rorid_to_line_number.txt	1.2 MB	plain text document	Thu 16 Feb 2012 02:58:54 PM CET
search_request.000.gz	660 bytes	Gzip archive	Thu 16 Feb 2012 01:02:28 PM CET
search_request.001.gz	634 bytes	Gzip archive	Thu 16 Feb 2012 01:02:34 PM CET
search_request.002.gz	619 bytes	Gzip archive	Thu 16 Feb 2012 01:02:41 PM CET
search_request.003.gz	635 bytes	Gzip archive	Thu 16 Feb 2012 01:02:50 PM CET
search_request.004.gz	624 bytes	Gzip archive	Thu 16 Feb 2012 01:03:08 PM CET
search_request.005.gz	631 bytes	Gzip archive	Thu 16 Feb 2012 01:03:23 PM CET
search_request.006.gz	634 bytes	Gzip archive	Thu 16 Feb 2012 01:03:28 PM CET
search_request.007.gz	643 bytes	Gzip archive	Thu 16 Feb 2012 01:03:34 PM CET
search_request.008.gz	640 bytes	Gzip archive	Thu 16 Feb 2012 01:03:39 PM CET
search_request.009.gz	644 bytes	Gzip archive	Thu 16 Feb 2012 01:03:44 PM CET
search_request.010.gz	643 bytes	Gzip archive	Thu 16 Feb 2012 01:03:50 PM CET
search_request.011.gz	666 bytes	Gzip archive	Thu 16 Feb 2012 01:03:57 PM CET
search_request.012.gz	633 bytes	Gzip archive	Thu 16 Feb 2012 01:04:07 PM CET
search_request.013.gz	644 bytes	Gzip archive	Thu 16 Feb 2012 01:04:12 PM CET
search_request.014.gz	671 bytes	Gzip archive	Thu 16 Feb 2012 01:04:17 PM CET
search_request.015.gz	662 bytes	Gzip archive	Thu 16 Feb 2012 01:04:21 PM CET

Figure 31 The result of the extraction process

When there are no more Jobs in the Job list the Harvesting process with HIT is finished. You should now have a folder structure with the root directory /opt/hit/ and the search_responses (compare Figure 31).

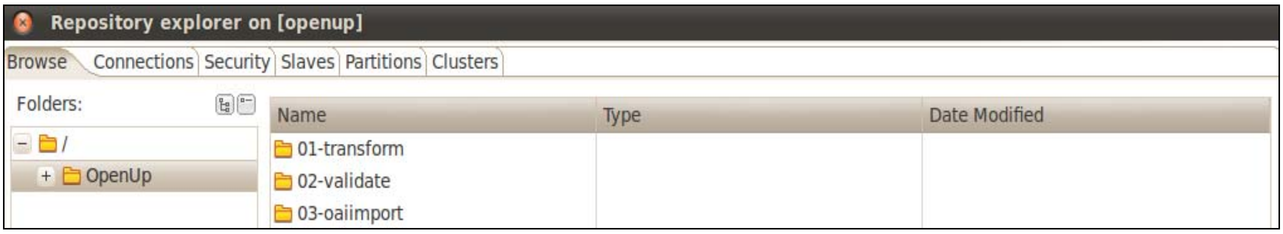
3 PENTAHO KETTLE (DATA TRANSFORMATION)

Pentaho Data Integration (PDI, also called *Kettle*) is the component of Pentaho responsible for the Extract, Transform and Load (ETL) processes.[14]

Pentaho Kettle is used to transform the ABCD records into correct ESE files. The complete process in Pentaho is categorized in three steps:

1. transform
2. validate
3. oai-import

This structure is also represented in the Pentaho repository (see Figure 32).



Name	Type	Date Modified
01-transform		
02-validate		
03-oaiimport		

Figure 32 Repository structure in Pentaho

Before we are starting our Jobs and Transformation it is useful to understand the database structure behind Pentaho.

3.1 Databases

Figure 33 shows the database "etl" with the four tables "Biocase_Harvest_to_ESE", "Biocase_Harvest_to_ESE_result", "Biocase_Harvest_to_ESE_tasks" and "BGBM_Media_URLS". You can see the fields for each table listed in Figure 33.

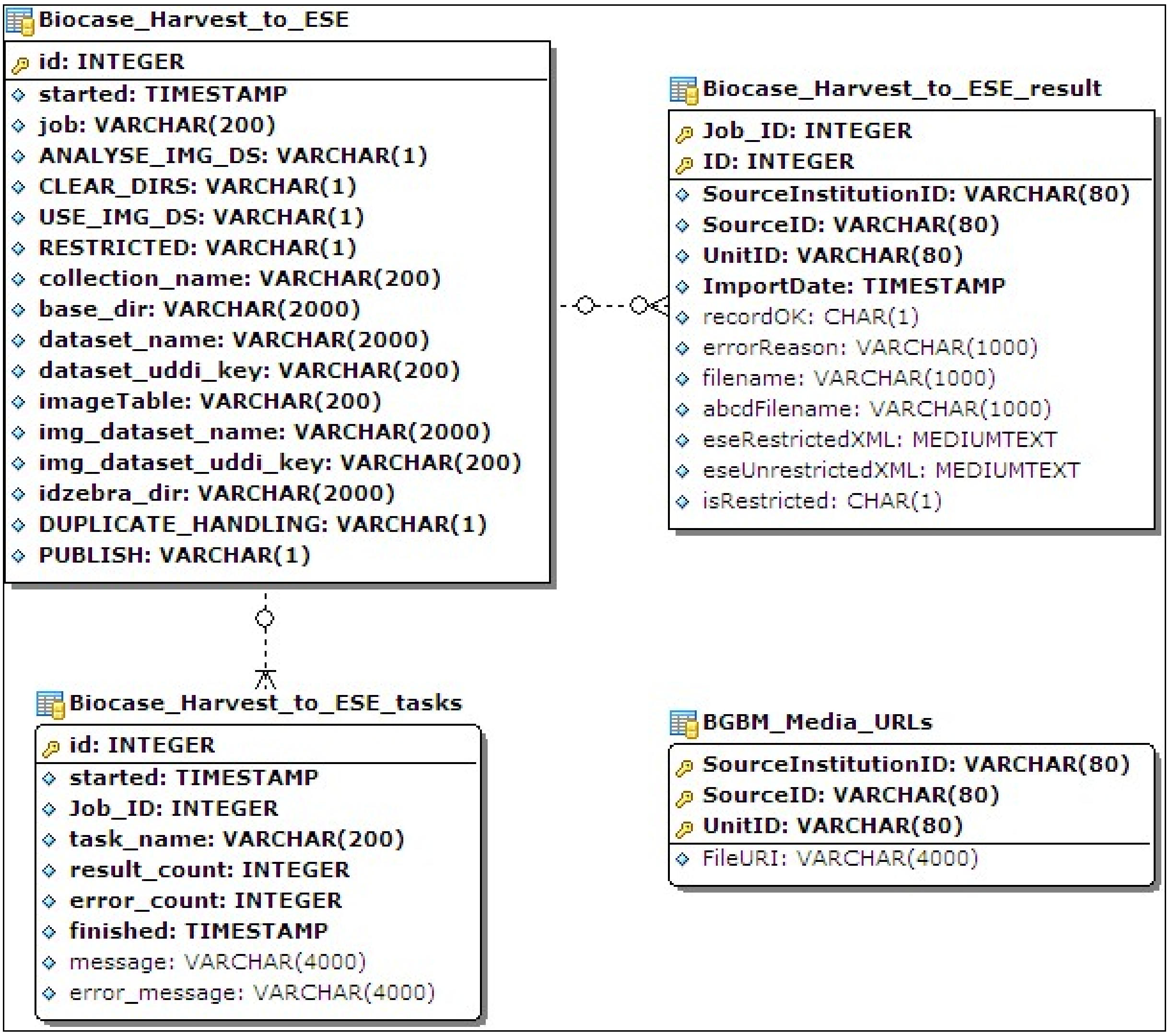
All the Jobs in Pentaho are documented in the table "**Biocase_Harvest_to_ESE**". You will later see that you have to change Job parameters before starting transforming the data. These parameters are all saved in "Biocase_Harvest_to_ESE".

