

ECP-2008-DILI-518001

BHL-Europe

Prototypes of deduplication tool and bibliographic database system for monographs and serials

Deliverable number	<i>D2.2</i>
Dissemination level	<i>Restricted</i>
Delivery date	<i>23 February 2010</i>
Status	<i>Final</i>
Author(s)	<i>Henning Scholz, Boris Jacob, Wolfgang Koller, Michael Malicky, Dennis Zielke</i>



eContentplus

This project is funded under the *eContentplus* programme¹,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

0 Document History

0.1 Contributors

Person	Partner
Henning Scholz	MfN
Boris Jacob	FUB-BGBM
Wolfgang Koller	NHMW
Michael Malicky	LANDOE
Dennis Zielke	UBER

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
17 February 2010	Henning Scholz, Boris Jacob, Wolfgang Koller, Michael Malicky, Dennis Zielke	0.1	First draft after finishing the concept
23 February 2010	Henning Scholz	1.0	Final version incorporating all feedback by partners and BHL

0.3 Reviewers and Approvals

This document requires the following reviews and approvals. The final revision and approval of this document took place during the WP2 day of the last project meeting in Berlin (23 February 2010).

Name	Date	Version
BHL-Europe consortium	23 February 2010	1.0

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
BHL-Europe consortium	17 February 2010	0.1
BHL- Europe consortium	01 March	1.0

Table of contents

0	DOCUMENT HISTORY.....	2
0.1	Contributors	2
0.2	Revision History.....	2
0.3	Reviewers and Approvals.....	2
0.4	Distribution	2
1	PURPOSE.....	4
2	BACKGROUND AND CONTEXT.....	4
3	DEDUPLICATION	5
3.1	Requirements	5
3.1.1	Overview	5
3.1.2	Defintion of a duplicate	5
3.1.3	Deduplication and further steps.....	6
3.2	Available deduplication procedures	8
3.2.1	BHL serial union catalogue	8
3.2.2	Levenshtein algorithm.....	10
3.2.3	GBV approach.....	10
3.3	Alternative concepts for deduplication.....	10
4	BIBLIOGRAPHIC DATABASE.....	11
4.1	Structure and specifications	11
4.2	Available technical solutions of the GRIB prototype.....	13
4.2.1	Biodiversity Heritage Library Serials Union Catalogue.....	13
4.2.2	GVK of the Common Library Network.....	16
4.2.3	Metadata repository of page images	19
4.3	Work plan towards BHL-Europe D2.3.....	19
5	CONTENT ANALYSIS STATUS.....	20

1 Purpose

To provide an overview of the available prototypes for managing bibliographic data in BHL-Europe, their specifications and limitations, and the plan to change, adapt and combine the prototypes in one live system: the Global References Index to Biodiversity (GRIB).

2 Background and Context

BHL-Europe will manage the acquisition, digitisation, and hosting of the digitised material contained in European institutions. The project will manage the process by which each institution digitises its biodiversity material and ensure that this is done efficiently and effectively. The efficient coordination and management of ‘commitments to digitise’ is vital (some of the material is available in several different locations in Europe) and unnecessary duplication should be avoided.

To eliminate duplication during the scanning process, several tools and databases – including technical solutions for duplication control – will be developed to analyse the content and support the management of the scanning initiatives of each partner. Existing tools are insufficient for the scale required by BHL-Europe. Thus, the adaptation and enhancement of these tools and databases are of great importance. Eventually, a bibliographic database will contain information on monographs and serials that have been scanned in the past and are available in the format defined in the MoU of BHL-Europe. Those are to be included in EUROPEANA and BHL-Europe. In addition, information will be provided about biodiversity literature in the process of being digitised and about which partner is responsible for scanning that material. Eventually, the database will contain information on all relevant literature that needs to be scanned, and will identify the partner who will be responsible for providing the material in the future. If there is no consortium partner with some of the critical content, appropriate content holders will be identified and encouraged to join the network to provide this content. This system ensures that every content provider and even potential content providers can check before starting the digitisation to see if the material is already in the process of being digitised. The local digitisation processes can be planned accordingly and duplication is reduced to a minimum, ensuring the effective use of the local resources available in each partner institution. The entire bibliographic database system with content management and deduplication functionalities is called “Global References Index to Biodiversity” (GRIB).

The development of GRIB is part of the IT objective of WP2 of BHL-Europe. This is a prerequisite to focus on the management objective of WP2 at a later stage of the project. GRIB is a database of biodiversity literature that indicates:

- a) monographs and serials that are relevant for the biodiversity community (i.e. the library catalogue records of the BHL-Europe partners)
- b) the distribution of this relevant literature in the partner libraries
- c) the portion that is already available in digital form (BHL Portal)
- d) the portion that is in the process of being digitised (BHL-Europe partners)
- e) the portion for which plans have been created for digitisation

f) the responsibilities for content contribution (bidding process).

During the last BHL-Europe General Meeting in Prague (Nov 2009) we agreed to adapt the existing BHL Serials Union Catalogue as a starting point for GRIB. By adding more catalogue data from our European partner libraries (including monographs) and by improving and adding functionality to the system, we are able to provide a working solution that fulfils the above mentioned requirements for GRIB. It was also agreed to evaluate other available solutions and coordinate with related projects and activities to avoid duplication of efforts. In that context we currently cooperate with the GBV (Common Library Network), a public non-profit institution building the GBV union catalogue for more than 400 German libraries. We evaluated the BHL Serials Union Catalogue and the GBV union catalogue as GRIB platforms and enhanced both systems based on our requirements. The results are presented below together with the work plan for D2.3.

3 Deduplication

3.1 Requirements

3.1.1 Overview

The biodiversity literature collections in Europe are spread across a large number of institutions. Inevitably, there is much duplication of content in our partner libraries. While each national government will be expected to fund scanning in their own country, it will be critical that we avoid scanning the same material in multiple locations. Duplication would waste European financial resources and deliver a very confusing set of Web sites for the user. However, besides the avoidance of duplicate scanning, the identification and management of existing duplicates in digital collections is another important aspect. Currently, BHL-Europe is managing more than 70,000 digital items with a number of duplicate records.

The process of deduplication has to take all this into account. We need first of all technologies that are able to recognise duplicates in databases and in library catalogues. On the other hand, we need criteria and guidelines what to do with duplicate records and what level of duplication is acceptable. In that context the definition of a duplicate is also not obvious. Eventually, we need a balance of cost/benefit of duplication vs. cost/benefit of deduplication. This was extensively discussed in BHL-Europe over the last months. It is, however, not only the problem of BHL-Europe. Many other projects have the same or at least very similar problems and are also discussing technologies and criteria. BHL-Europe will hopefully benefit from these ongoing discussions. In the following, we will provide some more details of the various aspects of duplication and deduplication in order to build the framework for our technological solutions.

3.1.2 Definition of a duplicate

If we want to deduplicate databases and catalogues, we first need to have a definition for duplicates. This is important to identify thresholds for accepting a record as a duplicate entry in a catalogue. We have collected a number of possible definitions and discussed those in the consortium and with BHL:

- Two scans of the same printing, the same edition, the same metadata.

