ECP-2008-DILI-518001

BHL-Europe

Live BHL-Europe system, with distributed storage and management and appropriate tools for the continued development of services and ingress of multilingual content

Deliverable number	*D3.9*
Dissemination level	*Public*
Delivery date	*22 May 2012*
Status	*Final*
Author	*Chris Sleep*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

1 Document History

1.1 Contributors

Person	Partner
Henning Scholz	MfN
Graham Higley, Chris Sleep, Lola Obajuluwa	NHM
Jiri Frank	NMP
Lee Namba, Zheng Li, Zhaoyu Li	ATOS
Andreas Kohlbecker	FUB-BGBM
Wolfgang Koller	NHMW

1.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
19 April 2012	Henning Scholz	0.1	ToC and draft components
18 May 2012	Chris Sleep	0.2	Draft of the Pre-Ingest guide
20 May 2012	Lola Obajuluawa	0.3	Draft compiled version including all developer information
21 May 2012	Chris Sleep	0.4	Draft cleaned up and amended
22 May 2012	Chris Sleep	1.0	Final version

1.3 Reviewers

This document requires the following reviews and approvals.

Name	Date	Version
BHL-Europe PCO	22 May 2012	1.0

1.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
BHL-Europe consortium	22 May 2012	1.0

2 Table of Contents

3 Purpose and document structure

The aim of this document is to describe the final BHL-Europe system, with components of this document aimed at the users of the BHL-Europe system; including information covering the use of the BHL-Europe Pre-Ingest tool by Content Providers, and tutorial information covering the BHL-Europe Portal. The remainder of the document concerns specific implementation details for the system components deployed for the final BHL-Europe system.

In comparison to D3.5 which described the concepts of the BHL-Europe system and also introduced the components needed within the system, this document details the technical background with detailed information by and for developers. The key components and implementations in this document were initially described in D3.7, which reported the state of technology development at that time. Further development and integration has since been carried out to deliver the system as recorded here.

The tools and components implemented were selected with the intent of long term continued usage and development, making use of those tools which have substantial community adoption and support. For example making use of Drupal CMS for the portal, Fedora commons for the archival storage, Apache Solr for search indexing and so on. Those entirely bespoke components which have been developed by BHL-Europe have been developed using an open source model, and are committed to a public repository, GitHub, with an open source licence applied, to support future enhancement possibilities.

4 Tutorials for the BHL-Europe Portal

The main purpose of the BHL-Europe portal is to bring together important search and information retrieval functionalities to help scientists and academics navigate within the digital literature which is crucial for their research. These functionalities are presented within a very user friendly and attractive interface, which will also be suitable for a wider public as well. As development of the portal nears completion – it is an appropriate time to take a look at the portal structure and functionalities.

This section provides an introductory tutorial demonstrating the basic functionalities of the portal. A more detailed step by step tutorial is also prominently displayed in the navigation menu on each page of the portal. Whereby the user is provided with useful information on the portal and how best to use each of the functionalities. The tutorials are text blocks explaining the functions on specific pages, and each portal page has its own tutorial.

The portal is separated into several areas with specific functions, linked together. The main areas are: front page, advanced search, browse, result list, bibliographic page and content viewer.

4.1 Front page

The front page consists of several sections with the simple search as the focal function. The main roles of the front page are (1) to briefly introduce the portal to users and (2) to enable users to do simple searches for literature. For casual or general users the front page acts as a means of attracting their attention, to explore the portal further. The front page sections and functions are described in the next sections.

4.1.1 Simple search

The simple search function lets the user search within the most used search categories: title, author, year (date of publication) and scientific name. The portal default setting automatically checks the first three of these search categories, and the user needs to check the scientific name search category if they wish to search using that function. All the user needs to do is type in their search term. Several web services are integrated in the simple search: these include autocomplete search, fuzzy search and taxon finder for a scientific name query. In addition to this simple search, an advanced search option is also available, with a direct link to it from the simple search section.

4.1.2 Content highlights

One of the eye-catching features of the front page is the self-rotating carousel. This carousel demonstrates some content highlights, randomly generated from the content highlight archive whenever one visits the page. Clicking on one of the books in the carousel will lead the visitor to its bibliographic page.

4.1.3 Help and tutorial

The rotating hint bar at the base of the front page provides the user with useful information on the portal and gives them hints on how to effectively use its functionalities. A link to the BHL-Europe tutorial is also prominently displayed in the navigation menu on each page of the portal. The tutorials are text blocks explaining the functions on specific pages, and every

portal page has its own tutorial. The front page tutorial also includes a link to the video tutorial, where several use cases are demonstrated to help show users how to effectively get the best from using the portal.

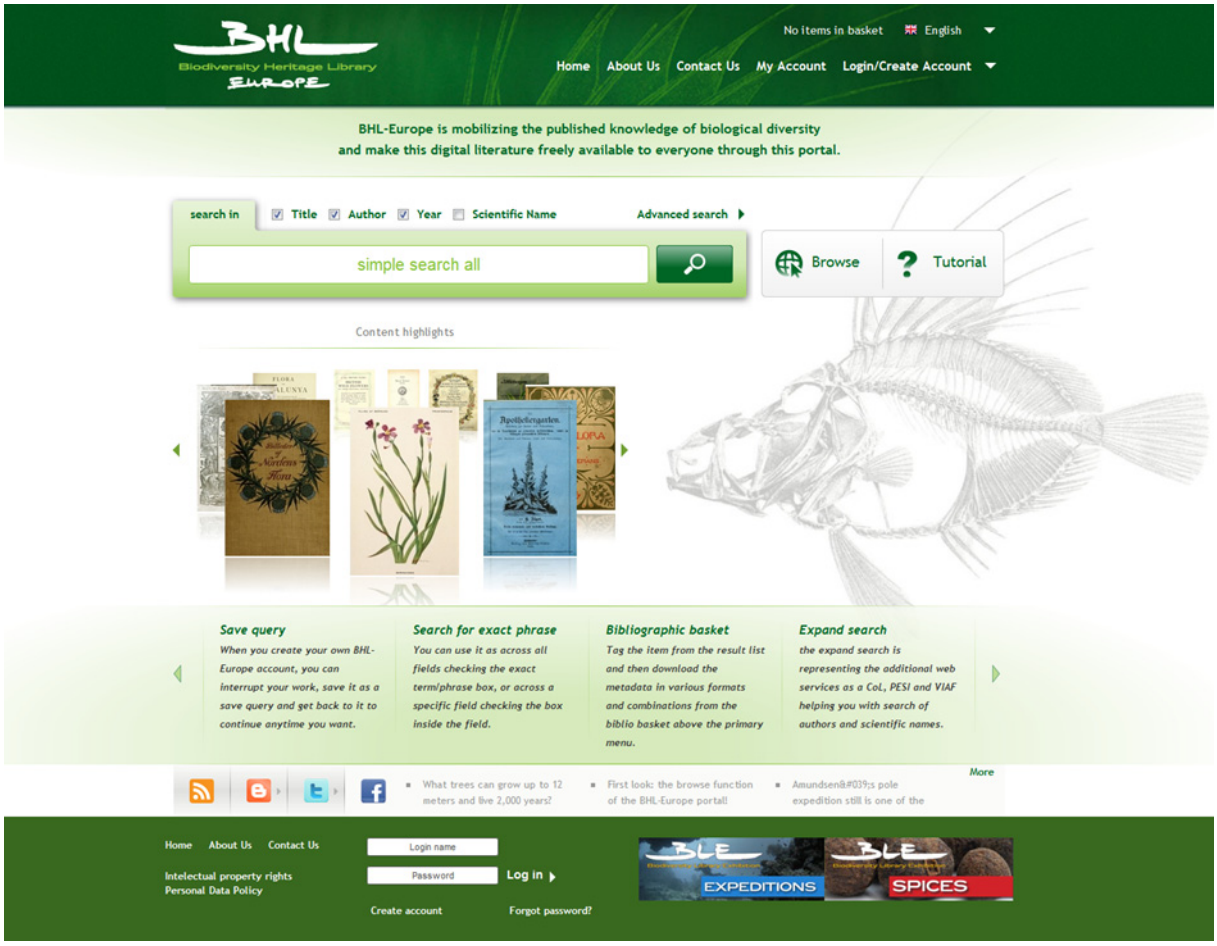

Figure 4-1: Front Page.

4.1.4 Social media bar

Finally, there's the social media bar, which displays the latest 3 posts on the BHL-E blog, Twitter and Facebook. This way, users get a direct view of what we've been up to – and can easily click through to subscribe, like or follow. Users can also subscribe to the BHL-Europe RSS-feed to keep up to date.

The front page has a slightly different template and wireframe for navigation from the other pages in the portal, but all the other portal pages maintain a uniform for navigation. Under the top banner with the logo and primary menu there is a **Navigation menu** in the center. This menu includes 3 links: Search (which is the advanced search), Browse and Tutorial. The tutorial link is for the specific page tutorial. This menu could be expanded or collapsed clicking on the three dots at his bottom. That's it for the front page. The next part is the result list.

4.2 Result list

The result list appears after performing a simple search, advanced search or a browse (see the separate sections below for further information on these functions). The results list is an easily navigated interface, with functions that help the user get precise search results. The result list is structured in two columns and several sections, which are described below.

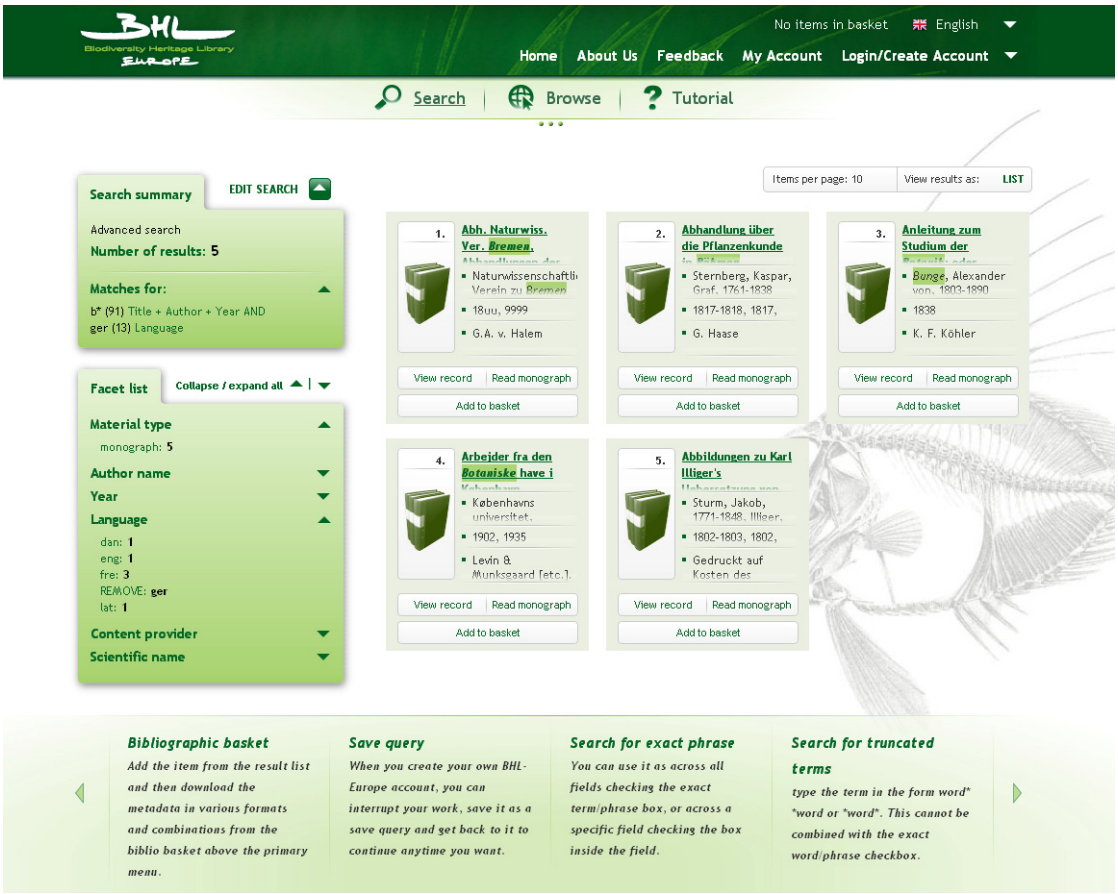

Figure 4-2: Result List.

4.2.1 Right Column

The bigger right column shows the result items as Journal/Series, Volumes, Articles and Monographs. Each item type has its own icon representing the specific content type and its number according to its position in the list. The metadata of each item is structured in 4 categories which depend on the content type. For a monograph, for example, these are: Title, Author, Year and Publisher. The title is always interactive and clicking on it will retrieve the bibliographic page for this item.

Below the item metadata there are three icons which enable the user to: go to the bibliographic page for that item, read the item in the content viewer, or tag the item and send it to the tagging basket.

The tagging basket menu with number of items with the number of selected items in it, is displayed on each portal page above the primary menu. Users can download metadata for tagged books in various formats, e.g. Summary, MODS, Endnote, Bibtex and OLEF or even

combine which information they prefer to download from selected books. There is of course the option to download them all.

Items can be easily removed/untagged either directly from the result list or from the basket itself.

There are two display modes to structure the list of results: the list view and the table view. The list view is a classical view sorting the items in lines. Each item has the full title displayed, no matter how long it is. The table view is designed to display more results on one screen (see picture above). Items are in blocks of the same size; three next to each other is the maximum.

The text is limited to three text lines for each data field. If the text is longer, which is normally the case for titles; you can see the whole title by moving the mouse over it. The number of items on one page for both table and list view can be set between minimum 6 and maximum 30, and you can always switch between the list and table view.

Relevance sorting is the default setting for the result list, and it can be sorted alphabetically by title if the search parameter is set to equal.

4.2.2 Left Column

The smaller left column includes two sections. The first contains the search summary with the number of items matched in the result and the search string. The search string contains the term searched, along with the search parameter and the number of matches for each individual search term in the index. So if you search for several terms across several parameters the search summary shows how many results there are for this complex search string in total, but it does not mean that the individual search terms do not appear more frequently in the index, if you were to search for them individually.

Every search parameter is entered on one line for a clearer overview and to enable the opportunity to remove or add more search parameters easily. To change the search parameters you can click on the edit search button in the search summary block and edit the search or browse settings (depending on from where the search began). The advanced search section or browse section appears above the result list.

One of the most useful (and amazing) tools on the portal is the facet list. The facet list is the second section in the left column and contains the following facet categories: Material type, Author, Year, Language, Content provider and Scientific name. The main function of the facet list is to allow users to filter the result list by combining selections in categories and help to find exactly what they are looking for. Each facet selection adds a new request at the end of the search string.

You can remove the filter by clicking on the minus button in the search block or by clicking on “REMOVE” directly on the category in the facet list. You can remove the requests independent of order. Each change will affect the search result and result list items. Most categories are self-explanatory, but it is important to mention specifically the “scientific name” category. Expanding this category will show a long list including genera and species names or synonyms for names found in the content OCR of the items in the result list. All results in categories in the facet list are sorted alphabetically or chronologically. You can expand or collapse individual categories or all of them together.

4.3 Advanced search

When doing an advanced search via the BHL-Europe portal, users can use Boolean operators and other structured search characters, like wildcards, brackets, forward and end word truncation ... These functions are described on the hint bar on the lower part of the advanced search page. For instance, users are able to search for an exact phrase or search for a range of years. More functions are described in the below sections.

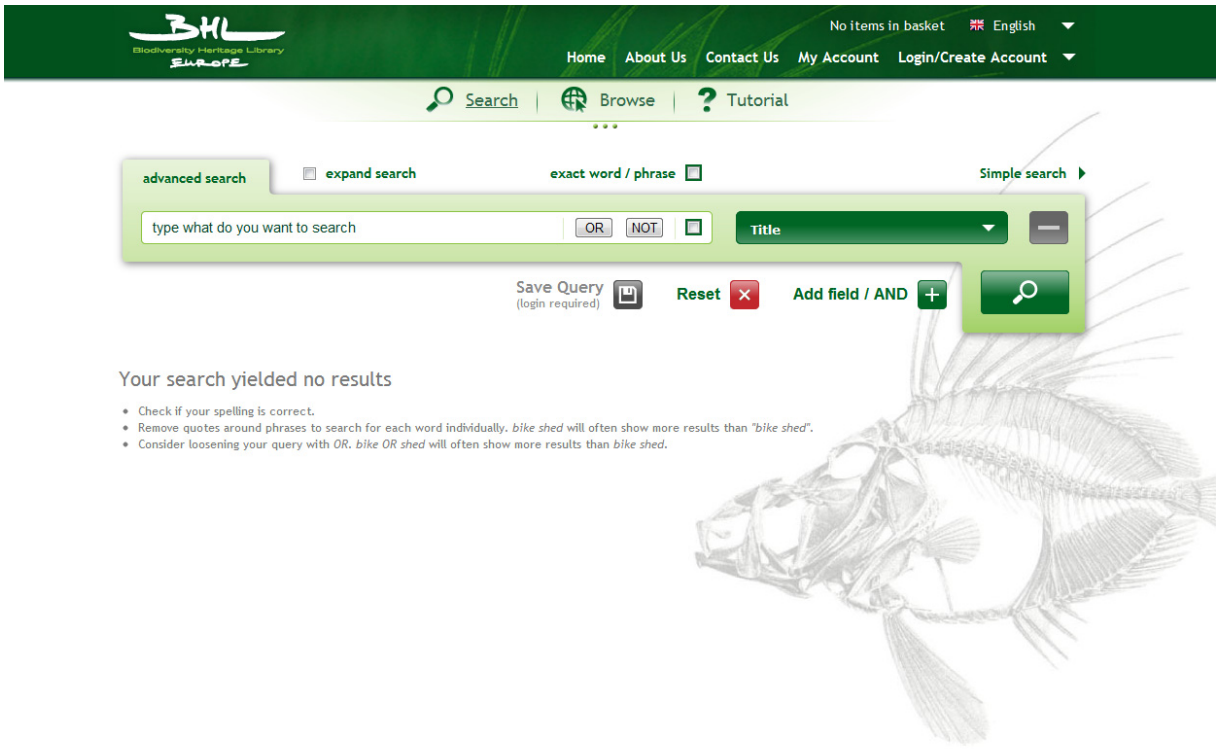

Figure 4-3: Advanced Search (1)

4.3.1 Metadata fields/categories

The advanced search allows users to search in a specific data field or search category and combine them. First and foremost is the title field, which searches across titles for all content types (Journal/Series, Volumes, Articles and Monograph). The other fields include Author, Year, Language, Place of publication, Scientific name, Common name etc. There are 11 search categories which are extended in combination with the facet list function in the result list.

