ECP-2008-DILI-518001

BHL-Europe

Content analysis & management status report 2 (metadata, page numbers, content providers)

Deliverable number	*D2.7*
Dissemination level	*Restricted*
Delivery date	*30 April 2011*
Status	*Final*
Authors	*Melita Birthälmer*

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
Graham Hardy	RBGE
Lesa Ng	RBGE
Michael Malicky	LANDOE
Michaela Hierschläger	LANDOE
Fritz Gusenleitner	LANDOE
Boris Jacob	RMCA

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
15 March 2011	Melita Birthälmer	0.1	First draft
15 April 2011	Melita Birthälmer	0.2	Second draft: incorporation of available and provided information until 15.04.2011
18 April 2011	Henning Scholz	0.3	Feedback
18 April 2011	Melita Birthälmer	0.4	Incorporation of feedback
19 April 2011	Melita Birthälmer	0.5	Incorporation of GRIB section Final Draft for external reviewer and consortium
28 April 2011	Melita Birthälmer	1.0	Incorporation of review

0.3 Reviewers

This document requires the following reviews and approvals.

Name	Position	Date	Version
Constance Rinaldo	Academic Librarian	27 April 2011	0.5

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
Consortium	19 April 2011	0.5
External Reviewer	19 April 2011	0.5

1 Table of Contents

2 Purpose

This document aims to give an overview of the content specific progress within BHL-Europe after the second year of the project. A first overview was given in the Content analysis and management status report 1 after one year project runtime. This report gives updated information with reference to the “Content analysis & management status report 1” D2.4.

3 Background

BHL-Europe supports European institutions in their digitisation efforts, manages the acquisition of biodiversity content in Europe and hosts digitised biodiversity content.

Chapter 4 gives an overview of the content management within BHL-Europe. First of all, the current status of the Global References to Biodiversity Index (GRIB) is given in Chapter 4.1. Detailed information about BHL-Europe content providers including content acquisition and an outlook for the third project year is given in Chapter 4.2.

Chapter 5 addresses content analysis of the BHL-Europe content. The process of establishing access for the BHL-Europe content providers to the BHL-Europe server and the actual process of uploading content is described in Chapter 5.1. Information about the underlying content of BHL-Europe is presented in Chapter 5.2. BHL-Europe acts additionally as a content aggregator for Europeana and therefore Chapter 5.3 shows the availability of BHL-Europe content within Europeana. Furthermore, the relation between content in BHL-US and European cultural heritage and an estimation of published content within the domain of biodiversity in the European Union is given in Chapter 5.4 and Chapter 5.5. The content analysis closes with an outlook for the third project year.

Although this document is a deliverable within WP2, it also contains information on WP3 outcomes as they impact WP2 outcomes, particularly in the area of content analysis.

4 Content Management

Content Management is an important task within BHL-Europe and comprises the management of the scanning process, the coordination of and communication with the content providers and data providers, and the content acquisition of new content for BHL-Europe. Within BHL-Europe we differentiate between data and content provider. A data provider is a partner of the BHL-Europe project providing only metadata (e.g. library catalogue) to BHL-Europe for the Global References Index to Biodiversity. A content provider is in contrast a partner of the BHL-Europe project providing content and metadata to the BHL-Europe repository and system.
In order to facilitate communication with our content and data providers a mailing list (bhl-e.cp@lists.hu-berlin.de) has been set up during the second project year. This mailing list is used for updating BHL-Europe content providers and announcing important information to our content providers.

4.1 Global References Index to Biodiversity (GRIB)

The Global References Index to Biodiversity (GRIB) is the joint literature index from the projects Biodiversity Heritage Library for Europe (BHL-Europe) and European Distributed Institute of Taxonomy (EDIT), developed together with the head office of the Common Library Network GBV (VZG). The GRIB incorporates bibliographic data from library catalogues provided by natural history institutions that are part of these projects. While importing library catalogues to the GRIB, duplicates are merged together. The GRIB lists the libraries where publications can be found, links records to publications already digitised and offers the possibility to indicate publications to be digitised that are not yet accessible online. Thus the GRIB helps content providers in their planning to select relevant biodiversity literature for digitisation.

4.1.1 Data status of the GRIB

The GRIB holds 325,805 de-duplicated records from the library catalogues of FUB-BGBM, MfN, NAT and NHM, as well as data from the Biodiversity Heritage Library. The actual number of de-duplicated records in the GRIB can be retrieved via http//:grib.gbv.de using the search term “ppn = 0?”. Table 1 gives an overview of the library that has been created from the bibliographic data provided by consortium members and imported into GRIB. The GRIB contains 426,421 records in total, of which 100,616 have been identified as duplicates and merged. 40,239 records have been marked as possible duplicates by the system.
The table also shows the amount of data delivered but awaiting import to GRIB, and also information on those catalogues and databases yet to be harvested from partners.

