ECP-2008-DILI-518001

BHL-Europe

Prototype of Web-database for content management and collection analysis

Deliverable number	*D2.3*
Dissemination level	*Restricted*
Delivery date	*3 May 2010*
Status	*Final*
Authors	*Henning Scholz, Boris Jacob, Wolfgang Koller, Jakob Voss*

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.

0 Document History

0.1 Contributors

Person	Partner
Henning Scholz	MfN
Boris Jacob	MfN
Wolfgang Koller	NHMW
Jakob Voss	VZG

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
16 April 2010	Henning Scholz	0.1	Initial draft version
30 April 2010	Jakob Voss	0.2	Details on BHL Scan List
30 April 2010	Wolfgang Koller	0.2	Details on GRIB digitisation management
02 May 2010	Boris Jacob	0.3	Notes and questions for further review
03 May 2010	Boris Jacob	1.0	Final version incorporating reviews

0.3 Reviewers

This document requires the following reviews and approvals.

Name	Position	Date	Version
Andreas Kohlbecker	EDIT Activity 5.8 Leader, FUB-BGBM	03 May 2010	0.3
Alexander Struck	Information Engineer, Ovid Technologies GmbH	03 May 2010	0.3
Dennis Zielke	BHL-Europe WP3 Team member, UBER	03 May 2010	0.3

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
Andreas Kohlbecker	02 May 2010	0.3
Alexander Struck	02 May 2010	0.3
Dennis Zielke	02 May 2010	0.3

1 Table of Contents

2 Purpose

To provide an overview of the two available web-databases for content management and collection analysis for managing bibliographic data by BHL-Europe content providers – the *BHL Scan List* and the further developed *GRIB prototype* – and to show their specifications and the plan to combine the systems into one live system: the Global References Index to Biodiversity (*GRIB*).

3 Background

BHL-Europe will support the digitisation and manage the acquisition and hosting of the digitised material contained in BHL-Europe partner institutions. The project will assist in the process by which each institution digitises its biodiversity material so it is done efficiently and effectively. Several tools and databases are being established by BHL-Europe to analyse the content and support the management of the scanning initiatives of each partner.

Two tools for content management and collection analysis exist right now. One is the *BHL Scan List*, hosted and maintained by the NHMW in Vienna. The other one is the *GRIB prototype*, developed, hosted and maintained by the VZG (Head Office of the Common Library Network GBV).

The *BHL Scan List* originates from the *Biodiversity Heritage Library Serials Union Catalogue* (aka *Serial Bidlist*) developed at NHM London. For BHL-Europe D2.2 this system was mirrored at NHMW, and changed to fit the needs of BHL-Europe, becoming the prototype of the *BHL-Europe Union Catalogue*. This prototype is now transformed into a life system by NHMW replacing the old *Serial Bidlist*. The *BHL Scan List* from now on serves both BHL-Europe and BHL content providers. The *BHL Scan List* is intended as an interim solution. As soon as the development of the bibliographic database *GRIB* is finished, its purpose will be further evaluated.

In the meantime and as described in BHL-Europe D2.2 as well, work on the prototype of the bibliographic database *GRIB* continued in cooperation with the VZG. For D2.3 the prototype had been enhanced by a digitisation management tool (API and widget). The *GRIB* will at a certain stage incorporate the necessary functionalities of the *BHL Scan List*. It will be further developed within the BHL-Europe technical architecture to meet the needs of the BHL-Europe partners, BHL and EDIT (see Fig. 7 Current architecture diagram and D3.5).

4 BHL Scan List

The *BHL Scan List* http://bhl.nhm-wien.ac.at/scanlist/ is a tool for BHL-Europe and BHL member institutions to coordinate their scanning efforts. It is capable of the basic functionalities like automatic and manual deduplication, coordinating the digitisation process and handling serials and monographs.

In addition, the *Schema Mapping Tool*[2] can be used to semi-automatically ingest new catalogues into the list. A basic deduplication routine based on ISSN matching is already implemented in a stored MySQL procedure.

