Global References Index to Biodiversity (GRIB)
- Technical Manual -

Version: Draft 0.7 (Nov. 30th 2010)
Dissemination Level: Restricted
Contact: boris.jacob@mfn-berlin.de

1 Introduction

1.1 EDIT, BHL-Europe and the GRIB

The Global References Index to Biodiversity (GRIB[1]) is supposed to serves two groups of users: scientists and 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.

The GRIB, a joint effort of EDIT/ViTaL and BHL-Europe, built by the Head Office of the Common Library Network GBV[2] (VZG[3]) allows users to browse and search deduplicated titles held by libraries of member institutions of the European Distributed Institute of Taxonomy (EDIT[4]) and the Biodiversity Heritage Library for Europe (BHL-Europe[5]), as well as click through to digital full texts from the Biodiversity Heritage Library (BHL[6]). In this sense the GRIB serves as a union catalogue of Natural History Museums and Botanical Gardens and a subject related virtual library for Biodiversity. The GRIB has also been enhanced by Digitisation Management Tool (DMT) assisting in digitisation projects by showing the taxonomic literature that is a) already available in digital form, b) in the process of being digitised, c) for which plans have been created for digitisation and d) which users indicated are important to be digitised in the future.

Steps so far:

1. December 2009: VZG sets up a demonstrator with data from AIT at a test database at http://kavia0.gbv.de/DB=1.83
2. Set up of the new and actual database in March 2010 for EDIT C5.145: “Import of test sets of selected libraries consisting of a) bibliographic data from the library catalogues and b) licence information on subscribed digital literature. Launch of a prototype of the references index.”[7]
3. April 2010: BHL-Europe D2.3: “Prototype of Web-database for content management and collection analysis.”
4. July 2010: EDIT C5.146: ”Import of freely available bibliographic metadata of digital resources from BHL to allow linking.”[8]
5. July 2010: EDIT C5.147: “Implementation of mechanism to nominate literature items for digitisation.”[9]
6. August 2010: EDIT C5.106: “Implementation of an im-/export data interface to and from CDM.”[10]
7. November 2010: Integration of the GRIB into the EDIT Community Single Sign-On

[1] http://grib.gbv.de

[2] http://www.gbv.de/vgm/index?lang=en

[3] Verbundzentrale des gemeinsamen Bibliotheksverbundes http://www.gbv.de/vgm/info/biblio/01VZG/