4.3.2 Add a field

To combine searches across several metadata fields, users can click on the plus button of the Add field function and add another search field. The number of search fields is not limited and you can combine fields to get the most precise search result. The default relationship between added search fields is the Boolean operator “AND”. Clicking a minus button at the right side of the field will remove the specific search field. It’s important to remember that

searches can also be improved when viewing the result list by making use of the facet list or editing the search.

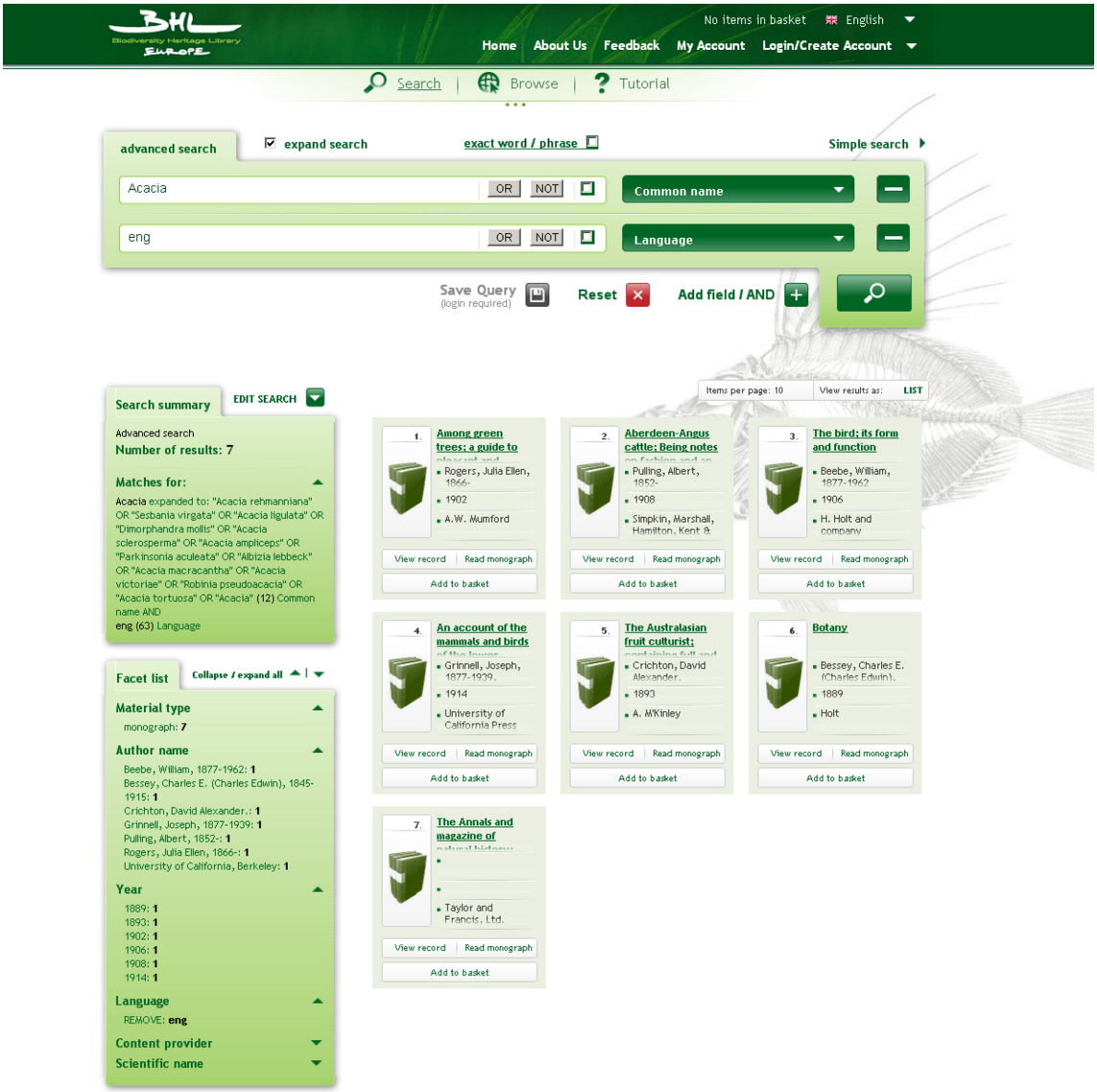

Figure 4-4: Advanced Search (2)

4.3.3 Exact search

The functionality of the exact term/phrase check box is pretty much self-explanatory. This function can be used as a global parameter to enable exact searching across all fields, or it can be specified for a specific search field by checking the check box inside the field.

4.3.4 Expand search

Checking this box enables you to expand your search to the data in the Catalogue of Life (CoL), the Pan-European Species directories Infrastructure (PESI), the Virtual International Authority Files (VIAF) and the Zeitschriftendatenbank (ZDB). These services help with searches for authors, scientific names or common names by using external name databases.

This means, for example, that if a user is searching for a particular author e.g. Darwin, the expand search enables them to find the results where Darwin's name appears in various formats, e.g. full name, part name, surname and initial, or even translated into different languages. This function allows the user to get the most complete list of results for their search term. For scientific names the web services also looks for synonyms and common names.

With this function users can search for scientific names via common names and the search will retrieve books where only the scientific name for the searched common name appears. This functionality also applies to taxonomic synonyms especially useful for those cases where the taxon names in the historic literature are no longer valid and are treated as synonyms of the accepted taxon name. The expand search function works as a global parameter, but affects only the metadata fields author, scientific name and common name.

4.3.5 Reset

Reset is a self-explanatory function. Clicking on the reset button will clear the whole search string.

4.3.6 Save query

If you are logged in you can save your search parameter under a specific name and load it when you would like to continue your search. The advanced search will be automatically prefilled by the search parameter. The number of search queries is not limited. Save queries are stored under the "View" in my account after login.

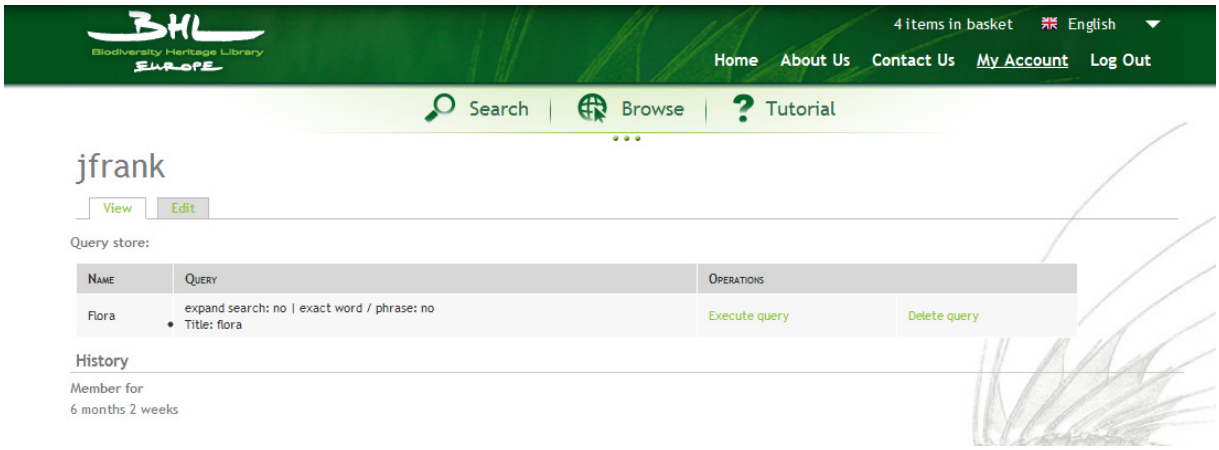

Figure 4-5: Save Query

4.4 Browse

The browse function is for users who are not sure what they are searching for or who want to scan the content available in the portal. Browse allows users to scroll through results from the five most used search categories: Title, Author, Journal title, Year and Content provider. The browse function is closely related to the advanced search and in fact it is a predefined advanced search function.

Browse by Title, **Author** and **Journal title** functions are linked to a Roman alphabet keypad. Browsing using these categories retrieves results where the letter selected are the first letters in the authors' names and first indexed title or journal title word listed.

4.4.1 Browse by year

The "browse by year" option contains two parts. The top part is represented by a time line where users can define the start and end of the time period. If the period is less or equal to one hundred years, the "Use the interval" button on the right activates and by clicking on it users are able to search within this particular time period. If the range is larger than one hundred years, the button is inactive to avoid getting a large set of results. The second part consists of decade buttons. Here the time line in the top part generates decade buttons for each decade. The timeline points are also by decade. The time line only generates decade buttons where content is available. The browse by year option allows the user to use several options to define a decade or range of years.

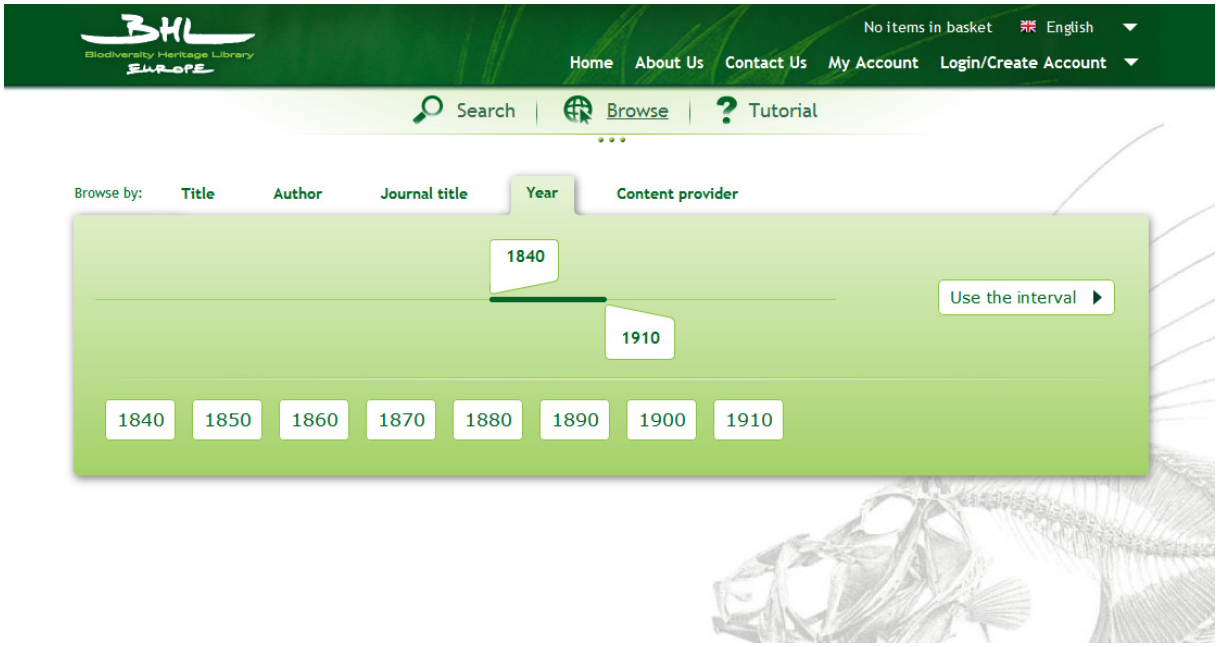

Figure 4-6: Browse by Year

4.4.2 Browse by content provider

This option shows the abbreviations of the content provider institutions and allows the user to search the content provided by a specific content provider.

The result list appears after you do any browse option. The words in titles or authors metadata fields are highlighted according to my browse parameter, to show the user why the exact

content related to my browse and displayed in the result list. The user can further refine the results using the facet list. When they start using the facet list, the search system in the background automatically switches into the advanced search option and by editing the search the user can adapt the specific search options. If users do not use the facet list, clicking on the edit search will make the Browse option appear so that users can continue browsing.

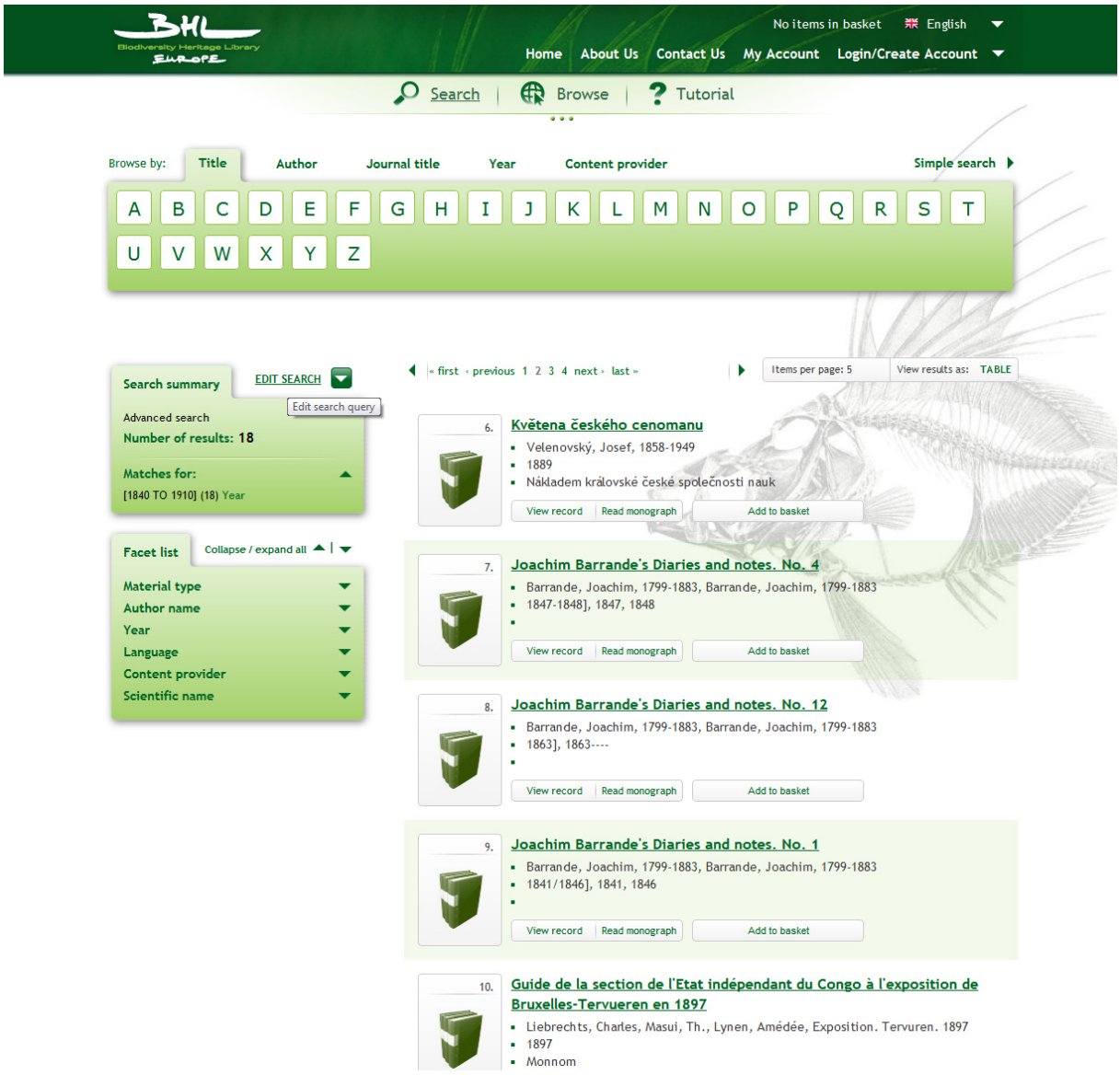

Figure 4-7: Browse and Result List

4.5 Bibliographic page

This page is a detailed description of the digital item and is reached by clicking on the item title or the View record button in the results list, the item title in the tagging basket, or the item title in the content viewer. The bibliographic page is structured in two main sections. The most prominent part of the biblio page is the item title at the top of the right hand section.

Figure 4-8: Bibliographic Page

4.5.1 Right section

This section could also be called the metadata section because it includes information about the item in various metadata formats. There are six available formats depending on underlying metadata, namely: Summary, Abstract, MODS, Endnote, Bibtex and OLEF. The **Summary** format consists of various basic metadata fields, customized for each content type and is of most use to the widest group of users; it is also the default format.

Closely related to the summary format is the **Abstract**, which appears, if it has been submitted by the content provider. Both the summary and abstract can be downloaded directly in .txt format. The remaining four formats are **MODS**, **Endnote**, **Bibtex** and **OLEF.** These formats are, of most value for specialists, developers, scientists or librarians. These formats give these users information on API creation, references and data mapping. The metadata in each of these formats can be directly downloaded by users. The tagging basket also provides another way of downloading metadata in these more specialist formats.

4.5.2 Left section

This section varies depending on the material type looked at. Above this section is a step back button connecting back to the results list. At the top of the section two types of information are displayed, one is an interactive link showing the content provider of the specific item and leading to a browse of this content provider, and the second is the ID for the item. Every item has its specific identification code. In the center of this section is a thumbnail or icon showing the material-type. All material types physically scanned and provided in the portal have thumbnails. These represent the monographs, articles and volumes which are not parsed into articles. Content which is not scanned as a single part is mainly journals/series and volumes which are parsed into separated articles. In the lower part of the right section is a smaller green block which has three main functions: 1. link to the digital item in cases where there is a hierarchical relationship between journal title and component volumes, or book series and component volumes; 2. select and read the item; 3. download digital items where the item is an unparsed volume, article or monograph. These physical items can be downloaded in various formats, namely PDF, OCR and JPG.