Browsing and searching the records within the *BHL Scan List* can be done without logging in. For bidding on items, a valid account is required. You can also log in as a test user:

Username	**eurcon**
Password	**risiga35**

Logging in (see Fig. 1 Login Form) can be done using the "Login" link in the left menu or at http://bhl.nhm-wien.ac.at/scanlist/index.php/users/login. The user will be asked to enter username and password. After filling in the form and clicking on the "Login" button, the user will be logged in.

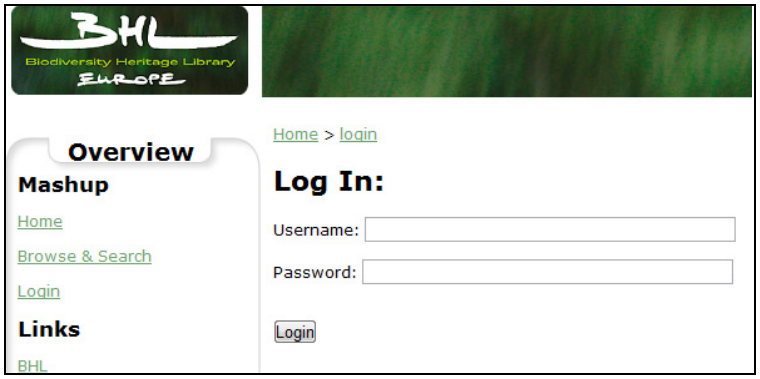

Fig. 1 Login Form

4.1 Searching for metadata records

Searching for a certain metadata record can be done using the "Browse & Search" functionality of the *BHL Scan List*. Once clicked, a list with all titles within the list will open. Using the filter functions, the database can be searched (see Fig. 2 Search Filter).

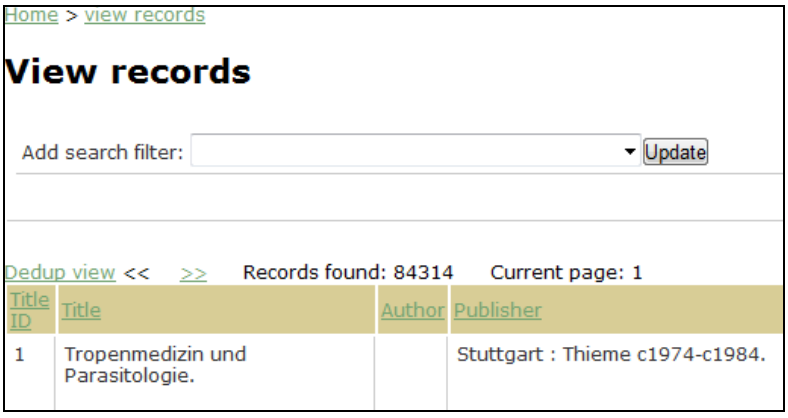

Fig. 2 Search Filter

[2] Schema Mapping Tool: http://code.google.com/p/schema-mapping-tool/

To add a criteria, simply select the field to search on (in the example below it is the “title” field). Once selected the page will automatically reload and a new line with an input box for the criteria will appear. Inside the input box the title (or part of the title) can be entered (see Fig. 3 Title Searching).