[4] http://www.e-taxonomy.eu/

[5] http://www.bhl-europe.eu/

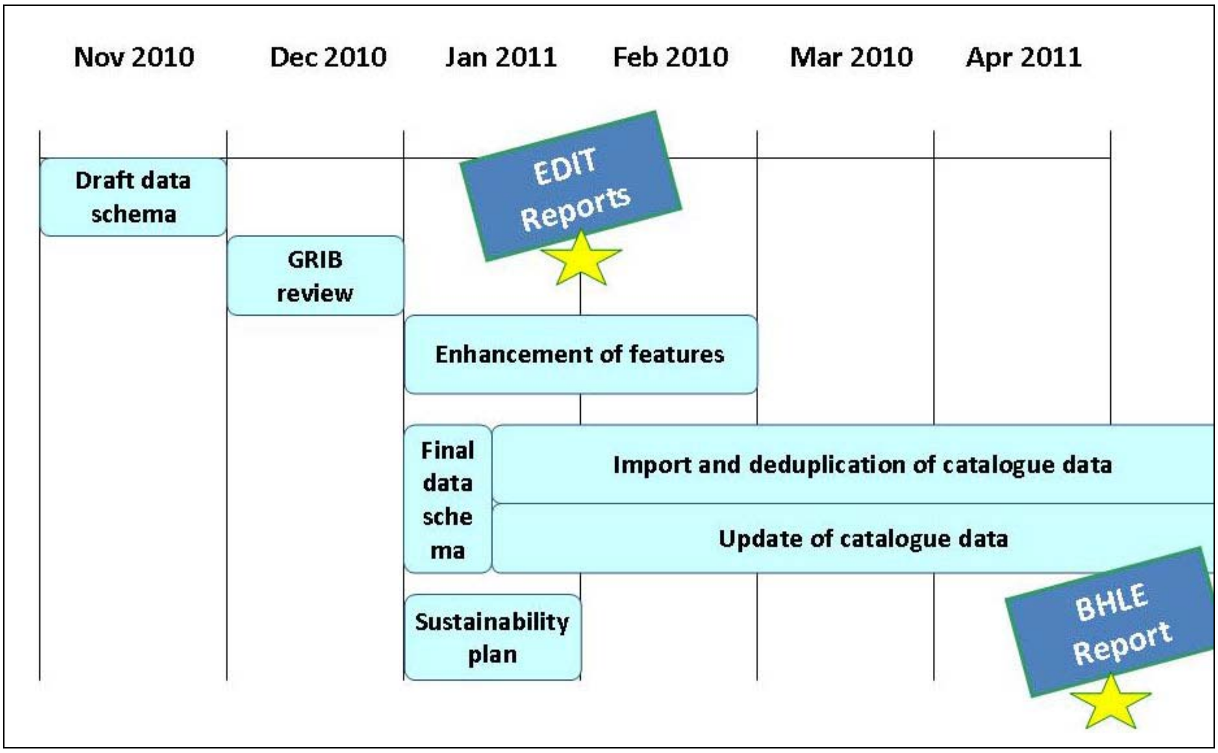
[6] http://www.biodiversitylibrary.org/

[7] http://wp5.e-taxonomy.eu/blog/files_edit_wp5/C5.145_2010_03_31_Import%20of%20test%20sets%20of%20selected%20libraries.pdf

[8] http://wp5.e-taxonomy.eu/blog/files_edit_wp5/C5.146_2010_08_14_Import%20of%20freely%20available%20biblioghraphic%20metadata.pdf

[9] http://wp5.e-taxonomy.eu/blog/files_edit_wp5/C5.147_201_07_30_Implementation%20of%20mechanism%20to%20nominate%20literature%20items%20for%20digitisation.pdf

[10] http://wp5.e-taxonomy.eu/blog/files_edit_wp5/C5%20106_2010_08_31_Implementation%20of%20an%20im-export%20data%20interface%20to%20and%20from%20CDM.pdf



1.2 Technology partner VZG

The VZG is the operating and service centre of the German library network Gemeinsamer Bibliotheksverbund (GBV), a public non-profit institution, funded by seven German federal states and the Stiftung Preußischer Kulturbesitz. It runs the Pica Central Library System (CBS) from OCLCPICA, which is the basis of The Common Union Catalogue (GVK[11]) of more than 400 GBV member libraries.

The Union Catalogue (GVK) is freely accessible via Internet. It contains more than 25,7 million title records with over 52, million holding records from all GBV libraries. Also included are the serial holdings being relevant for interlibrary loan of all large scientific German libraries of the German Union Catalogue of Serials (ZDB) and all subito supplying libraries of Germany and Austria. It is also possible to search for monographs, journals, articles, congress proceedings, microforms, electronic documents, music, maps, software etc.

The CBS software runs under UNIX and uses Sybase as RDMS. All software has been written in ANSI C/C++ and is Posix2 compliant. Although the software can run in principle on any UNIX platform, SUN Solaris is the recommended platform that is used by all current implementations. This homogeneous hardware environment improves the quality of second and third line support. Sybase as an RDMS is primarily used for administrative data such as library and ILL data, user data, the log records that are created for each database transaction and for all statistical information. For the bibliographic database Sybase is used as record manager. Bibliographical data are stored in one or more Sybase records; the application software performs all manipulations on a tag and sub-field level. The indexes to the bibliographic database are stored directly in a UNIX file in order to gain performance in searching. Nevertheless the indexes are maintained in real time.

[11] http://gso.gbv.de/xslt/DB=2.1/LNG=EN/

1.3 User roles

Three user roles are defined in the GRIB context right now, 1st Content User, 2nd technology User and 3rd Administrators.

