ECP-2008-DILI-518001

BHL-Europe

Final and enhanced Web-database for content management and collection analysis

Deliverable number	*D2.5*
Dissemination level	*Restricted*
Delivery date	*30 April 2011*
Status	*Draft*
Author(s)	*Boris Jacob*

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
Boris Jacob	RMCA

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
18 Mar 2011	Boris Jacob	0.1	First draft: ToC with short descriptions
05 Apr 2011	Boris Jacob	0.2	Second draft
27 Apr 2011	Boris Jacob	0.3	Third draft incorporating comments
27 Apr 2011	Boris Jacob	0.4	Fourth draft incorporating comments
	Boris Jacob	1.0	Final version incorporating reviews

0.3 Reviewers

This document requires the following reviews and approvals.

Name	Position	Date	Version
Constance Rinaldo	Academic Librarian	27 Apr 2011	0.4

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
Melita Birthälmer, Henning Scholz	18 Mar 2011	0.1
Melita Birthälmer, Henning Scholz	05 Apr 2011	0.2
Melita Birthälmer, Henning Scholz	27 Apr 2011	0.3
BHL-Europe consortium	27 Apr 2011	0.4

1 Table of Contents

2 Purpose

This document describes the functionalities of the Global References Index to Biodiversity (GRIB) as the final and enhanced web-database for content management and collection analysis within the Biodiversity Heritage Library for Europe project (BHL-Europe). It includes and updates information of previous deliverables and documentations. For comments or questions regarding the GRIB, please send an e-mail to: grib@bhl-europe.eu

3 Background

3.1 Introduction to the GRIB

Natural history museums and botanical gardens collectively hold the majority of the world's published knowledge on the discovery and subsequent description of biological diversity. BHL-Europe assists in making these holdings available online and one tool to do so is the Global References Index to Biodiversity[2] (GRIB).

The GRIB is the joint literature index from the projects Biodiversity Heritage Library for Europe[3] (BHL-Europe) and European Distributed Institute of Taxonomy[4] (EDIT). It was build and is maintained together with the Head office of the Common Library Network GBV (VZG[5]).