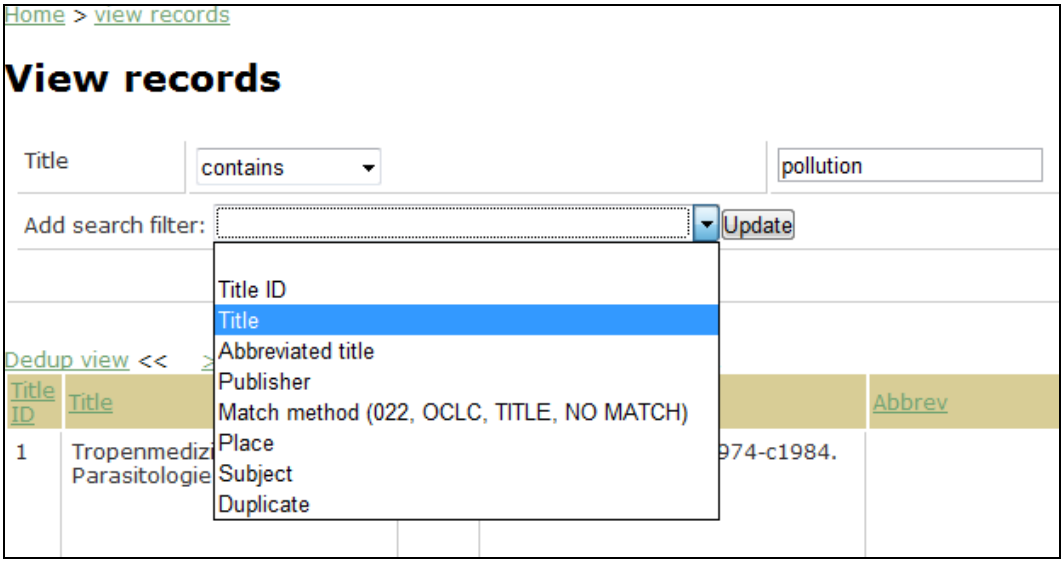

Fig. 3 Title Searching

Once the search string has been entered, a click on the “Update” button will filter the list.

Different criteria can be combined, even the same criteria can be repeated (as an example, both a title-filter for “pollution” and for “environment” can be added). The entered search strings are not case sensitive.

4.2 Bidding on an item

Once the correct item has been found, a bid can be placed on it to inform all other users that the institution will scan this item (or parts of it).

To place a bid, the detail view for an item has to be opened. To do so, the “more” link in the “Actions” column of the item-list has to be clicked.

Inside the detail view additional information for this item can be found (like the ISSN). Below the item details, a list of holdings is displayed. The holdings list contains an entry for each institution owning a copy of the item. This is to provide an overview of who has what (see Fig. 4 Holdings Information).

Holdings information

Owner	BHL location	Local ctrl	Holdings	Subject	e-access?	Match basis	OCLC
	AMNH	5900905_b108110x	QL353.I61.1-20 1939-1959. 1-20 1939-1959	Animals Nomenclature.		TITLE	5900905_b108110x
	FIELD	000037834	FIELD MUSEUM MAIN LIB. QL353 I5-o1-20(1939-1959)	Zoology Nomenclature Nomenclature.		TITLE	(OCoLC)5900905
	HAR	000150768	v.1-20(1939-1959) 1.1 MH-Z QL353 .I62x			TITLE	(OCoLC)05900905
	NHM	(Sirsi)_a53662	 1939-1954 [P] ; 1939-1959 [T] [Z] ; 1943-1959 [E].Palaeontology Library holdings incomplete. Zoology missing vol. 1A (1943-1955)			TITLE	
	SIL	128720	mamm mammref QL353 .I612v.1 (1943-1947) fish fishmai QL353 .I612v.1=opinions 1-16 moll mollref QL353 .I612v.2:sect.A (1939-1945) ent entref QL353 .I612v.2:sect.A:no.134-160 (1939-1945) invz invzmai QL353 .I612v.2:pt.12-1720-2123-2830-51 (1943-1947) rept reptref QL353 .I612opinions 240-455490-565 (1954-19571957-1965) mamm mammref QL353 .I612v.2 sect.A (1939-1945) ent entref QL353 .I612v.2:no.161-181 (1945-1947) fish fishmai QL353 .I612v.2:spc.9:pt.30a-v.5=opinions 134-282 mamm mammref QL353 .I612v.2 sect.B ent entref QL353 .I612v.3-9 (1944-1954) fish fishmai QL353 .I612v.7 mamm mammref QL353 .I612v.3-4 fish fishmai QL353 .I612v.8-9=opinions 284-317	Animals Nomenclature		TITLE	(OCoLC)ocm05900905

Fig. 4 Holdings Information

To actually bid on an item, the bid information has to be checked. It can be found below the holdings information.

There are two types of bids: the “partial” and the “full” bid. A partial bid means that only parts will be scanned (for example volume 1 to 12 out of 49). A full bid on the other hand means, that the entire serial will be scanned (attention: serials contain several volumes, most likely not all volumes will be scanned. Placing a full bid on a serial indicates that all volumes will be scanned).

Once it is clear what parts will be scanned, either clicking on the “Bid for all” or “Partial bid” will open the bid interface (see Fig. 5 Add Bid).