- The Content User (Scientist, European Citizen) can use the GRIB's portal and open web services to search, retrieve and save bibliographic information. If the Content User is a Scientist, he can register as such and nominate literature to be scanned.
- The Technology User (EDIT, BHL-Europe and BHL Content Provider) can do all of the above. Most importantly he provides data to be imported into the GRIB to the Content Administrator. Once the technology is in place he can also manipulate data inside the GRIB including manual decision on possible duplicates and using the Digitisation Management Tool in a more extensive way.
- The Content Administrator (the MFN, represented through EDIT/ViTaL and later BHL-Europe WP2), is responsible for disseminating information about the GRIB to the Content and Technology user, negotiating with the System Administrator possibilities of advancing the GRIB. He is also, together with the System Administrator, responsible for managing the import of bibliographic data into the GRIB.
- The System Administrator (the VZG) is responsible for advancing and maintaining the GRIB and for importing data into the GRIB.

2 Import, match and merge of library catalogues

2.1 Workflow

1. Data to be imported
    1. Initial export and supplement exports in the standard format of each and every library (e.g. MARC or flavours of MARC).
    2. Library catalogue data of biodiversity related literature.
    3. Information on subscribed journals.
    4. Information on freely available electronic fulltexts.
2. Send data to the MFNs FTP account.
3. The VZG takes the data and mappes it into Pica+
4. During the match&merge process the data will be run against the GRIB database for duplicate detection and solution. Three possible szenarios:
    1. No match, the dataset will be imported.
    2. A possible duplicate candidate (b-nova). This has to be decided manually.
    3. A duplicate, the two datasets will be merged

2.2 Deduplication

The prototype consists of library metadata from FUB BGBM, MFN, NAT, NHM and BHL. First the MARC 21 data from Naturalis has been imported. Then the data from FUB BGBM and MFN, coded in the German metadata standard MAB2, where matched with the Naturalis data. After this import, there where 156.370 datasets in the index and 5.804 titles have been merged, with 2.742 possible merging candidates being identified. Data from NHM and BHL where imported later on. The result of the BHL import was, that 85.115 title records where imported, 4.832 of those have been identified as duplicates (3.196 are duplicates with titles from the NHM; 1.316 from NNM; 320 from FUB BGBM). Altogether the GRIB now holds 325.805 Titles and Articles.

In this prototype the deduplication has been 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). [12]

→ New import, match and merge process in January 2011.
→ Adjusting the deduplication mechanism and reporting for D2.5 & D2.7 in April 2011.

2.3 Data schema and data representation in the GRIB

The Pica Data-Model for bibliographic items consists of three levels:

- Level 0: All relevant bibliographic (meta-)data describing the item, e.g. author, titel, year of publication …
- Level 1 : Bibliographic information about the item which is only useful for the holding institution, e.g. belonging to the borrowable stock …
- Level 2: All information concerning the physical unit (of the item), e.g. shelf mark ...

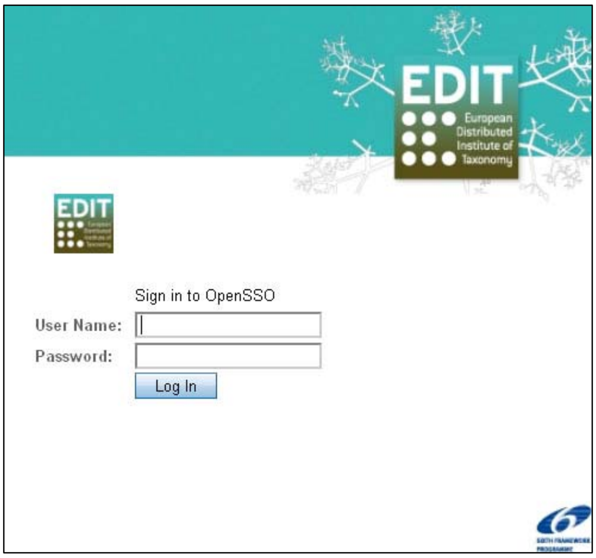
3 Website

3.1 Sign-On

The Sign-On system described in D2.3[13] has been replaced by a single sign on system using the EDIT infrastructure[14] (EDIT CSSO). EDIT is using a SAML2 based openSSO Identity provider (idP) and the VZG installed a Shibboleth Service Provider (SP)[15].