All the correct finished ESE records are saved in the table "**Biocase_Harvest_to_ESE_result**". It contains the transformation results.

In the table "**Biocase_Harvest_to_ESE_tasks**" all tasks per Job (transform, validate, oai-import) are saved. It shows also the error message if something goes wrong during the transformation.

Finally there is the table "**BGBM_Media_URLs**" where all media data sources (images) are saved. It has the function of a lookup table.

[14] http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial 17 Feb. 2012.



Biocase_Harvest_to_ESE
- id: INTEGER
- started: TIMESTAMP
- job: VARCHAR(200)
- ANALYSE_IMG_DS: VARCHAR(1)
- CLEAR_DIRS: VARCHAR(1)
- USE_IMG_DS: VARCHAR(1)
- RESTRICTED: VARCHAR(1)
- collection_name: VARCHAR(200)
- base_dir: VARCHAR(2000)
- dataset_name: VARCHAR(2000)
- dataset_uddi_key: VARCHAR(200)
- imageTable: VARCHAR(200)
- img_dataset_name: VARCHAR(2000)
- img_dataset_uddi_key: VARCHAR(200)
- idzebra_dir: VARCHAR(2000)
- DUPLICATE_HANDLING: VARCHAR(1)
- PUBLISH: VARCHAR(1)

Biocase_Harvest_to_ESE_result
- Job_ID: INTEGER
- ID: INTEGER
- SourceInstitutionID: VARCHAR(80)
- SourceID: VARCHAR(80)
- UnitID: VARCHAR(80)
- ImportDate: TIMESTAMP
- recordOK: CHAR(1)
- errorReason: VARCHAR(1000)
- filename: VARCHAR(1000)
- abcdFilename: VARCHAR(1000)
- eseRestrictedXML: MEDIUMTEXT
- eseUnrestrictedXML: MEDIUMTEXT
- isRestricted: CHAR(1)

Biocase_Harvest_to_ESE_tasks
- id: INTEGER
- started: TIMESTAMP
- Job_ID: INTEGER
- task_name: VARCHAR(200)
- result_count: INTEGER
- error_count: INTEGER
- finished: TIMESTAMP
- message: VARCHAR(4000)
- error_message: VARCHAR(4000)

BGBM_Media_URLs
- SourceInstitutionID: VARCHAR(80)
- SourceID: VARCHAR(80)
- UnitID: VARCHAR(80)
- FileURI: VARCHAR(4000)

Figure 33 Structure of the “etl” database with four tables

3.2 *Creating a folder structure*

The three main folders have sub-folders representing the countries from the content providers. When you have harvested a new datasource, you have to create a hierarchically correct folder in every of the three main folders.

Remember our example datasource we have harvested before – “Sahlberg” from the University of Finland. First of all you must create a new country folder named “Finland” in the main folders “transform”, “validate” and “oaiimport” (see Figure 34).

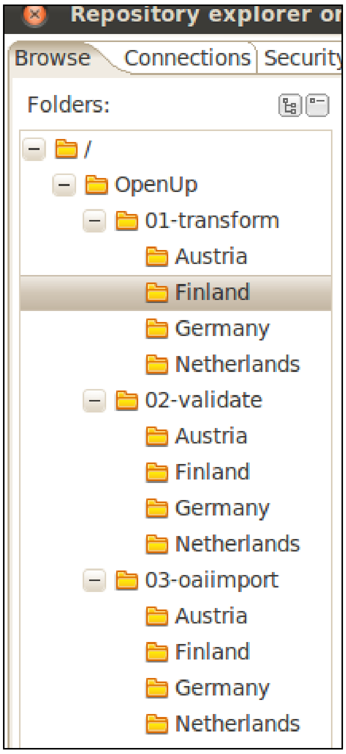


Figure 34 Creating a "Finland" folder in every category

As you can see there are already a few countries in every category. The Transformations or Jobs do never change, no matter which datasource is processed. So you can just copy an existing Job and save it as a new one. The only thing you have to adapt before starting the Jobs are the Job Parameters.

Before that you have to create three Jobs – one for every category. The names of the Jobs are consistent. The transform-Job is named after the collection name (see Figure 35 for our example "Sahlberg").

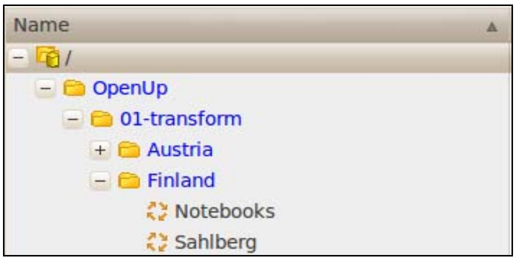
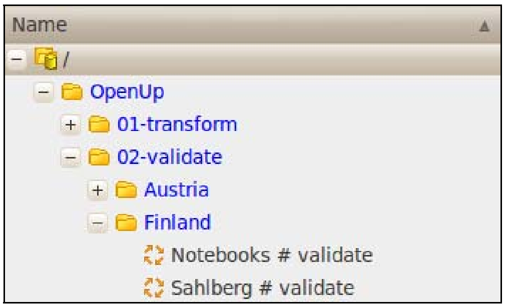


Figure 35 Two Finnish Jobs in the transform category

In the "validate" folder the Jobs are named after the collection plus the word "validate". Between the collection name and "validate" you have to type the symbol "#" (see Figure 36).

Important: The three Jobs for one datasource must have the same name (the collection name). Everything behind the "#" symbol is ignored by the system.