So every item independent of the material type has its own bibliographic page, but the difference is in the thumbnail, the possibility to download the item and the structure of the item summary. If the user clicks on “Read selected item”, it leads to the final key component of the portal – the content viewer.

4.6 Content viewer

The content viewer provides the user with several functions and possibilities to display selected books and articles.

4.6.1 View types

The basic view is the **one-page view** which shows just a single page. These single pages scroll vertically.

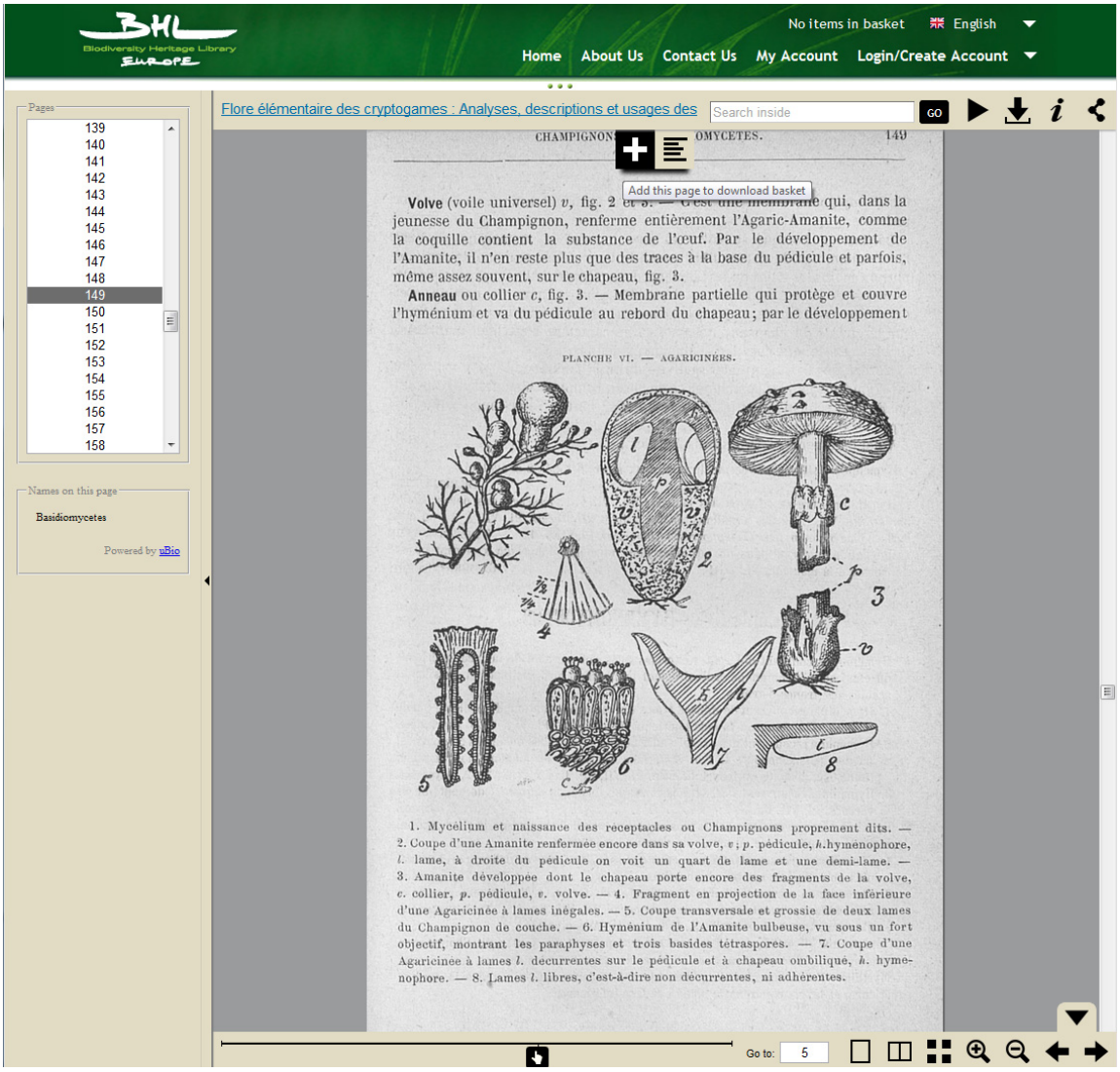

Figure 4-9: Content Viewer one-page view

The **two-page view** opens the book or article with two joined pages and looks more like a printed book, the user can even turn the virtual pages from right to left by clicking on the page. Pages can also be turned up by using the pages pane or the page scale (see below for more information about this).

Figure 4-10: Content Viewer Search & two-Page View

A very user friendly view for quick navigation within a book or selection of illustrations is the **thumbnail view** which shows the page thumbnails in a table with 6 pages per row. The number of rows depends on the number of pages. In each view it is possible to **zoom in** or **out**. Each type of view is represented by icons in the lower menu to the right of the page scale.

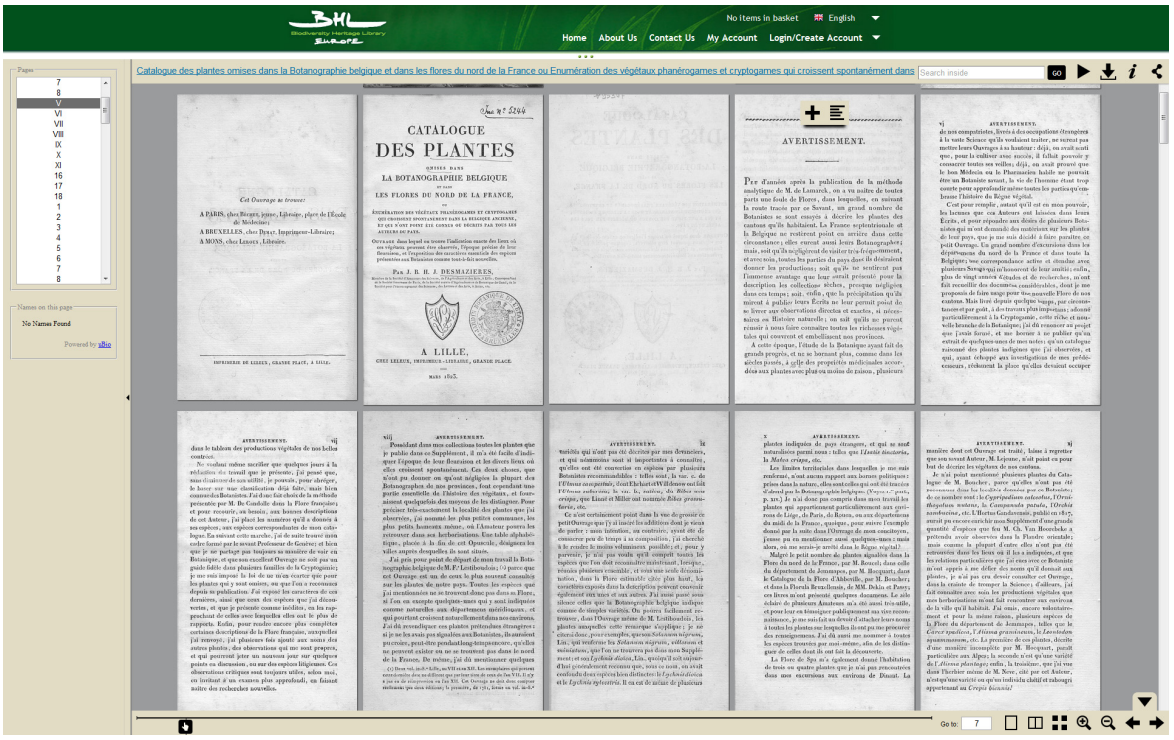

Figure 4-11: Thumbnail View

4.6.2 Navigation

There are several options for navigating within the content. In the left column, are two sections, which can be hidden by clicking on the arrow in the middle of the screen. The "Pages" section at the top provides a **Table of contents** function. Depending on the book data, users are able to look at a display of the pages in a book with page descriptions, such as title page, chapter or figure. In cases where such data is not specified this function displays a scroll down list of page numbers corresponding to the physically scanned pages in the book open in the viewer. The number of pages is connected to the view and display of the physical page depending on which type of view is set up. Another navigation tool is the **page scale** in the lower menu. Moving the cursor on this scale turns the pages depending on the view type and shows the page info (e.g. the number) corresponding to the Table of contents. To the right of the page scale is the small frame "**Go to**", allowing users to go to a specified page number. The lower menu is the same as the left block and can be hidden to enlarge the content screen. The pages can also be turned using the **left or right arrow** on the left side of the lower menu or on the keyboard. The last means of navigation is to turn the pages directly in the content screen by clicking on the pages in two-pages view or using the mouse scroll log in one-page or thumbnail view.

4.6.3 Download

When the cursor is on the page, independent of the view type, two icons on top of the page will appear. One icon shows as "+" and the other displays several lines. The plus icon adds pages to the basket, which means that it enables users to add images of pages to the **download basket.** By clicking the plus icon, it is changed to a minus icon and by operating this users are able to remove selected pages from the download basket. The basket icon is in the top right position in the content viewer menu and shows the number of pages which have been added. Clicking on it makes the light box view appear, giving several possibilities on how to download selected pages. Pages are displayed under the "**Your Basket**" menu and can easily be removed from the basket.

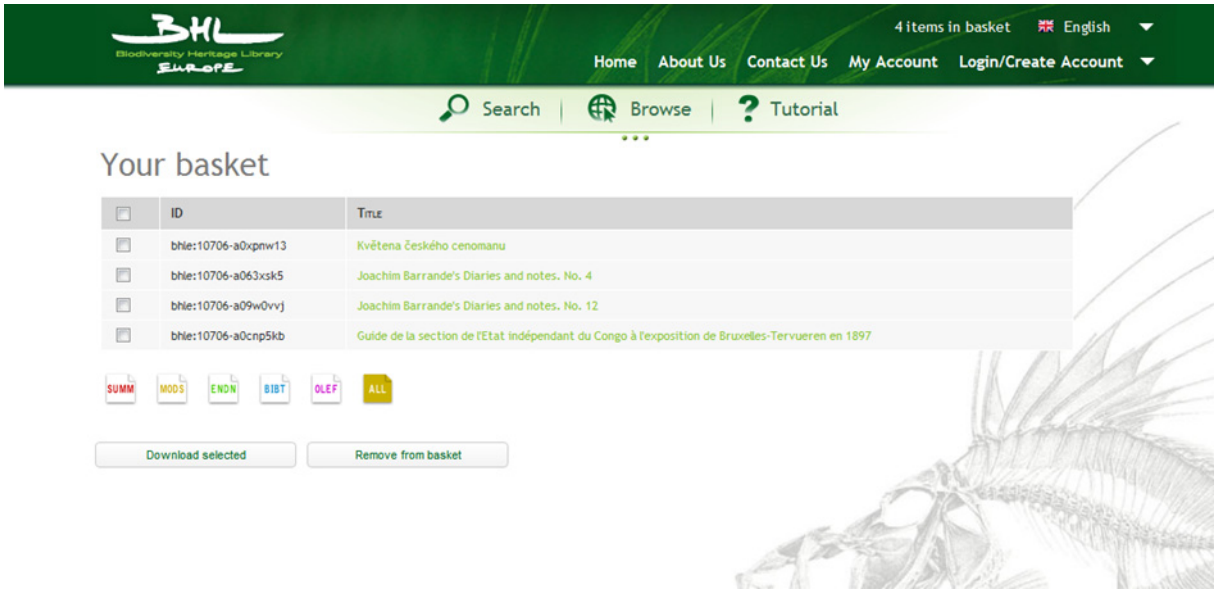

Figure 4-12: Bibliographic Basket

It is possible to download the page in three formats, PDF, OCR or JPEG and in various resolutions in the case of PDF or JPEG. To complete the download of pages users need to enter an email address to deliver the download link to, and click on the button marked "**Download**".

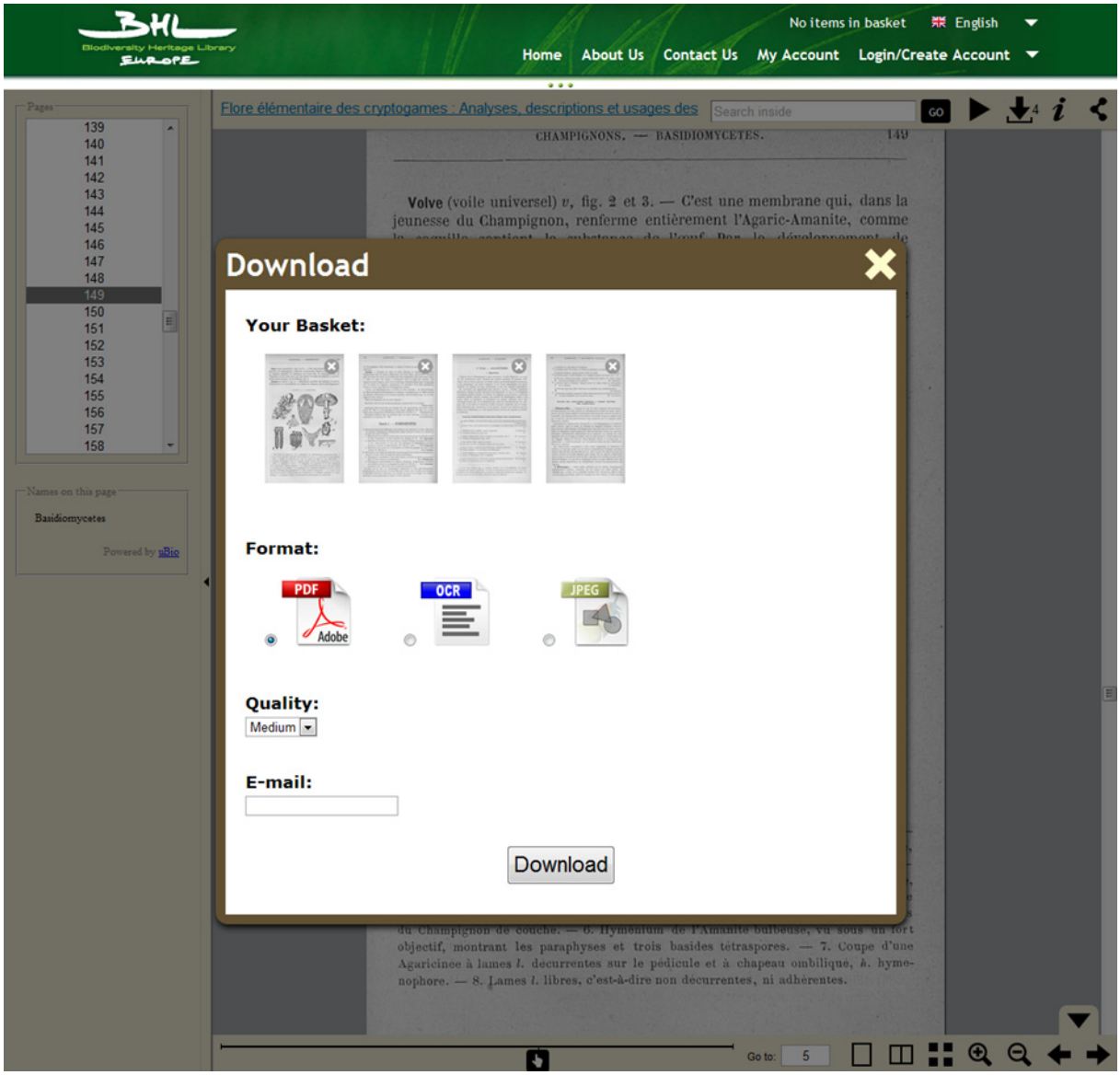

Figure 4-13: Content Viewer Basket

In case the user wants to download the whole book without clicking on each page in the content viewer, they can click on the book title, which will lead to the biblio page, where it is possible to download an entire book in various formats. The download option in the content viewer mainly focuses on specific selected pages as references, contents list, and of course for the illustrations. The pages can also be downloaded by using the actual browser options. The second icon on the page with lines links to the OCR text of the page.

4.6.4 Taxon finder

The lower section in this left column shows recognized taxa on the page being looked at in the content viewer. This section is powered by uBio web services.

4.6.5 Search

This function is related to the other navigation functions but is more specific. Search in the content viewer allows searching within the OCR text of the book displayed. The search frame is in the upper menu. If the term searched retrieves positive results, the occurrences of the search term in the text are indicated as reversed orange tear drop on the page scale. Moving the mouse over this tear drop shows the search term and the exact page where it is mentioned. The best view type for this function is one-page view. The term cannot be highlighted directly in the page, because the pages are displayed as pictures. This function is still on our wish list.

5 Technical Implementation

The technical development carried out by the BHL-Europe team is based around the application of best practice standards and open tools. The technical architecture of the BHL-Europe System is based on the Open Archival Information System (OAIS) reference model.

This architecture is described below with reference to the tools selected for the implementation and specific installation and configuration data as needed to for those tools to function as required.

5.1 OAIS Components.

In addition to the core OAIS elements, BHL-Europe developed two modules that fulfil the roles Provider and Consumer. The provider role is covered by a **Pre-Ingest** module and the Consumer is represented by a **Portal** module.

The **Portal** is the user interface for all BHL-Europe users. Search and retrieval is done within this component. The user is able to search for books via a simple, one field search box and an advanced search based on various metadata fields like author, publication date or title. Details of the search functionality came from the use cases developed and the results of the user requirement survey. Additional ontology services help the users to find books by enhancing taxonomic and common name. The retrieval provides the possibility to browse through the books of the result set with a book reader application and to download whole books, serials or just parts of a book like an article. The portal is Drupal based.

The **Access** component is the API that the portal and 3rd party applications use to pull information from the BHL-Europe system. It handles the provision of dissemination information packages (DIPs). DIPs are the downloadable items displayed at the Portal. Access is used to distribute the system's stored items throughout the web to other OAIS systems. This mechanism is used to support distributed storage and replication.

Data Management is used to store search and retrieval information about the items stored in Archival Storage. Data Management can be seen as the index Access will work with, when a search request is conducted. The information stored in Data Management is called Descriptive Information (DI). The main search index is based on Solr.