User	Password	Group	Institution
user.one	12345678	1	MuseumOne
user.two	12345678	1	MuseumTwo
user.three	12345678	1	LibraryOne
user.four	12345678	1	LibraryTwo
user.five	12345678	1	GardenOne
user.six	12345678	1	GardenTwo

In relation to the User Roles two groups of users have been defined:
User Group 1 = Librarian, User Group 2 = Scientists



Six test user accounts have been implemented, which are valid until 31st of December 2010. Personalised accounts will be handed out from January 2010 on. User details can be changed at https://idp.e-taxonomy.eu/opensso/UI/ a password reset is possible at http://dev.e-taxonomy.eu/trac/wiki/OpenSSOUser#ForgottenPasswords

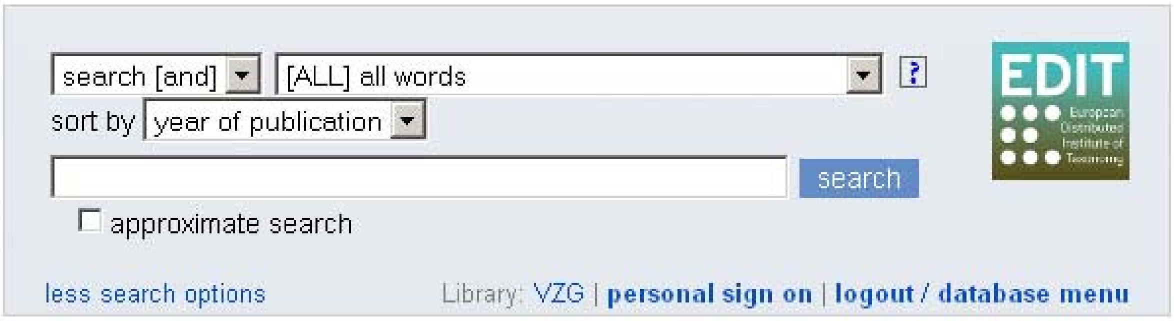
At the moment the user has to log on via http://bhleurope.gbv.de/ scrolling down to “Authentification (Shibboleth)” and clicking on “Login”. After typing in the credentials you will be directed to the GRIB homepage at

[12] See BHL-E D2.2: GBV approach of deduplication
[13] Scientists using the password “PUBLIC” and librarians using the password “EDIT”
[14] See http://dev.e-taxonomy.eu/trac/wiki/CSSO
[15] See http://dev.e-taxonomy.eu/trac/wiki/ShibbolethSP2InstallDebianLenny

http://grib.gbv.de/. That means you should log into the system prior to searching the GRIB. Going back to http://bhleurope.gbv.de/ you can see your log in status and what information has been send by the IdP.

→ The next steps are to place a log in link on the GRIB homepage at grib.gbv.de.
→ Instead of showing the name who changed the DST inside the Digitisation Management Widged, for privacy reasons the Institution will be shown.

