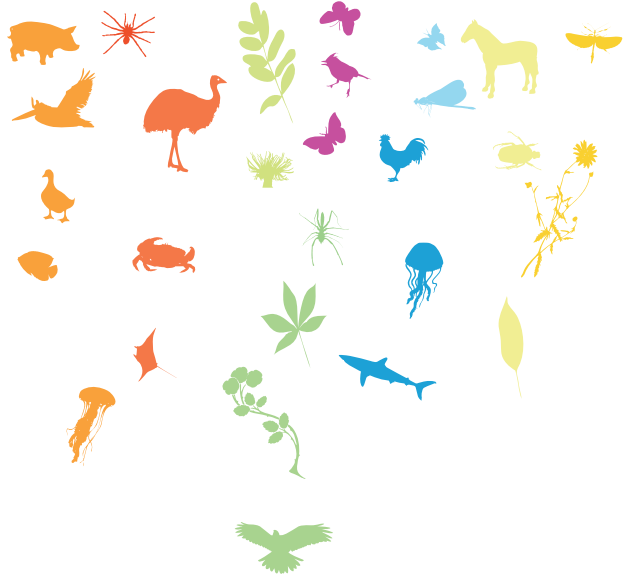




9 788883 032950



Tools for Identifying Biodiversity: Progress and Problems

Pier Luigi Nimis and Régine Vignes Lebbe (eds.)

2010



BHL-EUROPE: Biodiversity Heritage Library for Europe

Jana Hoffmann, Henning Scholz

Abstract — The Biodiversity Heritage Library for Europe (BHL-Europe) is an EU funded project making available biodiversity literature over various platforms, e.g. a multilingual portal for search and retrieval, the Global References Index to Biodiversity (GRIB), and Europeana. BHL-Europe brings together European digital biodiversity content already available and will assist in future scanning activities. BHL-Europe is asking the scientific community to contribute to the improvement of the BHL-Europe portal functionalities by giving feedback and participating in BHL-Europe workshops or to provide information about unexplored repositories of digital biodiversity content.

Index Terms — Biodiversity Heritage Library for Europe (BHL-Europe), digital library, biodiversity literature, taxonomy, taxonomic intelligence, optical character recognition (OCR).



1 INTRODUCTION

The lack of access to the published biodiversity literature is still a challenge in the day-to-day business of taxonomists or researchers dealing with biodiversity-related questions. In the past, only libraries of large and renowned institutions such as universities, natural history museums or botanical gardens housed specific literature indispensable for taxonomic work. Collecting relevant literature on a certain group of organisms was time consuming, cost-intensive and required loans of books or even a visit to the respective institution. Today, quick and easy access to digital literature is more and more important to facilitate scientific work. However, digitisation of literature is expensive and requires a lot of additional work on making the content available for extensive search and retrieval. Furthermore, there are still major problems with right holders, thus limiting the range of content freely available on the internet. For scientists it is of high importance to have a sustainable infrastructure they can rely on with a simple and quick mechanism to search for bibliographic information and free access to digital content of high quality. This is especially true for scientists working in developing countries with limited access to literature in general. As taxonomy is an ‘accumulative’ science it relies more than other

The authors are with the Museum für Naturkunde, Leibniz Institute for Research on Evolution and Biodiversity at the Humboldt University Berlin, Invalidenstr. 43, 10115 Berlin, Germany. E-mail: jana.hoffmann@mfn-berlin.de, henning.scholz@mfn-berlin.de.

disciplines on a complete record of literature on a group of organisms of interest and has a stronger focus on historic publications. Moreover global availability of digital content of biodiversity literature is also important for training students and early career scientists and helps promoting the importance of taxonomy as a discipline. Additionally, enhancement of availability of biodiversity literature for a wider public raises awareness of the importance of protecting our planets biodiversity.

Over the last few years a large number of library resources for taxonomists have been made available online - including virtual libraries and search engines as well as digital libraries. Since 2007, numerous libraries in the UK and USA are digitising their holdings of biodiversity literature and making them available on the internet. Today, BHL - the Biodiversity Heritage Library (<http://www.biodiversitylibrary.org>) is certainly the largest digital library for taxonomists offering free access to more than 30 million pages of historical biodiversity literature (as of July 2010) via the internet originating from 12 major natural history museum libraries.