- Two scans of different printing, the same edition, same or different metadata.
- Two scans of different printing, the same edition, same or different metadata, different quality (e.g. resolution of images, missing pages and plates, colours of the scans).
- Two scans of the same work but different metadata.
- Two scans of the same work but different metadata quality.
- Two scans of the same work but different editions.
- Two scans of the same work but different translations.
- Two scans of the same printing, the same edition, the same or different metadata but from a special collection
 - with different marginalia and hand-written annotations, e.g. by Darwin or Ernst Haeckel
 - with hand-coloured plates (unique editions)
 The problem with these items is that their distinctive feature (e.g. marginalia) is usually not obvious in the metadata.
- Two metadata records referring to different digital objects of the identical physical text.
- Two differing metadata records referring to the same digital object.

It is currently not decided yet, which of the above definitions of duplicates is applicable to BHL-Europe to what extent. This also has to do with the expectations of the target users. Librarians and taxonomists, for example, might have a different view on that. Even among taxonomists the opinions might differ. In regard to two scans of the same printing in different scanning quality some taxonomists might find b/w copies with 300 dpi sufficient for their purpose while others are asking for 600 dpi colour copies. If both versions are available, the user has the chance to find and use the different versions.

The definition of duplicates also needs to take the quality of the metadata into account. One example is already mentioned above: If the metadata don't show the special character of certain items (e.g. marginalia), it is impossible to identify those items in a database. Therefore, we first need to know all our data (metadata, content) better before specifying criteria in more detail. These information about our data will come during the test ingest process when we start working with the data in detail.

3.1.3 *Deduplication and further steps*

In our understanding deduplication is the process in a bibliographic data system by which it is established whether multiple bibliographic records do or do not refer to the same book or journal. Thus, we need a tool or technology that is able to (semi)automatically identify duplicates in our system effectively to support the work of the librarians and reduce the human (librarian) intervention to a minimum.

The definition of a duplicate does not necessarily influence the development or implementation of deduplication technologies. However, it sets the requirement for the technology to be flexible enough to work with different definitions. By tuning the algorithms and moving the threshold the technology must be able to be more sensitive or less sensitive to whatever level of duplicate we define as being relevant.

Once a deduplication procedure is established, we can think about how to handle the duplicate records and how to avoid duplication in the future. Currently a large number of duplicates are

already present in the BHL Portal. Two main procedures are obvious how to handle the duplicate items:

- choose one item and delete or at least suppress or mark the duplicates
- find a way to present duplicates appropriately

The first option would save us costs in managing and saving these duplicates. However, as the scanning has already been done, the added value of finding and using those duplicates might be more beneficial than saving the money for storage and management of that digital item. Having the various definitions of duplicates and the expectations of the users in mind, it is very dangerous to follow this option. If, for example, the quality criteria is applied to the item, it may be beneficial for some users to read the b/w copy as this copy loads faster in the book viewer than the high quality colour copy of the same publication. As the costs for storage are decreasing over time, it is a risk to loose value by eliminating assumed duplicates. Therefore, BHL-Europe is currently working on appropriate mechanisms to find duplicate records and present them appropriately in the search results.

One mechanism handling duplicate records after they have been identified as such is merging them. In a bibliographic data system with multiple bibliographic records referring to the same book or journal, precedence is established for a "preferred record" based on a set of criteria (source of cataloguing, completeness, etc.). Then selected data elements from the non-preferred records may be added to the preferred record to establish a master record. The non-preferred records are either deleted or suppressed or marked clearly.

In the context of deduplication and record merging it also makes sense to mention the FRBRization of catalogues. Following the Functional Requirements for Bibliographic Records, FRBRization brings together sets of works, expressions, or manifestations, rather than having a display that shows all of the different editions of the work. One example developed by members of our larger BHL community is already working with a related concept for the Trove project (<http://trove.nla.gov.au/>). Currently, discussions are ongoing what work concept is acceptable for BHL and BHL-Europe and how this is going to be implemented.

The presentation of duplicates (grouping, merging, FRBRization) both needs to work for the scanned items and the library catalogue records. This would help us identifying items that are already present in our digital repository. In a next step the librarian would be able to decide based on a list of criteria whether an additional scanning of a duplicate is required or not. That list of criteria is aligned with definitions of a duplicate. It may be useful to scan duplicates under the following assumptions:

- The same work, but different edition.
- The same work with different marginalia or hand-written annotations.
- The same work with hand-coloured plates.
- The same work, but different book/processing quality (e.g., paper quality, binding quality, cut of margins, missing pages, missing plates).
- The same work, but different scanning quality (e.g., image resolution, colour).
- The same work, but different metadata (e.g., missing page or plate numbers)
- The same book scanned as part of a large scale project (it might be easier to scan the entire collection instead of checking for duplicates and sorting them out).

This list is not complete yet, but will be continued over the next weeks. It is, however, important to balance those criteria with other considerations. The librarian has to answer, for

example, the question, whether the value of a duplicate following these criteria is greater compared to the value of a new item that has not been scanned at all. The technologies we are using and going to implement has to take these aspects into account. A solution for the last question would be a bidding function for the users, to ask explicitly for items to be scanned. If the users definitely need a duplicate, it has value to scan and deliver this item. The librarian and user interaction using the technology will help to set our scanning priorities for the future as efficient as possible.

3.2 Available deduplication procedures

3.2.1 *BHL Serials Union Catalogue*

For the BHL Serials Union Catalogue, the deduplication procedure is based on the ISSN and the titles (Figure 1). Every dataset is compared with each other without ranking or prioritise one over the other. For the numerical comparison only normalised ISSN are used. The comparison based on titles will only be made if the ISSN matching don't give exact matches. The title matching follows this procedure:

- 1) Exact string match? If none,
- 2) Subfield a only exact string match (where subfield a starts a at the start of the title field)? If none,
- 3) Longer 245 subfield a contains the shorter 245 (where subfield a starts a at the start of the title field)? If none,
- 4) Take significant keywords from the 245 subfield a of both records to be matched. Where there are more than significant two keywords per record, match where the keywords are present in both records. If none,
- 5) No match.

If either the ISSN or the title procedure for matching was positive, a canonical record is prepared. If the matching was negative (no match), a new record is created and inserted in the database as it is.

The above procedure is very time consuming. Therefore, we implemented a faster merging procedure for the entries based on the ISSN using a stored procedure inside MySQL. As ISSN numbers are very unique, they are very well suited for deduplication procedures of serials. A comparison of data already present in the database found 27 out of 30 serials of LANDOE in the database of 80,000 records. The three missing serials are very new and not present in the list to date. However, a number of serials have no ISSN number. Thus, an automatic identification is not possible. The mentioned ISSN matching procedure is therefore, not accurate enough for complete deduplication, but provide a very good base for quick and automatic deduplication. It is helpful to reduce the human interaction to a minimum and can be at least applied as a first step in a deduplication process.