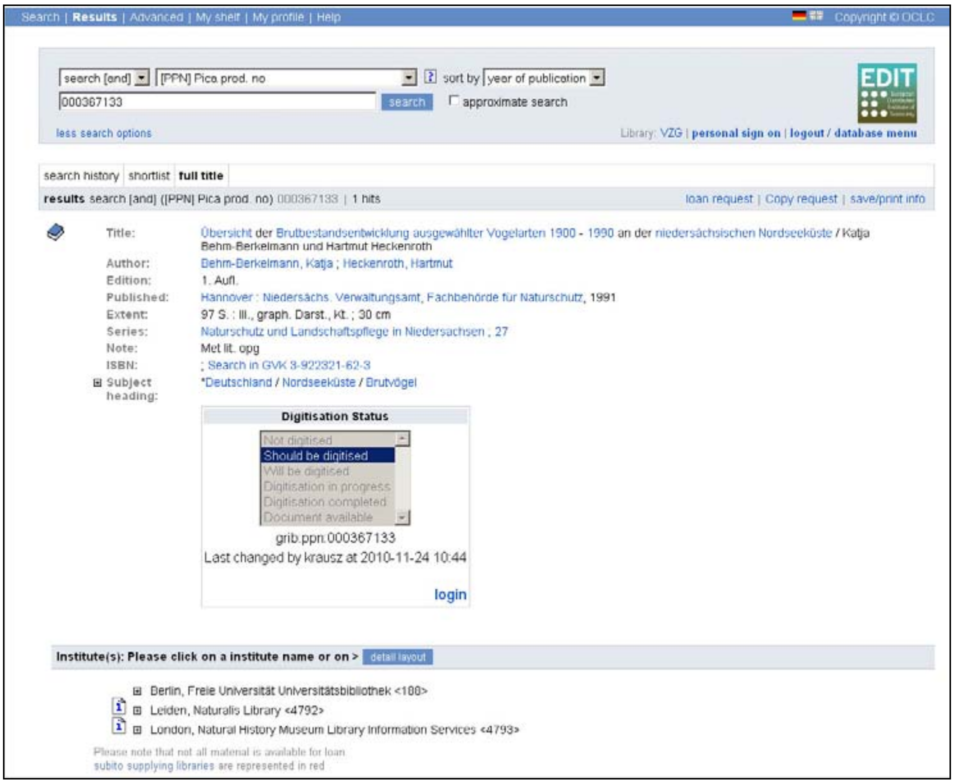
3.2 Examples for special Search



Search[16] for…	Enter search term...
All data sets	ppn=0?
Only “DST = document available” (i.e. data set with link to electronic full text)	“searchterm” dst=8305
Only without “DST = document available” (i.e. data sets without link to electronic full text)	“searchterm” -dst=8305
Data sets with “DST = not digitised” (i.e. data sets without electronic full text and without indication of being scanned in the future)	dst=8300

[16] For standard search see http://bit.ly/GSOhelp

3.3 Examples of data representation



Search | Results | Advanced | My shelf | My profile | Help
Copyright © OCLC
search [and]
[PPN] Pica prod. no
sort by year of publication
search
less search options
EDIT

000367133

□ approximate search

Library: VZG | **personal sign on** | **logout / database menu**

search history | shortlist | **full title**

results search [and] ([PPN] Pica prod. no) 000367133 | 1 hits — loan request | Copy request | save/print info

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: ; Search in GVK 3-922321-62-3
⊞ Subject heading: *Deutschland / Nordseeküste / Brutvögel

Digitisation Status

Not digitised
Should be digitised
Will be digitised
Digitisation in progress
Digitisation completed
Document available

