ECP-2008-DILI-518001

BHL-Europe

Content analysis & management status report 1 (metadata, page numbers, content providers)

Deliverable number	*D2.4*
Dissemination level	*Public*
Delivery date	*28 April 2010*
Status	*Final*
Authors	*Henning Scholz, 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
Henning Scholz	MfN
Boris Jacob	MfN
Wolfgang Koller	NHMW
Michael Malicky	LANDOE

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
15 Apr 2010	Henning Scholz	0.1	First draft
15 Apr 2010	Boris Jacob	0.2	Second draft
16 Apr 2010	Henning Scholz	0.3	Third draft for internal and external review
28 Apr 2010	Henning Scholz	1.0	Final version incorporating internal and external reviews

0.3 Reviewers

This document requires the following reviews and approvals.

Name	Position	Date	Version
Manuela Schulz	Academic Librarian	24.04.2010	0.3

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
Manuela Schulz	16 April 2010	0.3
BHL-Europe consortium	16 April 2010	0.3
BHL-Europe consortium	3 May 2010	1.0

1 Table of Contents

2 Purpose

This document is to give an overview of the content specific progress of the project since it started in May 2009. It is an update of the information on the content of BHL-Europe given in the Description of Work (Annex 1 of the BHL-Europe Grant Agreement).

3 Background and content providers

The content providers within the consortium will create a critical mass of high quality digital content representing the biodiversity domain. The content providers have been selected on the basis of their ability to contribute key biodiversity and taxonomic literature. Although an accurate definition of the key biodiversity and taxonomic literature is still in discussion, large natural history museums and botanical gardens are considered to have large collections of primary taxonomic literature comprising all kingdoms of life.

The content will not be restricted by proprietary third-party rights or any other constraints, which would limit its use in an open access environment using Creative Commons http://creativecommons.org/ licences. The digital content must either be in the public domain or else the content contributors must have permission from IP owners to give open access under Creative Commons.

The initial focus on public domain material is not a limitation of the project because systematic biology depends more than any other natural science upon historic literature. Another reason to focus on historical literature is that many old and important monographs are themselves inherently very rare, fragile or in need of conservation. This makes “hands on” access very difficult. This project will substantially reduce the need for handling of these rare and valuable materials.

Currently, BHL-Europe has 17 content providers involved as consortium members in the project (see also Figure 1):

1) Natural History Museum (NHM), UK
2) Narodni muzeum (NMP), CZ
3) Georg-August-Universitat Gottingen Stiftung Offentlichen Rechts (UGOE), DE
4) Land Oberosterreich (Oberoesterreichische Landesmuseen) (LANDOE), AT
5) Hungarian Natural History Museum (HNHM), HU
6) University of Copenhagen (The Natural History Museum of Denmark) (UCPH), DK
7) Stichting Nationaal Natuurhistorisch Museum Naturalis (NAT), NL
8) National Botanic Garden of Belgium (NBGB), BE
9) Royal Museum for Central Africa (RMCA), BE
10) Royal Belgian Institute of Natural Sciences (RBINS), BE
11) Bibliotheque nationale de France (BnF), FR
12) Museum national d'histoire naturelle (MNHN), FR
13) Consejo Superior de Investigaciones Cientificas (CSIC), ES

14) Royal Botanic Garden Edinburgh (RBGE), UK
15) Smithsonian Institution (SIL), US
16) Missouri Botanical Garden (MOBOT), US
17) Helsingin yliopisto, University of Helsinki, Viikki Science Library (UH-Viikki), FI