With the next re-imports the GRIB will allow for faceted searching for different forms of publications (e.g. monographs, serials, electronic publications), however, this distinction is not implemented in the current data version in the GRIB. As of this date, a total of 426,421

items have been imported and de-duplicated and 325,805 unique records are available within the GRIB.

Nr.	Name	Abbreviation	Available data	Delivered data	Imported data into the GRIB	Number of records
1	Stichting Nationaal Natuurhistorisch Museum Naturalis	NAT	1 library catalogue (1)	1 library catalogue	1 library catalogue	79,451
2	Natural History Museum	NHM	1 library catalogue	1 library catalogue	1 library catalogue	97,052
3	Freie Universität Berlin	FU-BGBM	1 library catalogue	1 library catalogue	1 library catalogue	82,723
4	Museum für Naturkunde Berlin	MfN	1 library catalogue	1 library catalogue	1 library catalogue	82,080
5	Biodiversity Heritage Library	BHL-US	1 repository database	1 repository database	1 repository database	85,115
6	National Botanic Garden of Belgium	NBGB	1 library catalogue	1 library catalogue		
7	Royal Botanic Garden Edinburgh	RBGE	1 library catalogue	1 library catalogue		
8	Smithsonian Institution	SIL	1 library catalogue	1 library catalogue		
9	Hungarian Natural History Museum	HNHM	1 library catalogue 1 article database	1 library catalogue 1 article database		
10	Royal Belgian Institute of Natural Sciences	RBINS	1 library catalogue	1 library catalogue		
11	Humboldt-Universität zu Berlin	UBER	1 repository database	1 repository database		
12	Helsingin yliopisto	UH-Viikki	1 library catalogue	1 library catalogue		

Nr.	Name	Abbreviation	Available data	Delivered data	Imported data into the GRIB	Number of records
13	Nationaal Herbarium Nederland	NHN	1 library catalogue [(1)]	1 library catalogue		
14	Consejo Superior de Investigaciones Cientificas; Museo nacional de Ciencias Naturales	CSIC	1 library catalogue 1 archival catalogue			
15	Museum and Institute of Zoology, Polish Academy of Sciences	MIZPAS	1 library catalogue			
16	Museum National d'Histoire Naturelle	MNHN	1 library catalogue 1 repository database			
17	Università degli Studi di Firenze	MSN	1 library catalogue 1 repository database			
18	Narodni muzeum	NMP	1 library catalogue			
19	Royal Museum for Central Africa	RMCA	10 library catalogues			
20	Georg-August-Universität Göttingen	UGOE	1 library catalogue 1 repository database [(2)]			
Total amount of imported records						**426,421**

Table 1: Available library catalogues provided by content providers & status of import to the GRIB (18.04.2011)

[(1)]*…NAT provides a second library catalogue from the Nationaal Herbarium Nederland. This catalogue is listed under Nr. 13.*
[(2)]*… the AnimalBase repository catalogue is available from UGOE.*

4.1.2 Delivery and update of the GRIB data

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

1. Delivery of the data by the data provider
2. Conversion of the delivered data to Pica+ format
3. De-duplication and merging of the records

While method and frequency of delivering the data in step 1 varies on the partners' side, there is a fixed process on the side of the VZG in step 2 and 3.
A data provider can provide bibliographic data to the VZG by manually exporting the data or by providing a Z39.50 or an OAI-PMH interface for automatically harvesting by the VZG.
Once the VZG receives the data in the format of the providing institution, it has to be mapped manually to the GRIB database format Pica+, which is the basis of the "Match&Merge" process and thus de-duplicates records. During the de-duplication process different results can occur:

- A match is detected and the new imported data is stored together with the already existing data.
- A new imported item is detected as a possible match and is kept distinct but marked for further manual checking.
- No match is detected and all imported records are kept independently in the GRIB.