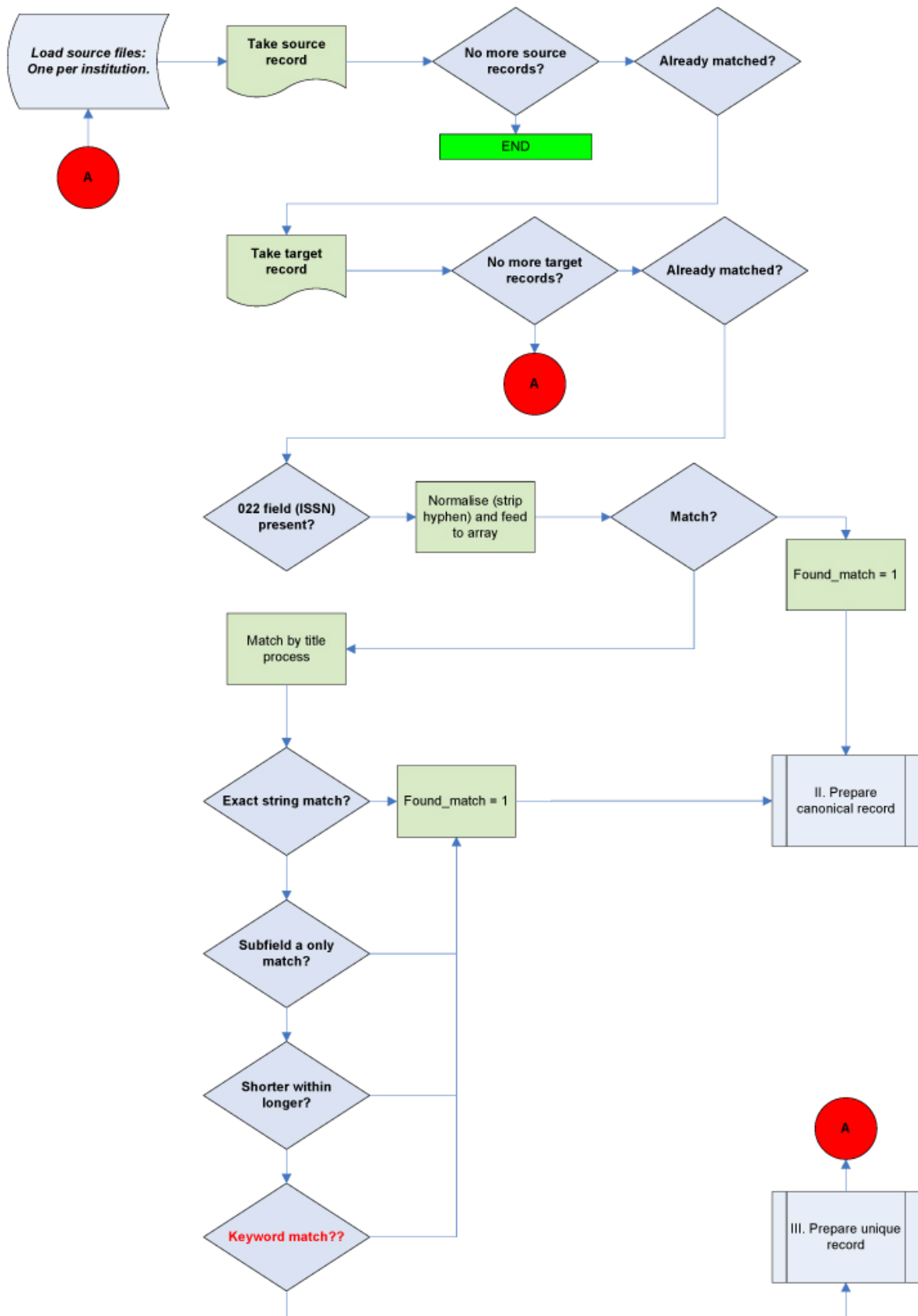


Figure 1. Matching procedure established for the BHL Serials Union Catalogue.

3.2.2 *Levenshtein algorithm*

During the first months of the project, our former WP2 leader had been testing the Levenshtein algorithm for identifying duplicates in databases through string matching. For testing, data from BHL and Internet Archive were available. In a first step it was validated if the metadata and volume information (for serials) were appropriate for processing. During this pre-ingest, the cleaned records were transformed into an internal generic XML-schema. The number of records was limited to 5,000 for testing. A process was created to weight and organise the metadata. At the end of this process, an archive was created containing the generic MARC records separated by bibliographic types (monographs, serials, etc). This was the final step of the data preparation for deduplication using the Levenshtein algorithm.

The Levenshtein algorithm compares two strings and identify the number of operations to transform one string into another. To transform a string „North Western Journal“ into the string „South Western Journal“ that already exists in the database requires two steps. Depending on the threshold for similarity of strings, the string would be either a child of an existing string or established as a new string (parent) in the database. For the matching procedure, the title (MARC 245a) was used during the testing phase and a threshold of 80% was used. Below 80%, a new title was considered as a new (parent) record, above 80% a new title was considered as a child record of an existing record in the database. Unfortunately, our WP2 leader left the project before all test runs were finished and before the software was released as a prototype to be used by BHL-Europe.

3.2.3 *GBV approach*

The system implemented by the GBV for BHL-Europe D2.2 consists of library metadata from BGBM, MFN, NAT and NHML. First the MARC 21 data from Naturalis is imported. Then the data from BGBM and MFN, coded in the German metadata standard MAB2 (Maschinelles Austauschformat für Bibliotheken), will be matched with the Naturalis data.

In this prototype the deduplication will made through a simple matching procedure using the International Standard Book Number (ISBN), the International Standard Serial Number (ISSN), identification numbers i.e. from the German national library and two algorithms: the author-tilte key 4-4 (Autor-Titelschlüssel: ats 4-4) and the title key 4-2-2-1 (Titelschlüssel: tsl 4-2-2-1).

Ats is a key consisting of the first 4 letters of the surname of the author and the first 4 letters of the title. If the author surname has less than 4 letters, a blankspace is added. Example: “Karl May: Winnetou” leads to ats: may winn.

Tsl consists of the first 4 letters of the title entry word, the first 2 letters of the second title word, the first 2 letters of the third title word and the first letter of the fourth title word. Example: “David Herbert Lawrence’s Poetry” leads to tsl: davihelap

3.3 **Alternative concepts for deduplication**

In order to evaluate options for deduplication procedures we also evaluated possible subcontractors that have working system ready for use. One company that seemed to be suitable is SSL (System Simulation, <http://www.ssl.co.uk/>). This company has experience in developing matching algorithms. We submitted a sample of data for them for testing and evaluation purposes. This sample set consisted of data from the serial bidlist.

Method

- Unmatched title records extracted from current serials mashup (37,264 rows) of which some would be genuinely unique and others that had failed to match under the previous system.
- 17 sample records for matching extracted and expected results via manual investigation (the latter not supplied to SSL).
- SSL import to database and run their matching process and return results which are manually checked.

Results

Of a sample 17 rows supplied, 44 matches were expected (inclusive of the original record being searched for).

- SSL matched 36 records (81.8%).
- SSL matched falsely to an additional 14 records (extra 31.8% false matches)
- SSL did not find matched for 8 records (18.2%)
- SSL matched 1 record twice (duplicate) (2.2%)

Conclusion

BHL has worked on the philosophy that >60% is a good result for text pattern matching. Therefore 81% is excellent. We recommend some tweaking to reduce additional false matches is performed. These often occur where

1. the publisher differs between the titles (they are separate titles)
2. subfield b of the 245 title field (sub-title) is clearly different and should be recognized as such.

4 Bibliographic database

4.1 Structure and specifications

Two types of records are identified to be represented in the GRIB: monographs and serials. Monographs have an author, whereas serials are tagged on a title level. Ultimately, serials are broken down to an article level with author and title information, but this is not the standard approach of BHL currently.

There are two groups of users identified: scientists/the public and librarians. Scientists/the public – the primary user of the content – will search for literature via the BHL Portal or services like ViTaL, the Virtual Taxonomic Library of the EDIT project. Also vifabio, a metasearch engine across biology resources including BHL, funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) would be an appropriate tool for this user group. Before that we need a system for the librarians of the partner institutions to actually manage the content acquisition process in BHL-Europe.