2 OVERVIEW OF BHL-EUROPE

In 2009 the European Commission launched a new project 'BHL-Europe - Biodiversity Heritage Library for Europe' (<http://www.bhl-europe.eu>) within the framework of the *eContentplus* program. This project will run for 36 months through April 2012. The BHL-Europe consortium consists of 28 partner institutions (natural history museums, botanical gardens, libraries, right holders and companies) including 26 European institutions and two American institutions representing BHL (US). BHL-Europe aims among others at (1) supporting existing digitisation initiatives with best practice guides, for example, (2) facilitate and enable the initiation of new scanning initiatives, and (3) bringing together existing digital content scattered all over Europe in a number of libraries and natural history institutions. Currently, 18 out of the 28 consortium partners of BHL-Europe are active contributors to the corpus of digital resources. This corpus of more than 100,000 monographs and serial volumes in April 2012 will eventually be available on three platforms (Fig. 1) (1) a multilingual BHL-Europe-Portal for search and retrieval for scientists and public users, (2) the Global References Index to Biodiversity (GRIB), and (3) Europeana.

The technical architecture of BHL-Europe is based on the Open Archival Information System (OAIS) reference model. It is the backend of the multilingual portal for managing content ingestion, archival and delivery of the digital objects (Fig. 2). A prototype of the new portal will be available in fall 2010, but the final system is expected for the end of the project in April 2012.

BHL-Europe and EDIT (<http://wp5.e-taxonomy.eu/>) are building the Global References Index to Biodiversity (GRIB), a database generated from the partner libraries catalogues and completed with content management and deduplication functionalities, that eventually refers to all of the worlds published biodiversity literature. This will enhance the possibilities of search and retrieval of digital literature for taxonomists significantly. It will also assist librarians in the process of scanning planning. A GRIB prototype is working already and the final system is expected to be finished in spring 2011.

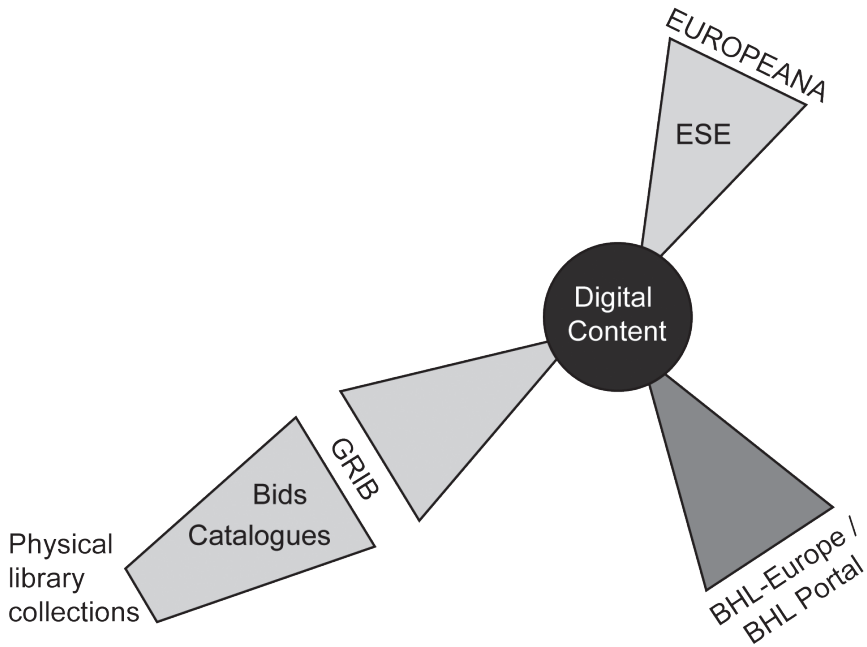


Fig. 1 – The BHL-Europe users (taxonomists, general public) will mostly access the content either through the BHL-Europe / BHL Portal or Europeana (ESE = Europeana Semantic Elements). The major access route for the librarians managing the scanning process is the Global References Index to Biodiversity. It is composed of the catalogue records of the physical library collections.

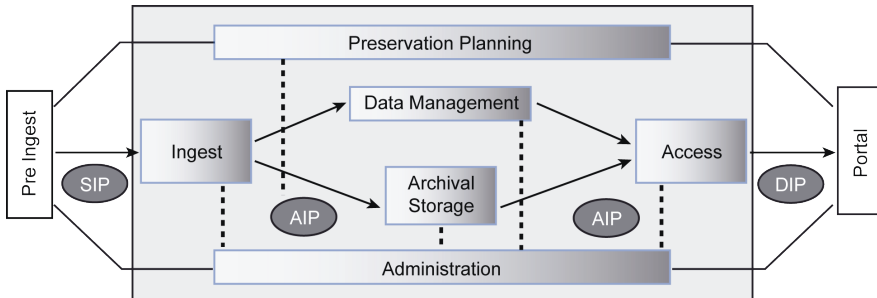


Fig. 2 – High-level overview of the OAIS components (grey box) with BHL-Europe Pre Ingest and Portal. The OAIS reference model differentiates between three kinds of information objects. The SIP, Submission Information Package, is being sent in by the data producers (content providers), the AIP, Archive Information Package, is preserved in the Archival Storage, and the DIP, Dissemination Information Package, is provided to the consumers.