Fig. 5 Add Bid

In the bid form, a start and an end date can be selected. This is important for serials so that it is clear what parts will be scanned. Inside the Notes field, the volumes should be noted (for example “Volume 1-12”). If the holdings should contain a gap or can not be scanned for any other reason, it should be noted inside the Exceptions field (for example “gap Volume 8-9”).

Once all information has been entered, the bid can be added using the “Submit bid” button. Once clicked the bid will be saved in the database and will be visible for all other users.

The *BHL Scan List* also helps keeping track of the scanning process itself. The digitisation status in the bid information list (on the item detail page) indicates if the scanning already has been done or if it is still pending. Once the scan has been processed, the bid should be updated to reflect the change. This can be done by editing the bid and changing its status to “Complete”.

4.3 View existing bids

To get an overview of the existing bids, it is possible to show a list of them. By clicking on “View existing bids” a list of all bids in the system can be displayed. The list contains information about the bidding institution, the status of the bid and the title of the item the bid was placed on (see Fig. 6 List of existing bids).

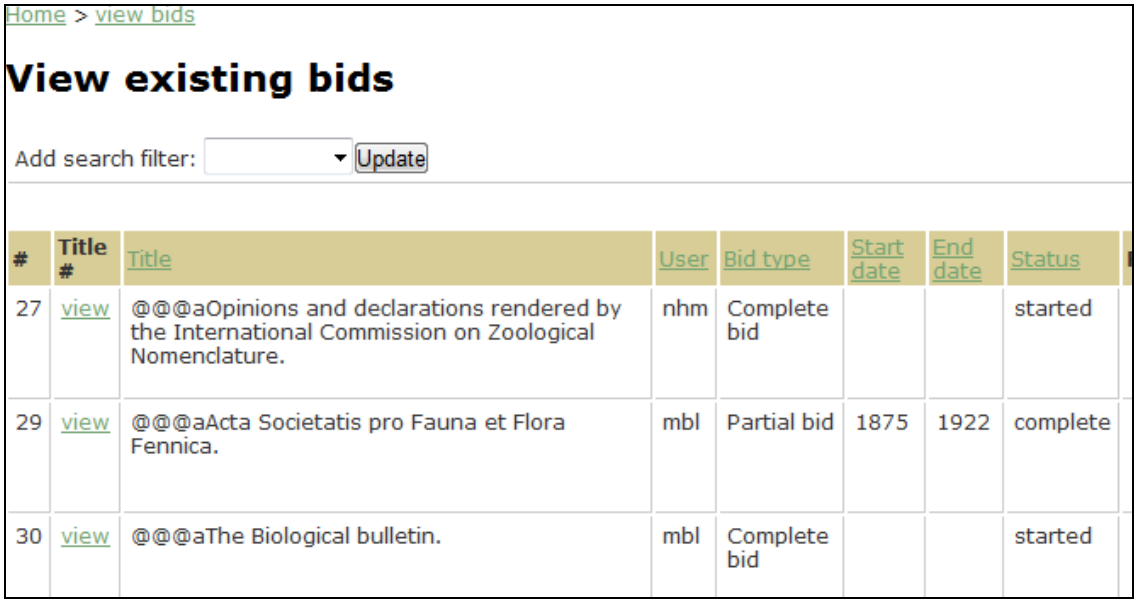

Fig. 6 List of existing bids

The list can be filtered using the same method as described above. However, the available search criteria are different. By using the search criteria it is possible to display who has scanned what etc.

By adding a “User ID” search filter, it is possible to show a list of my own bids by adding my username as search string (for details see above).

5 GRIB prototype

The Head Office of the Common Library Network GBV (VZG) creates the *GRIB prototype* which is located at a test database at http://kavia0.gbv.de/DB=1.83/ (automated log in as user: VZG).

The *GRIB* is part of the BHL-Europe technical architecture (see Fig. 7 Current architecture diagram, and D2.2 & D3.5 for details) and communicates with other parts in different, to be specified ways. Those related to the content provider are:

1) Catalogue building,
2) Scanning planning,
3) ID-Reporting and – 4) being an internal content provider process –
5) Pre-ingest.