GRIB consists of three components (for the integration of GRIB in BHL-Europe see Fig. 2):

- (1) The catalogue records of the content provider and the bids related to these records on the level of item and title.

- (2) The index to the bibliographic information, identified by author and title, combined with the relevant catalogue records.
- (3) The digital page images (BLOBS = Binary Large Objects) and all their derivatives and related metadata.

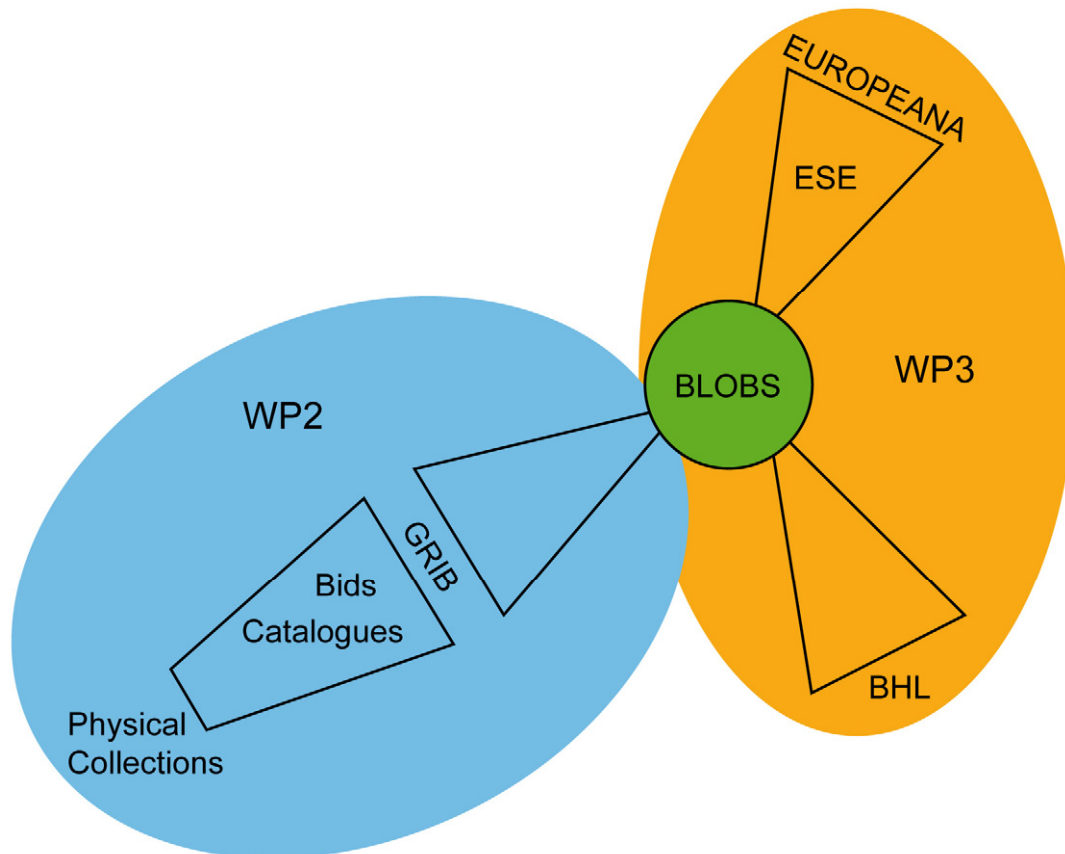


Figure 2. The Global References Index to Biodiversity is composed on the catalogue records of the physical collections, the index and the link to the BLOBS. It is an important component of WP2 of BHL-Europe and provides one access route for the digital images files in the repository. It is currently thought to be the access route of the librarians managing the content hosted by BHL. The actual content integration of the BLOBS will be done by WP3. The BHL-Europe users (taxonomists, general public) will access the BLOBS either through the BHL Portal or Europeana. For them, WP3 also improves the functionality of the BHL Portal and connects the BHL-Europe content to Europeana.

The catalogue records

These are the library catalogues of all partner libraries in the project, not only those with digital page images. This is to get a complete as possible overview of the relevant literature for the biodiversity community. These library catalogues usually have no deduplication tools in place, so all editions and multiple language titles are going to be harvested by GRIB. If an item is identified to be digitised by BHL-Europe, this will be also documented in the library catalogues and included in the scanning workflow of the partner according to the local requirements and processes.

The index

This is the link from the library catalogue records to the digital page images in the BHL-Europe repository. Index records are created for each unique catalogue record. These index records are displayed when searching for a title. If identical titles (duplicate records) are identified by the system, these catalogue records are merged under one single index record and are displayed when checking the details of that index record. When looking for “The origin of species” the result would be one hit, which is the index record. By clicking on that record all catalogue records would be displayed, with the library information, the original spelling of the title, the copies and editions in the partner libraries, the bids on that record, the digitisation status, and the link to the BLOBS of that record.

The digital page images (BLOBS)

These are the files stored either in the BHL-Europe repository or in the repository of the BHL-Europe content provider. For BHL-Europe it is ideal to store the original page image files, all derivatives and the related metadata. In some cases it may be required by the content provider to not share their original page image files but only the derivatives with BHL-Europe for storage in a central repository. The actual access model per partner is defined with the MoU that is signed on a bilateral base between the content provider and the BHL-Europe project.

4.2 Available technical solutions of the GRIB prototype

4.2.1 Biodiversity Heritage Library Serials Union Catalogue

This solution is based essentially on the Biodiversity Heritage Library Serials Union Catalogue developed at NHM London (see Figure 3). This system has proven to work for the management of the serial digitisation process. MARC exchange format records were received from the institutions, matched and merged and made available on the web. Partners are able to log in and bid against records of journals which they will be digitising in order to minimise duplication of effort. For this prototype, two catalogues of our partners were included in the modified version (BHL-Europe Union Catalogue), which are NHMW and LANDOE.

The technology used was Apache, MySQL and PHP scripting. The PHP part is based on cakePHP (a rapid application development framework using the Model-View-Controller standard). Therefore the system is not limited to Apache & MySQL but rather it can support several environments and database backends (for the detailed requirements of cakePHP see <http://book.cakephp.org/view/28/Requirements>).

Current system functionality includes:

- Marc exchange format records (minimum required fields stipulated)
- Records matched and merged through batch php scripting (see above).
- Merged set made available over the Web and fully searchable/sortable
- Authentication (1 per institution)
- Ability to bid to scan part or all of a title’s holdings at a per title level.
- Ability to manually merge records which the scripting did not detect
- Merging the entries based on the ISSN using a stored procedure inside MySQL.
- Ability (at NHM) to link to record and download to our local scanning management system.

The existing system has some weaknesses that are mainly related to the performance (speed) and search/filtering functionality, among others are:

1. Matching criteria for automated merge is too crude and needs to be refined in conjunction with input from the BHL library community.
2. Automated merge process is too cumbersome and non-optimized and needs to be integrated into a discrete module. It is anticipated this module will be used several times as partners join or non-partners submit data.
3. A parallel and synchronizable database is required to allow automated merge to take place without affecting performance in the live database.
4. There are issues with display of diacritics which affect searchability.
5. A keyword search is required to allow for more effective search and manual merging.
6. The table structure needs to be optimized in order to speed up the whole system (including creating new indexes).