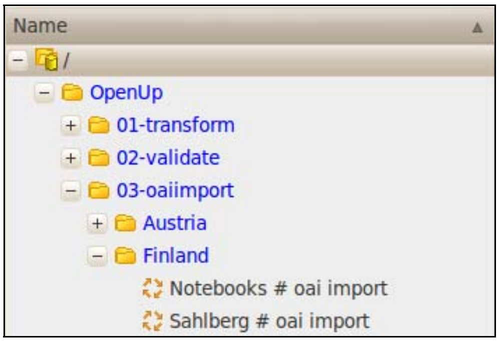


Name

- /
  - OpenUp
    - 01-transform
    - 02-validate
      - Austria
      - Finland
        - Notebooks # validate
        - Sahlberg # validate

Figure 36 The Finnish Jobs in the validate category

Now one Job is missing for the oai-import. Again the name of the Job has the same collection name followed by # oai import (see Figure 37).



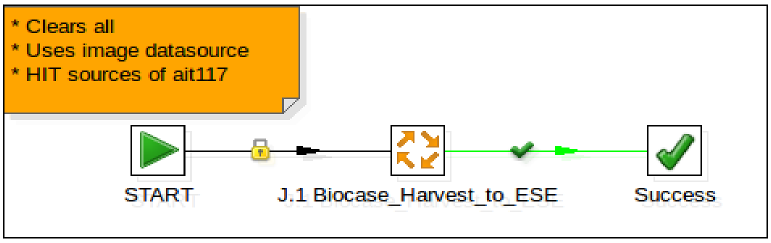
Name

- /
  - OpenUp
    - 01-transform
    - 02-validate
    - 03-oaiimport
      - Austria
      - Finland
        - Notebooks # oai import
        - Sahlberg # oai import

Figure 37 The Finnish Jobs in the oaiimport category

3.3 01-transform

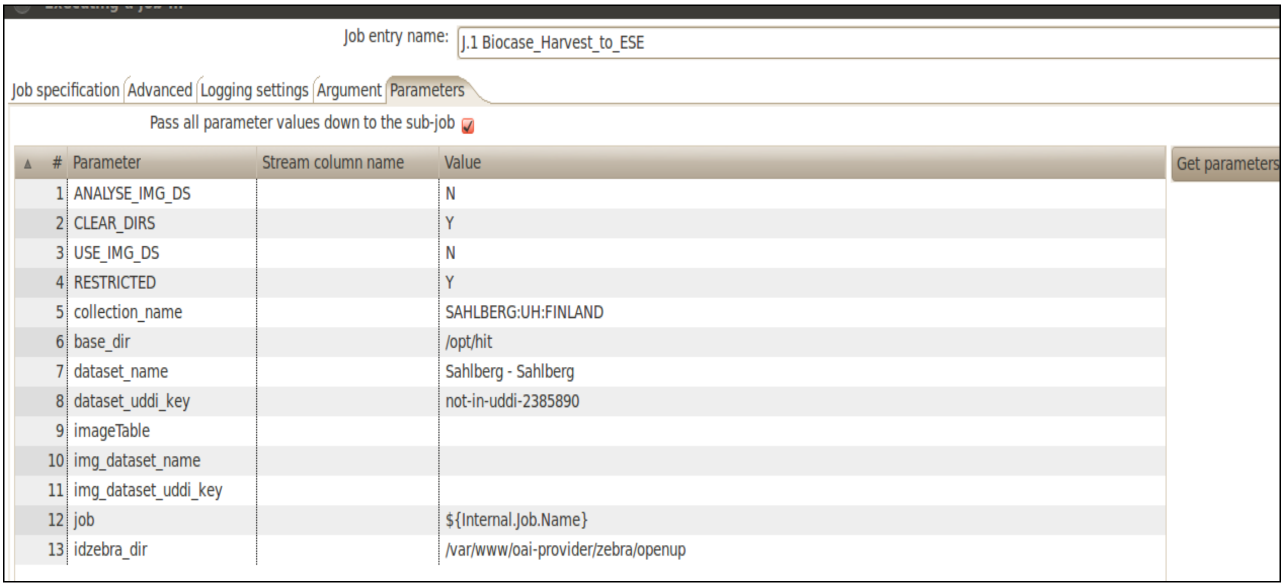
In the “Sahlberg” example the Job in the “transform” directory looks like shown in Figure 38.



* Clears all
* Uses image datasource
* HIT sources of ait117

Figure 38 The Job “Sahlberg” in the transform category

You can open the Job parameters by double-clicking on the orange Job icon in the middle and go to the last tab called “Parameters” (see Figure 39).



Job entry name: J.1 Biocase_Harvest_to_ESE

Pass all parameter values down to the sub-job ☑

#	Parameter	Stream column name	Value
1	ANALYSE_IMG_DS		N
2	CLEAR_DIRS		Y
3	USE_IMG_DS		N
4	RESTRICTED		Y
5	collection_name		SAHLBERG:UH:FINLAND
6	base_dir		/opt/hit
7	dataset_name		Sahlberg - Sahlberg
8	dataset_uddi_key		not-in-uddi-2385890
9	imageTable		
10	img_dataset_name		
11	img_dataset_uddi_key		
12	job		${Internal.Job.Name}
13	idzebra_dir		/var/www/oai-provider/zebra/openup

Figure 39 Parameters of the Job "Sahlberg"

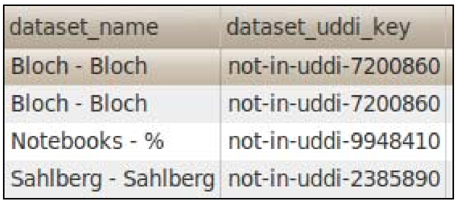
In Figure 39 the Parameters are already filled in correctly. First of all you have to define if the collection is "Restricted" or "Unrestricted" (Parameter number 4). You can do this by typing Y (for Yes, it is restricted) or N (for No, it is not restricted = unrestricted) in the correct value field.

Parameter number 5 is the collection identifier. It has always the same pattern: COLLECTION_NAME:CONTENT_PROVIDER:COUNTRY

Please do only use capital letters. As you can see the collection identifier of our example collection "Sahlberg" is SAHLBERG:UH:FINLAND

Parameter number 6 shows the base directory /opt/hit you have defined in the installation process of the HIT harvester (compare Figure 8).

The "dataset_name" and the "dataset_uddi_key" (Parameter 7 and 8) are taken from the SQL database "Biocase_Harvest_to_ESE" (see Figure 40, compare Figure 33).

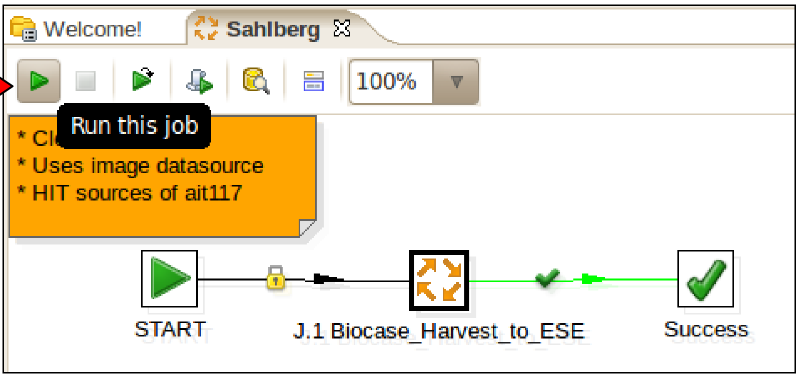


dataset_name	dataset_uddi_key
Bloch - Bloch	not-in-uddi-7200860
Bloch - Bloch	not-in-uddi-7200860
Notebooks - %	not-in-uddi-9948410
Sahlberg - Sahlberg	not-in-uddi-2385890

Figure 40 The columns "dataset_name" and "dataset_uddi_key" in "Biocase_Harvest_to_ESE"

Parameter 12 is the variable ${Internal.Job.Name}. Therefore it is important that the three Jobs for one collection have the same name. The last Parameter "idzebra_dir" is the zebra directory.

