


Project no. 018340

**Project acronym: EDIT**

**Project title: Toward the European Distributed Institute of Taxonomy**

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: "Global Change and Ecosystems"

# C5.150 Design of mechanism to update bibliographic data from the libraries into the references index

Due date of component: Month 57
Actual submission date: Month 59

Start date of project: 01/03/2006 Duration: 5 years

Organisation name of lead contractor for this component: 30 MFN

Revision: Final

| **Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)** | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | X |
| **PP** | Restricted to other programme participants (including the Commission Services) | |
| **RE** | Restricted to a group specified by the consortium (including the Commission Services) | |
| **CO** | Confidential, only for members of the consortium (including the Commission Services) | |

# C5.150 Design of mechanism to update bibliographic data from the libraries into the references index

The EDIT/ViTaL index of references (Global References Index to Biodiversity, called GRIB[1]) is 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 updated less often than data from the Biodiversity Heritage Library (BHL[2]) because the Internet Archive (BHLs scanning partner) is ingesting new data into BHL much more frequent, than the library catalogues change their data.

The method of updating depends on the technological infrastructure of the partner (if the partner system supports the Z39.50[3] protocol or an OAI-PMH[4] interface for example). As well as on the data they provide:

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

The mechanism to update bibliographic data into the references index consists of three major steps: 1) The delivery of the data, 2) its conversion and 3) the deduplication and merging of the records. 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[5] in step 2 and 3.

## 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.

a) Manual export of library catalogues and additional bibliographic data.

One way to provide library catalogue data is by manual exporting and uploading it to an FTP-Server at the Museum für Naturkunde Berlin (MFN), from where it is taken by the VZG for further processing.

b) Harvesting library catalogue data via Z39.50.

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.

c) Harvesting repository data or library catalogue data via OAI-PMH.

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.

1 http://grib.gbv.de/
2 http://www.biodiversitylibrary.org/
3 See http://www.loc.gov/z3950/agency/
4 See http://www.openarchives.org/OAI/openarchivesprotocol.html
5 Head office of the Common Library Network GBV (Gemeinsamer Bibliotheksverbund)

## 2. Conversion of the data

Once the data has reached the VZG (either by taking it from the ftp-account from the MFN or by harvesting it directly from the partners), the process of updating the index is always the same. The data is coming from the providers in different formats (like MARC 21[6] or flavours of MARC, MAB[7], as a spreadsheet or an XML file). It is mapped to the database format Pica+ and stored temporarily into a database (called Signaldatenbank), from where it is further processed.

## 3. Updating the GRIB-database

From the Signaldatenbank the data is transferred into the GRIB (database no. 1.68) and thus updated by a process called Match & Merge. During the matching process different results can occur.
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 data it is the hyperlink to the related electronic full text. This new information is merged with the existing record.
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 upon by a librarian.
The last possibility is that a record is new to the GRIB. It will then be imported into the GRIB as a entirely new title.

6 See http://www.loc.gov/marc/bibliographic/
7 See http://www.d-nb.de/standardisierung/formate/mab.htm