The bidlist uses an own table-structure for storing only the required data-fields (like ISSN, title, author, subject, etc.). Basically there are three important tables for the base functionality bid list: “bibs”, “holdings” and “bids” (the tables “users”, “groups”, “groups_permissions” and “permissions” are used by the authentication system “othauth” which is a standard module from cakePHP. For more details see <http://bakery.cakephp.org/articles/view/othauth-0-5-documentation>). Both “holdings” and “bids” are referencing the “bibs” table using a foreign-key called “bib_id” (target “bibs.id”).

The “bibs” table contains the basic metadata for an item entry. Some of the fields are named like their corresponding MARC fields (e.g. “022”, which is ISSN). The other fields are normally self explanatory, like “author”, “title”, “subjects”, etc. The “bibs” table basically contains all metadata which doesn’t (or better shouldn’t) change for the same item.

The “holdings” table contains all metadata required for tracking the holdings of an institution for a given item. Most of the columns are again named like their corresponding MARC fields (e.g. “035”, System Control Number). The most important (non MARC) fields in this table are:

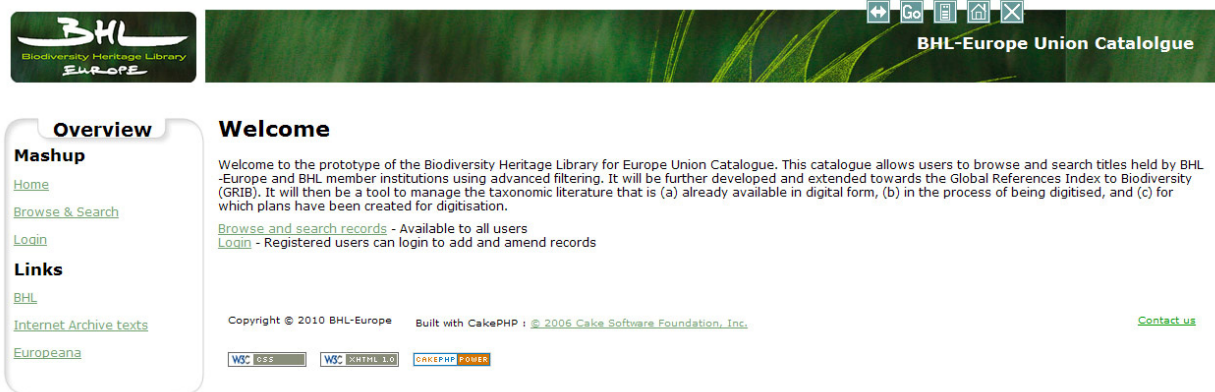
- “bib_id”: required to reference to an entry inside the “bibs” table.
- “hol_x” entries (there are four of them, hol_1 to hol_4): should contain the information for what parts of an item the institution is holding.
- “place”: should contain the local identifier for the institution (e.g. “MfN” for “Museum für Naturkunde”).

The “bids” table contains the bids from individual institutions for an item. It offers the ability to store information about what institution is digitizing what part of a series. The most interesting fields for this table are:

- “startdate”: Start date for digitization (e.g. 1978).
- “enddate”: End date for digitization (e.g. 1995).
- “excepts”: Exception for digitization (e.g. 1980-1981). Can contain exceptions if there are gaps in the holdings for an institution.
- “user_id”: ID of user who made this bid. Important to identify the institution.

Integration into the later BHL-Europe system can be easily done by using some simple mappings, as the bidlist doesn’t require many input fields. Another possibility to integrate it into the BHL-Europe infrastructure will be to directly access the data of the BHL-Europe

system (which should be the preferred method in order to avoid duplicate data dumps). This could be done by modifying the controller classes of the bidlist application.



Overview

Mashup

[Home](#)

[Browse & Search](#)

[Login](#)

Links

[BHL](#)

[Internet Archive texts](#)

[Europeana](#)

Welcome

Welcome to the prototype of the Biodiversity Heritage Library for Europe Union Catalogue. This catalogue allows users to browse and search titles held by BHL-Europe and BHL member institutions using advanced filtering. It will be further developed and extended towards the Global References Index to Biodiversity (GRID). It will then be a tool to manage the taxonomic literature that is (a) already available in digital form, (b) in the process of being digitised, and (c) for which plans have been created for digitisation.

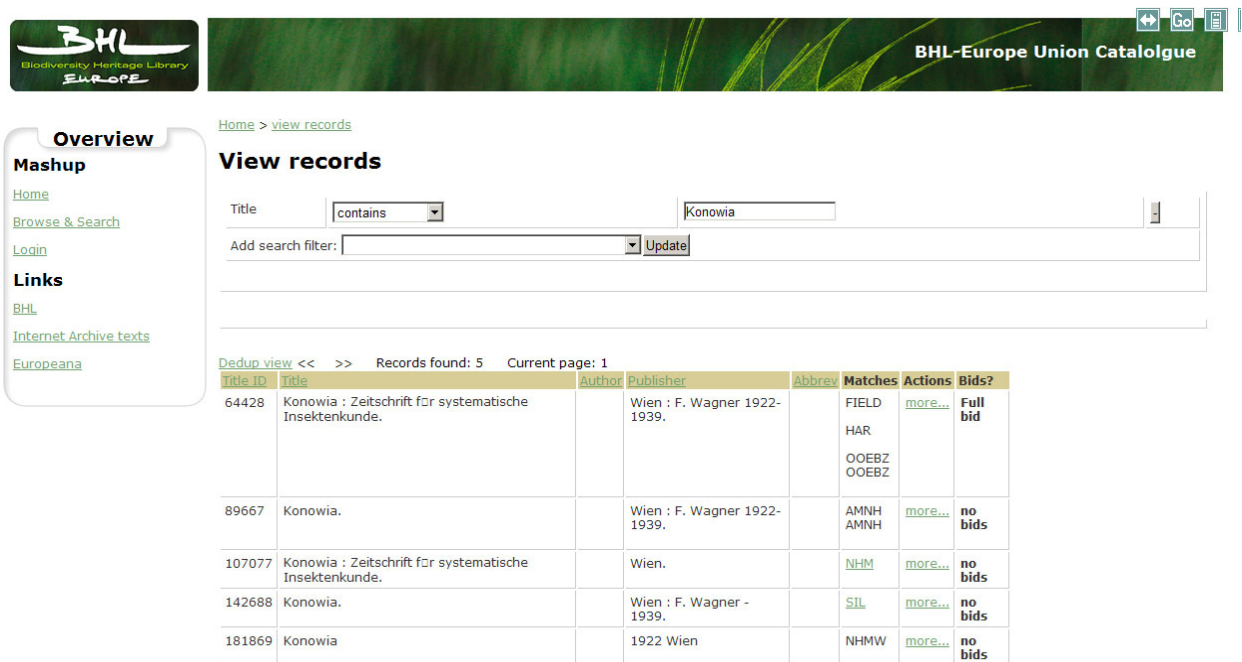
[Browse and search records](#) - Available to all users

[Login](#) - Registered users can login to add and amend records

Copyright © 2010 BHL-Europe Built with CakePHP : @ 2006 Cake Software Foundation, Inc. [Contact us](#)

W3C CSS W3C XHTML 1.0 CAKEPHP POWER

Figure 3. Welcome screen of the BHL-Europe Union Catalogue based on the BHL Serials Union Catalogue (<http://bhlserialist.test.www.nhm.ac.uk/>).