When everything has been filled in correctly you can click "OK" and then start the Job by clicking on the "Play" symbol (see Figure 41).

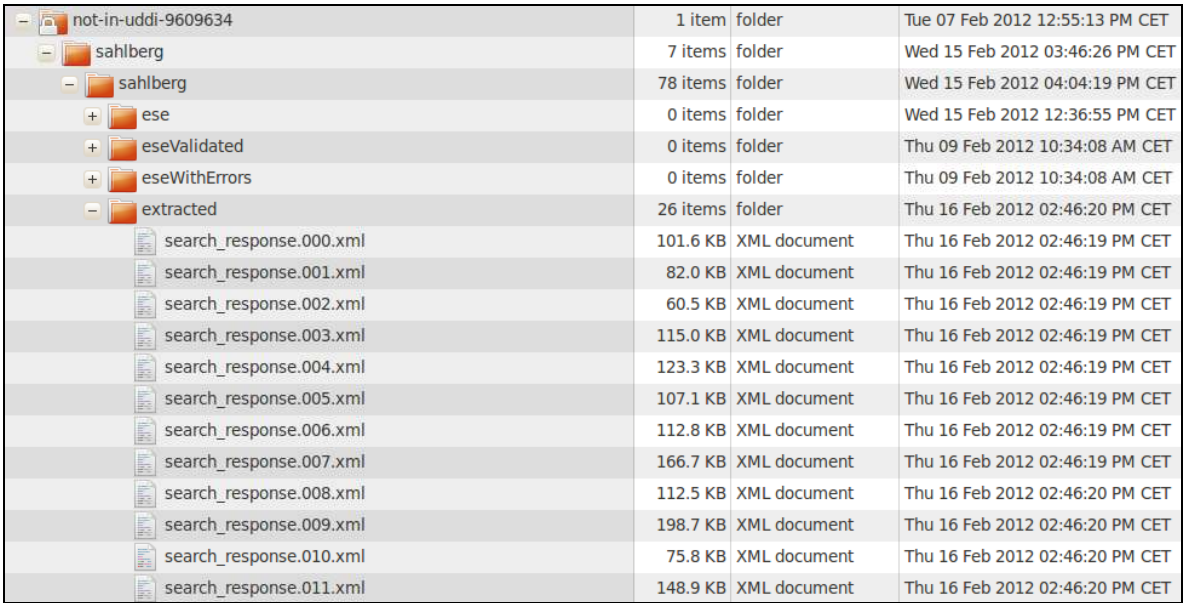


* Cl
* Uses image datasource
* HIT sources of ait117

Figure 41 Starting the Job “Sahlberg”

You must not start the “validate” Job before the “transform” Job is finished. It is very important to keep the order 01-transform, 02-validate, 03-oaiimport.

The results of this first Job are XML files in ABCD format in the folder “extracted” (see Figure 42).

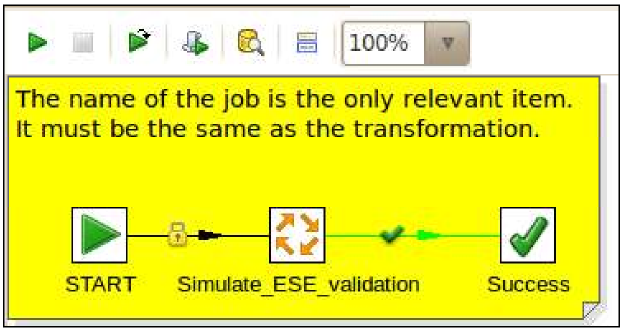


− not-in-uddi-9609634	1 item	folder	Tue 07 Feb 2012 12:55:13 PM CET
− sahlberg	7 items	folder	Wed 15 Feb 2012 03:46:26 PM CET
− sahlberg	78 items	folder	Wed 15 Feb 2012 04:04:19 PM CET
+ ese	0 items	folder	Wed 15 Feb 2012 12:36:55 PM CET
+ eseValidated	0 items	folder	Thu 09 Feb 2012 10:34:08 AM CET
+ eseWithErrors	0 items	folder	Thu 09 Feb 2012 10:34:08 AM CET
− extracted	26 items	folder	Thu 16 Feb 2012 02:46:20 PM CET
search_response.000.xml	101.6 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.001.xml	82.0 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.002.xml	60.5 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.003.xml	115.0 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.004.xml	123.3 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.005.xml	107.1 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.006.xml	112.8 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.007.xml	166.7 KB	XML document	Thu 16 Feb 2012 02:46:19 PM CET
search_response.008.xml	112.5 KB	XML document	Thu 16 Feb 2012 02:46:20 PM CET
search_response.009.xml	198.7 KB	XML document	Thu 16 Feb 2012 02:46:20 PM CET
search_response.010.xml	75.8 KB	XML document	Thu 16 Feb 2012 02:46:20 PM CET
search_response.011.xml	148.9 KB	XML document	Thu 16 Feb 2012 02:46:20 PM CET

Figure 42 ABCD records in the folder “extracted” after running the “Sahlberg” Job

3.4 02-validate

When the first Job is finished one can open the Job “Sahlberg # validate” you have created before and start running this Job (see Figure 43).



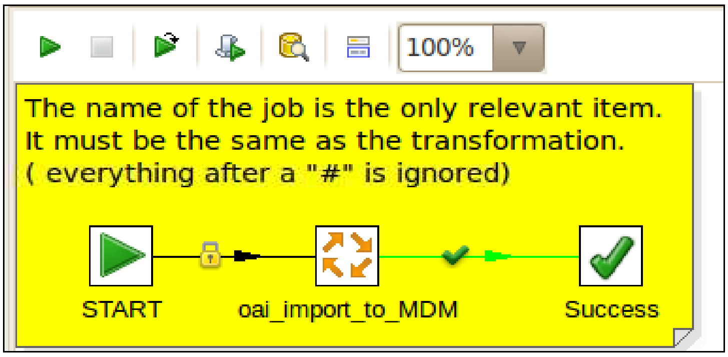
The name of the job is the only relevant item.
It must be the same as the transformation.

Figure 43 The Job "Sahlberg # validate"

This Job simulates ESE validation by copying the records in the "ESEvalidated" directory.

3.5 03-oai-import

Finally open the Job "Sahlberg # oai import" in the 03-oaiimport directory (see Figure 44). Start this Job after the "validate" Job is finished.