At the present time, at over 40,000, the total number of identified possible duplicates is significant. Plans are in place to adapt the de-duplication process over the next few weeks, to ensure that duplicates are found at a higher rate. The amount of time spent mapping the original data to the Pica+ format, and identifying duplicates, is dependent on the quality of data supplied by each data provider.
In general, one can say, that the better the data format and the richer the provided data, the better are the results. Good formats in this sense are standard bibliographic data formats like MARC, different flavours of MARC, Pica, MAB or even DC.
However, in the majority of cases for the import to the GRIB received data is not in a standard bibliographic data format. The duration for the de-duplication process is reduced if adequate fields for identifying a record as unique are provided, such as "ISBN" or "OCLC number". If unique fields like that are not included in the data, the de-duplication algorithm uses the "Title" of the record which might not identify a new imported record as a duplicate because of variations in spelling, for instance. Additionally, items with the same or very similar title might erroneously be marked as duplicates. Thus it must be marked for further manual checking. The quite low quality of available data slows down the process of de-duplications and thus the final import and availability of the data within the GRIB. At the time of writing, data meeting a higher standard has been imported or is scheduled at a higher priority for import to the GRIB. As the duration of import depends on the quality of data provided, a detailed plan cannot be given. However, it is planned to finish the import of those catalogues already provided and re-import the ones that are already in the GRIB by the beginning of the third project year. The VZG also updated the central library system software recently and this process also slowed down the import of new data during the last weeks.

4.1.3 Hosting and maintenance of the GRIB

MfN and VZG have an agreement for the VZG to host and maintain the GRIB infrastructure, as well as updating the index with information from BHL-Europe, EDIT and BHL partners until the end of February 2020. The GRIB is based on the OCLC PICA-CBS (Central Library System) hosted by the VZG. It is accessible via standard interfaces (WWW, Z39.50, SRU, internal XML, OAI) and the VZG aims to make it accessible 24 hours, 7 days a week. The library catalogue data submitted by partners is to remain within the GRIB. The copyright of the original data is retained by the institution supplying the data. A physical export of the GRIB is only possible in agreement with the MfN.

Detailed information about functionalities and technical specifications of the GRIB can be found in the deliverable “Final and enhanced Web-database for content management and collection analysis” D2.5.

4.2 Content providers

The following chapter gives an overview of primary content providers since the beginning of the project. Content providers were acquired during the first and second project years. The process of content acquisition within BHL-Europe is described and an outlook for the third project year is given.

4.2.1 Primary content providers

BHL-Europe content providers aim to provide high quality digital content representing the biodiversity domain. The content providers within the consortium have been selected on the basis of their ability to contribute key biodiversity and taxonomic literature on as wide a range of organisms as possible. Thus national botanic gardens and natural history museums within European Union partner states are content providers because of the richness of their library collections.

From the beginning of the project BHL-Europe has had 17 primary content providers involved as consortium members which are listed below:

Nr. of CP	Name	Abbreviation	Country
1	Natural History Museum	NHM	United Kingdom
2	Národní museum	NMP	Czech Republic
3	Georg-August-Universität Göttingen Stiftung Öffentlichen Rechts	UGOE	Germany
4	Land Oberösterreich (Oberösterreichische Landesmuseen)	LANDOE	Austria
5	Hungarian Natural History Museum	HNHM	Hungary
6	University of Copenhagen (The Natural History Museum of Denmark)	UCPH	Denmark

Nr. of CP	Name	Abbreviation	Country
7	Stichting Nationaal Natuurhistorisch Museum Naturalis	NAT	Netherlands
8	National Botanic Garden of Belgium	NBGB	Belgium
9	Royal Museum for Central Africa	RMCA	Belgium
10	Royal Belgian Institute of Natural Sciences	RBINS	Belgium
11	Bibliothèque nationale de France	BnF	France
12	Muséum national d'Histoire naturelle	MNHN	France
13	Consejo Superior de Investigaciones Científicas	CSIC	Spain
14	Royal Botanic Garden Edinburgh	RBGE	United Kingdom
15	Smithsonian Institution	SIL	United States of America
16	Missouri Botanical Garden	MOBOT	United States of America
17	Helsingin yliopisto University of Helsinki Viikki Science Library	UH-Viikki	Finland

Table 2: List of primary content providers at the beginning of the project

4.2.2 Additional acquired content providers

First year of the project

BHL-Europe aims to acquire further content providers to increase the amount of digital content significantly during the project lifetime and thus make as much content as possible available to numerous groups of users such as scientists, decision makers and the European population. New content providers have joined the project and will contribute content to BHL-Europe.

During the first year of the project two institutions entered the project as new content providers.

Nr. of CP	Name	Abbreviation	Country
18	Humboldt-Universität zu Berlin	UBER	Germany
19	Universitätsbibliothek Bielefeld	UB-Bielefeld	Germany

Table 3: Additional content provider after first year of project lifetime

The *Humboldt Universität zu Berlin (UBER)* was already part of the consortium with the role of a disseminator and digital library expert and has added the role of a content provider.
The *Universitätsbibliothek Bielefeld (UB-Bielefeld)* has been acquired as a new content provider for BHL-Europe.