Archival Storage is the component where all digital representations of the books are stored. This includes the metadata and all images of the scanned pages. The data stored in Archival Storage is called Archival Information Package (AIP). By using only AIPs it is possible to recreate the whole system.

The **Ingest** component is responsible for taking Submission Information Packages (SIP), extracting the descriptive information and creating AIPs from the SIPs and submitting them to Data Management and Archival Storage, respectively. It is also responsible for validating and monitoring the status of the Ingest procedures.

Pre-Ingest represents a provider of the Open Archival Information System (OAIS). It is used by the content providers (libraries, digitization centres, etc.) to prepare the data for ingest. This is done on a directory basis and applies taxonomic ontology services to enrich the metadata. Pre-Ingest will push the AIP forward to Ingest once it is created.

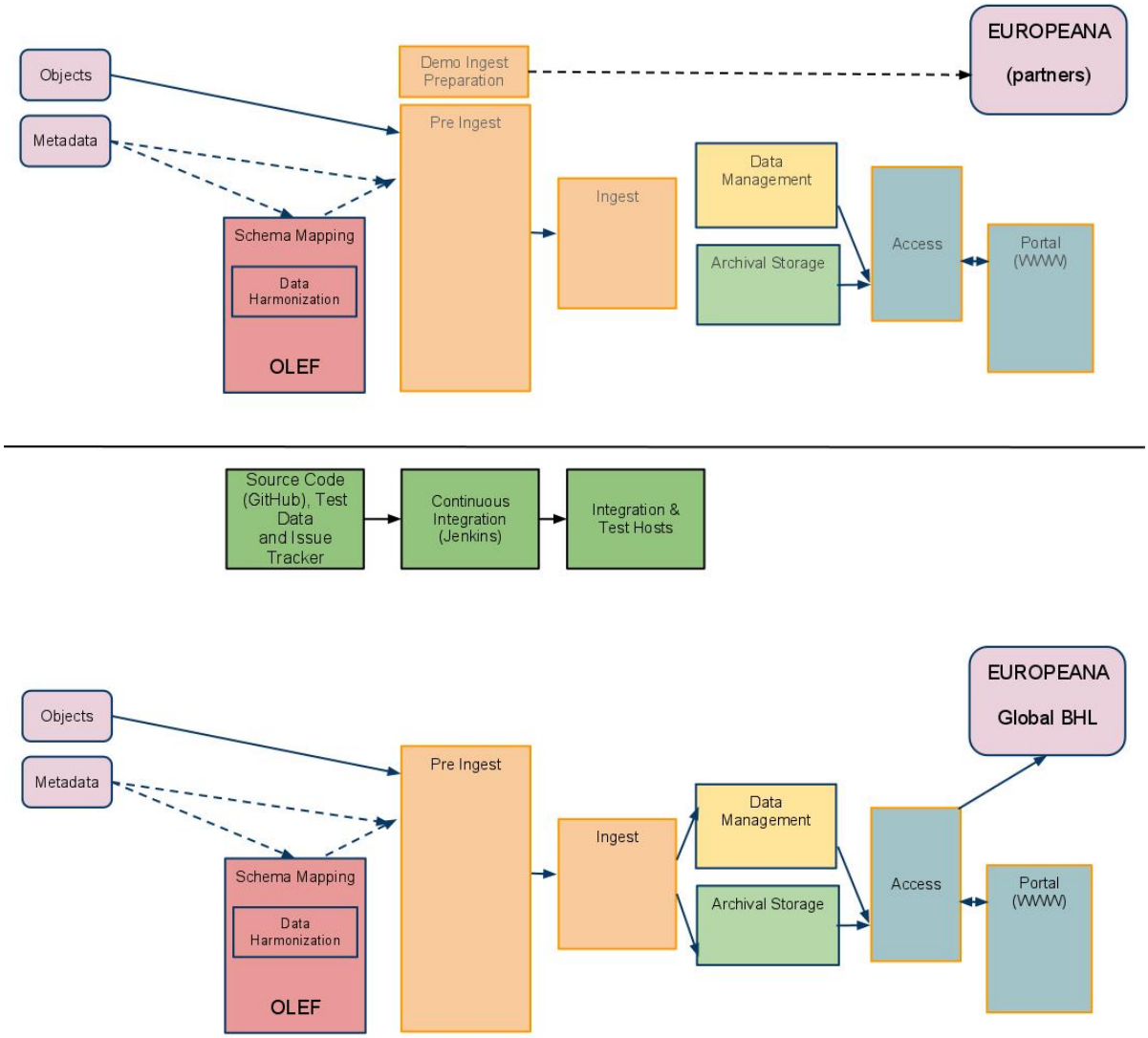

Figure 5-1: Simplified version of the BHL-Europe architecture as described in more detail in D3.5 and D3.7. The upper image illustrates the continuous integration process supported by GitHub and Jenkins; the lower image illustrates the final architecture implemented in the live BHL-Europe system.

5.2 OAIS Component Implementation

5.2.1 Pre-Ingest

The Pre-Ingest process is one of the more complex components developed for BHL-Europe. Pre-Ingest happens right before Ingest but this implementation also overlaps some aspects of the OAIS Ingest. As such, the BHL-Europe Pre-Ingest harmonisation brings together the submitted data into the final Archival Information Package, which the Ingest component delivers to Archival Storage.

The functionality required for the BHL-Europe Pre-Ingest tool were not readily available in the community toolkits available during the development of the project, thus the Pre-Ingest was developed from the ground up.

The Pre-Ingest tool integrates the BHL Schema Mapping Tool (SMT), a component designed to transform the various source document metadata to the Open Literature Exchange Format (OLEF), presenting a web interface to Content Providers to push content harmonisation forward through the various steps required to prepare for ingest (see D3.7 for details[1]).

5.2.1.1 Schema Mapping Tool Installation

 # Login as adminuser
 # Change into dev environment, opt, pre-ingest folder
adminuser@bhl-mandible:~$ cd /mnt/nfs-demeter/dev/opt/pre-ingest/

 # Create smt-cli directory
adminuser@bhl-mandible:/mnt/nfs-demeter/dev/opt/pre-ingest$ mkdir smt-cli

 # Upload latest smt-cli (binary distribution) using SFTP to /mnt/nfs-demeter/dev/opt/pre-ingest/smt-cli/

 # Test if SMT is working correctly
adminuser@bhl-mandible:/mnt/nfs-demeter/dev/opt/pre-ingest$ cd smt-cli/
adminuser@bhl-mandible:/mnt/nfs-demeter/dev/opt/pre-ingest/smt-cli$ java -jar SMT-CLI.jar

Running a test conversion
adminuser@bhl-mandible:/mnt/nfs-demeter/dev/opt/pre-ingest/smt-cli$ java -jar SMT-CLI.jar -cm 5 -m c -if /mnt/nfs-demeter/upload/webroot/RMCA/Annals_RMCA_Zoology/serie_1_fishes_reptiles_amphibians/vol_1/Boulenger_Poissons\ nouveaux/poissonsnouveau00boulRMCA.xml -of olef.xml

5.2.1.2 Pre-Ingest Tool Common tasks

To add a Content Provider Account:

- connect to database admin tool, MySQL workbench[2] is recommended.

[1] http://www.bhl-europe.eu/de/publikationen/dokumente/key-components-documented-for-output-of-d35-eg-bhl-europe-portal-ocr-demonst

[2] http://www.mysql.com/products/workbench/

- apply template for user creation from https://github.com/bhle/bhle/blob/master/pre-ingest/ingest/docs/database/create_user.sql (make sure to edit and use the correct database, set the username to match the CP upload account name and set the content home to match the provider's upload directory

5.2.2 Ingest

The Ingest component is responsible for the workflow and submission of Archival Information Packages to Data Management and Archival Storage. It is also responsible for validating the SIP and monitoring the status of the Ingest procedures. To meet this need, a batch tool was constructed, using the Spring batch[1] framework, making use of the ActiveMQ message broker[2] to handle messaging between components.

Springbatch Job Manager

The Ingest Tool is a Maven project consisting of two modules:

- ingest-core: basic framework of ingest, including batch ingest based on Spring Batch, messaging based on Spring Integration
- ingest-web: integrate ingest-core with Spring Batch Admin

common tasks

- maven install/package: (is not recommended to manually test/package/install this project because it contains a list of properties, please refer to Jenkins job for ingest)
- deploy: ln -s {ingest_web_project}/dist/ingest-web.xxx.war /var/lib/tomcat6/webapp/ingest.war
- undeploy: adminuser@bhl-int1: cd {ingest_web_project} ; mvn tomcat:undeploy -Dpath=ingest

location

- Github: https://github.com/bhle/bhle/tree/master/ingest
- web-application: /var/lib/tomcat6/webapp/ingest.war
- log: /var/log/ingest/ingest.log, /var/log/tomcat6/catalina.out

dependencies

- Pre-ingest Tool
- Fedora Repository

modifying config files

- {ingest_core_classpath}: ingest.properties
    1. The original file is placed in {ingest_web_project}/src/main/resources/ingest.properties, all environment-dependent properties are masked as ${xxxx}

[1] http://static.springsource.org/spring-batch/

[2] http://activemq.apache.org/

2. Maven filters masked properties for each compile, e.g. mvn test -Dfedora.client.username = fedoraAdmin. (Incomplete properties will cause breakdown)
3. Overriding properties with system properties in runtime is done by Spring, for more information, please refer to {ingest_core_project}/src/main/resources/META-INF/spring/env-context.xml

- {ingest_core_classpath}: logback.xml
    1. ${log.path} is filtered by Maven during build

- {ingest_classpath}: environment-test/prod.properties
    1. Platform-specific configuration files. These files contains the same list of parameters in Jenkins.
    2. Triggered by: mvn clean package -DskipTests -Denvironment=test (or -Denvironment=prod)

how to test if it's working:-

- Every build by Jenkins ensures passing all test cases.
- For testing of batch ingest, please go to http://bhl-int.nhm.ac.uk/ingest/jobs/batchIngestJob, fill job parameters with (of course the 'activation' step will fail because a000test does not exist)

    1. GUID=bhle:a000test
    2. URI=file:///{location_of_AIP_folder_path}

What if batch purge

- mvn exec:java -Dexec.mainClass="com.bhle.ingest.util.QueryAndPurge" -Dexec.args="{query}"
- {query}: e.g. *bhle*, bhle-10706-a000test*
- Purge all ingested SIP from Pre-ingest:

    1. mvn exec:java -Dexec.mainClass="com.bhle.ingest.util.QueryAndPurge" -Dexec.args="*10706*" (every object contains the GUID bankID 10706)

proxy information

- http://bhl-int.nhm.ac.uk/ingest
- http://bhl-test.nhm.ac.uk/ingest

5.2.3 Archival Storage

This entity provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing routine and special error checking, providing disaster

recovery capabilities, and providing AIPs to Access. The Fedora Commons Repository[1] was selected to implement this functionality.

5.2.3.1 Fedora

common tasks

- start: adminuser@bhl-test2: /opt/archival-storage/fedora/tomcat/bin/startup.sh
- stop: adminuser@bhl-test2: /opt/archival-storage/fedora/tomcat/bin/shutdown.sh
- Administration Browser:

http://bhl-<environment>.nhm.ac.uk/fedora/admin

  - Host: bhl-<environment>.nhm.ac.uk
  - Port: 80
  - Context: fedora

- Simple Search in Fedora:

http://bhl-<environment>.nhm.ac.uk/fedora/search

location

- FEDORA_HOME: /opt/archival-storage/fedora
- config files (/opt/archival-storage/fedora/server/config):
  - akubra-llstore.xml
    - Description: This config file describes the rules and location of Fedora Low Level Storage. Now, there are three location for placing objects, which are objectStore, shortTermDatastreamStore, longTermDatastreamStore.
  - logback.xml
    - Description: This config file describes how Fedora makes logs.
  - jaas.conf
    - Description: This config file describes the mechanism of authentication. By default, Fedora uses fedora-users.xml to authenticate. Optional choice can be Islandora Filter, LDAP.
  - fedora-users.xml
    - Description: This config file describes users and roles for authentication, and this is also one part of Fedora Security Layer, more details please refer to /opt/archival-storage/fedora/server/config/jaas.conf
  - fedora.fcfg
    - Description: This is the general config file of Fedora, including database connection, service deployment, etc.

- logs: /var/log/fedora/

- data: /mnt/nfs/<environment>/data/archival-storage/fedora

[1] http://fedora-commons.org/

- When you receive 401 HTTP status code, it may due to the following security policies.
    - /opt/archival-storage/fedora/data/fedora-xacml-policies/repository-policies
        - deny-apim-if-not-localhost.xml
        - deny-inactive-or-deleted-objects-or-datastreams-if-not-administrator.xml

dependencies

- mysql : table <environment>_as_fedora
- tomcat (embedded in Fedora)

modifying config files

- Every time a config file is modified, Fedora needs to be restarted.

how to test if it's working

- http://bhl-<environment>.nhm.ac.uk/fedora

login information if necessary

- NONE

proxy information

- http://bhl-<environment>.nhm.ac.uk/fedora

installation

- Create database for fedora in the MySQL database service.
- Download binary installer fcrepo-installer-3.4.2.jar in http://www.fedora-commons.org/software/repositoryadminuser@bhl-test2: java -jar fcrepo-installer-3.4.2.jar
- List of selection:
    - Installation type: custom
    - Fedora home directory: /opt/archival-storage/fedora
    - Fedora administrator password:
    - Fedora server host: [default: localhost]
    - Fedora application server context: [default: fedora]
    - Authentication requirement for API-A: [default: false]
    - SSL availability: false
    - Servlet engine: [default: included]
    - Tomcat home directory: [default: /opt/archival-storage/fedora/tomcat]
    - Tomcat HTTP port: [default: 8080]
    - Database: mysql
    - MySQL JDBC driver: [default: included]
    - Database username: [redacted]
    - Database password:[redacted]
    - JDBC URL: jdbc:mysql://DBServer/<environment>_as_fedorauseUnicode=true&characterEncoding=UTF-8&autoReconnect=true
    - JDBC DriverClass: [default: com.mysql.jdbc.Driver]
    - Enable FeSL AuthN: [default: true]

- Enable FeSL AuthZ: [default: false]
- Policy enforcement enabled [default: true]
- Low Level Storage: [default: akubra-fs
- Enable Resource Index: true
- Enable Messaging: true
- Deploy local services and demos: false

5.2.3.2 Fedora Low Level Storage

Often abbreviated "LLStore", the *Low Level Storage Interface* is a critical component of Fedora. It stores and provides access to the authoritative copy of all digital object XML (FOXML) and datastreams managed by a Fedora repository.

common tasks

- compile: mvn package
- location
- Github: https://github.com/bhle/bhle/tree/master/archival-storage/fedora/llstore
- **akubra-llstore.xml: /opt/archival-storage/fedora/server/config**

dependencies

- Fedora

modifying config files

1. In akubra-llstore.xml, modify the paths for objectStore, shortTermDataStreamStore and longTermDataStreamStore
2. Purge all objects in Fedora Repository, and restart Fedora

how to test if it's working

- Check whether the corresponding folder in akubra-llstore.xml (normally,/mnt/nfs/test/data/archival-storage/fedora/data) contains data ingested.

login information if necessary

- NONE

proxy information

- NONE

Installation

1. Purge all objects in Fedora Repository, then shut down the server;
2. Place akubra-mux-0.3.jar and bhle-llstore-0.0.1.jar in ${FEDORA_HOME}/tomcat/webapps/fedora/WEB-INF/lib;
3. Replace akubra-llstore.xml with the one in the install package, and modify the store paths and DataStream IDs according to your needs;
4. Restart the server.

How it works

A subclass of org.akubraproject.mux.AbstractMuxConnection overrides the getStore method to provide BlobStore according to the keywords of DataStream IDs in akubra-llstore.xml.

And the filesystem storage is reused from akubra-fs (simple filesystem implementatio) and akubra-map (wraps an existing BlobStore to provide a blob id mapping layer) without any modification. Therefore, all the path mappings for objects and datastreams are still based on MD5 mapping**.**

5.2.4 Data Management

Data Management is the OAIS entity that contains the services and functions for populating, maintaining, and accessing a wide variety of information. For BHL-Europe this includes three main components.

Searching for BHL-Europe data is implemented by the Apache Solr[1] search platform. This was selected to enable the full-text search, hit highlighting, faceted search and other features required for the BHL-Europe portal, and to enable scaling as required for service growth. Solr is seeded with metadata harvested from the data archived in Fedora by the Fedora Generic Search Service (Gsearch)[2]. Islandora[3] has been selected to provide the asset management services for ongoing maintainance of the BHL-Europe archived assets.

5.2.4.1 Solr

common tasks

- start: adminuser@bhl-int1: sudo /etc/init.d/jetty start
- stop: adminuser@bhl-int1: sudo /etc/init.d/jetty stop
- status: adminuser@bhl-int1: sudo /etc/init.d/jetty status
- reindexing solr:
  - option 1) manually via shell script

```
HOST="bhl-int1"
FOXML_FOLDER=/mnt/nfs/dev/data/archival-
storage/fedora/data/objectStore
WEBAPP="fedoragsearch"
# fedora on 8080, solr on 8983
PORT="8080"

#
# clear the solr index
#
ssh $HOST "/usr/bin/sudo /etc/init.d/jetty stop"
rm -r /mnt/nfs/dev/data/data-management/solr/core/data/*
ssh $HOST "/usr/bin/sudo /etc/init.d/jetty start"
#
# trigger indexing
#
```

[1] http://lucene.apache.org/solr/

[2] https://wiki.duraspace.org/display/FCSVCS/Generic+Search+Service+2.2

[3] http://islandora.ca/

#

curl --user fgsAdmin:$GS_PASSWORD "http://$HOST:$PORT/$WEBAPP/rest?operation=updateIndex&action=fromFoxmlFiles&value=${FOXML_FOLDER}"

- option 2) run the script in /mnt/nfs/test/opt/scripts/post2solrByGsearch/post2solr.sh
- option 3) start jenkins job: http://bhle-dev-1.nhm.ac.uk/job/data-management/

location

- Github: https://github.com/bhle/bhle/tree/master/data-management/solr
- web-application: /mnt/nfs/dev/data/data-management/solr
- application-container:/usr/share/jetty/
- lib: /mnt/nfs/dev/opt/data-management/solr/lib
    1. /mnt/nfs/dev/data/data-management/solr/solr.xml
    2. /mnt/nfs/dev/opt/data-management/solr/core/conf (symlink to /home/adminuser/dev/opt/data-management/solr/core/conf)
    3. /etc/default/jetty
- logs: /var/log/jetty/${TIMESTAMP}.stderrout.log
- data: /mnt/nfs/dev/data/data-management/solr/core/data

dependencies

- gsearch: solr data home
- portal: solr URL

modifying config files

1. solr.xml: restart jetty
2. core/conf: reindex
3. /etc/default/jetty: restart jetty; if data home has changed: change fgsearch index.properties.