Since June 2010 BHL-Europe content is made accessible for a wider public via Europeana (<http://www.europeana.eu>), the virtual European library. More than 80,000 books are currently accessible in Europeana and this number will increase continuously while BHL-Europe is harvesting digital literature from its content providers.

3 BHL-EUROPE FOR TAXONOMISTS

3.1 TAXONOMISTS AS PORTAL USERS

A major goal of BHL-Europe is providing and facilitating open access to taxonomic literature for a number of target users (scientists, hobby scientists, students, teachers, environmental and conservation agencies, etc.). This will result in a multilingual access point for search and retrieval of biodiversity content through a robust biodiversity community portal with an open and distributed architecture and specific functionalities, as described above. Integrated web tools like taxonomic intelligence (TI) will facilitate search for taxon-specific biodiversity information and thus improve efficiency of research in biology and access to information for a wider public. The key components for achieving this goal are excellent portal tools and the improvement of optical character recognition (OCR). OCR processing of scanned pages is the base for extracting specific terms and taxon names out of the pages. However, this is still a major challenge as automatic OCR accuracy is still not sufficient for our purpose, especially for historical literature and texts in multiple languages.

BHL-Europe aims at a deep level of language integration in the indexing system (multilingual indexing). For taxonomists key data of searchable metadata are names of organisms, biological groups and provenance. However, names of taxa and groups of organisms are inconsistent, represent different and changing scientific concepts or are vernacular names. Thus it is a major task for BHL-Europe to improve existing taxon-recognition tools (TI) for an advanced search of taxonomic information.

3.2 TAXONOMISTS AS PARTNERS

BHL-Europe is building the portal and all associated services for the users to meet their needs and requirements. BHL-Europe has to understand and evaluate the requirements of the users and how they are going to use the results of the project. Therefore, a very close cooperation between the users and the project is essential to make the project a success.

BHL-Europe is targeting a large number of different users ranging from libraries over different types of scientists to the general public. A number of instruments are currently used or will be used for the user interaction to prioritise the technical and collection development plan:

(1) Web analytics will be used to quantify the use of the portal (visits, unique visitors, page views, referring sites, country coverage).

(2) Users are encouraged to drop feedback messages either using the BHL online discussion forum or using the online contact form. BHL also has an issue tracking system (Gemini) in place to collect user feedback.

(3) Face-to-face and virtual interactions between the BHL-Europe members are helpful to get important input, as the project includes a number of key users from different user groups (e.g. libraries, taxonomists). These users from within the project work together in Use Case Workgroups. Their major task is the development of use cases for the portal prototype and testing of the portal functionalities.

(4) Suggestions of actors and users of large international projects like EOL or Europeana that are setting priorities based on their experiences are taken into account.

(5) BHL-Europe considers developments in biodiversity informatics and networked scientific communication like TDWG developments or PLoS Biodiversity Hubs.

(6) Specific user evaluations will be carried out twice during the project. The results of the first online user survey analysing the demand and service elements of the project will be publicly available soon and will be fed into the BHL-Europe IT development plan.

(7) BHL-Europe offers training opportunities on how to use the portal and its functionalities as well as other BHL-Europe products, e.g. the GRIB, and will ask for feedback on possible improvements. A first workshop will be held during the BioSystematics 2011 in Berlin (<http://www.biosyst-berlin-2011.de/>).

(8) BHL-Europe is also present with talks and posters in numerous scientific conferences to personally discuss with the scientists and to attract new users.

All information collected from the user's side this way will form the basis for developing a comprehensive set of use cases for the BHL-Europe portal and leads to further improvements of the system infrastructure.

In the past, individual scientist or the scientific community could not influence the choice of biodiversity literature for major scanning activities. BHL-Europe is implementing a mechanism that will enable users/ scientists placing a scan request for a specific volume using the GRIB infrastructure. This will allow libraries/ content provider to set up a priority list for their scanning activities and making highly demanded literature available first. As an intermediate solution, BHL has implemented a scanning request form in their feedback system.

Another goal of BHL-Europe is to seek for new partners (content provider, right holders) that can potentially contribute open access digitised biodiversity literature to the overall BHL repository. Therefore, support from the scientific community is highly welcome in naming additional repository of digital content and bibliographic data or in discussing with rights holder to make their content freely available.

4 CONCLUSION

The Chinese Academy of Science and the Atlas of Living Australia have been joining BHL already and negotiations with organisations in other countries are underway to further extend the BHL network. All these projects will work

together sharing content, protocols, services and digital preservation practices and promote the idea of a Global Biodiversity Heritage Library.

ACKNOWLEDGEMENT

BHL-Europe is co-funded by the Community Programme *eContentplus*, which is gratefully acknowledged.