Overview

Mashup

[Home](#)

[Browse & Search](#)

[Login](#)

Links

[BHL](#)

[Internet Archive texts](#)

[Europeana](#)

[Home](#) > [view records](#)

View records

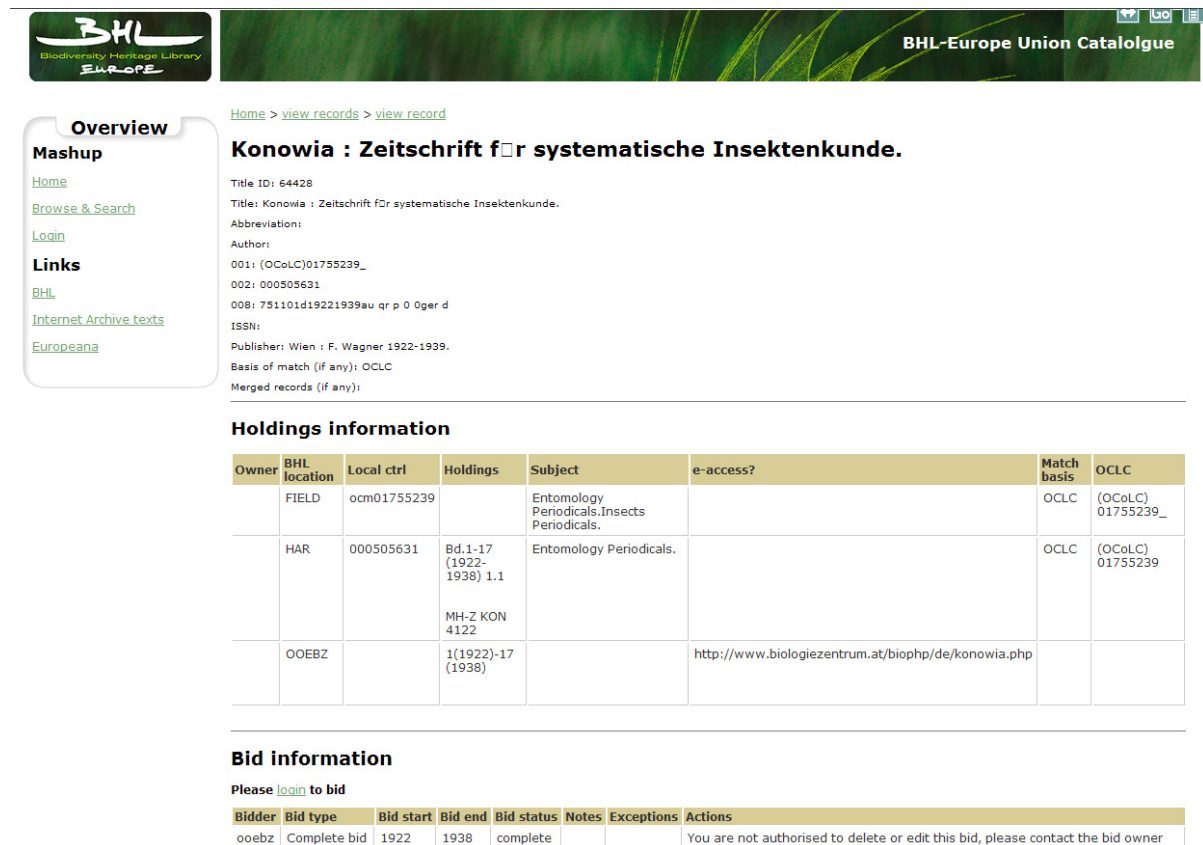
Title contains

Add search filter: Update

Dedup view << >> Records found: 5 Current page: 1

Title ID	Title	Author	Publisher	Abbrev	Matches	Actions	Bids?
64428	Konowia : Zeitschrift für systematische Insektenkunde.		Wien : F. Wagner 1922-1939.		FIELD HAR OOEBZ OOEBZ	more...	Full bid
89667	Konowia.		Wien : F. Wagner 1922-1939.		AMNH AMNH	more...	no bids
107077	Konowia : Zeitschrift für systematische Insektenkunde.		Wien.		NHM	more...	no bids
142688	Konowia.		Wien : F. Wagner - 1939.		SIL	more...	no bids
181869	Konowia		1922 Wien		NHFW	more...	no bids

Figure 4. Search result for one particular serial that is also present in BHL-Europe partner institutions (NHMW, LANDOE=OOEBZ).



BHL Biodiversity Heritage Library EUROPE

BHL-Europe Union Catalogue

Home > view records > view record

Konowia : Zeitschrift für systematische Insektenkunde.

Title ID: 64428
 Title: Konowia : Zeitschrift für systematische Insektenkunde.
 Abbreviation:
 Author:
 001: (OCoLC)01755239_
 002: 000505631
 008: 751101d19221939au qr p 0 0ger d
 ISSN:
 Publisher: Wien : F. Wagner 1922-1939.
 Basis of match (if any): OCLC
 Merged records (if any):

Holdings information

Owner	BHL location	Local ctrl	Holdings	Subject	e-access?	Match basis	OCLC
	FIELD	ocm01755239		Entomology Periodicals.Insects Periodicals.		OCLC	(OCoLC)01755239_
	HAR	000505631	Bd.1-17 (1922-1938) 1.1 MH-Z KON 4122	Entomology Periodicals.		OCLC	(OCoLC)01755239
	OOEBZ		1(1922)-17 (1938)		http://www.biologiezentrum.at/biophp/de/konowia.php		

Bid information

Please [login](#) to bid

Bidder	Bid type	Bid start	Bid end	Bid status	Notes	Exceptions	Actions
ooebz	Complete bid	1922	1938	complete			You are not authorised to delete or edit this bid, please contact the bid owner

Figure 5. Full record of Title ID 64428 also showing that OOEBZ has placed a full bid for this journal as it is already digitised by this partner.

4.2.2 GVK of the Common Library Network

The Common Library Network GBV (Gemeinsamer Bibliotheksverbund¹) is a public (i.e. non-profit) institution, funded by seven northern German federal states and the Stiftung Preußischer Kulturbesitz. It builds and hosts the union catalogue GVK (Gemeinsamer Verbundkatalog²), a bibliographic database, comprises the library holdings of the more than 400 GBV member libraries.

They run the proprietary Pica-CBS, an OCLC software based on Pica Search and Index. The GBV is an independent development partner of OCLC, i.e. they have technically and juristically full code access. Developments by the GBV are even reused by OCLC. The system is UNICODE/UTF-8 compliant.

The GBV can build the index with the functionalities we want, with our web interface, etc.

They can harvest/access library data via different protocols and ways: E-Mail, ftp active, ftp passive, CD, OAI Interface. A deduplication and merging mechanism will be implemented according to our specifications. The index will also be updated regularly. A scan request/bidding functionality as required by BHL-Europe does not exist, but can be build using the programming interface used for the shopping cart function.