The configuration of solr will effect Gsearch, please refer to /opt/archival-storage/fedora/tomcat/webapps/fedoragsearch/FgsConfig/fgsconfig-basic.properties

after modifying it, to update the configuration of gsearch run:

ant -f fgsconfig-basic.xml

how to test if it's working

only on localhost - http://bhl-int1:8983/solr/

proxy information

- http://bhl-test.nhm.ac.uk/solr

5.2.4.2 Jetty

Currently solr is installed in Jetty:
solr home: /mnt/nfs/<environment>/data/data-management/solr
core: used by Gsearch

install jetty:
apt-get install jetty libjetty-extra

configure jetty:
configuration in /etc/default/jetty

- set port to 8983
- allow any host to connect
- set JVM ARGS: -Dsolr.solr.home=/mnt/nfs/<environment>/data/data-management/solr

copy the solr war to /usr/lib/jetty/webapp/:

cp /mnt/nfs-demeter/apache-solr-3.3.0/dist/apache-solr*.war /usr/lib/jetty/webapp/solr.war

restart Jetty:
sudo /etc/init.d/jetty restart

5.2.4.3 GSearch

common tasks

- start: adminuser@bhl-test2: /opt/archival-storage/fedora/tomcat/bin/startup.sh
- stop: adminuser@bhl-test2: /opt/archival-storage/fedora/tomcat/bin/shutdown.sh
- reindex: http://bhl-<environment>/fedoragsearch/rest -> updateIndex -> updateIndex FromFoxmlFiles
- browse index: http://bhl-<environment>/fedoragsearch/rest -> browseIndex

location

- Github: https://github.com/bhle/bhle/tree/master/data-management/gsearch
- web-application: /opt/archival-storage/fedora/tomcat/webapps/fedoragsearch
- application-container: /opt/archival-storage/fedora/tomcat
- config files:
    1. /opt/data-management/gsearch/fgsconfig-basic.properties
- logs: /var/log/gsearch/fedoragsearch.daily.log

dependencies

- Fedora Commons
- Solr
- Ant

modifying config files

1. fgsconfig-basic.properties:
   adminuser@bhl-test2: ant -f /opt/data-management/gsearch/fgsconfig-basic.xml

2. It is recommended to delete all index in Solr (/mnt/nfs/test/data/data-management/solr/core/data) before reindexing.

how to test if it's working

- http://bhl-t<environment>/fedoragsearch/rest -> updateIndex, if there is no error message

Installation

- Version 2.3 from https://github.com/fcrepo/gsearch (git clone)
- Follow the instructions in FedoraGenericSearch/src/html/fedoragsearch-doc.html
- Move config files to data-management folder:

adminuser@bhl-test2: mv -r /opt/archival-storage/fedora/tomcat/webapps/fedoragsearch/FgsConfig/* /opt/data-management/gsearch

5.2.4.4 Islandora

common tasks

- start: adminuser@bhl-int1: sudo /etc/init.d/apache2 start
- stop: adminuser@bhl-int1: sudo /etc/init.d/apache2 stop
- restart: adminuser@bhl-int1: sudo /etc/init.d/apache2 restart

location

- web-application: /mnt/nfs/dev/data/data-management/drupal
- modules: /mnt/nfs/dev/data/data-management/drupal/sites/default/modules/

dependencies

- Fedora
- BookReader
- BookToolbox

modifying config files

- NONE

how to test if it's working

- connect a web browser to bhl-<environment>.nhm.ac.uk/datamanagement

Installation

- Download Islandora 11.2.0 in http://islandora.ca/11-2
- Required modules:
  - Islandora
  - Solr Search
  - XML Forms
  - Content Model Forms
  - Objective Forms
  - php_lib
  - Tabs
  - Book Solution Pack
- Unzip all modules and copy to /opt/data-management/drupal/sites/default/modules/
- Enable modules in Drupal Modules

- Modify /opt/data-management/drupal/sites/default/modules/islandora_solution_pack_book/book.ini to adapt Book Reader
    - Line 48: $viewer_url = variable_get('fedora_base_url', '') . '/get/' . $this->pid . '/islandora:viewerSdef/getViewer' . $qs;
    - Change to: $viewer_url = '/fedora/objects/' . $this->pid . '/methods/bhle-service:bookSdef/bookreader?ui=full#mode/2up';

5.2.5 Access

The OAIS Access entity contains the services and functions which make the archival information holdings and related services visible to Consumers. The components implemented for the BHL-Europe Access service include the BookReader, based on the Internet Archive BookReader[1], the JPEG 2000 Image Server Djatoka[2], used to render scalable images, and the Fedora OAI Provider service (based on ProAI [3]) for provision of data via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). These components are served from an apache tomcat[4] container.

5.2.5.1 Access Tool

Access is a Maven project consisting of six modules,

- access-core: basic framework of access, including batch derivative generated based on Spring Batch, messaging based on Spring Integration
- access-static: helper classes for Apache static URLs and Djatoka URLs
- access-bookreader: generates thumbnails and bookreader.json from BookReader (based on IA BookReader)
- access-download: generates full package of PDF, JPEG and OCR, also provides offline download
- access-oai: OAI-PMH provider based on Proai
- access-web: integrates all modules and provides RESTfull entries for each serivce

common tasks

- maven install/package: (is not recommended to manually test/package/install this project because it contains a list of properties, please refer to Jenkins job for ingest)
- deploy:
    - ln -s {access_web_project}/dist/access-web.xxx.war /var/lib/tomcat6/webapp/access.war
    - ln -s {access_oai_project}/dist/access-web.xxx.war /var/lib/tomcat6/webapp/access-oai.war
- undeploy: adminuser@bhl-int1: cd {ingest_web_project} ; mvn tomcat:undeploy -Dpath=access

location

- Github: https://github.com/bhle/bhle/tree/master/access

[1] http://openlibrary.org/dev/docs/bookreader
[2] http://djatoka.sourceforge.net/
[3] http://proai.sourceforge.net/
[4] http://tomcat.apache.org/

- web-application: /var/lib/tomcat6/webapp/access.war, /var/lib/tomcat6/webapp/access-oai.war
- log: /var/log/ingest/access.log, /var/log/ingest/access-oai.log, /var/log/tomcat6/catalina.out

dependencies
- Fedora Repository
- Djatoka
- ImageMagick

modifying config files
- Each module holds a property file.
  - access-core: access.properties, logback.xml
  - access-download: access-download.properties
  - access-oai: proai.properties, log4j.properties
  - access-static: access-static.properties
  - access-web: batch-default.properties
- {access_classpath}: environment-test/prod.properties
  - Platform-specific configuration files. These files contains the same list of parameters in Jenkins.
  - Triggered by: mvn clean package -DskipTests -Denvironment=test (or -Denvironment=prod)

how to test if it's working
- Every build by Jenkins ensures passing all test cases.
- BookReader: access/stream/a000test
- Download: Full package generation of PDF, JPEG and OCR; Offline download in BookReader
- Batch derivative generation: access/batch-admin/jobs/generateDerivatives, fill job parameters with:
  - PID=(choose a PID of a 'book' object)

proxy information
1. http://bhl-int.nhm.ac.uk/access
2. http://bhl-test.nhm.ac.uk/access

5.2.5.2 Static Files

Introduction
In order to increase performance pre-generated derivatives will be cached and served via an http server such as varnish.

URLs
http://<environment URL>/static/<guid>/<file>

Files For the bookreader:
http://<environment URL>/static/<guid>/jp2/<guid>_<page number>.jpg

Examples:

http://bhl-test.nhm.ac.uk/static/a0hhmgs6/a0hhmgs6_mods.xml <-- test environment
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_mods.xml <-- production environment
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_olef.xml
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_bibtex.bib
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_endnote.end
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_thumbnail.jpg <-- more details below
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_full_pdf.pdf
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_full_jpg.zip
http://www.bhl-europe.eu/static/a0hhmgs6/a0hhmgs6_full_ocr.txt

Thumbnails:
Thumbnails are not immediately generated when the item is ingested, which means we should request the generation of thumbnails at the first time. From the Portal side, this should be broken into two steps:
1. Hit http://www.bhl-europe.eu/static/<guid>/<guid>_thumbnail.jpg to see whether the thumbnail is available.
2. If not (return 404 NOT FOUND), then try http://www.bhl-europe.eu/access/ download/ <guid>/thumbnail

File Storage
The files will be stored on the fileshare. There is no limit to the number of folders in a folder in GPFS however in order to avoid too many sub-folders in one container we store the files in a pairtree.

Examples
/mnt/nfs/production/access/items/a0/hh/mg/s6/a0hhmgs6/ <--> http://www.bhl-europe.eu/access/items/a0hhmgs6

Files For the bookreader:
/mnt/nfs/production/access/items/a0/hh/mg/s6/a0hhmgs6/jpg/a0hhmgs6_00001.jpg

5.2.5.3 Access Tomcat

common tasks

- start: adminuser@bhl-test2:/$ sudo etc/init.d/tomcat6 start
- stop: adminuser@bhl-test2:/$ sudo etc/init.d/tomcat6 stop
- status: adminuser@bhl-test2:/$ sudo etc/init.d/tomcat6 status

how to test if it's working

- tomcat: http://bhl-test.nhm.ac.uk/access
- tomcat manager: http://bhl-test.nhm.ac.uk/access/manager/html

location

- web-application: /var/lib/tomcat6/webapps
- application-container:/usr/share/tomcat6/
- config files:
  - /etc/tomcat6/server.xml
  - /etc/tomcat6/tomcat-users.xml
  - /etc/default/tomcat6

- /usr/share/tomcat6/bin/catalina.sh
- logs:
  - tomcat: /var/log/tomcat6/catalina.out

dependencies
- NA

modifying config files
1. after modifying any config file: restart tomcat

login information if necessary
- tomcat manager: http://bhl-test.nhm.ac.uk/access/manager/html
  username=REDACTED

proxy information
- configured in the site /etc/apache2/sites-enabled/bhl-test
- http://bhl-test.nhm.ac.uk/access (tomcat)
  - ProxyPass /access http://bhl-test2.nhm.ac.uk:8090
    ProxyPassReverse /access http://bhl-test2.nhm.ac.uk:8090

Installation
- Tomcat
- apt-get install tomcat6 tomcat6-admin
- configuration in /etc/tomcat6/server.xml
  - Change http port from 8080 to 8090 to avoid conflict with fedora-tomcat
    <Connector port=”8090″ protocol=”HTTP/1.1″
  - Change shutdown port from 8005 to 8095 to avoid conflict with fedora-tomcat
    <Server port="8095" shutdown="SHUTDOWN">
- configuration in /etc/tomcat6/tomcat-users.xml
  - Add <role rolename="manager"/>
    < user username="REDACTED" password="REDACTED"
    roles="manager"/>

5.2.5.4 Djakota

Djatoka is a web application running on tomcat installed via apt-get install tomcat6.

common tasks
- start: adminuser@bhl-test2:/$ sudo etc/init.d/tomcat6 start
- stop: adminuser@bhl-test2:/$ sudo etc/init.d/tomcat6 stop
- status: adminuser@bhl-test2:/$ sudo etc/init.d/tomcat6 status

how to test if it's working
- djatoka: http://bhl-test.nhm.ac.uk/adore-djatoka/
  - for BaseURL: http://bhl-test.nhm.ac.uk/adore-djatoka/resolver
  - click on "Open image in new window" and an image of a map of Arizona and New Mexico should appear

location
- web-application: /var/lib/tomcat6/webapps
- application-container:/usr/share/tomcat6/
- config files:

- /opt/access/adore-djatoka/bin/env.sh
- /usr/share/tomcat6/bin/catalina.sh
- djatoka: /var/log/djatoka/djatoka.log

dependencies

- Tomcat (Access)
- modifying config files
  after modifying any config file: restart tomcat

login information if necessary

- NA

proxy information

- configuration in /etc/apache2/sites-enabled/bhl-test
  - ProxyPass /adore-djatoka http://bhl-test2.nhm.ac.uk:8090/adore-djatoka
    ProxyPassReverse /adore-djatoka http://bhl-test2.nhm.ac.uk:8090/adore-djatoka

Installation

- Djatoka
- Version 1.1 From http://sourceforge.net/projects/djatoka/
- Download Djatoka from
  http://sourceforge.net/projects/djatoka/files/djatoka/1.1/adore-djatoka-1.1.tar.gz
- untar to /opt/access/adore-djatoka/
- modify /opt/access/adore-djatoka/bin/env.sh
  - DJATOKA_HOME=/opt/access/adore-djatoka
  - KAKADU_HOME=$DJATOKA_HOME/bin/$PLATFORM
    export KAKADU_HOME
    JAVA_OPTS="$JAVA_OPTS -Xms256m -Xmx512m -XX:MaxPermSize=256m -Djava.awt.headless=true -Dkakadu.home=$KAKADU_HOME -Djava.library.path=$LIBPATH/$PLATFORM $KAKADU_LIBRARY_PATH"
    JAVA_OPTS="$JAVA_OPTS -Dlog.dir=/var/log/djatoka"
    export JAVA_OPTS
- modify /usr/share/tomcat6/bin/catalina.sh to call env.sh
  - add
    # ---
    . /opt/access/adore-djatoka/bin/env.sh
    # OS specific support. $var _must_ be set to either true or false.
- Deploy adore-djatoka-1.1/dist/adore-djatoka.war to /var/lib/tomcat6/webapps
- Set JAVA_HOME to sun JDK by modifying /etc/default/tomcat6
  - #JAVA_HOME=/usr/lib/jvm/openjdk-6-jdk
    JAVA_HOME=/usr/lib/jvm/java-6-sun

5.2.6 Portal

The BHL-Europe Portal is the primary end user consumer interface to the BHL-Europe system. The Portal is based on the Drupal CMS[1]

location

- Github: https://github.com/bhle/bhle/tree/master/portal/drupal

The GitHub repository holds both the custom BHL-Europe Drupal modules, plus database to enable auto-deployment via Jenkins.

[1] http://drupal.org/

5.3 Common Services

Separate from the OAIS components, the BHL-Europe Implementation requires a number of common services to support functionality. These include the common database engine, MySQL[1], source code versioning and issue tracking is delivered using GitHub[2], with the continuous integration tool Jenkins[3] used to automate deployment of code from the GitHub repository to the BHL-Europe integration environment. ActiveMQ was selected to manage message queuing between components.

5.3.1 MySQL Database

MySQL is installed on the common service host bhl-db1

Installation and maintenance is via apt-get

common tasks

- start: adminuser@bhl-db1:/$ sudo etc/init.d/mysq lstart
- stop: adminuser@bhl-db1:/$ sudo etc/init.d/mysql stop
- status: adminuser@bhl-db1:/$ sudo etc/init.d/mysql status
- create a new database:

mysql> create database `environment_component_toolname`;

- grant rights on a database to enable access from a remote host

mysql> grant all on `environment_component_toolname`.* to `REDACTED`@`IP.RANGE`;

how to test if it's working

- from the command line:

mysql> show databases;

- from an application (eg drupal):

Configure to connect to the database service bhl-db1. as the application user to confirm the application connectivity.

5.3.2 Active MQ & Stomp

Common tasks:

- start:adminuser@bhl-mandible:$ sudo /etc/init.d/activemq start
- stop:adminuser@bhl-mandible:$ sudo /etc/init.d/activemq stop
- restart:adminuser@bhl-mandible:$ sudo /etc/init.d/activemq restart

[1] http://www.mysql.com

[2] https://github.com/bhle

[3] http://jenkins-ci.org/

Manage mutiple activemq instances:

- /opt/activemq/apache-activemq-5.5.1/instance-<instance name>/bin/:$./instance-<instance name> start/stop/restart

Port information:
There are mutiple activemq instances running , providing JMS service to three versions of BHL: integration, test, production.
Each version uses different ports to communicate and admin:

Instance-integration:

- JMS message and fedora communication port:61616
- Stomp protocol port: 61613
- Admin port: 8161

Instance-production:

- JMS message and fedora communication port:61617
- Stomp protocol port: 61614
- Admin port: 8162

Instance-test:

- JMS message and fedora communication port:61618
- Stomp protocol port: 61615
- Admin port: 8163

Location:

- **Installation location:**/opt/activemq
- **Backup data location:**/opt/activemq/apache-activemq-5.5.1/data
- **Log location**:/opt/activemq/apache-activemq-5.5.1/data/activemq.log
- **Config files**:/opt/activemq/apache-activemq-5.5.1/conf
    - activemq.xml: This is the basic config file for activemq.The things can be configured including ports, transport connectors, network connectors, persistence providers & locations, etc.
    - log4j.properties: log configured file.