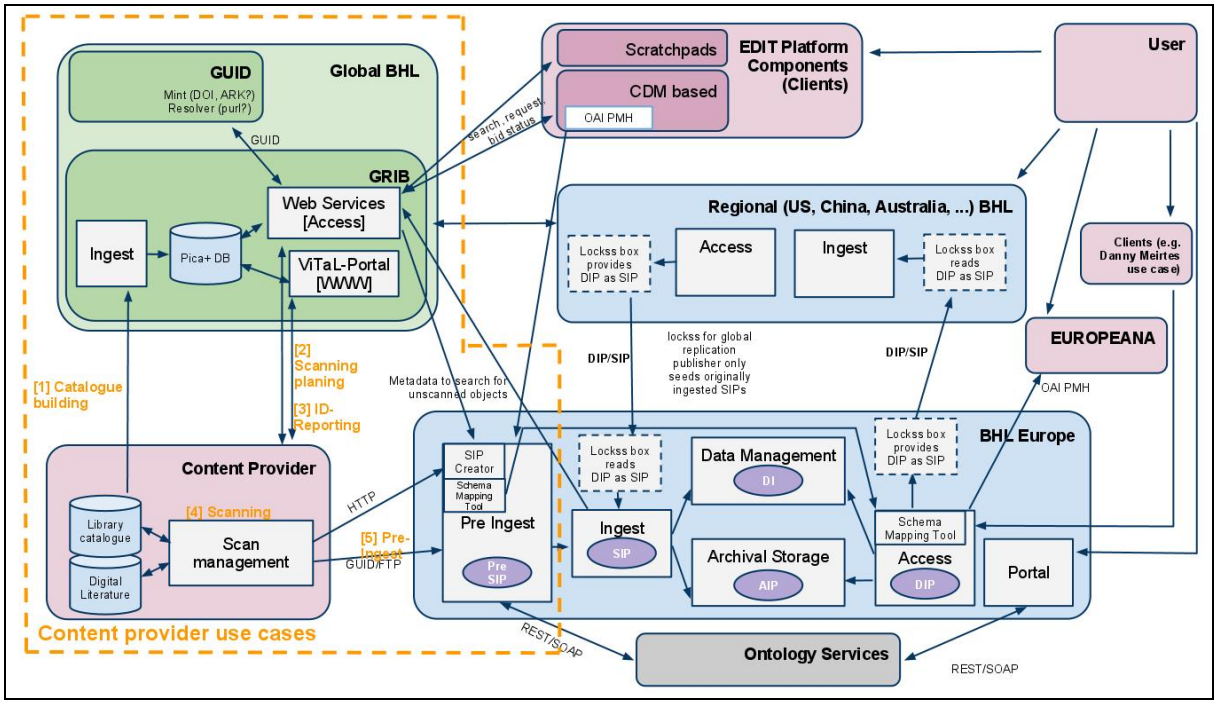

Fig. 7 Current architecture diagram & Content provider use cases

Once the bibliographic data from the BHL-Europe partner library catalogues have been send to the VZG, there is a standard method to ingest new metadata records into the record store (Pica+ DB). On ingest each record gets assigned at least one unique identifier (the Pica Production Number PPN) that is used to later access and modify it.

In addition, there are several REST-based APIs (Web Services) for the following use cases:

- Search and browse the record collection: Standard library catalogue user via ViTaL-Portal (WWW).
- Export selected metadata records in various formats: unAPI
- Interface search records with given criteria (for instance all records that have been marked for digitisation): SRU/SRW

5.1 Digitisation Management

Since introduced in D2.2 the *GRIB prototype* has been enhanced by a digitisation management tool consisting of a customized JSON-API to modify the digitisation status of a record and a JavaScript-based Widget (see Fig. 8 Digitisation management widget, e.g. http://ws.gbv.de/bhleuropa/_testbutton.php?ppn=001315714) that can be embedded in the catalogue user interface (see Fig. 9 Full title view with digitisation management widget) and in any other web application. Alternatively, applications like BHL-Europe OAIS, CDM based EDIT platform components or the EDIT Scratchpads, and content providers can directly access the JSON-API.

Fig. 8 Digitisation management widget

A bibliographic record within the GRIB prototype can have five different status at the moment: 1) Not digitised, 2) Should be digitised, 3) Digitisation in progress, 4) Digitisation completed and 5) Document available.