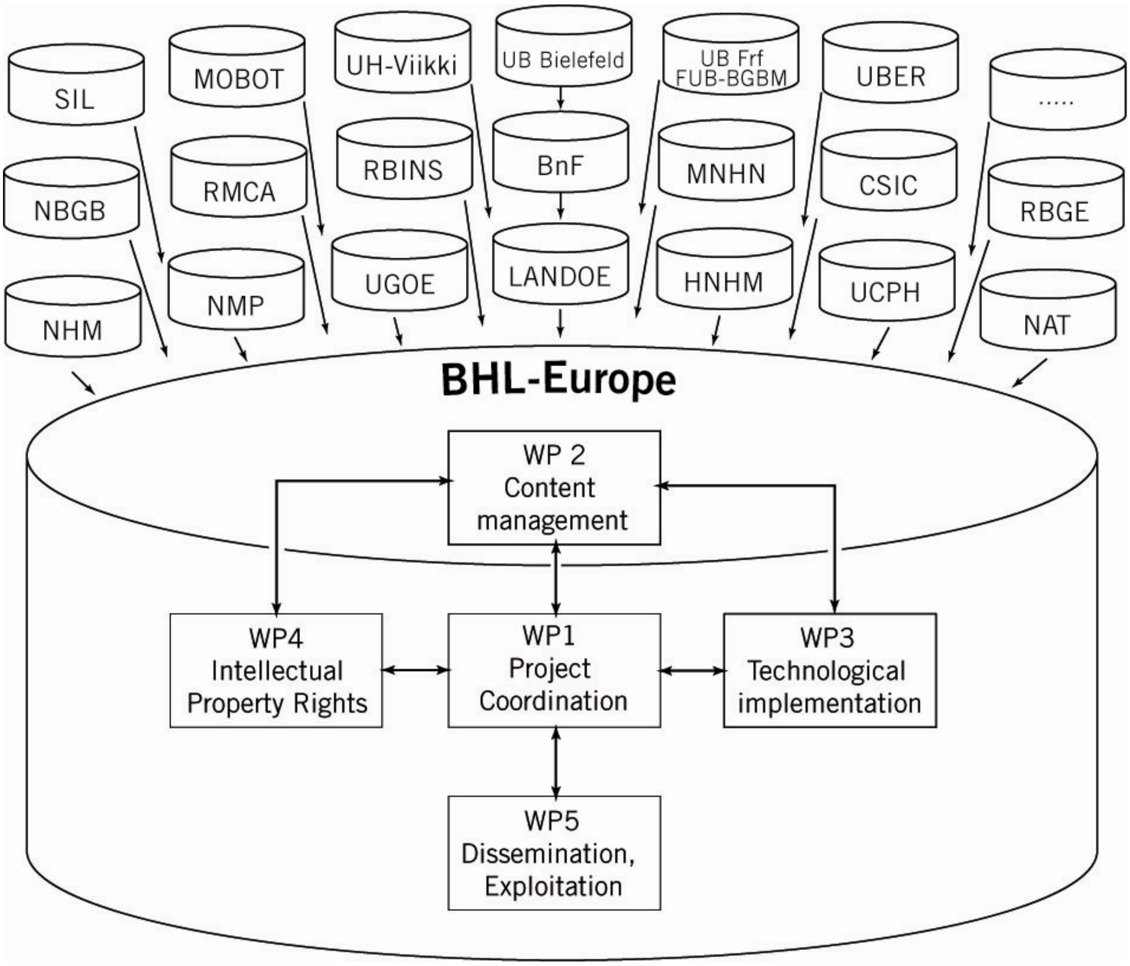

Figure 1. BHL-Europe internal organisation and the network of content providers today and in the future.

In addition to the above mentioned content providers, we expect new partners to provide further content during the project's lifetime. These new partners may be other important libraries in countries not yet engaged in the project or learned societies. These four institutions have partially signed the MoU and are currently expected to provide content to BHL-Europe:

1) Universitätsbibliothek Bielefeld, DE
2) Freie Universitat Berlin (FUB-BGBM), DE
3) Universitätsbibliothek Johann Christian Senckenberg, Frankfurt am Main, DE
4) Humboldt-Universitat zu Berlin (UBER), DE

For specific information on the digitisation projects of our content providers it is referred to the BHL-Europe newsletter, where we provide well illustrated details on individual projects and activities. The quarterly newsletters can be downloaded from our project Web site under http://www.bhl-europe.eu/newsletter.php. Newsletter #1 (July 2009) has articles on digitisation in the Czech Republic and BHL digitisation at the Natural History Museum London. Newsletter #2 (December 2009) has articles on activities in the Viikki Science Library, the Royal Belgian Institute of Natural Sciences and the digitisation at BnF for Gallica and Europeana. Newsletter #3 (March 2009) has an article about the digitisation activities in Goettingen. Further contributions will follow with the next newsletters.

4 Content analysis

BHL-Europe will support the digitisation and manage the acquisition and hosting of the digitised material contained in European 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. One of those tools, a database system with content management and deduplication functionalities, is called "Global References Index to Biodiversity" (GRIB). At the moment, the first prototype is ready for testing and usage for BHL-Europe consortium members only.

The GRIB infrastructure will harvest the library catalogues of institutions with libraries focused on taxonomic and biodiversity literature. These institutions are mainly natural history museums and botanical gardens, but also zoological, botanical and palaeontological libraries of universities, and national libraries. This common library catalogue of all these institutions will represent the vast majority of the literature relevant for the scientific community but also for other target users of the system like citizen scientists, students, teachers, policy makers and general interest audience.

Several tools and procedures will be used for building a priority list of books and journals to be included in BHL-Europe in digital form. We will analyse such major indexes as *Index Kewensis,* Sherbourne's *Index Animalium,* and Neave's *Nomenclator Zoologicus*. We are also working on mining other resources to determine those journals that have been most cited in the literature of species identification and description. Furthermore, we use the experiences of our partners in building the priority list. Most of the BHL-Europe partner institutions are natural history museums or botanical gardens thus a large number of experts in the biodiversity domain are available to select relevant content for BHL-Europe. Some of our partners already make use of this expertise in building local priority lists of journals and monographs (e.g. NBGB, RMCA, UH-Viikki). Furthermore, BHL colleagues in the USA are working on additional collection development strategies and criteria to prioritise items for digitisation. All priority lists will be included in the GRIB once this system is full functional. We expect this to happen towards the end of this year 2010. In addition to the joint efforts of our consortium to build priority lists of content for BHL-Europe, the GRIB will have the functionality to allow users to nominate content for digitisation.