The backup data, log and config files are located under:
/opt/activemq/apache-activemq-5.5.1/instance-<integration/production/test>

Dependencies:

- JDK
- Fedora (If it needs to be used as a bridge to send the messages from fedora)

Installation:
ActiveMQ is installed on the common service host bhl-mandible.
ActiveMQ isn't supported by the debian dist, it need to be installed manually. Now we are using version 5.5.1

Install location:/opt/activemq

Inststallation steps:

- wget http://labs.mop.com/apache-mirror//activemq/apache-activemq/5.5.1/apache-activemq-5.5.1-bin.tar.gz
- tar xzvf apache-activemq-5.5.1-bin.tar.gz
- adding startup:
  - $ ln -sf /opt/activemq/apache-activemq-5.5.1/bin/activemq /etc/init.d/
  - $ update-rc.d activemq defaults
- build default config file: $ /etc/init.d/activemq setup /etc/default/activemq

Test:

Using the following command to check whether port 61616 is up:

1. netstat -an|grep 61616

- **Setup Stomp**
- Open activemq.xml, add a connector to the broker using the stomp URL:

<transportConnectors>
<transportConnector name="stomp" uri="stomp://0.0.0.0:61613"/>
</transportConnectors>

Test running:

- Download php client from here:http://stomp.fusesource.org/download.html
- Unzip the package, put the Stomp.php and Stomp file into Examples
- Change the destination in first.php

$con = new Stomp("tcp://bhl-mandible.nhm.ac.uk:61613");

- Run first.php by using command: php first.php, you can see:

$Sent message with body 'test'
$Received message with body 'test'

To Setup mutiple instances

- cd /opt/activemq/apache-activemq-5.5.1/
- bin/activemq create instance-<instance name>
- cd instance-<instance name>
- chmod 755 bin/instance-<instance name>
- bin/instance-<instance name> setup ~/.activemqrc-instance-<instance name>
- change the port information in activemq.xml and jetty.xml under /opt/activemq/apache-activemq-5.5.1/instance-<instance name>/conf

Configure store-and-forward ActiveMQ message broker bridge between Fedora

This configuration allows Fedora to use the ActiveMQ to send update messages.

Steps:

- Create the Fedora activemq.xml file, put it in FEDORA_HOME/server/config
- <beans xmlns:amq="http://activemq.apache.org/schema/core">
- <!-- ActiveMQ JMS Broker configuration -->
- <amq:broker id="broker" useShutdownHook="false">
- <amq:managementContext>
- <amq:managementContext connectorPort="1093" createConnector="false"/>
- </amq:managementContext>

- <!-- Your remote broker, configured with failover -->
- <amq:networkConnectors>
- <amq:networkConnector name="fedorabridge" dynamicOnly="true" uri="static:(failover:(tcp://0.0.0.0:61616))"/>
- </amq:networkConnectors>
- <!-- The directory where Fedora will store the ActiveMQ data -->
- <amq:persistenceAdapter>
- <amq:amqPersistenceAdapter directory="file:/mnt/nfs/dev/opt/archival-storage/fedora/data/activemq-data/localhost/amq"/>
- </amq:persistenceAdapter>
- </amq:broker>
- <!-- Set this to prevent objects from being serialized when
- passed along to your embedded broker; saves some overhead processing -->
- <bean id="jmsConnectionFactory" class="org.apache.activemq.ActiveMQConnectionFactory">
- <property name="objectMessageSerializationDefered" value="false"/>
- </bean>

</beans>

- Change the java.naming.provider.url parameter in fedora.fcfg
- <param name="java.naming.provider.url" value="vm://localhost?brokerConfig=xbean:file:/mnt/nfs/dev/opt/archival-storage/fedora/server/config/activemq.xml"/>
- Drop the jars xbean-spring-3.4.3.jar and spring-context-2.5.6.jar files into the Fedora webapp WEB-INF/lib directory
- restart Fedora

Test running:

- Run the test case provideing here to establish the connection with the ActiveMQ server

```
public class TestActiveMQ implements MessagingListener{

 private String textPath = "./";

 MessagingClient messagingClient;
 public void start() throws MessagingException {
 Properties properties = new Properties();
 properties.setProperty(Context.INITIAL CONTEXT FACTORY,
 "org.apache.activemq.jndi.ActiveMQInitialContextFactory");

 properties.setProperty(Context.PROVIDER_URL, "tcp://activemqhost.domain:61616");
 properties.setProperty(JMSManager.CONNECTION_FACTORY_NAME, "ConnectionFactory");
 properties.setProperty("topic.fedora", "fedora.apim.*");
 messagingClient = new JmsMessagingClient("example1", this, properties, false);
 messagingClient.start();
```

```
}

public void stop() throws MessagingException {

messagingClient.stop(false);
}

public void onMessage(String clientId, Message message) {

String messageText = "";
try {
messageText = ((TextMessage)message).getText();
} catch(JMSException e) {
System.err.println("Error retrieving message text " + e.getMessage());
}
System.out.println("Message received: " + messageText + " from client " + clientId);
writeText("receiveMessage", "Message received: " + messageText + " from client " + clientId);
}

public void writeText(String textname,String date){

File filePath=new File(textPath);
if(!filePath.exists()){
filePath.mkdirs();
}
try {
FileWriter fw =new FileWriter(textPath+File.separator+textname);
fw.write(date);

if(fw!=null)
fw.close();
} catch (IOException e) {

e.printStackTrace();
}



}

public static void main(String[] argus){
TestActiveQ acq = new TestActiveQ();
try{
acq.start();
```

```
    }
    catch(Exception e){
    }

     }

    }
```

- The setup works if it prompts "connectionTo JMS- connected", and each time Fedora has update, this test client can receive the update message

5.3.3 GitHub

BHL Europe uses the GitHub service for configuration management of the project.

The project can be found here: https://github.com/bhle

To commit to the project you must have commiter's privileges.

The bhle repository contains directories for all of the components (pre-ingest, ingest, archival storage, portal, etc.).

This guide will go through the portal setup since it is the main shared development environment.
N.B. The following instructions only guide you through the normal workflow, for more flexible usage of git, please refer to http://www.kernel.org/pub/software/scm/git/docs/user-manual.html

- **Prepare your github account**
- Add your SSH key to GitHub: http://help.github.com/linux-set-up-git/
- **Getting the project**

cd to the directory where you want to clone the repository on your local machine for example /home/user then execute the command

$ git clone git@github.com:bhle/bhle

Go into the newly created directory

$ cd bhle

Import bhle/portal/drupal/database/bhle-portal.sql to create your instance of the portal drupal database. This script will create a database called "bhle-portal"

$ mysql -u rootusername -p bhle-portal < portal/drupal/database/bhle-portal.sql

Copy the reference settings file bhle/portal/drupal/www/sites/default/ref.settings.php to settings.php;

$ cp portal/drupal/www/sites/default/ref.settings.php
portal/drupal/www/sites/default/settings.php
In the settings.php modify the $database array with your database information.
In Apache create a virtual host or use a symbolic link for example named "portal" to the directory "portal/drupal/www", e.g.

Alias /portal "/home/bhle/portal/drupal/www"
<Directory "/home/bhle/portal/drupal/www">
Options Indexes FollowSymLinks
AllowOverride All
Order allow,deny
Allow from all
</Directory>
then restart the Apache server
Now you can test the portal with your web browser for example

- http://localhost/portal

- **Contribute to the project**

1. Every time you are ready to contribute to the project
$ git pull
$ git checkout -b your-own-branch-name-here

Now remember to update your mysql database from the database file pulled from GitHub!

Nw you are working in your own branch, which means the local master branch is unaffected. You can check if you are in the branch by

$ git status
This would print out the name of the branch in the first line and other information.

2. Contribute all want you to this project, and test your contribution until it is good enough to be pushed back to the repository

3. Dump your database (change the user name "root" accordingly to your local setup). Make sure you are in the root folder of your bhle git repository!

$ mysqldump -u root -p --ignore-table="bhle-portal.watchdog, bhle-portal.sessions" bhle-portal > portal/drupal/database/bhle-portal.sql
4. Add additional files to index

$ git add .

5. Commit your contribution
$ git commit -m "your commit message here"
If git is not committing your deletions or other changes you can alternatively use

$ git commit -a -m "your commit message here"

6. Synchronize the repository (within your own branch) again, in case any others update the remote master branch

$ git pull origin master
if pull successfully , go to step 7

if there are conflicts, you can do two things:
1) Decide not to merge. You can roll back to the recent commit (all modification will be lost).

$ git reset --hard HEAD
2) Resolve the conflicts (**recommanded**). The most easy way to do it is by using mergetool (depends on your favorite tool for merge)

$ git mergetool
$ git commit -m "your merge message here"

7. Switch to the local branch
$ git checkout master

8. Merge your own branch into the master
$ git merge your-own-branch-name-here

9. Push all your contribution back to the remote repository
$ git push

10. If you need to discard your own branch, type
$ git branch -d your-own-branch-name-here

Working behind a firewall

Please see https://github.com/blog/642-smart-http-support for more information.
configure the proxy:
$ git config --global http.proxy <PROXY-HOST>:<PORT>

clone via http
$ git clone http://git@github.com/bhle/bhle.git

switch your clone to http
replace ${user-name} by your git user name.
$ git remote set-url origin http://${user-name}@github.com/bhle/bhle.git

5.3.4 Jenkins

To ensures the BHL-Europe integration builds are reproducible, we use the Jenkins Continuous Integration service. Jenkins deploys each time a **clean build**, which is built fully from GitHub Source Code Control. All code for the deployment including third-party jars, build scripts, release notes, etc. must be checked into GitHub.

Login:
use can use the usual admin account to log into jenkins.

We install Jenkins through apt-get.

Installation

```
wget -q -O - http://pkg.jenkins-ci.org/debian/jenkins-ci.org.key | sudo
apt-key add -
sudo echo "deb http://pkg.jenkins-ci.org/debian binary/" >
/etc/apt/sources.list.d/jenkins.list
sudo aptitude update
sudo aptitude install jenkins
```

What does this package do?

- Jenkins will be launched as a daemon up on start. See /etc/init.d/jenkins for more details.
- The 'bhladmin' user is configured to run this service.
- Log file will be placed in /var/log/jenkins/jenkins.log. Check this file if you are troubleshooting Jenkins.
- /etc/default/jenkins will capture configuration parameters for the launch.
- By default, Jenkins listen on port 8082. Access this port with your browser to start configuration.

Running Jenkins behind Apache

mod_proxy
mod_proxy works by making Apache perform "reverse proxy" — when a request arrives for certain URLs, Apache becomes a proxy and further forward that request to Jenkins, then it forwards the response back to the client.
The following Apache modules must be installed:

```
a2enmod proxy
a2enmod proxy_http
```

```
ProxyPass         /  http://jenkins-host:8082/jenkins
ProxyPassReverse  /  http://jenkins-host:8082/jenkins
```

```
ProxyRequests     Off
```

```
# Local reverse proxy authorization override
# Most unix distribution deny proxy by default (ie /etc/apache2/mods-
enabled/proxy.conf in Ubuntu)
<Proxy http://bhl-mandible.nhm.ac.uk:8081/jenkins*>
  Order deny,allow
  Allow from all
</Proxy>
```

This assumes that you run Jenkins on port 8082.

Reference:
https://wiki.jenkins-ci.org/display/JENKINS/Installing+Jenkins+on+Ubuntu
https://wiki.jenkins-ci.org/display/JENKINS/Running+Jenkins+behind+Apache

6 BHL–Europe Source Code Licencing

All BHL-Europe source code and development is covered by the Modified BSD Licence[1].

This version allows unlimited redistribution for any purpose as long as its copyright notices and the license's disclaimers of warranty are maintained. The license also contains a clause restricting use of the names of contributors for endorsement of a derived work without specific permission.

```
Copyright (c) <year>, <copyright holder>
All rights reserved.

Redistribution and use in source and binary forms, with or without
modification, are permitted provided that the following conditions are met:
    * Redistributions of source code must retain the above copyright
      notice, this list of conditions and the following disclaimer.
    * Redistributions in binary form must reproduce the above copyright
      notice, this list of conditions and the following disclaimer in the
      documentation and/or other materials provided with the distribution.
    * Neither the name of the <organization> nor the
      names of its contributors may be used to endorse or promote products
      derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS"
AND NY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
IMPLIED ARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE
ARE DISCLAIMED. IN NO EVENT SHALL <COPYRIGHT HOLDER> BE LIABLE FOR ANY
DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES
(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR
SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER
CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT
LIABILITY, OR TORT(INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY
OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH
DAMAGE.
```

[1] http://en.wikipedia.org/wiki/BSD_licenses#3-clause_license_.28.22New_BSD_License.22_or_.22Modified_BSD_License.22.29)

7 Appendix

7.1 User Guide for the BHL-Europe Pre-Ingest Tool

7.1.1 Purpose

This document aims to give end user guidance how to process and ingest data to BHL-Europe. The Pre-Ingest tool described allows BHL-Europe services to complete preparation of uploaded content and bring it to a state ready for ingestion into the archive store; ready for presentation via the BHL-Europe portal. The target readers of this document are content providers.

7.1.2 Preliminary Steps

Some preliminary steps must be followed before data can be processed by the Pre-Ingest tool. These steps are documented more fully in the BHL-Europe file submission guidelines, mentioned here for completeness.

7.1.2.1 Prepare your content

The initial step before any document can be processed by BHL-Europe is for the content to be scanned, metadata prepared and file naming applied. Content must confirm to the specification detailed in the BHL-Europe file submission guidelines before it can be processed by the Pre-Ingest tool. If the content upload structure or data does not conform to these guidelines, then you will need to liaise with the BHL Europe team to determine what harmonisation scripts/steps need to be carried out to bring the content to a state of ingest readiness.

7.1.2.2 Content Upload

Image and metadata content must be uploaded to the BHL-Europe server according to the instructions provided in the File Submission Guidelines. The folder structure that this content is uploaded into will normally be reflected during the Pre-Ingest selection and activation step, unless significant harmonisation is required to further prepare those data.

When your content format is confirmed good for ingest, an account will be configured and login details will be provided by the BHL-Europe team to enable access to the Pre-Ingest tool.

7.1.3 Pre-Ingest Setup

7.1.3.1 Log on to the tool

The Pre-Ingest tool is accessible via the URL: http://bhl.nhm.ac.uk/preingest.

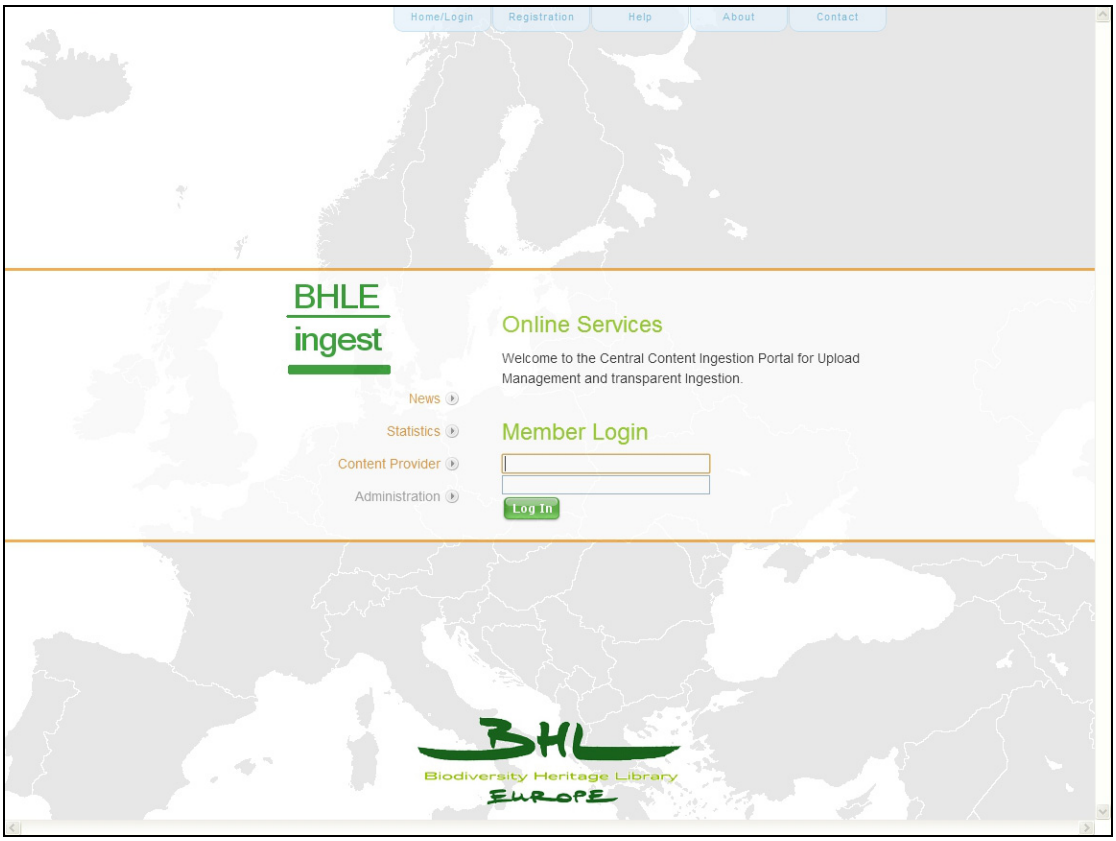

Figure 7-1: Pre-Ingest Logon

On logon, the main interface is displayed.

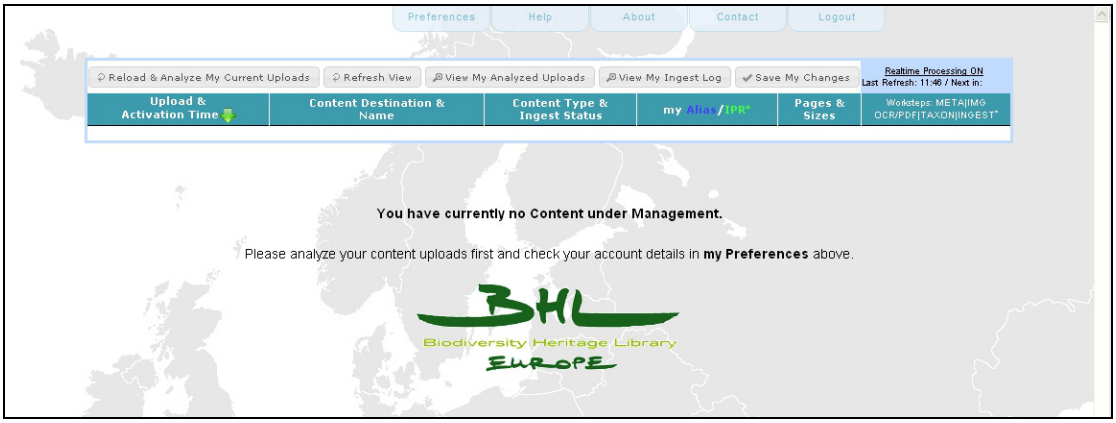

Figure 7-2: Main interface

7.1.3.2 Confirm your preference settings

On first connection, select the Preferences option from the top toolbar to ensure that all defaults are correct. Some of these are user editable, the others are configured by the BHL-Europe administrator.