Two further partners entered the project during the first year of the project and have been noted in the first "content analysis & management status report" D2.4. However, problems with funding of the scanning projects occurred and therefore these two institutions withdrew the offer to provide content to BHL-Europe. These two former content providers are listed below:

- Freie Universität Berlin (FUB-BGBM), Germany
- Universitätsbibliothek Johann Christian Senckenberg, Frankfurt am Main, Germany

Second year of the project

During the second year of the project 6 institutions agreed to act as new content providers for BHL-Europe.

Nr. of CP	Name	Abbreviation	Country
20	Università degli Studi di Firenze	MSN	Italy
21	Museum and Institute of Zoology, Polish Academy of Sciences	MIZPAS	Poland
22	Museum für Naturkunde Berlin	MfN	Germany
23	Institute of Paleobiology Polish Academy of Sciences	PAS	Poland
24	Gesellschaft für Biologische Systematik	GfBS	Germany
25	Landesbibliothekszentrum Rheinland-Pfalz - dilibri	dilibri	Germany

Table 4: Additional content providers after second year of project lifetime

The *Università degli Studi di Firenze (MSN)* and *the Museum and Institute of Zoology, Polish Academy of Sciences (MIZPAS)* were part of the consortium since the beginning of the project. Both already had the role of a domain and language expert and during the second year of the project added the role of content provider.
MSN worked together with our consortium partner and content provider UGOE in a collaborative project and prepared biodiversity literature for digitisation within the UGOE library. Literature - most of Italian origin - was transported from Florence to Göttingen in winter 2010 and will be made available to BHL-Europe through UGOE.
MIZPAS successfully submitted a digital repository project in 2010 with the start of project in 2011 and plan to start digitizing their resources beginning of summer 2011.
The *Museum für Naturkunde Berlin* was successful in raising funds for a book scanner. Preparation of a literature corpus to be scanned is currently in progress and it is assumed that MfN will provide content to BHL-Europe soon.
The Institute of Paleobiology of the Polish Academy of Science[2] has been acquired as a new content provider to BHL-Europe and will provide digital content from their already online available archive (http://www.app.pan.pl/archives.html) to BHL-Europe including future born digital publications.

[2] www.paleo.pan.pl

The *Gesellschaft für Biologische Systematik*[3] (GfBS) is a German society for biological systematics and will contribute all their publications older than two years to BHL-Europe.
Dilibri[4] is the digitised collection of regional studies about the federal state Rhineland-Palatinate and will provide biodiversity related content to BHL-Europe as a new content provider.

4.2.3 Content Acquisition

Attracting new content providers is and will be an ongoing process. For this purpose a document called “BHL-Europe in a nutshell” has been prepared in collaboration with WP4 and introduced during the Content Provider and Technical meeting in December 2010 in London. The document aims to be a very brief but informative document including the most important facts about BHL-Europe, benefits and IPR issues. "BHL-Europe in a nutshell" has been translated by our consortium into following languages:

- English
- Spanish
- French
- Russian

The English document is available in the Appendix within this document. All languages have been made available to the dissemination team for publication on the official BHL-Europe website[5] and have been uploaded to the BHL-wiki[6].

During the second year of the project consortium members took part in a new method for content acquisition. More precisely, the identified method takes advantage of our consortiums’ connections in order to attract new content providers within their own countries using the “BHL-Europe in a nutshell” document. This new method is designed to be sustainable to ensure the continued growth of BHL-Europe not only during the project time but also after the end of the financial aid from the European Union.
The new method was tested by our consortium partners RBGE in winter and spring 2010/2011. The process of using the “BHL-Europe in a nutshell”, feedback and lessons learned are described below.

[3] www.gfbs-home.de
[4] www.dilibri.de
[5] www.bhl-europe.eu
[6] https://bhl.wikispaces.com/Content+_Management

Process of using BHL-Europe in a nutshell for content acquisition

RBGE identified two ways of using the "BHL-Europe in a nutshell" document for the content acquisition:

- Send emails to known and unknown individuals within the biodiversity community of the own country – with and without prior telephone contact.

- Use in poster sessions at conferences.

Both ways are described in more detail hereafter.

1. *Send emails to known and unknown individuals within the biodiversity community of the own country*

    - Send a cover letter along with "BHL-Europe in a nutshell" to institutional staff with contacts within the biodiversity community to receive feedback and comments for prospective content providers that can be approached.

    - Prepare a review of your own imprints included in BHL-Europe to give an example of current content.

    - Prepare - if necessary - an introductory information sheet about your institution in regards to BHL-Europe

    - Send emails to secretaries of selected natural history organisations. Where no secretary can be identified send information to the general email contact for the society. Include following documents:
        i. Cover letter
        ii. BHL-Europe in a nutshell
        iii. Review of own imprints
        iv. Introductory information sheet about your institution - if necessary

    - Try to answer questions that arise from the content providers to give the prospective content provider an understanding of how the processes work for you within BHL-Europe, providing a real and practical example.