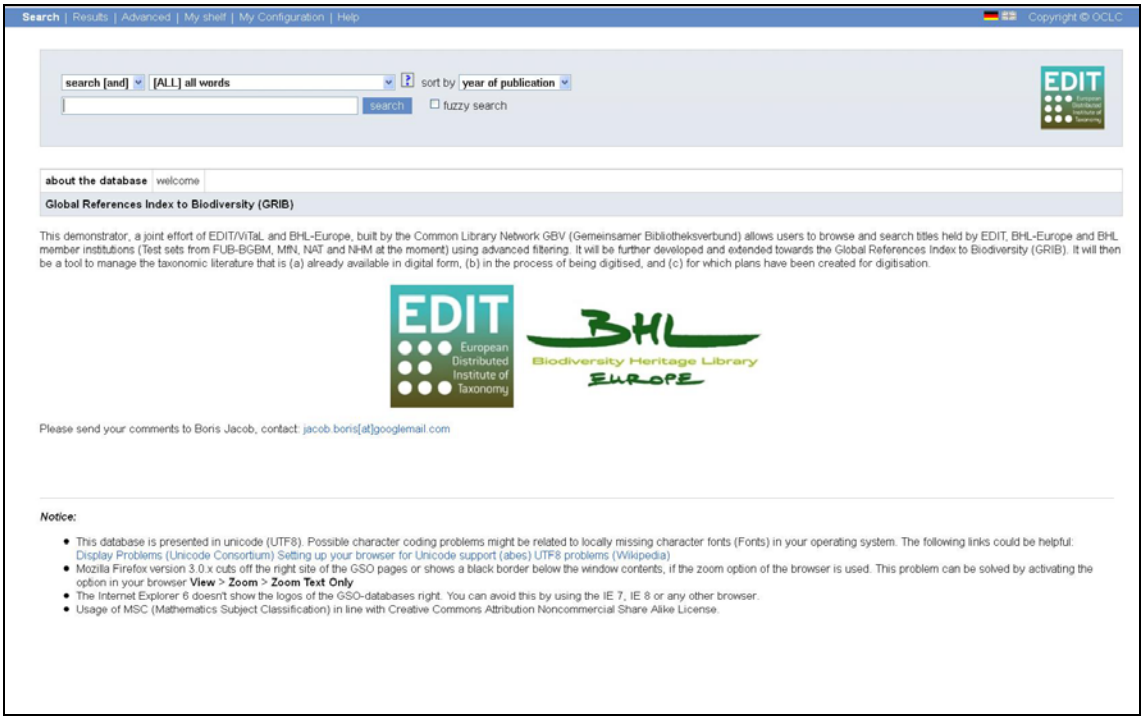

Fig. 1: Homepage of the Global Reference Index to Biodiversity: http://grib.gbv.de

[2] GRIB homepage: http://grib.gbv.de/LNG=EN/
[3] BHL-Europe homepage: http://www.bhl-europe.eu/
[4] EDIT homepage: http://www.e-taxonomy.eu/
[5] About GBV and VZG: http://www.gbv.de/vgm/info/biblio/01VZG/

The index incorporates bibliographic data from library catalogues from natural history institutions being part of these projects, and thus shows, where biodiversity related legacy publications are held in Europe. The data in the GRIB is de-duplicated which makes it easier to search and retrieve the publications. It also links through to already digitised publications found in the Biodiversity Heritage Library[6] (BHL-US) and later in the BHL-Europe portal. For those publications, that haven't been digitised yet, the GRIB offers the possibility to mark them for future digitisation.