The name of the job is the only relevant item.
It must be the same as the transformation.
(everything after a "#" is ignored)

Figure 44 The Job "Sahlberg # oai import"

When this is done the work with Pentaho Kettle is done. You should now have correct ESE records that can be controlled on the OAI-Provider-platform.

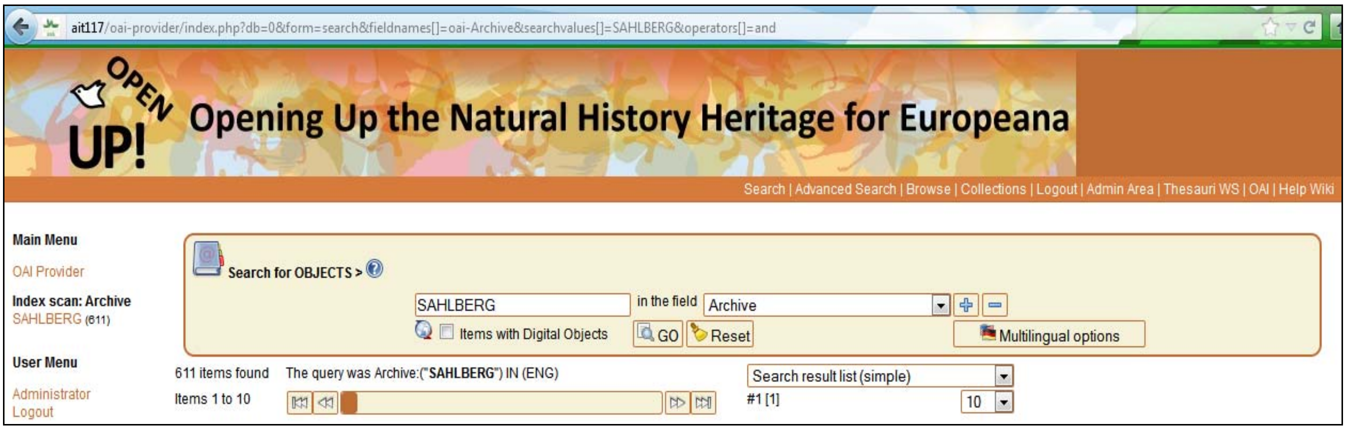
4 THE OAI-PROVIDER

You can reach it by typing [your domain][15]into your internet browser. You can now have a look on the ESE records that have been uploaded with the Pentaho Job “Sahlberg # oai import”.

To do this you can use the “Advanced search” or the “Browse” function.

4.1 Advanced search

You can simply type your query in the search box and click on “Go” (see Figure 45). Furthermore you can define in which field the search term should appear (see Figure 46).



ait117/oai-provider/index.php?db=0&form=search&fieldnames[]=oai-Archive&searchvalues[]=SAHLBERG&operators[]=and

Index scan: Archive
SAHLBERG (611)

Search for OBJECTS >

SAHLBERG in the field Archive

□ Items with Digital Objects

611 items found The query was Archive:("SAHLBERG") IN (ENG)

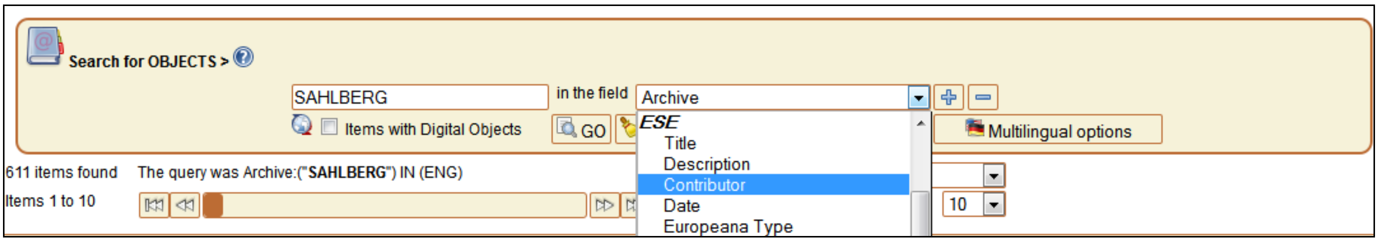
Items 1 to 10

Search result list (simple)

#1 [1] 10

Opening Up the Natural History Heritage for Europeana
Search | Advanced Search | Browse | Collections | Logout | Admin Area | Thesauri WS | OAI | Help Wiki
Main Menu
OAI Provider
User Menu
Administrator
Logout
GO
Reset
Multilingual options

Figure 45 Searching for “Sahlberg”



Search for OBJECTS >

SAHLBERG in the field Archive

□ Items with Digital Objects

611 items found The query was Archive:("SAHLBERG") IN (ENG)

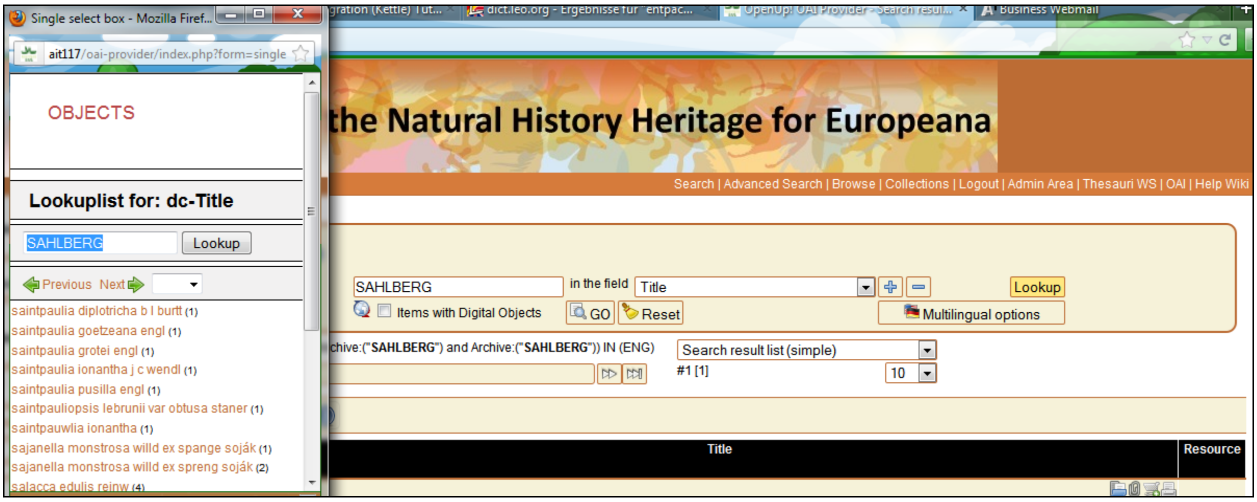
Items 1 to 10

10

Figure 46 Using the “in the field” search option

If you need help during your research you can use the “Lookup” function (see Figure 47).