    - Seek assistance from BHL-Europe WP2 lead - if necessary - for questions that arise.

    - Assist the prospective content provider with preparing to sign the MoU or connect the prospective content provider and WP2 lead for continuing negotiations.

2. *Use in poster sessions at conferences*

   - Print "BHL-Europe in a nutshell" for display at conferences, e.g. poster session.

   Although the main focus of attending conferences might not be acquisition of new content providers, it is an opportunity to publicise BHL-Europe, Europeana and your involvement with BHL-Europe. Consequently, this exposure might help indirectly with acquiring new content for the project and Europeana.

Feedback and lessons learned

Acquiring new content providers for BHL-Europe is not an easy task. RBGE sent out information about BHL-Europe to 32 prospective content providers within Scotland via email and received 3 responses, 2 asking for further information and one noting that the project sounded very interesting and he would forward the message to the society committee and put it on the agenda for the next meeting. Feedback from further telephone discussions centred around issues relating to IPR and the open access nature of the project. The feedback from advertising BHL-Europe in poster sessions was limited with only one request for further information. All requests have been followed up and to date there has been no response. At the time of writing, there were a few expressions of interest in the project, but unfortunately new content has not been secured for BHL-Europe. This shows that content acquisition is not a short-term process and much effort is needed to communicate with prospective content providers. The whole process from the beginning until signing the MoU might take several months.
The experience so far shows that there is a general support for the central aims of BHL-Europe - namely free access to biodiversity literature. However, there are two key problems that need to be overcome:

1. The matter of funding sources for the digitisation of potential new content.

2. Intellectual property rights (IPR) issues with potential content.

Taking each of these in turn, content is potentially available but funding for digitisation is not. In the current constrained financial climate it is increasingly more difficult to attract funds. The first problem might be avoided by first contacting only those institutions with digital content available on their own website. However, this makes it more difficult to attract those institutions for BHL-Europe as they already have their own online presence.
Questions relating to the issue of intellectual property rights, covered areas such as how do you identify, record and contact third party rights holders, and how much time does implementing and following through these processes take.

4.2.4 Outlook for the third project year

Content acquisition needs to be an ongoing process. Furthermore, building on experiences to date, this task needs to be taken up with additional effort by the whole project team not just the content management work package group, in order to increase the visibility of BHL-Europe to as broad a range of potential content providers. The process of communication needs to be simple and fast. The content management work package (WP2) has already taken time and effort to identify potential workers to operate within the area of content acquisition and to attempt to increase the effectiveness of work carried out in this important area.
During the last year of content acquisition it was noted that prospective content providers often request additional information for specific topics e.g. metadata requirements for BHL-Europe or licensing of content - IPR. Thus additional brief documents (1-2 pages) will be prepared about several topics (e.g. metadata, IPR, collection policy) using the “Delivery of the first version of the approved best practice guidelines and standards” D2.6 as a basis. These might be important for content acquisition to facilitate this ongoing task.
Further a list of potential content providers - with already available digitised content - including all necessary contact details has been prepared and will be used for content acquisition during the third project year. Following up communication with institutions that already have shown interest in BHL-Europe that may be willing to provide content is another subtask for content acquisition during the next project year.
Additionally, while working on content acquisition, it has been noticed, that the current MoU must be adapted for new content providers who are not currently part of the consortium. The current MoU includes links to documents (e.g. Description of Work, D2.1) that are not public and cannot be sent out to external persons. Furthermore legal consequences of the wording of the current MoU need clarification to prepare for the time after the end of the project. The revision of the MoU began at the end of the second project year and will be completed sometime in the third year of the project.

5 Content Analysis

Content within BHL-Europe will not be restricted by proprietary third-party rights or any other constraints which would limit its use in an open access environment using a Creative Commons licenses. The digital content must either be in the public domain, or else the content contributors must have permission from intellectual property owners to provide access under Creative Commons. In this part of the report we aim to give a detailed analysis of the freely available content provided by BHL-Europe content providers. The process of uploading content is described as well as the underlying content of BHL-Europe. BHL-Europe content that is available through Europeana is described in detail. Additionally the relationship between European cultural heritage and content within BHL-US and estimates of published biodiversity content within the European Union are presented. An outlook of the content management is given at the end of this chapter.