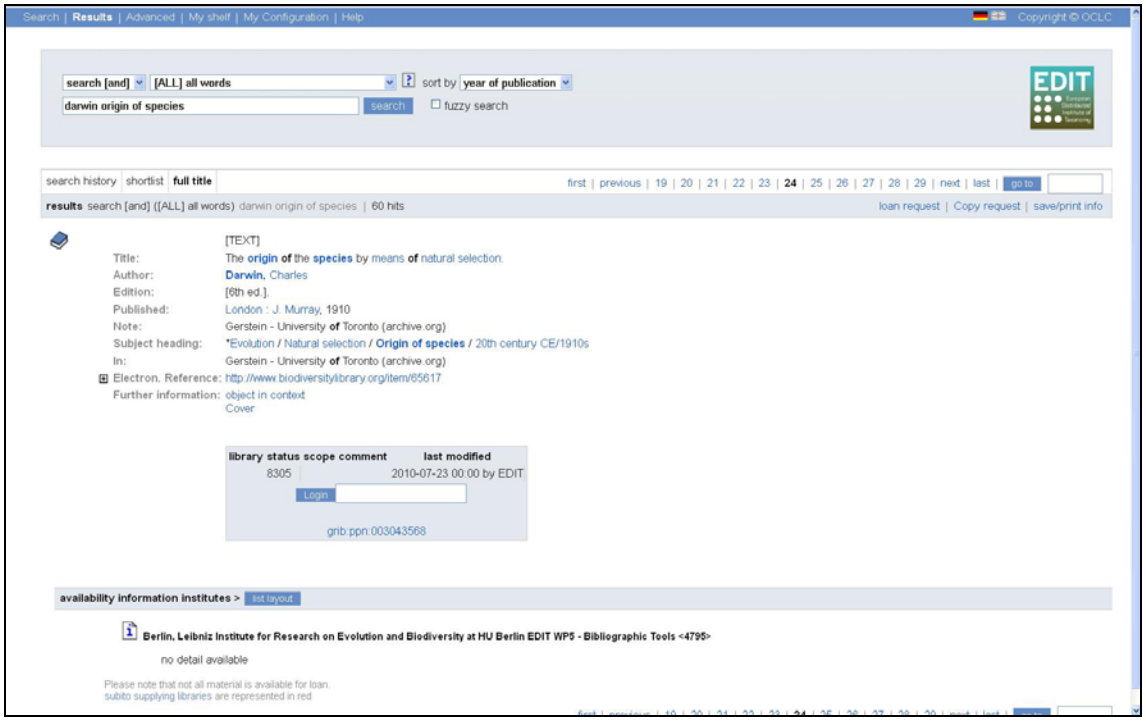

Fig. 2: Screenshot of a dataset in the GRIB

The GRIB is supposed to serve two groups of users: Content users[7]/Scientists and Technology users[8]/Librarians. Scientists in the need for online access to digitised literature can search the GRIB and nominate titles for digitisation, that are not available online yet. Librarians on the other hand can use the tool to indicate the digitisation status of a title, within their libraries digitisation workflow.

3.2 Content users/Scientists

Content users/Scientist can use the GRIB portal and open web services to search, retrieve and download bibliographic information. Unlike in the BHL-Europe portal, the user will not only find literature that is available online on BHL-US and BHL-Europe later on, but also references to publications held in European natural history institution libraries. This allows scientists to get relevant results in their specific domain.

[6] BHL homepage: http://www.biodiversitylibrary.org/
[7] See BHL-Europe DoW Version of 27/02/2009 p. 18ff.
[8] See BHL-Europe DoW Version of 27/02/2009 p. 19f.

3.3 Technology users/Librarians

Technology User/Librarians of BHL-Europe, EDIT and BHL partner institutions provide their bibliographic data to build the GRIB and use the index for content management and collection analysis in regard to their digitisation activities.

For each bibliographic item of the partner libraries the GRIB holds information on its digitisation status. In the process of planning scanning activities, librarians can check if a publication is already digitised or planned to be digitised by another library and thus plan their digitisation projects on a word-wide level to avoid duplicate scanning. A publication in the GRIB can either be:

1. not digitised yet,
2. nominated to be digitised by a scientist (to be activated in the future),
3. intended to be digitised by a librarian, or
4. already digitised and accessible in electronic form.

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.

Libraries, which are associated with one of the partner projects and wish to provide their data to the GRIB can send an email to grib@bhl-europe.eu.

3.4 Administration and maintenance of the GRIB

The administration of the GRIB is done by the MfN (via BHL-Europe WP2) regarding organisational aspects (i.e. conceptualising the index and updating the bibliographic database) and the VZG regarding technical aspects (implementation and maintenance of the GRIB, as well as de-duplication of the bibliographic data). The GBV is a public and non-profit institution, funded by seven northern German federal states and the Stiftung Preußischer Kulturbesitz. It builds and hosts the union catalogue GVK[9], a bibliographic database, comprises the library holdings of the more than 400 GBV member libraries. The GVK, as well as the GRIB, are based on the proprietary Pica-CBS, an OCLC software based on Pica Search and Index (PSI). The VZG is an independent development partner of OCLC, i.e. they have technically and juristically full code access.

Both sides agreed to keep the GRIB updated and maintained until the end of February 2020 at least.

The GRIB is accessible via Standard interfaces (WWW, Z39.50, SRU, internal XML, OAI). The provided library catalogue data stays inside the GRIB. The copyright of the original data is kept with the providing institution. A physical export of the GRIB is possible in agreement with the MfN.

4 Import and update of bibliographic data

The data inside the GRIB will be updated on a regular basis, the timing as well as the method depend on the partners providing the data. Data from a regular partner library catalogue for example has to be updated less often than data from the BHL-US because the Internet Archive (the scanning partner of BHL) is ingesting new data into BHL-US much more frequent, than the library catalogues change their data.

Although different possibilities of providing bibliographic data are possible, the best results in regard to further processing the data have been achieved via manual export in MARC format.

[9] GVK (Gemeinsamer Verbundkatalog) homepage: http://gso.gbv.de/xslt/DB=2.1/LNG=EN/

The method of updating depends on the technological infrastructure of the partner (if the partner system supports the Z39.50[10] protocol or an OAI-PMH[11] interface for example). As well as on the data they provide:

- Bibliographic data from library catalogues,
- Information on subscribed journals from the libraries, and
- Hyperlinks to electronic publications and related bibliographic data from open access repositories.