[15] for example http://localhost/oai-provider/index.php



Single select box - Mozilla Firef...

ait117/oai-provider/index.php?form=single

OBJECTS

Lookuplist for: dc-Title

SAHLBERG Lookup

Previous Next

saintpaulia diplotricha b l burtt (1)
saintpaulia goetzeana engl (1)
saintpaulia grotei engl (1)
saintpaulia ionantha j c wendl (1)
saintpaulia pusilla engl (1)
saintpauliopsis lebrunii var obtusa staner (1)
saintpauwlia ionantha (1)
sajanella monstrosa willd ex spange soják (1)
sajanella monstrosa willd ex spreng soják (2)
salacca edulis reinw (4)

gration (Kettle) Tut...
dict.leo.org - Ergebnisse für "entpac...
OpenUp! OAI Provider - Search resul...
Business Webmail
the Natural History Heritage for Europeana
Search | Advanced Search | Browse | Collections | Logout | Admin Area | Thesauri WS | OAI | Help Wiki

SAHLBERG in the field Title Lookup

☐ Items with Digital Objects GO Reset Multilingual options

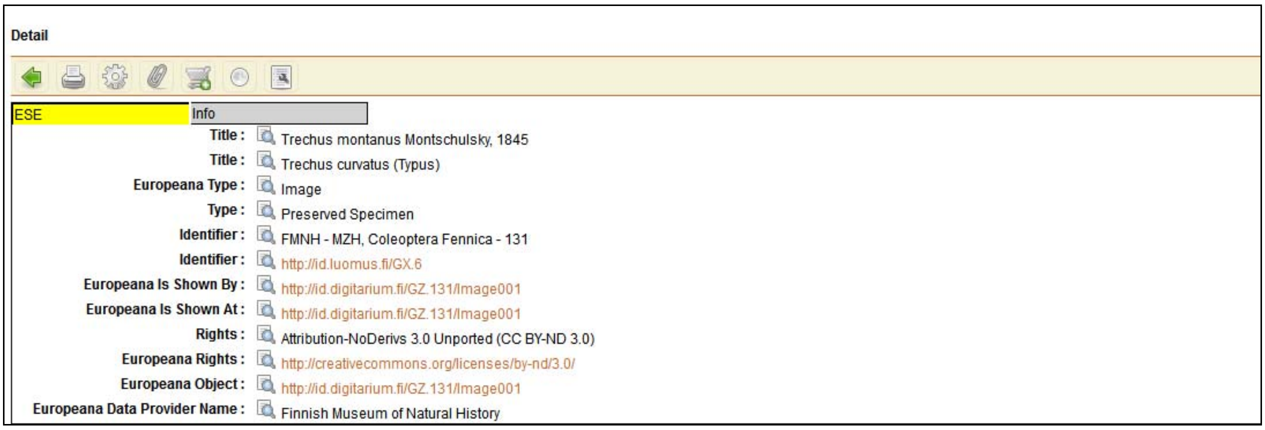
chive:("SAHLBERG") and Archive:("SAHLBERG")) IN (ENG)

Search result list (simple)

#1 [1] 10

Figure 47 Looking up titles of the collection "Sahlberg"

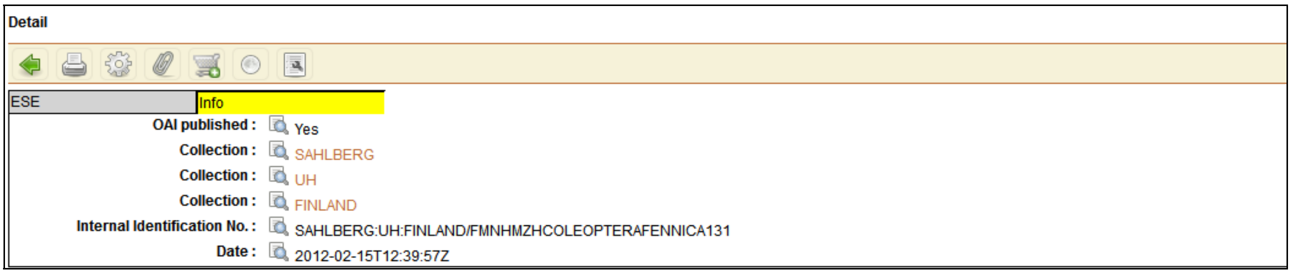
When clicking on one of the result records you see the ESE record with the different fields (see Figure 48). You can switch to the "Info" tab to control the collection information (see Figure 49).



Detail

Title : Trechus montanus Montschulsky, 1845
Title : Trechus curvatus (Typus)
Europeana Type : Image
Type : Preserved Specimen
Identifier : FMNH - MZH, Coleoptera Fennica - 131
Identifier : http://id.luomus.fi/GX.6
Europeana Is Shown By : http://id.digitarium.fi/GZ.131/Image001
Europeana Is Shown At : http://id.digitarium.fi/GZ.131/Image001
Rights : Attribution-NoDerivs 3.0 Unported (CC BY-ND 3.0)
Europeana Rights : http://creativecommons.org/licenses/by-nd/3.0/
Europeana Object : http://id.digitarium.fi/GZ.131/Image001
Europeana Data Provider Name : Finnish Museum of Natural History

Figure 48 Displaying the ESE record



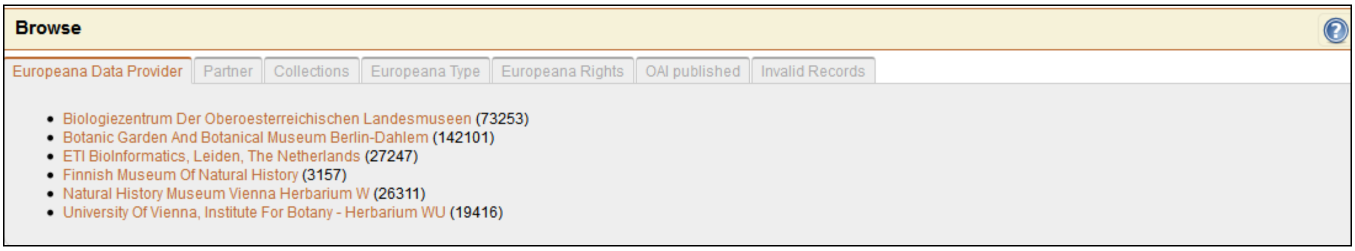
Detail

OAI published : Yes
Collection : SAHLBERG
Collection : UH
Collection : FINLAND
Internal Identification No. : SAHLBERG:UH:FINLAND/FMNHMZHCOLEOPTERAFENNICA131
Date : 2012-02-15T12:39:57Z

Figure 49 Displaying the collection information

4.2 Browse

You can use the "Browse" function to find records as well. As you can see in Figure 50 you can browse the records by Europeana Data Provider, Partner, Collections, Europeana Type, Europeana Rights, OAI published and Invalid Records. In brackets you can see the number of records.