5.1 Content upload

Managing data upload to the BHL-Europe server began in autumn 2010 in cooperation with WP3. Before content providers were able to upload content to the BHL-Europe server hosted at the Natural History Museum (NHM) in London requirements had to be reviewed. Functional and technical requirements were gathered and then the feasibility of meeting these requirements was discussed with NHM staff, in line with NHM infrastructure and security constraints. Simultaneously all content providers were asked to provide primary and backup contacts for the content upload within their institution. In order to coordinate the required information a wiki page[7] was established. According to the information provided by our content providers and the functional and technical specifications, NHM installed the FTPS server and created for each content provider an account and sent out logon details in November 2010. It turned out that some content providers had difficulties using a FTPS connection within their own institution. Therefore all content providers were asked during the Content Provider and Technical meeting in London to check their FTPS connection and report any problems. The above mentioned wiki page contains a list of all BHL-Europe content providers, general information about the FTPS account and a ‘How to use’ section. For those content providers who had problems with the primary FTPS solution a second solution has been provided.
Content providers are asked individually to upload content to the NHM server. First they upload test content to the NHM server including the scanned images and the corresponding metadata according to the file submission guidelines[8]. The test content is then checked by AIT and by NHMW. AIT verifies if the file submission guidelines have been followed and NHMW starts the data mapping to the OLEF schema creating a configuration for every content provider to reflect the characteristics of the content provider data. After NHMW and AIT approve the uploaded content, the content provider is advised to continue the content upload.

[7] https://bhl.wikispaces.com/BHL-Europe_Data_Upload

[8] https://bhl.wikispaces.com/BHL-E_WP3_PREINGEST

5.2 Underlying content of BHL-Europe

The following section shows the underlying content of BHL-Europe according to the information given by our content providers within the signed Memorandum of Understanding or the Description of Work and the quantity already available on the BHL-Europe server.

	Quantity according to MoU [pages]			Quantity on BHL-Europe server [pages] (1)
European content providers	**April 2010**	**April 2011**	**April 2012**	
NHM (3) (4)	2,382,713	2,782,713	3,182,713	2,414,879
NMP	300	2,000	3,000	3,080
UGOE	100,000	100,000	100,000	
LANDOE	500,000	600,000	650,000	
HNHM	23,000	44,000	50,000	7,777
UCPH		50,000	100,000	
NAT	86,500	88,500	90,500	1,537
NBGB	5,000	35,000	50,000	78
RMCA	15,000	50,000	50,000	26,668
RBINS		30,000	70,000	
BnF	150,000	150,000	200,000	226,715
MNHN	112,000	172,000	232,000	100,000
CSIC (2)	21,000	15,000	14,000	2,896
RBGE	25,211	25,211	25,211	25,776
UH-Viikki	5,985	12,985	19,985	11,455
UBER	12,200	12,200	12,200	
UB-Bielefeld	21,800	21,800	21,800	
PAS			10,800	
Delibri			4,400	
GfBS			3,157	
MSN (5)			55,216	
MIZPAS			2,000	
Total amount (European content providers)	3,460,709	4,191,409	4,946,982	2,820,861

BHL-US content providers				Quantity on BHL-US server [pages] (1) (6)
BHL-US (without NHM)	30,200,335	36,000,000	39,000,000	31,262,876
Total amount (European & US content provider)	33,661,044	40,191,409	43,946,982	34,083,737

Table 5: Underlying content of BHL-Europe

(1)… Counted on 12.04.2011
(2)… According to DoW
(3) … NHM provides content to BHL-Europe & BHL-US. To facilitate process BHL-Europe harvests NHM content from BHL-US.
(4) … Transfer of content to BHL-Europe is to date of writing in planning stage.
(5)… MSN provides content to BHL-Europe over UGOE. This content is not included in the stated page numbers for UGOE.
(6)… Data transfer from BHL-US server (Woods Hole) to BHL-Europe server (London) is in planning stage and scheduled for May 2011.

Content upload to BHL-Europe server started in December 2010. Content providers were and are individually asked to upload content to BHL-Europe. Content providers with no content in Europeana or no institutional portal in which content is already made available online and thus have less experience in providing content received the highest priority for approaching them. These content providers may need more time and assistance with the process of uploading content that consists of scanned images or born digital content and the corresponding metadata. Up-to-date information about the process of uploading, metadata mapping and pre-ingest can be found on the wiki[9], as this is an ongoing process.

More detailed information about the specific partners is given in the following paragraph.