The mechanism to update bibliographic data into the references index consists of two major steps:

1) The delivery of the data and
2) its conversion, de-duplication and merging.

While method and frequency of delivering the data in step 1) vary on the partners side, there is a fixed process on the side of the VZG in step 2).

4.1 Delivery of the data

The delivery of the data from the data provider to the VZG can take place in three different ways via manual export, Z39.50 harvesting, or OAI-PMH harvesting. The decision on which way to use depends on the providers infrastructure, the internal workflow and the kind of data they are providing.

1. One way to provide library catalogue data is by manual exporting and uploading it to the BHL-Europe's FTP-Server at the Natural History Museum (NHM), from where it is taken by the VZG for further processing. A GRIB folder is provided within the main directory of each content provider on the NHM server for this purpose.
2. Z39.50 is a common standard protocol for search and retrieval in library catalogues. The VZG will make use of it to harvest the library catalogue data. However, the information on subscribed journals are unlikely to be part of this data and will have to be exported manually.
3. Very few library systems provide an open OAI-PMH interface, but for open access repositories it is a standard method to provide access to their metadata. Where ever a system provides this interface, the VZG can harvest data from it.

4.2 Conversion, de-duplication and merging of the data

Once the data has reached the VZG, the process of updating the index is always the same. The data is coming from the providers in different formats (like MARC 21[12] or flavours of MARC, MAB[13], as an spreadsheet or an XML file). It is mapped to the internal GRIB database format Pica+ and stored temporarily into a database (called Signaldatenbank), from where it is further processed.

From the Signaldatenbank the data is transferred and thus updated into the GRIB by a process called Match & Merge. During the matching process different results can occur.

1. One is that a title is already represented by a bibliographic record in the GRIB. In that case any new information will be stored in the already existing record. In case of a library catalogue record the most valuable information are the holdings information, whereas in case of an import of BHL-US data it is the hyperlink to the related electronic full text. This new information is merged with the existing record.

[10] See http://www.loc.gov/z3950/agency/
[11] See http://www.openarchives.org/OAI/openarchivesprotocol.html
[12] See http://www.loc.gov/marc/bibliographic/
[13] See http://www.d-nb.de/standardisierung/formate/mab.htm

2. There is also the possibility that during the matching process a title is identified as a possible duplicate. In that case the record is marked but not merged yet and has to be reviewed and decided on by a librarian.
3. The last possibility is, that a record is new to the GRIB. It will then be imported into the GRIB as a whole new title.

5 Access via Website

The chapter explains the access to the GRIB functionalities from the WWW-interface. It is structured from easy to specific, starting with the simple search for the general use and ending with special functionality for the librarian users.

5.1 Simple Search

Simple search is the standard interface when accessing the GRIB at http://grib.gbv.de/LNG=EN/. It provides a search field next to the "search" button to a search term or phrase, three drop down menus for search options and a "fuzzy search" check-box.

Fig. 3: Simple search interface on the GRIB homepage

5.1.1 Search field

A search term is a sequence of characters without spaces (you cannot use the characters &, |, ~, (), / unless they are preceded by a backslash, \. For example: ups \& downs). If you want to search for a specific phrase, you must use quotation marks, for example: "Darwin species". The search engine is not case sensitive.

Above the search field you will see three drop-down menus in the search bar with which you can modify a search operation.

5.1.2 Search method

In the left drop-down menu you can select one of three search methods:

- Search (or): The standard selection is the search method OR. A search operation with two or more search terms shows all titles that contain at least one of the search terms. For example: darwin OR origin searches for all titles that contain either darwin or origin or both.
- Search (and): In the menu you can also select the AND search method. A search operation of two or more search terms shows the titles that contain all the search terms entered. For example: darwin AND origin searches for all titles that contain both darwin and origin.
- Browse: With the selection Browse you can scan the index for the search terms you have entered.