grib:ppn:000367133
Last changed by krausz at 2010-11-24 10:44

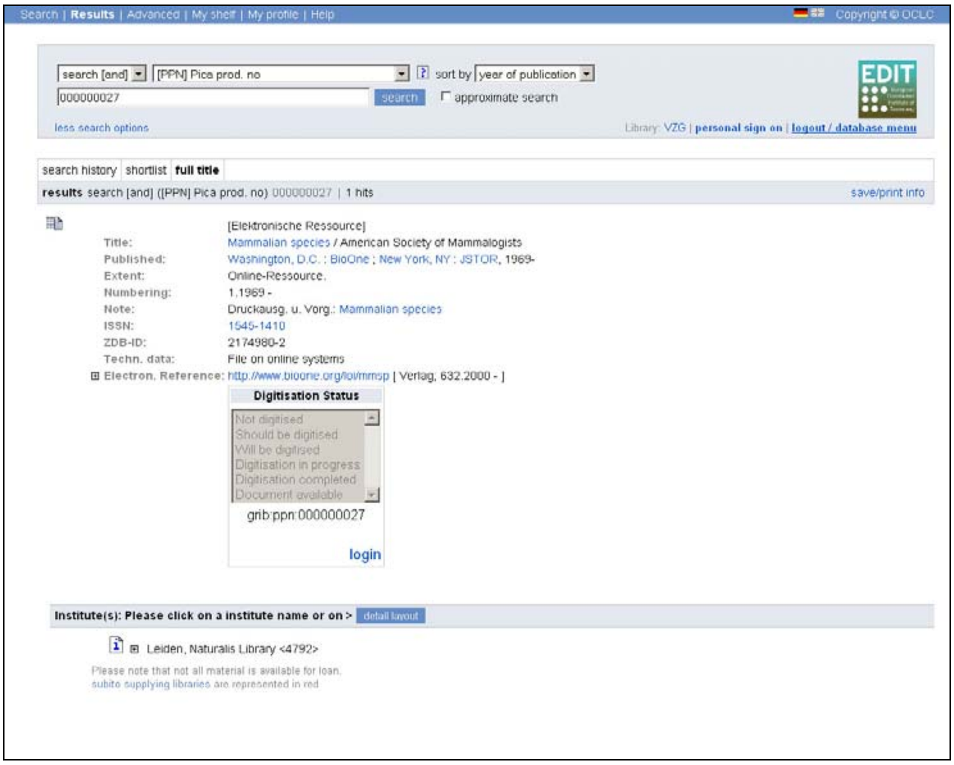
login

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

⊞ Berlin, Freie Universität Universitätsbibliothek <188>
⊞ Leiden, Naturalis Library <4792>
⊞ London, Natural History Museum Library Information Services <4793>

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

Create an index of bibliographic references (Deduplication) and allow users of the index of references to nominate literature items for digitisation (Example: grib:ppn:000367133)[17]



Search | Results | Advanced | My shelf | My profile | Help
Copyright © OCLC
search [and]
[PPN] Pica prod. no
sort by year of publication
search
less search options
EDIT

000000027

□ approximate search

Library: VZG | **personal sign on** | **logout / database menu**

search history | shortlist | **full title**

results search [and] ([PPN] Pica prod. no) 000000027 | 1 hits — save/print info

[Elektronische Ressource]
Title: Mammalian species / American Society of Mammalogists
Published: Washington, D.C. : BioOne ; New York, NY : JSTOR, 1969-
Extent: Online-Ressource.
Numbering: 1,1969 -
Note: Druckausg. u. Vorg.: Mammalian species
ISSN: 1545-1410
ZDB-ID: 2174980-2
Techn. data: File on online systems
⊞ Electron. Reference: http://www.bioone.org/loi/mmsp | Verlag, 632.2000 - |

Digitisation Status

Not digitised
Should be digitised
Will be digitised
Digitisation in progress
Digitisation completed
Document available

grib:ppn:000000027

login

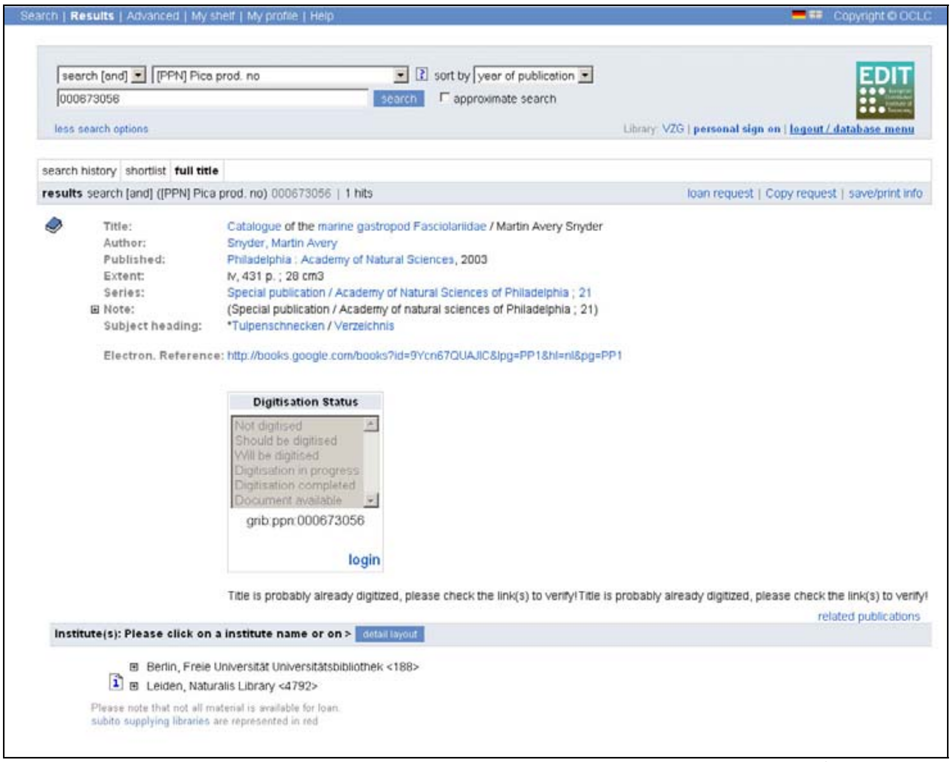
Institute(s): Please click on a institute name or on > detail layout