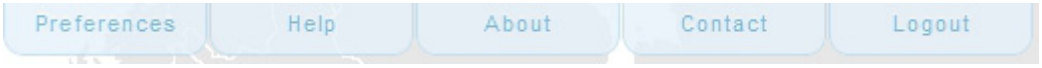

Figure 7-3: Toolbar

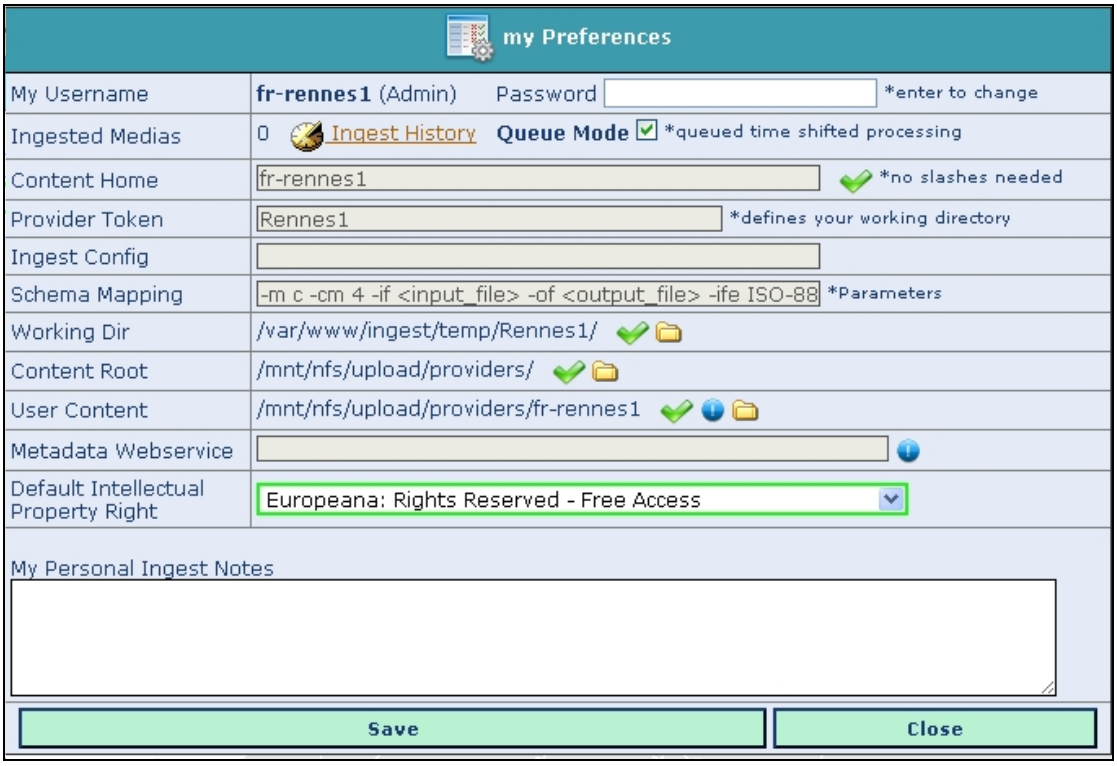

Figure 7-4: Preferences

The content home, provider token, ingest configuration and schema mapping and metadata webservice definitions are controlled by the BHL-Europe administrator.

The end user has direct control over the account password plus the Queue Mode setting.

Queue Mode (when selected) will run the image preparation, OCR and taxonomic identification steps of the workflow (described further below) as background processes, independent of the login session. It is recommended that Queue Mode is selected by default under normal operations.

When Queue Mode is unchecked, those processes run directly connected to the login session, and the output will be seen on screen. For any document with a significant number of pages, this will mean that all processing is tied to the currently selected workstep, and any interruption to that process (such as selecting another step or document to prepare) will override it. Queue Mode is useful for occasional debugging and smaller documents.

Figure 7-5: Password and Queue Mode

The end user has direct control over the default intellectual property setting. This is applied if there is no IPR declaration embedded in the metadata for a document, and can be overridden on an item by item basis.

Figure 7-6: Intellectual Property Right Default

7.1.3.3 Analyze and select content for ingest

The next step is to select the item of uploaded content that is to be prepared for ingest.

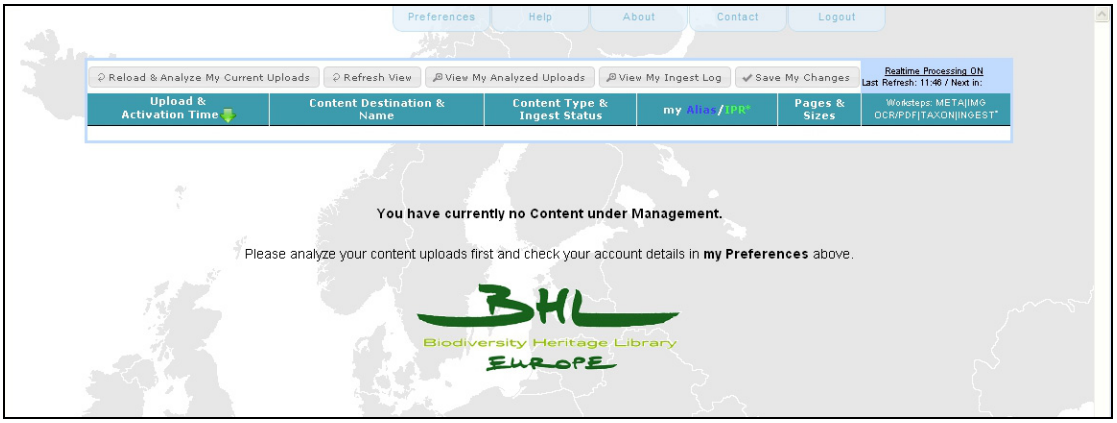

Figure 7-7: Select the Reload & Analyze My Current Uploads button.

Figure 7-8: Reload and Analyze

This will open the Content Analyzer window to display the uploaded items potentially available for processing.

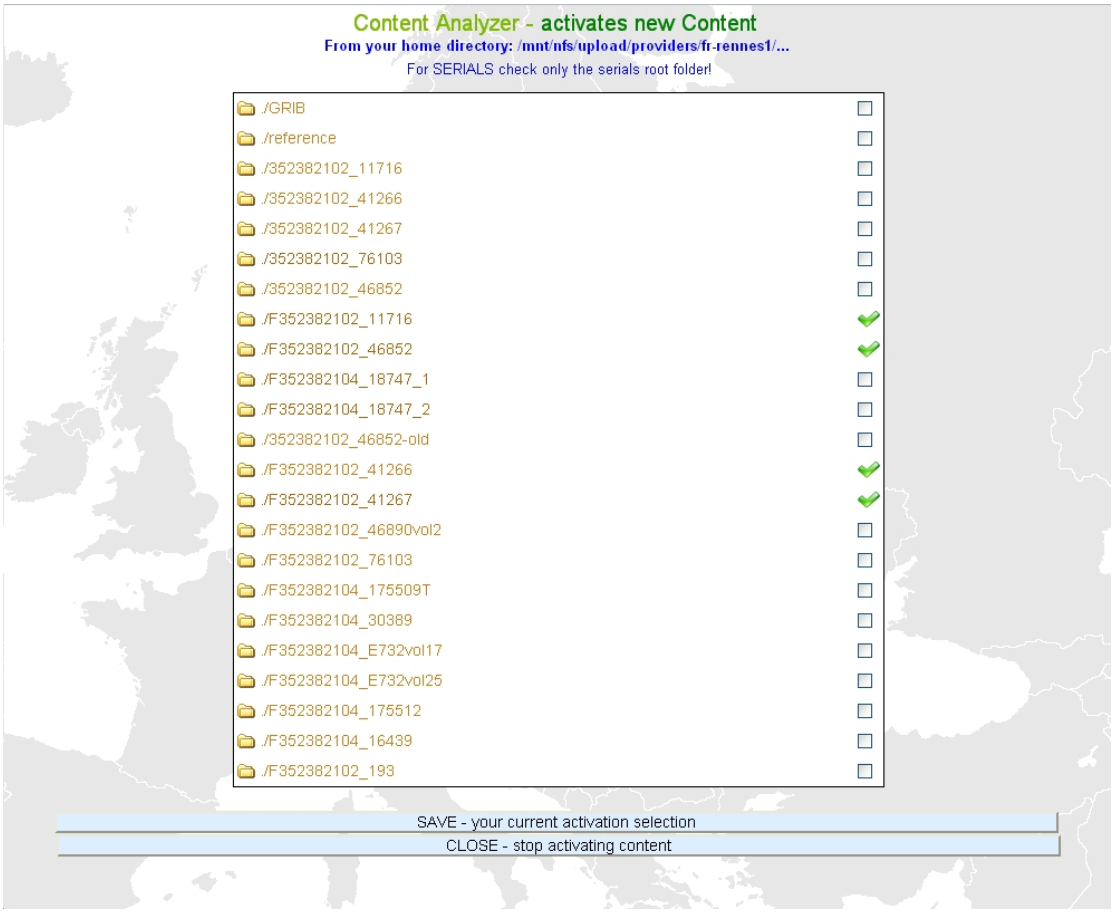

Figure 7-9: Content Analyzer

Navigate to the content to ingest by selecting the appropriate folder and check the checkbox.

If your content is organised into subfolders, selecting the folder will then show the content inside the folder, from where you will need to check the Activate Content Root/File option for the item that matches the desired document.

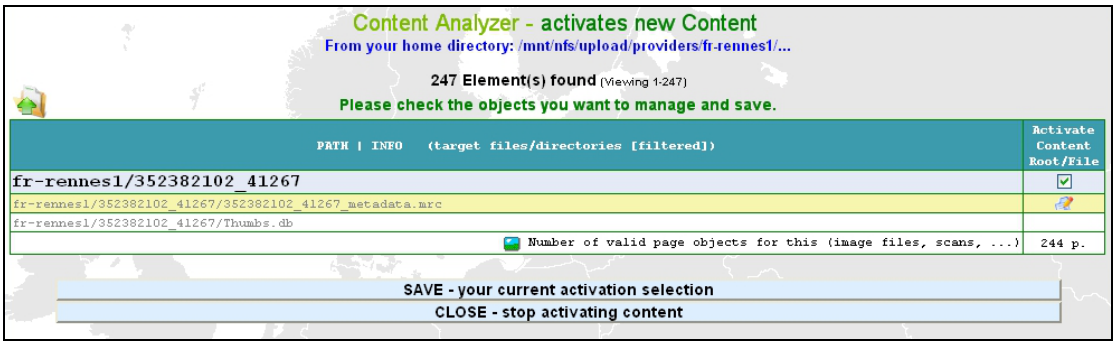

Figure 7-10: Activate content item

Once selected, the Save current activation selection will mark the item selected.

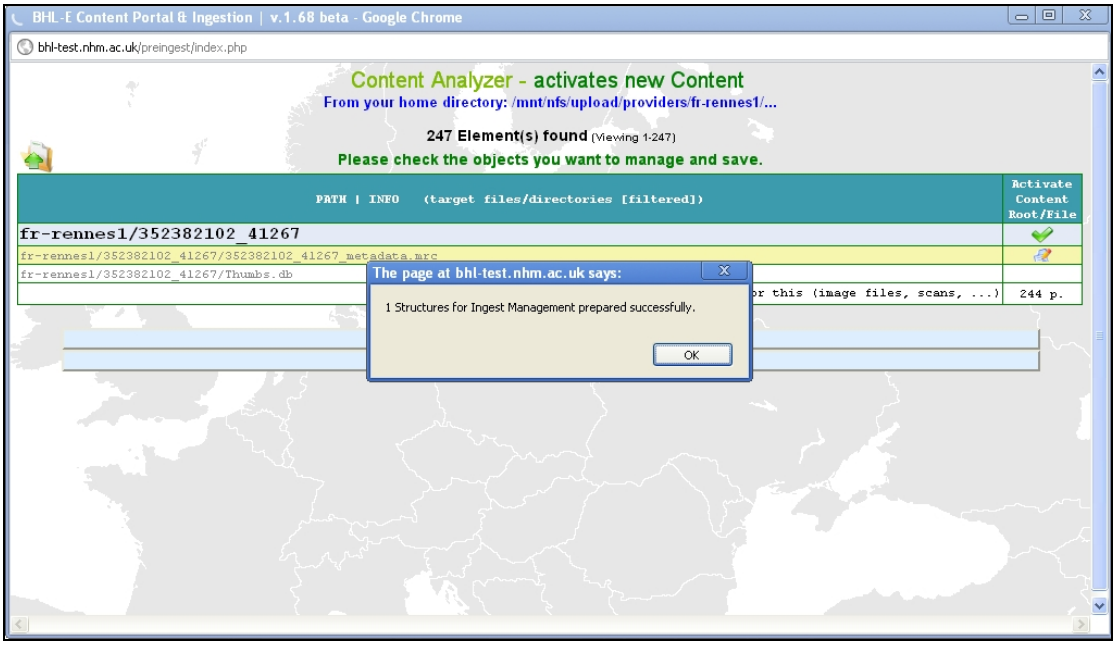

Figure 7-11: Activated content item

More content can then be selected if desired. Once all content is selected, the Close option will refresh the main interface to show readied content.

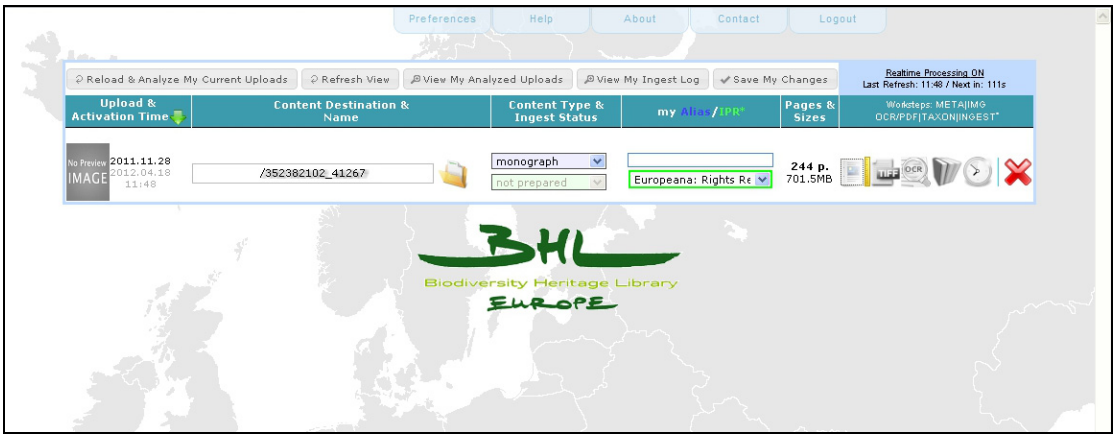

Figure 7-12: Main interface - with ready content

Each item has a set of data associated, shown on the main interface. These can be changed if needed for individual items if the automatic values are not correct.

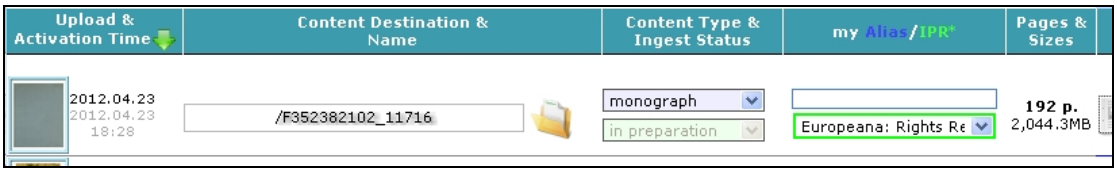

Figure 7-13: Pre-Ingest content item data

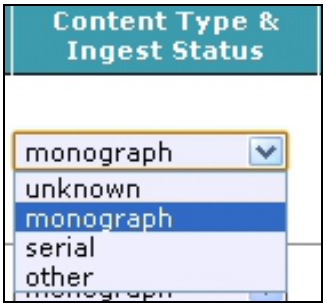

Figure 7-14: Content Type setting

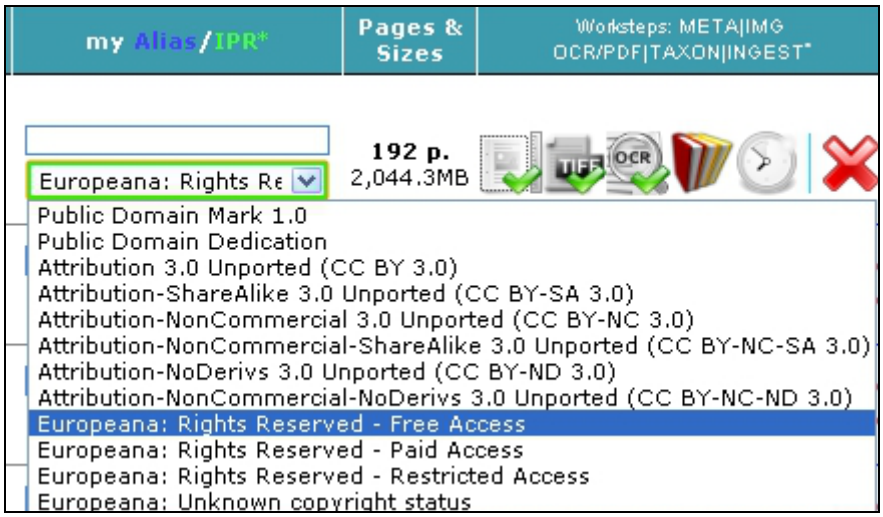

Figure 7-15: Individual item rights

After amending any of these data, make sure to store the new information with the 'Save My Changes' button.