5.1.3 Search key

In the middle drop-down-menu you can select a search key with which you can further modify a search operation. There is a whole list of search keys, here are some examples:

- [ALL] all words without full text: Select it if you want to search different parts simultaneously, such as title, keyword and summary.
- [TIT] title (keywords): Select it if you want to search the titles of books, magazines, conference proceedings or magazine articles or other material.
- [PER] person/author: Select it if you are searching for a specific author.
- [ISB] ISBN (books): Select it if you have the International Standard Book Number for a specific publication.
- [ISS] ISSN (periodicals): Select it if you have the International Standard Serial Number for a specific periodical.
- [PPN] Pica prod. No: A unique identifier for all records within the GRIB
- [DST] Digitisation Status: You can search the Digitisation status using numerical code or keywords:
  - 8300 = "\Not Digitised"[14]
  - 8301 = "Should be digitised"
  - 8302 = "Will be digitised"
  - 8305 = "Document available"

5.1.4 Sorting the results

With the right drop-down-menu you can sort the results of your search operation by:

- Year of publication: By default titles are sorted by year of publication. This means that the most recent publications are then displayed first.
- Relevance: You can also sort publications by relevance. This means they are displayed in sequence of importance. Titles that are most to the point are displayed first. The search engine analyses the contents of the title using a combination of the following factors: Frequency (the number of times a search term appears in a title), Compactness (the relative length of the titles found) and Reverse work frequency (terms that seldom appear in the entire database weigh more heavily).

5.2 Advanced Search

Besides entering a query in the simple search interface, you can also use the advanced search screen by clicking on the "Advanced" link or going to http://grib.gbv.de/advanced_searchfilter

The screen consists of three parts: An extended search interface, a checklist for material selection, and a checklist for library selection.

[14] "Not" is a stop-word as has to be deactivated by a preceded backslash, \.