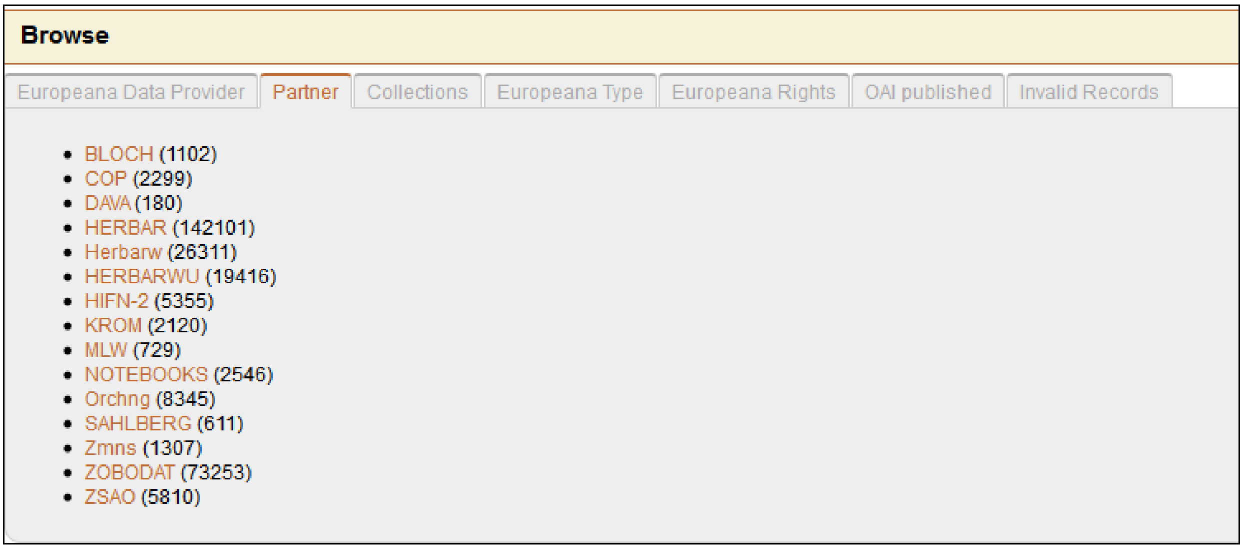


Browse

- Biologiezentrum Der Oberoesterreichischen Landesmuseen (73253)
- Botanic Garden And Botanical Museum Berlin-Dahlem (142101)
- ETI BioInformatics, Leiden, The Netherlands (27247)
- Finnish Museum Of Natural History (3157)
- Natural History Museum Vienna Herbarium W (26311)
- University Of Vienna, Institute For Botany - Herbarium WU (19416)

Figure 50 Browsing the records

If you are looking for the "Sahlberg" records for example, you can either click on the "Finnish Museum of Natural History" or going to the "Partner" or "Collections" tab (see Figure 51).



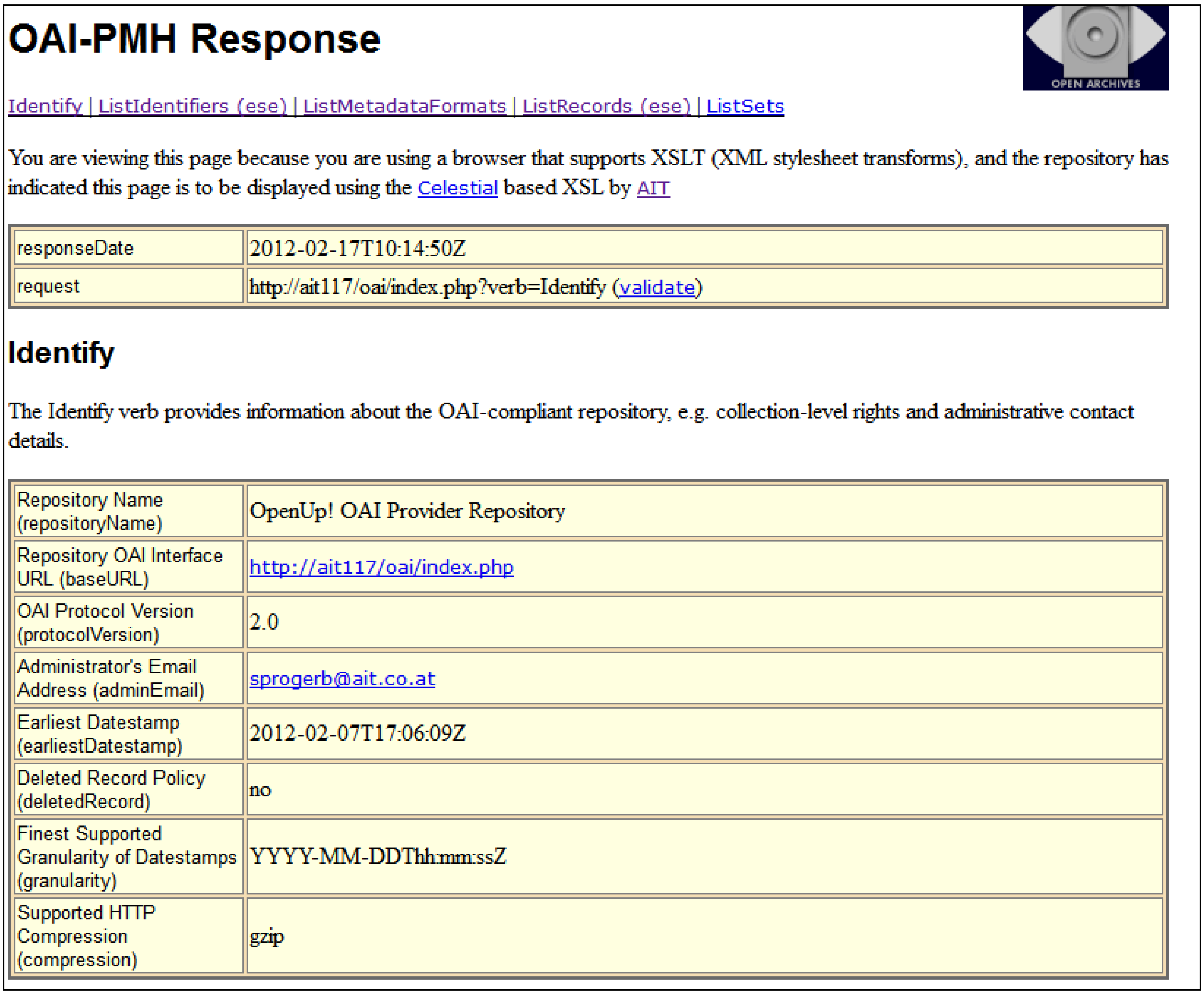
Browse

- BLOCH (1102)
- COP (2299)
- DAVA (180)
- HERBAR (142101)
- Herbarw (26311)
- HERBARWU (19416)
- HIFN-2 (5355)
- KROM (2120)
- MLW (729)
- NOTEBOOKS (2546)
- Orchng (8345)
- SAHLBERG (611)
- Zmns (1307)
- ZOBODAT (73253)
- ZSAO (5810)

Figure 51 Browsing the partner "Sahlberg"

4.3 OAI

When clicking on "OAI" in the upper right corner the following window opens (see Figure 52).