⊞ Leiden, Naturalis Library <4792>

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

Provide information on subscribed digital literature resources (Example: grib:ppn:000000027)[18]

[17] http://grib.gbv.de/SET=2/TTL=1/CMD?ACT=SRCHA&IKT=12&SRT=YOP&TRM=000367133



000673056

□ approximate search

results search [and] ([PPN] Pica prod. no) 000673056 | 1 hits

Title: Catalogue of the marine gastropod Fasciolariidae / Martin Avery Snyder
Author: Snyder, Martin Avery
Published: Philadelphia : Academy of Natural Sciences, 2003
Extent: iv, 431 p. ; 28 cm3
Series: Special publication / Academy of Natural Sciences of Philadelphia ; 21
⊞ Note: (Special publication / Academy of natural sciences of Philadelphia ; 21)
Subject heading: *Tulpenschnecken / Verzeichnis

Electron. Reference: http://books.google.com/books?id=9Ycn67QUAJIC&lpg=PP1&hl=nl&pg=PP1

Digitisation Status
Not digitised
Should be digitised
Will be digitised
Digitisation in progress
Digitisation completed
Document available
grib:ppn:000673056

Title is probably already digitized, please check the link(s) to verify! Title is probably already digitized, please check the link(s) to verify!

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

Link to freely available digital resources (Example: grib:ppn:000673056)[19]

3.4 Digitisation Management Widget[20]

The tool consists of a customized JSON-API to modify the digitisation status of a record via web service and a JavaScript-based English language widget that is embedded in the GRIB user interface and can be implemented in other web applications like CDM based EDIT platform components or the BHL-Europe portal.

Digitisation Status	Status No.	Status Id
Not digitised	1	8300
Should be digitised	2	8301
Will be digitised	3	8302
Digitisation in progress	4	8303
Digitisation completed	5	8304
Document available	6	8305

Values for the Digitisation Status

Depending on the group of user you belong to, you can change the digitisation status in different ways Scientists can nominate literature to be scanned by changing the digitisation status to “Should be Digitised”, whereas Technology user can choose to set the status to “Should be digitised”, “Will be digitised”, “Digitisation in progress”, “Digitisation completed” and “Document available” (See Table 1).

→ Suggestion to reduce the status to the following four: Not digitised, Should be digitised, Will be digitised, Document available.
→ Change of wording to scan management?
→ Discussion on how many Widgets to use per data set.

[18] http://grib.gbv.de/SET=2/TTL=1/CMD?ACT=SRCHA&IKT=12&SRT=YOP&TRM=000000027
[19] http://grib.gbv.de/SET=2/TTL=1/CMD?ACT=SRCHA&IKT=12&SRT=YOP&TRM=000673056
[20] See EDIT C5.147 “Implementation of mechanism to nominate literature items for digitisation.”