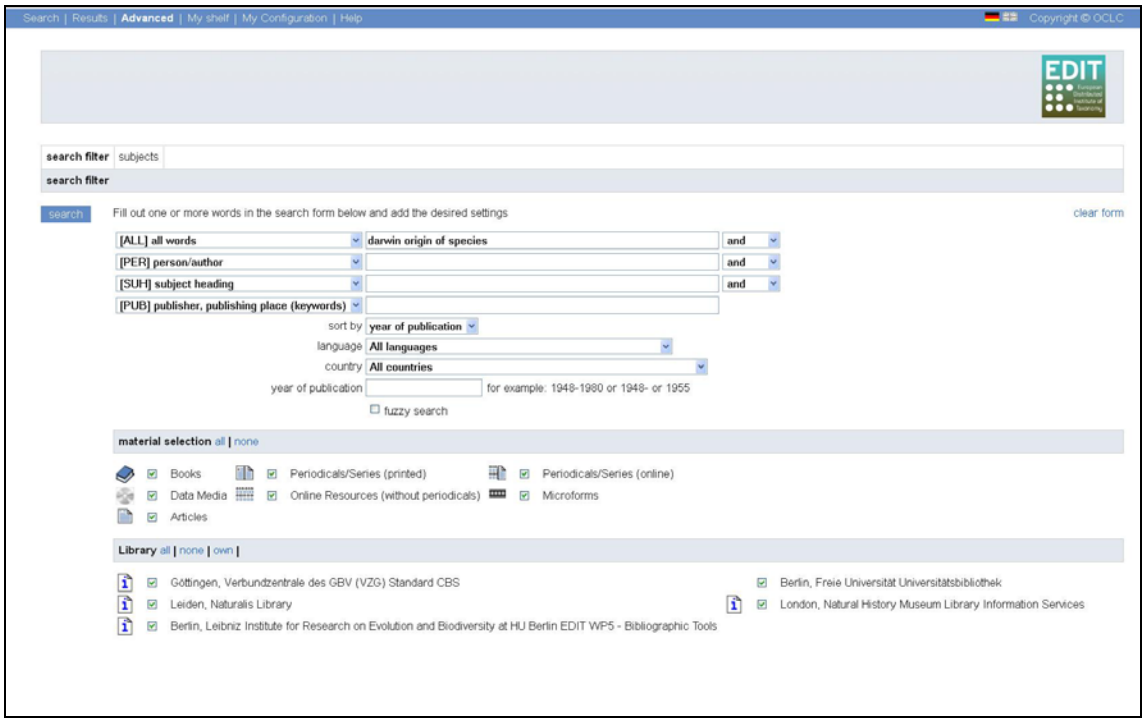

Fig. 4: Screenshot of the advanced search functionalities

5.2.1 Extended search interface

In the extended search interface you can search with the same methods and for the same search keys like in the simple search, but it allows for variations by providing multiple search fields and menus.

5.2.2 Checklist for material selection

The checklist shows the kind of materials that are referenced in the GRIB. By checking one or more boxes you can narrow down your search to variants of the following materials: Books, Periodicals/Series (printed), Periodicals/Series (online), Online Resources (without periodicals), Microforms, and Articles.

5.2.3 Checklist for library selection

The checklist shows all libraries, which provided bibliographic data to the GRIB. By checking one or multiple boxes you can narrow down your search to the holdings of one or more specific libraries.

5.3 Digitisation management Widget

The Digitization Management tool consists of a customized JSON-API to modify the digitisation status of a record via web service (see chapter 6.2 Digitisation management via JSON API) and a JavaScript-based English language widget that is embedded in the GRIB user interface.

There are four different digitisation status (DST) to be activated in a digitisation process. If there is no electronic document related to a bibliographic record, this record gets the status “Not Digitised”. In that case a registered scientist will be able to set the status to “should be digitised” and/or a librarian who plans to digitise that particular holding can set it to “Will be digitised”. Once a digitised version of that holding is accessible in the GRIB, the status will be set to “Document available”.