Depending on the role of the user, there are different possibilities to mark a record within the five different status. If you log in as content provider you can change the status within the given five possibilities. Logged in as a scientist, you can only change the status to: 1) Not digitised and 2) Should be digitized. This is to give an example, the status and the role management will be adapted after clarification of the workflow.

Log in as content provider	Username: **EDIT**
Log in as scientist	Username: **PUBLIC**

The JSON-API is a simple REST-API which means that all parameters are provided as URL parameters added to a base URL. The only exception is authentication which is currently based on a session cookie until the user management framework has been defined. The URL parameters are:

- Id: A unique record identifier. Up to now only the "PPN" id as assigned on ingest is supported. It must be prefixed by "bhleuropa:ppn:", for instance id=bhleuropa:ppn:1234 to refer to the record with PPN identifier "1234". The same identifier can also be used in the unAPI API to get single metadata records.
- Status: The new digitization status. Must be a value from a predefined list. This status can also be used to search for records in the user interface and via SRU/SRW.
- Callback: A JavaScript callback method name.

The response format is a simple JSON object with the following fields:

- Id: The unique record identifier
- Status: The digitization status of the record
- Options: A list of digitization status which the current user is allowed to set
- Error: An error message (only set if the action failed). Additional parameters and response content may be added on request.

After final clarification of the workflow the API may also support setting an URL of the digitised document and showing the user which has last modified a digitisation status.

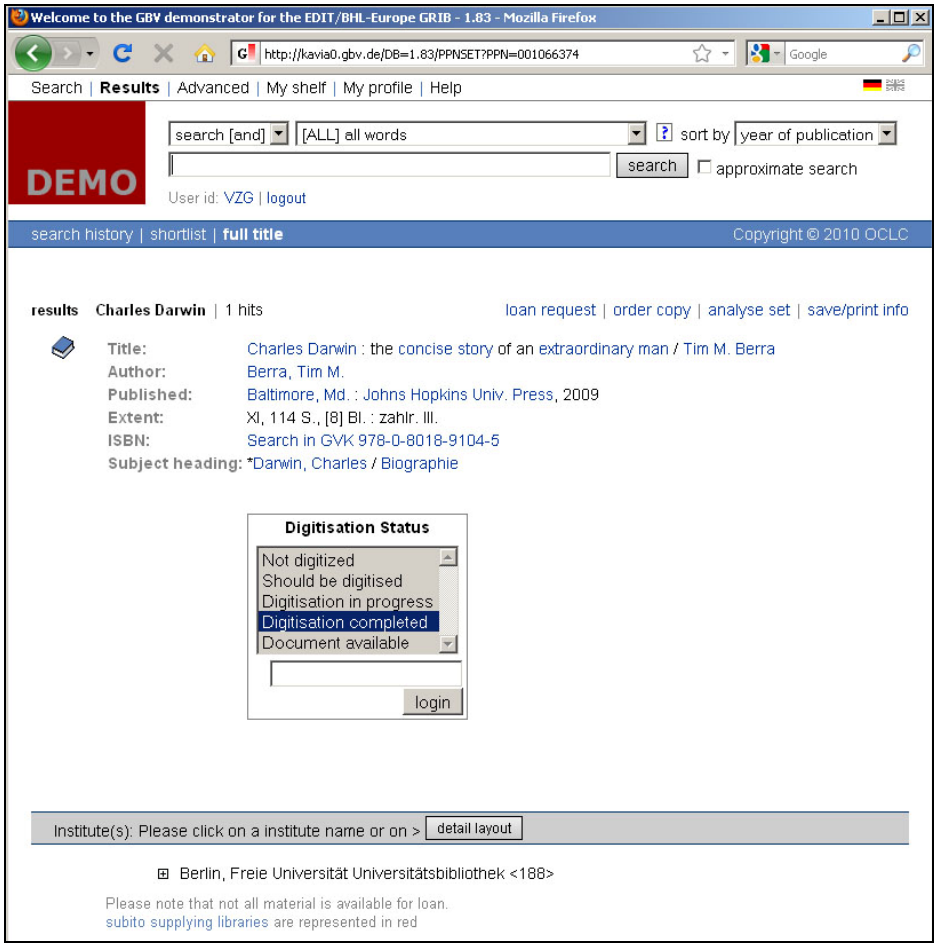

Fig. 9 Full title view with digitisation management widget

The Widget can be installed by adding two pieces of HTML to any given website (see e.g. Fig. 9 Full title view with digitisation management widget): First you must include the Widget as JavaScript:

```
<script type="text/javascript" src="WIDGETURL"></script>
```