4 Web-Services

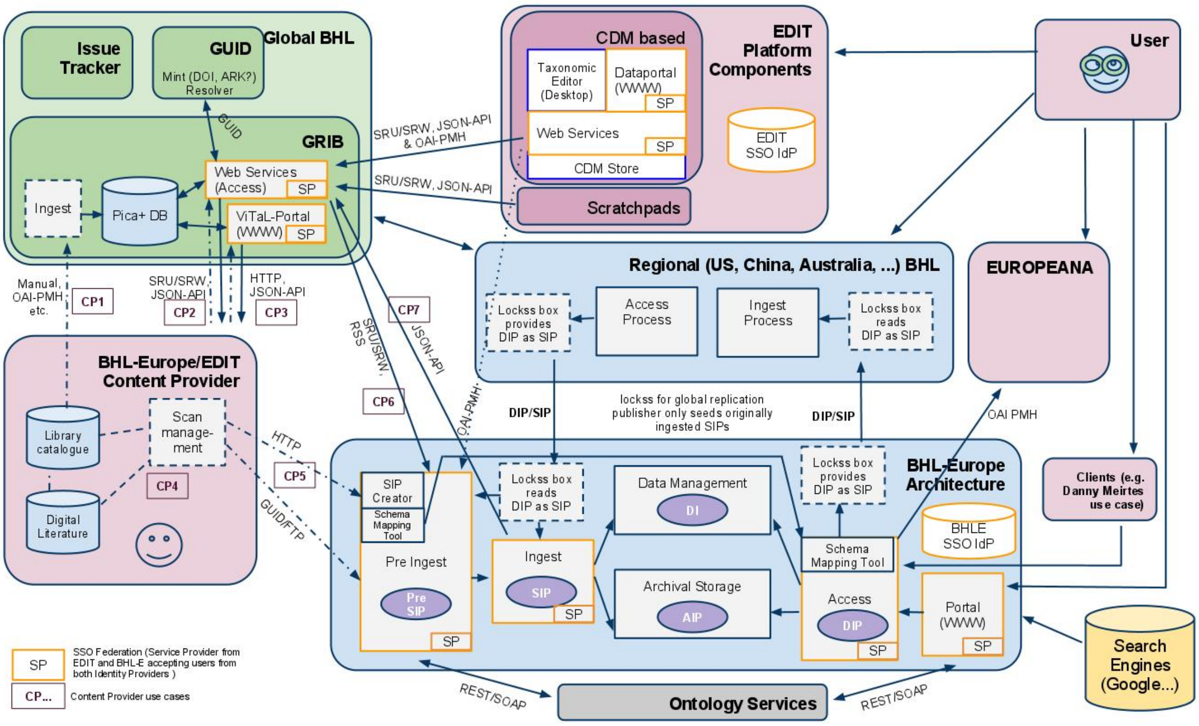


Fig. 7 BHLGlobalArchitectureDiagram

4.1 Search/Retrieval via URL (SRU)

For exporting data the GRIB offers a Search/Retrieval via URL interface (SRU) which allows 3rd party tools like the CDM based EDIT platform components or BHL-Europe to find data in the GRIB and retrieve those is several standard formats. On the other hand the GRIB can harvest and import data from the CDM store via the Open Archives Initiative Protocol for Metadata Harvesting[21] (OAI-PMH).

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[22] – 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 imported from the GRIB in the following standards: MARC21, DC (Dublin Core), PICA, PICA short, UNIMARC and UNIMARC short.

The GRIB is still considered at a prototypical stage, in that the data consists of the test sets (see C5.145) together with data derived from BHL (see C5.146), and, that the data is mapped within the database to the Pica fields alone and not to DC fields yet (i.e. search queries should be done on the PICA fields, not on the DC fields).

→ New import of library data in January 2011

[21] http://www.openarchives.org/OAI/openarchivesprotocol.html
[22] http://www.loc.gov/standards/sru/

4.2 UnAPI

An open unAPI has been implemented to retrieve single data sets in different formats, e.g.:

- unapi.gbv.de/?id=grib:ppn:002513064&format=pp (Pica+ format)
- unapi.gbv.de/?id=grib:ppn:002513064&format=ris (RIS)
- unapi.gbv.de/?id=grib:ppn:002513064&format=dc (DublinCore)

4.3 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:000766127"}

The URL parameters are:

- Id: A unique record identifier. For now only the PPN id as been assigned on ingest with the prefix “grib”.
- Status: The new digitisation status must be a value from a predefined list[23] (See Table 1 “Status Id”).
- 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 digitisation status of the record.
- Options: A list of digitisation status which the current user is allowed to set.

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

→ Generate access possibility via API key for services who want to address the JSON API in the future.

[23] See Table: Values for the Digitisation Status