MfN (Nr. 22, Table 4) is not included in Table 5 because the number of pages is not available at this time. MfN is preparing to scan, but page estimates are not possible at the moment.
NHM is a content provider for BHL-Europe as well for BHL-US ((3), (4)). To facilitate the process of providing content for NHM and avoiding duplication of work, BHL-Europe harvests NHM content from BHL-US. At the time of writing, transfer of this content to BHL-Europe is at the planning stage. NHM contributed content has been included in the statistics as it is already available online on BHL-US.
MSN digitised zoological works in collaboration with our partner UGOE. 430 volumes had been delivered to Göttingen, but many journal volumes were not digitised for conservation reasons, and a few titles have been recently digitised by BHL-US partners. In total, 274

[9] https://bhl.wikispaces.com/Content+_Management

volumes (55,216 pages) have been digitised and will be uploaded by UGOE to BHL-Europe. The 55,216 pages are not included in the stated page numbers for UGOE in Table 5.

5.2.1 BHL-Europe ingest

Once content is delivered to BHL-Europe, it needs to be ingested into the BHL-Europe portal. First of all, mapping the metadata to the BHL-Europe schema needs to be completed for the pre-ingest stage and this is done within WP3 by NHMW. The following table gives a summary of the current status of this metadata mapping for content providers. After the preparation for mapping is finished, WP3 - more exactly the pre-ingest team - can ingest the available content to BHL-Europe.

Content Provider	Mapping status	Pre-ingest status
RBGE	finished	planned for Pre-ingest
RMCA	finished	planned for Pre-ingest
CSIC	finished	planned for Pre-ingest
UH-Viikki	finished	planned for Pre-ingest
BnF	finished	
NMP	finished	
NBGB	in progress, communication with CP	
MNHN	in progress, communication with CP	
NAT	in progress	
HNHM	in progress, communication with CP	

Table 6: Mapping and Pre-ingest status of BHL-Europe content providers

5.2.2 Interconnected repositories

In the summer of 2009, interconnection has been shown by WP3 to be technically possible for the following partners by mapping metadata from various BHL-Europe content providers and first ingesting content to a test server and then ingesting content to the portal prototype.

Nr. of CP	Content Provider
1	NAT
2	UGOE
3	RBGE
4	UH-Viikki
5	BnF
6	LANDOE
7	MNHN
8	NMP
9	NBGB

Nr. of CP	Content Provider
10	RBINS
11	CSIC
12	MOBOT
13	BHL-US
14	NHM
15	UBER
16	UB-Bielefeld

Table 7: Interconnected repositories within the portal prototype

5.3 Content available through Europeana

BHL-Europe acts as a content aggregator for Europeana and therefore provides the available biodiversity material not only through the BHL-Europe portal, but also through Europeana. To ensure information is up-to-date in Europeana a monthly re-harvesting has been scheduled.
At the time of writing, content from the following partners has been harvested by Europeana.
Europeana counts available content in items, while BHL-Europe counts pages. Therefore Table 8 shows the quantity of content available using the Europeana indicator (items) as well as the BHL-Europe indicator (pages).

Content Provider	Europeana indicator [items]	BHL-Europe indicator [pages]
LANDOE	3,511	540,062
NAT	3,516	96,604 (1)
UBER	60	12,200
UBBI	1634	20,449
UCPH	28	1,516
UH-Viikki	30	11,455
RBGE	183	25,776
RMCA	52	26,668
CSIC	9	2,896
NHM	4,170	2,414,879
BHL-US	83,402	31,262,876
Total amount	96,595	34,415,381

Table 8: BHL-Europe content in Europeana (28.04.2011)

(1)… Estimated numbers of volumes and corresponding pages given within the MoU have been used as data set for calculation of the mean value of pages per volume for each content provider if number of pages could not be counted. Calculation table is included in the Appendix.

According to Table 5, BHL-Europe will provide 43,946,982 pages to Europeana by the end of the project. By the end of the second project year BHL-Europe already has provided 78.3 % of the expected content to Europeana.

5.4 European cultural heritage and BHL-US

The relationship between content in BHL-US and European cultural heritage has already been explained in detail in the updated first content analysis & management status report - D2.4 provided in July 2010. However, the most important aspects are also stated within this report to demonstrate the necessity of a close collaboration with BHL-US.

European cultural heritage is distributed all over the world and is not only available within Europe. Biodiversity heritage literature is part of our European cultural heritage and is also scattered worldwide. For example, the Marine Biological Laboratory (MBL), founded 1888, is the oldest private biological laboratory in the United States. Since 1920, 56 Nobel Prize winners have been associated with MBL during their careers. Of these awardees, 22 were European citizens who spent part or all of their careers at MBL working for European institutions. Their careers often include responses to mid-20th century forces for which working at MBL was often part of a path leading to freedom from the upheaval caused by historical events. August Krogh, Otto Meyerhof, Salvador Luria and Albert Claude are only a few names of Nobel Prize awardees of European background that worked at MBL, and extended European cultural heritage outside of Europe. Hence, European researchers have travelled to the MBLWHOI (Marine Biological Laboratory Woods Hole Oceanographic Institution) Library for many years to read and study literature of European origin and also to enrich the libraries during their time in residence. It is known that some of these researchers worked in the MBLWHOI Library because publications of European origin were present there, but no longer present in Europe.
By collaborating with BHL-US, we are able to repatriate this European heritage literature back to Europe via BHL-Europe and Europeana and consequently enable European citizens to access this literature easily.
An estimation of the total BHL-US corpus in June 2010 showed that ~ 53 % of all pages available to that date were of European origin (published in Europe).