Second you can add one or more HTML elements of the CSS class "bhleuropa-digitize" that have a record identifier in the "title" or in the "value" attribute. Examples:

```
<span class="bhleuropa-digitize" title="bhleuropa:ppn:1234"></span>
<button class="bhleuropa-digitize" value="bhleuropa:ppn:1234">digitize status</button>
```

The elements are then automatically replaced by an instance of the Widget.

5.2 Collection analysis

Besides the SRU/SRW interface there is a possibility to search records that have been marked within the digitisation management via the *GRIB* Web-search interface with the following search terms: Not digitised “dst 8300”, Should be digitised “dst 8301”, Digitisation in progress “dst 8302”, Digitisation completed “dst 8303”, Document available “dst 8303”.

The standard functionality to analyse metadata records within the GRIB prototype works for: 1) Material code (Fig. 10 Analyzing all 156.370 metadata records for material code), 2) Language code and 3) Country code. It is accessible via the link “analyse set” in the top right corner of each metadata record or search result list. If you search the *GRIB* with the term “ppn=00?” you retrieve a list of all existing metadata records, which you can analyse the same way. Please keep in mind that right now the *GRIB prototype* is fed with a test set of metadata records only.

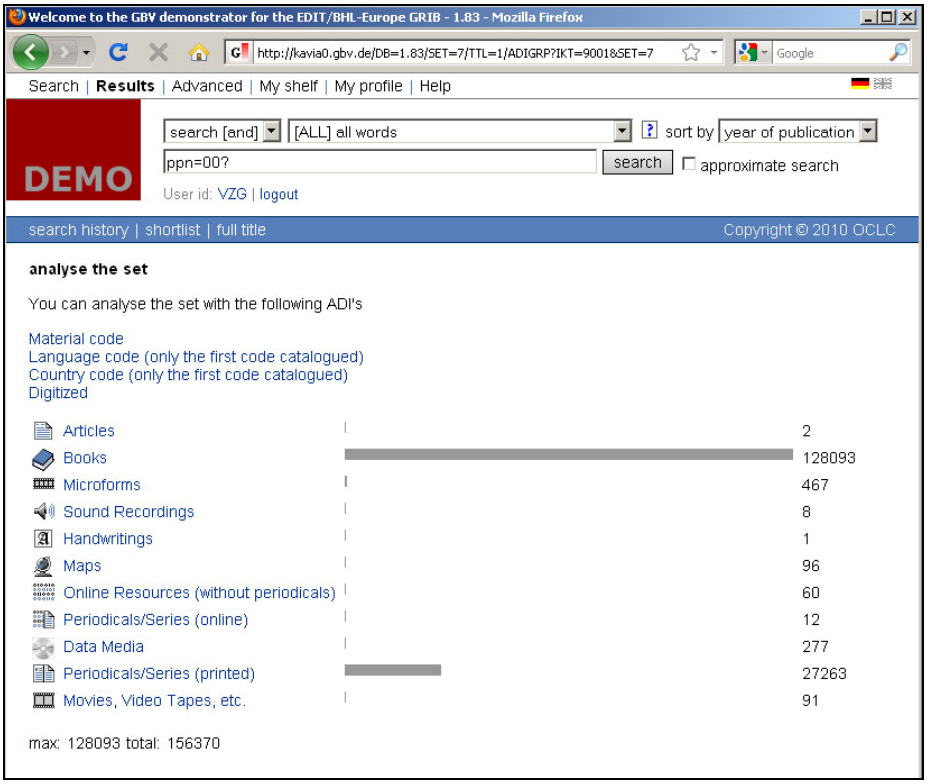

Fig. 10 Analyzing all 156.370 metadata records for material code