OAI-PMH Response

Identify | ListIdentifiers (ese) | ListMetadataFormats | ListRecords (ese) | ListSets

You are viewing this page because you are using a browser that supports XSLT (XML stylesheet transforms), and the repository has indicated this page is to be displayed using the Celestial based XSL by AIT

responseDate	2012-02-17T10:14:50Z
request	http://ait117/oai/index.php?verb=Identify (validate)

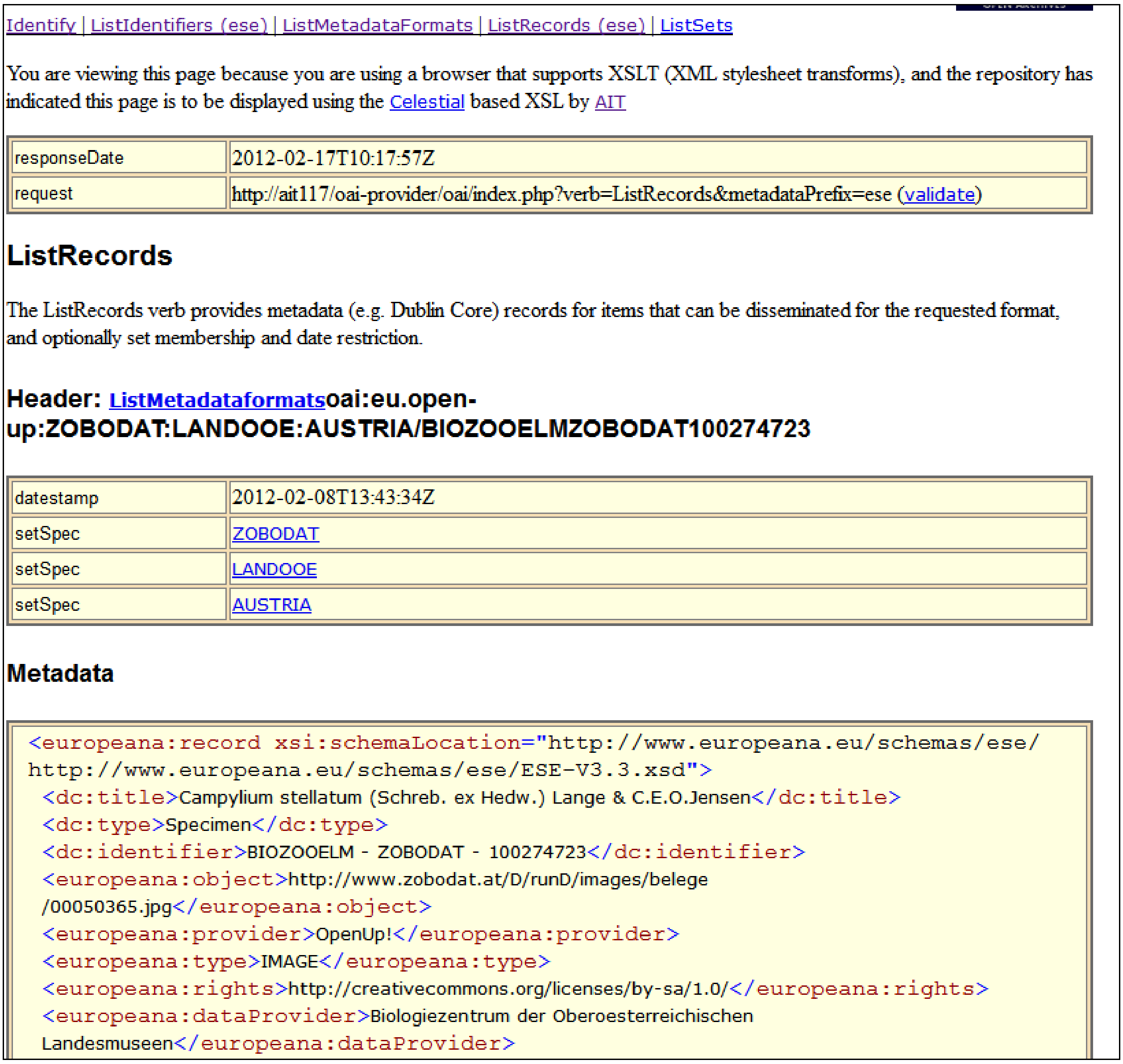
Identify

The Identify verb provides information about the OAI-compliant repository, e.g. collection-level rights and administrative contact details.

Repository Name (repositoryName)	OpenUp! OAI Provider Repository
Repository OAI Interface URL (baseURL)	http://ait117/oai/index.php
OAI Protocol Version (protocolVersion)	2.0
Administrator's Email Address (adminEmail)	sprogerb@ait.co.at
Earliest Datestamp (earliestDatestamp)	2012-02-07T17:06:09Z
Deleted Record Policy (deletedRecord)	no
Finest Supported Granularity of Datestamps (granularity)	YYYY-MM-DDThh:mm:ssZ
Supported HTTP Compression (compression)	gzip

Figure 52 OAI-PMH Response

There you can choose between “Identify”, “ListIdentifiers (ese)”, “ListMetadataFormats”, “ListRecords (ese)” and “ListSets”. Figure 53 shows an example record after clicking on “ListRecords (ese)”.



Identify | ListIdentifiers (ese) | ListMetadataFormats | ListRecords (ese) | ListSets

You are viewing this page because you are using a browser that supports XSLT (XML stylesheet transforms), and the repository has indicated this page is to be displayed using the Celestial based XSL by AIT

responseDate	2012-02-17T10:17:57Z
request	http://ait117/oai-provider/oai/index.php?verb=ListRecords&metadataPrefix=ese (validate)

ListRecords

The ListRecords verb provides metadata (e.g. Dublin Core) records for items that can be disseminated for the requested format, and optionally set membership and date restriction.

Header: ListMetadataformatsoai:eu.open-up:ZOBODAT:LANDOOE:AUSTRIA/BIOZOOELMZOBODAT100274723

datestamp	2012-02-08T13:43:34Z
setSpec	ZOBODAT
setSpec	LANDOOE
setSpec	AUSTRIA

Metadata

```
<europeana:record xsi:schemaLocation="http://www.europeana.eu/schemas/ese/
http://www.europeana.eu/schemas/ese/ESE-V3.3.xsd">
  <dc:title>Campylium stellatum (Schreb. ex Hedw.) Lange & C.E.O.Jensen</dc:title>
  <dc:type>Specimen</dc:type>
  <dc:identifier>BIOZOOELM - ZOBODAT - 100274723</dc:identifier>
  <europeana:object>http://www.zobodat.at/D/runD/images/belege
  /00050365.jpg</europeana:object>
  <europeana:provider>OpenUp!</europeana:provider>
  <europeana:type>IMAGE</europeana:type>
  <europeana:rights>http://creativecommons.org/licenses/by-sa/1.0/</europeana:rights>
  <europeana:dataProvider>Biologiezentrum der Oberoesterreichischen
  Landesmuseen</europeana:dataProvider>
```

Figure 53 Clicking on “ListRecords (ese)”

5 LIST OF REFERENCES

Hit. http://code.google.com/p/gbif-indexingtoolkit/ 17 Feb. 2012.

Pentaho. http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation 17 Feb. 2012.

6 LIST OF FIGURES