The work on the GRIB is still in progress. Therefore, we need an interim solution to manage the digitisation process and help with the analysis of available content. This interim solution is the BHL Scan List as an extension of the Biodiversity Heritage Library Serials Union Catalogue developed at NHM London for the BHL project. The current functionality of the BHL Scan List includes:

- Marc exchange format records (minimum required fields stipulated)
- Records matched and merged through batch php scripting (see above)
- Merged set made available over the Web and fully searchable/sortable
- Authentication (1 per BHL and BHL-Europe partner institution)
- Ability to bid to scan part or all of a title's holdings at a per title level
- Ability to manually merge records which the scripting did not detect
- Merging the entries based on the ISSN using a stored procedure inside MySQL
- Ability to link to record and download to our local scanning management system

Currently, the BHL Scan List has 84,314 records. As numerous duplicates are still not merged, the number of serials is difficult to estimate. However, 30,742 items in 2554 serial titles are currently available digital via the BHL Portal (12 volumes per serial average). 2,891 titles are currently in progress of being digitised (1,193 complete bids, 1,698 partial bids). Further analysis are necessary to estimate the number of serials that still needs to be digitised based on the current corpus of serials in BHL and BHL-Europe partner libraries.

As we are still building the infrastructure for an effective content analysis, we need to have an overview of the available corpus of biodiversity literature to be scanned by national partners and BHL. This also helps us building a priority list of literature for digitisation programmes. Therefore, we did some calculations and estimations of that corpus. We based our calculations on real numbers of available serials from Austria, good estimates of German language biodiversity literature and estimates from BHL. Based on the number of population figures we extrapolated the page number estimates of German language biodiversity literature for the EU. Based on these first estimates, we have a range from 88,000,000 pages up to 166,375,000 pages of natural history literature published within the EU until today. Towards the end of the current year we expect more detailed estimates after we harvested actual catalogue records into the GRIB.

5 Metadata

In order to provide a sustainable system within BHL-Europe, the metadata needs to be defined properly. Information which may be requested, but has no definition in our metadata scheme, cannot be displayed or used. Therefore, the discussion of metadata standards and schemas is a continuous tasks since the start of the project in May 2009. We include a number of BHL-Europe target users and experts as members of the BHL-Europe consortium in these discussions: taxonomists, librarians, ICT specialists, and others. In August 2009, BHL-Europe ratified a first document on metadata fields to be used by BHL-Europe (non-public Deliverable D2.1: Catalogue of content holder requirements (quality, quantity, accessibility, standards and specifications of content and metadata)). Based on the D2.1 fields, several available standard schemes (like ONIX, METS, MARCXML, etc.) were evaluated in depth. All of those schemas are able to contain the basic information from D2.1.

During a project meeting in Vienna (March 2010), further requirements were defined by the content providers and users. An overview list of BHL-Europe high-level use cases was developed together with the ICT experts of the projects. The ongoing refinement of each of these preliminary use cases will possibly reveal the need for new metadata fields (or crossing out of existing fields) and provide information on the necessary obligation for the fields (mandatory, recommended, optional). An important fact for creating the list of required metadata fields is the separation into three main areas: administrative, descriptive and structural metadata.

It was decided during the meeting to continue with the proposed BHL-Europe METS profile and the list of BHL-Europe metadata fields as described in D2.1 and further elaborated in the BHLEMasterDataElementList. Both D2.1 and the BHLEMasterDataElementList are non-public documents. The final list of metadata elements, however, will be available to the public in the second year of the project. In the last meeting it was suggested to further refine/extend the list by metadata elements on article level. The current list of metadata fields will be used for the BHL-Europe prototype. The results of these tests will point at possible shortcomings and will affect the final design of the BHL-Europe metadata scheme.

6 Page and volume numbers

BHL-Europe is counting it's volume of content in two units: volume and pages. BHL-Europe is providing page level access to all the content harvested by us. Therefore, every content provider is required to provide individual page images and not only PDF or multipage TIFF files representing a volume. We will have a book viewer on our portal as this is the way BHL currently is exposing the content to the user (see http://www.biodiversitylibrary.org). Europeana, however, is counting individual books as objects to be searched and displayed via http://www.europeana.eu. One object is represented with a thumbnail on the portal and it makes no sense displaying every single page present in our repository. That would make search and retrieval via Europeana very difficult and is not a user friendly approach. Therefore, we need to know the page and volume numbers contributed by our partners.

As we have the volume numbers from our partners and use them for our ingestion planning, we will use volume numbers herein and in following reports like this one. Page numbers are also a little more abstract and library users are probably more used to count in books than in pages. This makes the comparison of the quantity of content easier. However, as a rule of thumb, one volume can be calculated with 350 pages average.

Currently, all content providers have a total of more than 46,000 books to contribute to the BHL-Europe project. This number is continously increasing as scanning projects are still in progress. Taking all the partners of our network into account, we expect about 80,000 volumes to be available via BHL-Europe and Europeana in about 18 months time. This number is not taking into account the contribution of additional network partners that will join BHL-Europe in the near future.