¹ <http://gso.gbv.de/xslt/DB=2.1/LNG=EN/>

² <http://gso.gbv.de/xslt/DB=2.1/LNG=EN/>

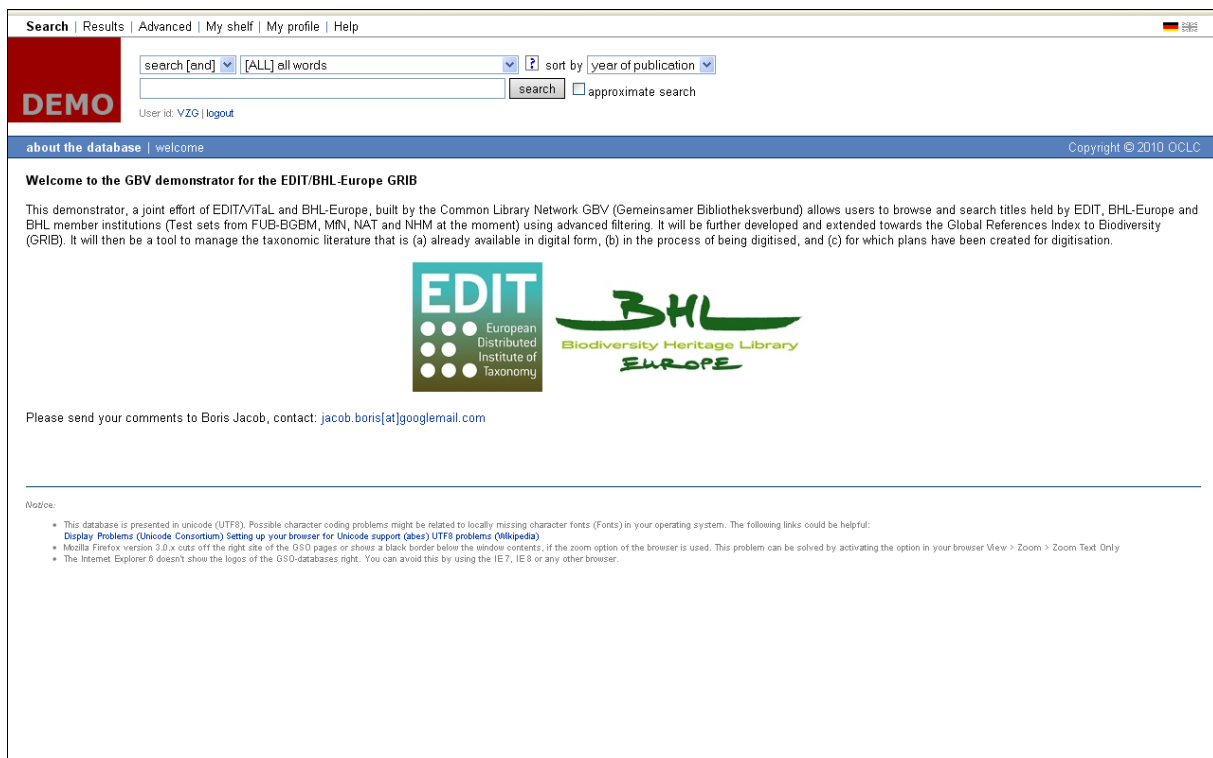
BHL-Europe will have full writing access to the data to help with data enhancement and also writing access to the system to help with IT development work (e.g. web interface development). The copyright of the metadata will be kept by the original rights holder.

Once build, the index can be accessed and even fully harvested via OAI-PMH. This allows BHL-Europe to reuse the data for backup or presentation in other systems. In that sense, the GBV can provide actually two products: one is the full infrastructure (hosting, functionalities), the second is the service (data enhancement incl. merging). An integration with the scanning workflow suite Goobi¹ is possible.

Altogether the positive aspects are a low risk, because the technology is ready, the system is fast and can deal with large numbers of data and users. Another advantage of the GBV system is the sustainability: GBV is well funded, well established in the library community and is sustainable in its own. The use of GBV technologies would also save BHL-Europe development resources. By saving these resources we can use them to help our partners with data enhancement / semantic enrichment to better meet the needs of our users.

It is not worked out in detail yet, how and to what extent we can integrate the GBV technologies with the entire BHL-Europe system. As the GBV software is not open source, we still need to develop, adapt and implement deduplication tools for the BHL-Europe Portal itself in a separate step.

In conclusion, getting the required features into the GRIB platform build on GBV technologies is not a matter of technical feasibility but of system definition. We are confident that the GBV system can meet all needs BHL-Europe has for GRIB. Therefore, we have decided to continue the cooperation with GBV and build D2.3 based on this platform.



Search | Results | Advanced | My shelf | My profile | Help

search [and] [ALL] all words sort by year of publication

search approximate search

User id: VZG | logout

about the database | welcome Copyright © 2010 OCLC

Welcome to the GBV demonstrator for the EDIT/BHL-Europe GRIB

This demonstrator, a joint effort of EDIT/ViTaL and BHL-Europe, built by the Common Library Network GBV (Gemeinsamer Bibliotheksverbund) allows users to browse and search titles held by EDIT, BHL-Europe and BHL member institutions (Test sets from FUB-BGBM, MN, NAT and NHM at the moment) using advanced filtering. It will be further developed and extended towards the Global References Index to Biodiversity (GRIB). It will then be a tool to manage the taxonomic literature that is (a) already available in digital form, (b) in the process of being digitised, and (c) for which plans have been created for digitisation.

EDIT European Distributed Institute of Taxonomy

BHL Biodiversity Heritage Library EUROPE

Please send your comments to Boris Jacob, contact: [jacob.boris\[at\]gmail.com](mailto:jacob.boris[at]gmail.com)

Notes:

- This database is presented in unicode (UTF8). Possible character coding problems might be related to locally missing character fonts (Fonts) in your operating system. The following links could be helpful: [Display Problems \(Unicode Consortium\)](#) [Setting up your browser for Unicode support \(besz\)](#) [UTF8 problems \(Wikipedia\)](#)
- Mozilla Firefox version 3.0.x cuts off the right side of the GSD pages or shows a black border below the window contents. If the zoom option of the browser is used. This problem can be solved by activating the option in your browser View > Zoom > Zoom Text Only
- The Internet Explorer 6 doesn't show the logos of the GSD-databases right. You can avoid this by using the IE7, IE9 or any other browser.

Figure 6. GVK BHL-Europe Union Catalogue: <http://kavia0.gbv.de/DB=1.83/SET=1/TTL=1/LNG=EN/>.

¹ <http://www.carpet-project.net/en/tools-and-services/page/8/carpet/production-level-goobi>

Search | **Results** | Advanced | My shelf | My profile | Help DE

DEMO User id: VZG | [logout](#)

search [and] [ALL] all words [?] sort by year of publication
ppn 0? approximate search

search history | **shortlist** | full title Copyright © 2010 OCLC

results search [and] ppn 0? | 156370 hits [analyse set](#) | [save/print info](#)

1. Flora Europaea check-list and chromosome index
Moore, David Moresby. - Cambridge [u.a.] : Cambridge Univ. Press, 1982
2. Revue générale des écrits ... ; 2 ; 1789
/ Pulteney, Richard. - 400 S.
3. [Einführungsh.]
/ Jaeger, Wolfgang. - 40 S.
4. Revue générale des écrits ... ; 1 ; 1789
/ Pulteney, Richard. - VI, 386 S.
5. Anatomie der blütenlosen Pflanzen : Bakterien, Algen, Pilze, Flechten, Moose und Farnpflanzen
Gerlach, Dieter. - Stuttgart : Franckh, 1962
6. Geologie von Europa ; Bd. 2, Das außeralpine Westeuropa, Bd. 4, Arthropoda: Insecta ; Morphology and systematics (Elateroidea, Bostrichiformia, Cucujiformia partim)
/ Bubnoff, Serge von. - 2010
7. Esperiana ; 15
2010
8. Mycotheca Graecensis ; Fasc. 23. = (Nr. 441 - 460) ; Dupla Graecensia Fungorum (2010, Nr. 101 - 200)
/ Scheuer, Christian. - 2010
9. Spezielle Zoologie ; 2 ; Wirbel- oder Schädeltiere
/ Angermann, Renate. - 2. Aufl. - 2010
10. Rhododendrons in India
Sastry, A. R. K.. - [s.l.] : BS Publ., 2010

Word	Type	Count
0?	[[PPN] Pica prod. no]	[156370]

first | previous | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | next | last |

Figure 7. Search for all 156,370 test data currently in the GVK BHL-Europe Union Catalogue.