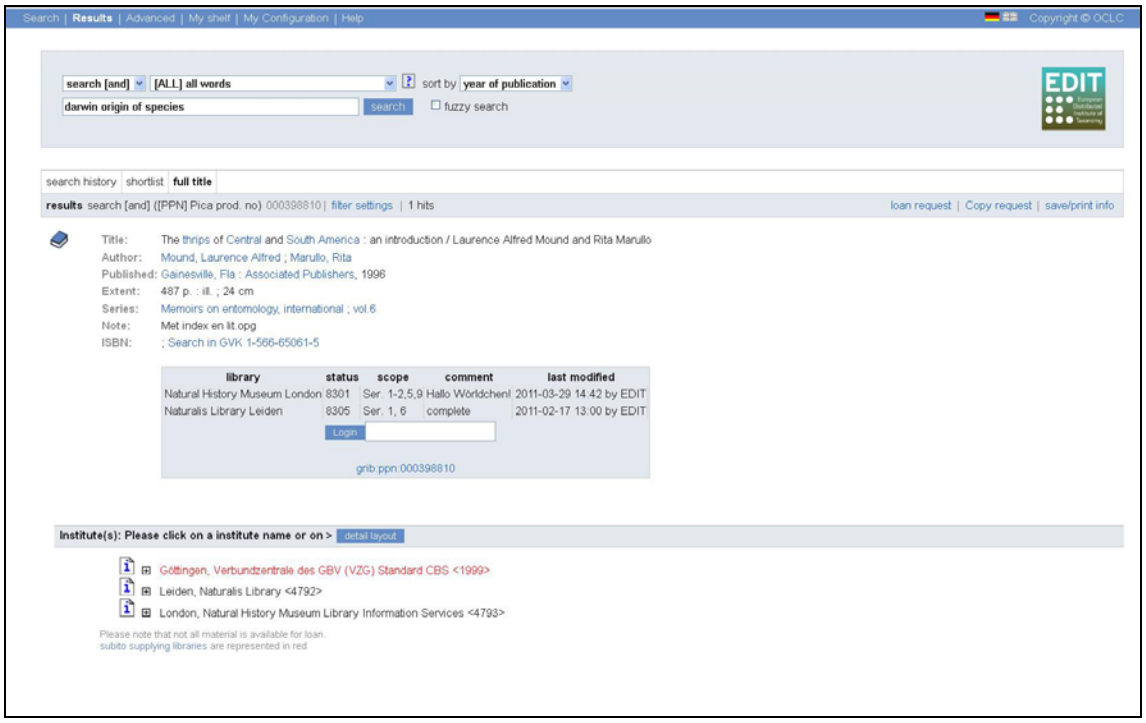

Fig. 5: Screenshot from the digitisation management widget

Librarians receive their sign-on details once their bibliographic data is incorporated into the GRIB. They can only set the status for holdings of their library, however, but it will be shown in the super record. While setting the DST to "will be digitised", the librarian has to indicate if he/she plans to digitise the holding in complete or just in part. This is especially important for a journal title. "Complete" is to be chosen, if the librarian plans to scan its entire volume of the title (note: not the complete existing title). If a "Partial" digitisation is planed, the librarian can indicate the exact holdings in the "Volume Articulation".

6 Access via Web-Services

This chapter shows the possibilities to access search and digitisation management functions of the GRIB via Web Services.

6.1 Search/Retrieval via URL (SRU)

For exporting data the GRIB offers a Search/Retrieval via URL interface (SRU) which allows 3rd party tools to find data in the GRIB and retrieve those in several standard formats.

The SRU interface is accessible at http://grib.gbv.de/sru/. Besides the standard search fields for Author, Title and Year – the specifications can be found at the Library of Congress[15] – there are some customized fields to be searched and retrieved:

- The Pica Production Number (PPN), a unique identifier for every set of bibliographic records within the GRIB.
- The Digitisation Status (DST), indicating the current stage of an item in the digitisation workflow

From a technical perspective the GRIB is functioning, data can already be retrieved from the GRIB in the following standards: MARC21, DC (Dublin Core), PICA, PICA short, UNIMARC and UNIMARC short.

[15] http://www.loc.gov/standards/sru/

6.2 Digitisation management via JSON API

The digitisation management functionalities could also be addressed via the JSON-API which is a simple REST-API, i.e. all parameters are provided as URL parameters added to a base URL. Due to the change of the sign on system it does not work at the moment. Allowing services to access the JSON API could be done via an API key.

1. Checking on the digitisation status.
   Request: http://bhleurope.gbv.de//api?id= ppn:000766127
   Response:
   {"options":[],"status":"8300","modifiedby":"EDIT","id":"bhleurope:ppn:000766127"}
2. Changing the digitisation status.
   a. Login in as group librarian:
      Request:
      http://bhleurope.gbv.de//api?id=ppn:000766127&login=1&user=EDIT
      Response: {"options":["8300","8301","8302","8303","8304","8305"],
      "user":"EDIT","id":"bhleurope:ppn:000766127"}
   b. Changing the status:
      Request: http://bhleurope.gbv.de//api?id=ppn:000766127&status=8301
      Response:
      {"options":["8300","8301","8302","8303","8304","8305"],"status":8301,
      "modifiedby":"EDIT","user":"EDIT","id":"bhleurope:ppn:000766127"}
   c. Logging out and looking at the new status:
      Request: http://bhleurope.gbv.de//api?id=ppn:000766127&logout=1
      Response:
      {"options":[],"status":"8301","modifiedby":"EDIT","id":"bhleurope:ppn:0007
      66127"}

The URL parameters are:

1) Id: A unique record identifier. For now only the PPN id as been assigned on ingest with the prefix “grib”.
2) Status: The new digitisation status must be a value from a predefined list[16] (See Table 1 “Status Id”).
3) Callback: A JavaScript callback method name.

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

4) Id: The unique record identifier.
5) Status: The digitisation status of the record.
6) Options: A list of digitisation status which the current user is allowed to set.

Error: An error message (only set if the action failed).

- information concerning the physical unit (of the item), e.g. shelf mark ...

[16] See Table: Values for the Digitisation Status