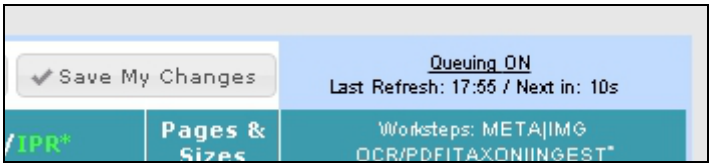

Figure 7-16: Save changes

7.1.4 Content Item Worksteps

7.1.4.1 Overview

To prepare each content item, a set of sequential work steps then needs to be set running.

- The next workstep will be in colour to denote it as ready to run. Those steps to follow will be greyed out until ready. Each must be selected in turn for a content item.
- When queued mode is active, the selected work step icon will colour cycle while the process is still underway. This means that processing will not be interrupted should the end user disconnect and reconnect later. Once the background job has completed, the icon will be in static colour again.

To complete the step click on it once more – this function enables review/debugging by temporarily suspending Queue Mode and rerunning the step if any errors have occurred.

- Once the step is complete, it will be marked with a green tick, and the next workstep will be readied.

7.1.4.2 Mapping the Metadata to OLEF

The first work step processes the uploaded document metadata and transforms to the Open Literature Exchange Format, OLEF used by BHL-Europe services.

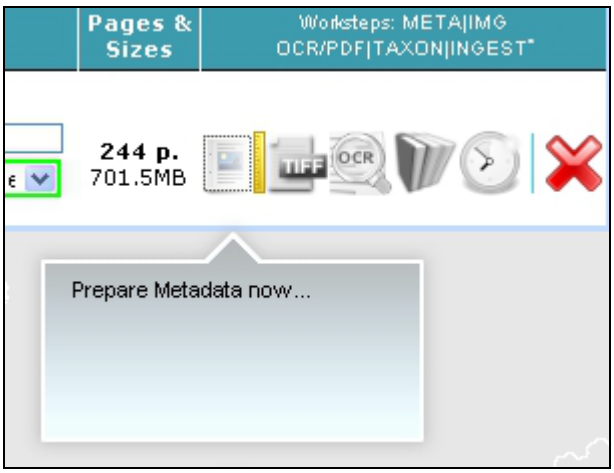

Figure 7-17: Metadata Mapping

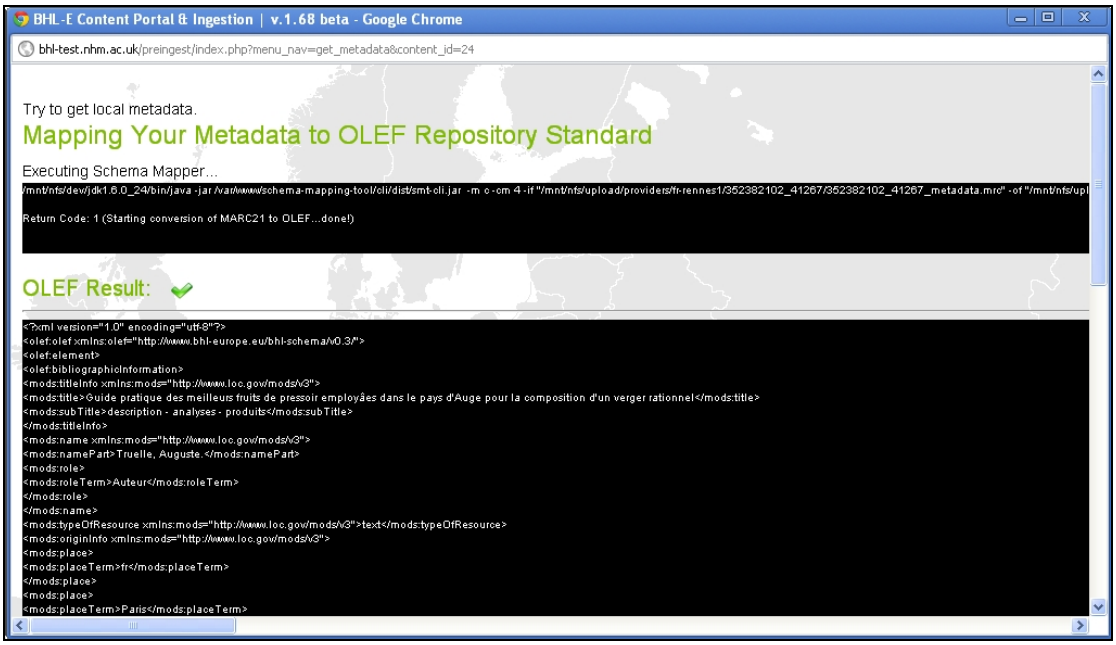

Figure 7-18: Mapping in progress

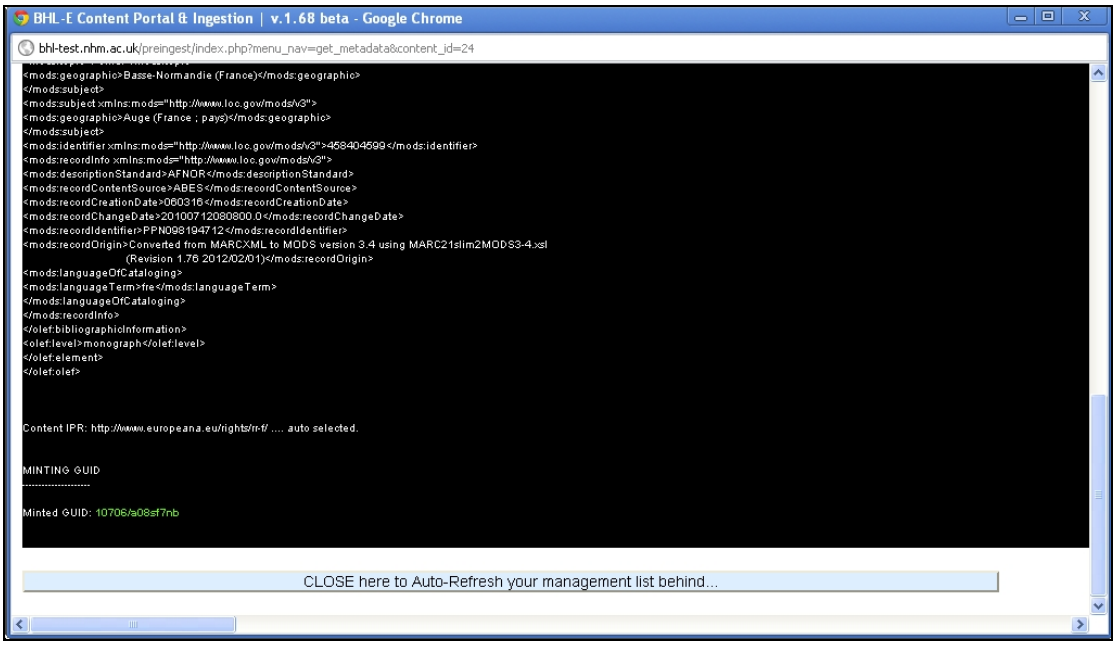

Figure 7-19: Mapping Complete

7.1.4.3 Prepare Tiff images

Once metadata is prepared, archival tiff images need to be checked or generated.

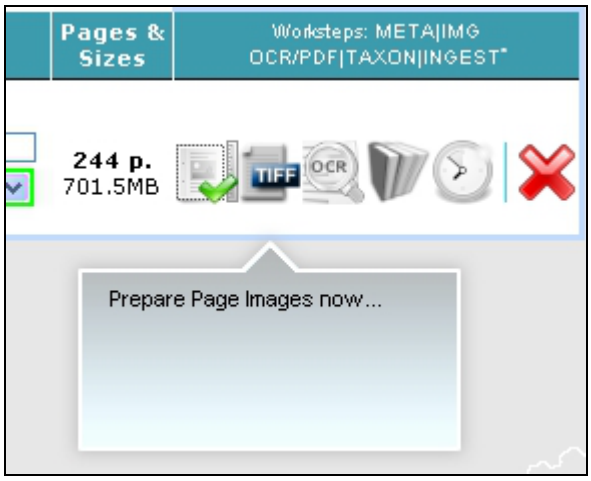

Figure 7-20: Tiff Preparation

Where an upload includes tiff files directly, this preparation will complete quickly. PDF uploads will take some time to generate the tiffs; in this case queued processing will allow the end user to queue a batch of documents for background processing.

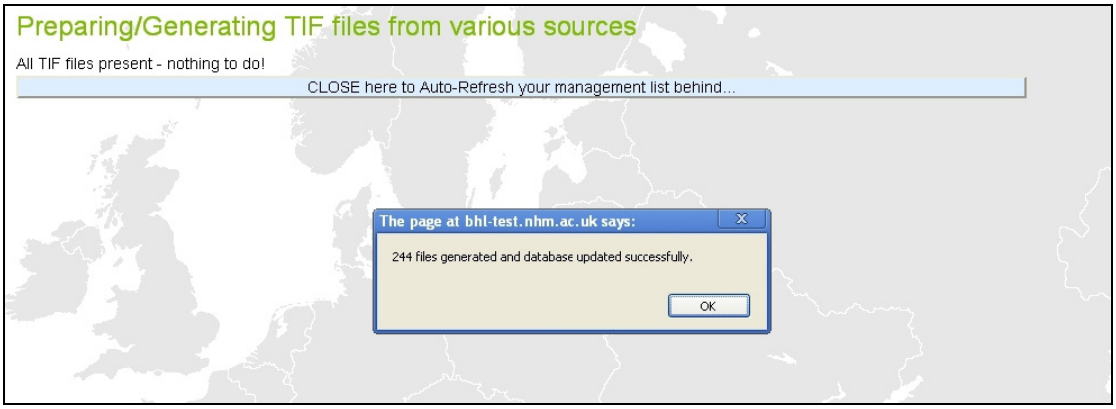

Figure 7-21: Tiff Preparation Complete

7.1.4.4 Generate OCR

The next workstep passes the tiff images into an OCR process to extract text. (pdf text will be directly extracted).

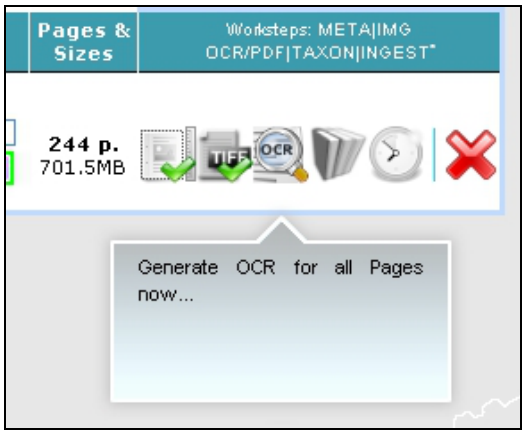

Figure 7-22: OCR generation

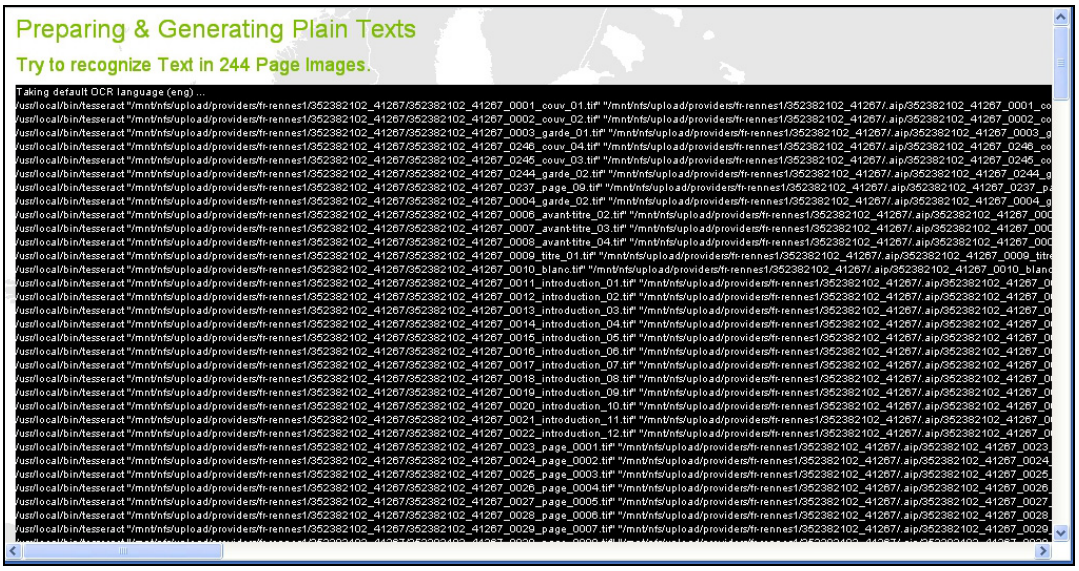

Figure 7-23: OCR queuing

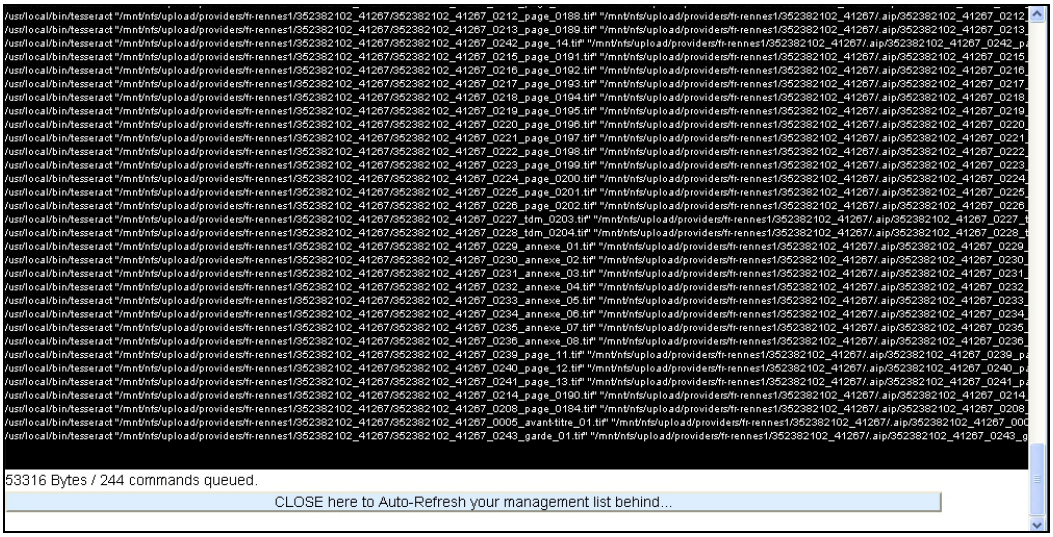

Figure 7-24: OCR queued

7.1.4.5 Prepare Taxonomic Information

The taxonomic preparation step uses the uBio service (www.ubio.org) to process the OCR text and extract recognised names. These names are written into the OLEF metadata.

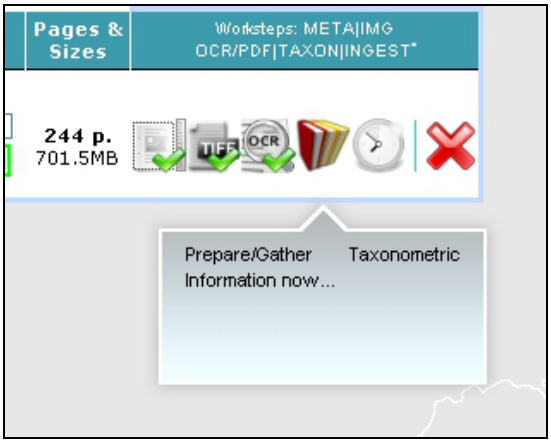

Figure 7-25: Prepare Taxonomic information

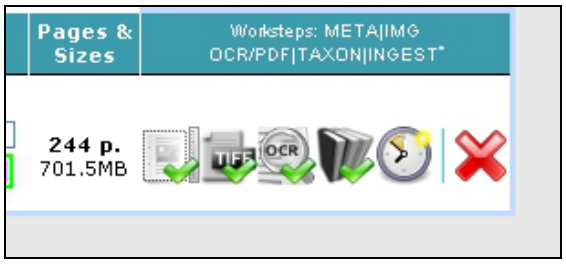

Figure 7-26: Taxonomic preparation complete

7.1.4.6 Send for Ingestion

The final work step is to consolidate the final metadata for the object, and submit the completed preparation for archiving by the Fedora commons archive component for ingest to the archive.

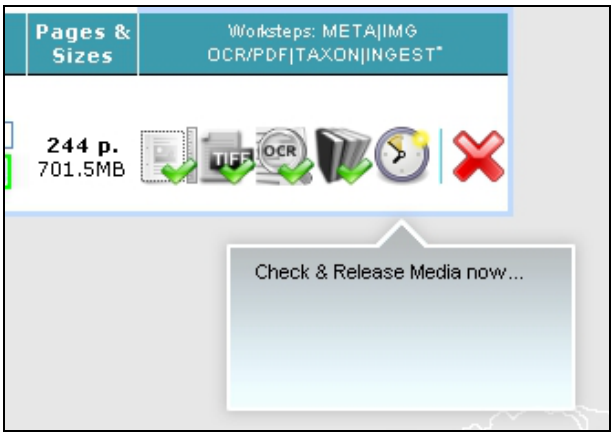

Figure 7-27: Ingest

7.1.5 Post Ingest Processes

The following process steps are automated, and mentioned here for completeness.

7.1.5.1 Indexing

On completion of the ingest to the archive, the document will be automatically indexed by the Fedora GSearch tool, at this point the document metadata will be visible in the BHL-Europe portal and can be found via searches.

7.1.5.2 Derivative Generation

In parallel with ingest to the BHL-Europe archive, an automated batch process generates a number of derivative formats of the object, including downloadable variants and page image forms needed by the BHL-Europe content viewer to enable reading online. This generation is not immediate, and will run in the background. Once complete, the content viewer will be able to render the document and downloads will be available.