Search | **Results** | Advanced | My shelf | My profile | Help DE

DEMO User id: VZG | [logout](#)


search [and] [ALL] all words [?] sort by year of publication
ppn 0? approximate search

search history | shortlist | **full title** Copyright © 2010 OCLC

results **Übersicht der Brutbestandsentwicklung ausgewählter Vogelarten 1900 - 1990 an der niedersächsischen Nordseeküste** | 1 hits [loan request](#) | [order copy](#) | [analyse set](#) | [save/print info](#)

 PPN: 000367133
Title: Übersicht der Brutbestandsentwicklung ausgewählter Vogelarten 1900 - 1990 an der niedersächsischen Nordseeküste / Katja Behm-Berkelmann und Hartmut Heckenroth
Author: Behm-Berkelmann, Katja ; Heckenroth, Hartmut
Edition: 1. Aufl.
Published: Hannover : Niedersächs. Verwaltungsamt, Fachbehörde für Naturschutz, 1991
Extent: 97 S. : Ill., graph. Darst., Kt. ; 30 cm
Series: Naturschutz und Landschaftspflege in Niedersachsen ; 27
Note: Met lit. opg
ISBN: 3-922321-62-3

Subject heading: *Deutschland / Nordseeküste / Brutvögel

 [Zitierlink](#) [Über den Zitierlink können Sie diesen Titel als Lesezeichen ablegen oder weiterleiten]

Institute(s): Please click on a institute name or on >

Berlin, Freie Universität Universitätsbibliothek <188>
 Leiden, Naturalis Library <4792>

Please note that not all material is available for loan.
subito supplying libraries are represented in red

Figure 8. Deduplicated titles with holdings information in the GVK BHL-Europe Union Catalogue.

4.2.3 Metadata repository of page images

At <http://bhl.ait.co.at>, a DISMARC port was created for BHL-Europe with a first draft of an appropriate metadata set with the same base functionality. Metadata from various BHL-Europe content provider institutions were mapped against the MODS schema and inserted into the demo server. To date the test server contains more than 40,000 records from 14 institutions. The mapping of this data was carried out using tools already developed during the DISMARC-project or using ETL-tools (extract-transform-load) available in the open source domain (Pentaho). The same technologies will be used to map this data into an ESE (Europeana Semantic Elements) compliant schema.

The BHL-Europe version of the DISMARC port has currently two functions. First, it serves as a test environment for the data of our content provider. Thus, it helps to identify and refine the requirements for data quality and metadata structure. This is important for the development of the German prototype as these data are related to the actual page images for the portal. The work on the data using the test environment is also important to refine the specifications for the GRIB development. The second function of the test environment is the link to Europeana. As BHL-Europe has not finished the set up of its own hardware infrastructure, we are working with the AIT infrastructure to prepare our data for the Rhine release of Europeana this summer.

4.3 Work plan towards BHL-Europe D2.3

From 1 March 2010, the existing GRIB prototype build by GBV will be enhanced and extended to have the functionality we need for D2.3. We need to have a database of the taxonomic literature that indicates (a) the portion that is already available in digital form, (b) the portion that is in the process of being digitised, and (c) the portion for which plans have been created for digitisation.

In a first step, BHL metadata will be deduplicated and merged with the library metadata implemented into the GBV GRIB solution from BGBM, MfN, NAT and NHM. That means all of the literature occurring both in those four libraries and in BHL are then represented in the GRIB with metadata as well as a link to the BHL metadata and thus to the BHL Portal. Those metadata entries within the GRIB with a link to BHL are the portion of content that is already available in digital form.

In general those metadata entries within the GRIB without a link to BHL are the portion for which plans have been created for digitisation. This is due to the fact, that GRIB has the aim to show all of the literature relevant to biodiversity.

This general perspective will be specified by scan requests done by users on the one hand and by a bid from libraries to scan a certain publication on the other hand. A scan request does not necessarily lead to a process of scanning. That one is started by the bid of the library. Literature that is bidden on thus is the portion that is in the process of being digitised.

The full functionality (scan requests, bidding, etc.) will be provided through a web interface. It will thus allow to support the analysis of domain content and management of the scanning process.

All the development and implementation work mentioned above will be done in close cooperation with our partner libraries and technical development team (WP3) to meet their requirements. A meeting is scheduled for end of March with all content providers involved to further review GRIB and specify the system.

5 Content analysis status

The complete content analysis and management status report (metadata, page numbers, content providers) for the first year is expected as D2.4 in April 2010. In this section, I would like to highlight some examples to illustrate the ongoing work in this area to date. This mainly focuses on new content already connected to BHL and BHL-Europe and the analysis of required content.

The colleagues from UH-Viikki establish a RefShare database to identify items for scanning: <http://www.refworks.com/refshare/?site=014941135929600000/RWWS4A1009351/074281243853638000>. The RefShare database is based on the local library catalogue, and the team in Helsinki came up with a set of search criteria, for instance a Finnish publisher, subject area etc. Then two people from the library transferred the records to RefWorks and the folder was shared for commenting. Primarily certain researchers of the campus and the Finnish Museum of Natural History were asked to comment. The current list is extensive, so some items on the list will probably not be digitised in near future.

The colleagues at RMCA and NBGB are investigating also items for scanning operations in cooperation with their local scientists. Each of them established a wiki to collect books and journals that the scientists need to see online in priority in their field and which they have not found in digital format yet. The items were added to the wiki and it was searched and verified if this item really is not available online.

The RMCA wiki is accessible under: http://193.190.223.46/wiki_ext/index.php/List_RMCA

The NBGB wiki is accessible under: http://193.190.223.46/wiki_ext/index.php/List_NBGB

In October 2009, BHL finished the ingestion of biodiversity related content of the entire Internet Archive corpus from non-member BHL libraries into the BHL Portal. This increased the BHL corpus by 33,000 volumes to now more than 70,000 volumes of literature. As IPR and licensing questions still needs to be discussed with those libraries, this new content is not available for BHL-Europe and Europeana now, but hopefully before the end of BHL-Europe.

One important aspect of WP2 is to attract new content providers (Task 2.3.2). BHL-Europe has to enlarge its network of content provider from 16 to 30 at the end of the project. To date we are negotiating with the university library Bielefeld. This institution is running a project focused on the digitisation of German language journals from the 18th and 19th century (<http://www.ub.uni-bielefeld.de/diglib/aufklaerung/index.htm>), including biodiversity content (<http://www.ub.uni-bielefeld.de/diglib/aufkl/naturforscher/naturforscher.htm>). The Humboldt University Berlin, one of our project partner but no content provider so far, is also running a repository with very valuable biodiversity content (rare material), e.g. <http://edoc.hu-berlin.de/ebind/mfn/keller1-2005-Mn01674331/XML/index.xml>. We have to figure out how to integrate this content into BHL. Eventually, the colleagues from Madrid (CSIC) are approaching their partner institutions to provide content for BHL-Europe. The Real Jardin Botánico is the first of these institutions that has shown interest in sharing their biodiversity content with BHL-Europe. Further details of this sharing are still in discussion.