Continent	Amount of content [%]
Europe	53.2
North America	44
South America	0.2
Australia	0.8
Africa	0.2
Asia	1.3
Oceania	0.3

Table 9: Origin of BHL-US corpus (June 2010)

This high level of European content within BHL-US shows the importance of this content for BHL-Europe: in making BHL-US data and content interoperable with European content providers and managing the content acquisition process over all partners in cooperation with BHL-US, our European content providers avoid duplicating scanning efforts. Thus, collaborating with BHL-US also has the important economic consequence of saving tax money for Europeans. This aspect of the partnership should not be underestimated. At the time of writing there are 34 million pages available from BHL-US. Among these pages it is estimated that 18.1 million pages are of European origin.

5.5 Content availability in Europe

This section aims to give an estimate of available natural history content within BHL-Europe. The numbers presented are an extrapolation based on published content in Austria. Our consortium partner LANDOE has calculated the amount of natural history content already published and available for digitisation planning in Austria.

Table 10 shows the calculated numbers of available biodiversity content in Austria.

Type of material	Available content [pages]
Digitised Serials	500,000
Serials published, not digitised yet	700,000
Monographs published, not digitised yet [(1)]	1,000,000 - 1,500,000
Total amount	**2,200,000 - 2,700,000**

Table 10: Estimation of available biodiversity content in Austria

[(1)]… based on numbers from NHMW. Lower estimate of published biodiversity monographs is 3,000; upper estimate of monographs is 5,000. Assumption: mean value of pages for a monograph is 300 pages.

In Austria published natural history content ranges from 2.2 to 2.7 million pages. These estimated numbers and the population figures of Austria, Germany and Europe are used in the following calculation to estimate the published natural history content within Europe.

Nowadays, Austria has ~ 8 million inhabitants and published 2.2 - 2.7 million pages of natural history content. Applying this ratio to Germany, which has ~80 million inhabitants, the amount of published natural history content should be ten times higher and thus ranging from 22 - 27 million pages published in Germany. The European Union has a population of ~500 million which is ~5.5 times the population of Austria and Germany together. Therefore it is estimated that natural history content published within the European Union ranges from 133,100,000 -166,335,000 pages.

Region	Population [million]	Amount of pages
Austria	~8	2,200,000 - 2,700,000
Germany	~80	22,000,000 - 27,000,000
Austria & Germany	~90	24,200,00 - 29,700,000
European Union	~500	133,100,00 - 166,335,000

Table 11: Estimation of available natural history content available within the European Union

Further estimates made by our consortium partner LANDOE show that within the domain of earth sciences there is additionally a minimum of one million pages and within the domain of agriculture a minimum of 1.9 million pages available for digitisation for serials within the current borders of Austria.

These results indicate that a huge amount of biodiversity content is available and only a small share of this content can be made available within the BHL-Europe project time. Most of this content exists only as print versions and still needs to be digitised.

5.6 Outlook for the third project year

During the third project year all content providers will be asked to upload their remaining content to BHL-Europe as soon as possible to ensure that all content will be ingested to BHL-Europe and Europeana. All less experienced content providers have been approached during the second half of the second project year. During the first half of the third project year, partners with online content will be approached as well. Content that is already online should be more quickly and easily ingested into BHL-Europe.

Detailed plans for ingestion into Europeana and the BHL-Europe portal cannot be made in advance as this depends on the progress of content upload from the content providers' side as well as on the progress of the development of all portal modules. All content providers are asked simultaneously to provide content and early fall 2011 will be the deadline for finishing uploads of already digitised and thus available content. As soon as a content provider partly or completely finishes the upload, the BHL-Europe ingestion team will be asked to ingest the content according to their internal workload. Up-to-date information on the current status of each content provider can be found on the wiki[10].

[10] https://bhl.wikispaces.com/Content+ Management

Appendix

A: BHL-Europe in a nutshell - English version

BHL-Europe in a nutshell

BHL-Europe vision: Build a Digital Open Access Library for Biodiversity Literature.

- **What is BHL-Europe?**

The Biodiversity Heritage Library for Europe (BHL-Europe) is an ambitious and innovative European Commission funded program, started on 1 May 2009. BHL-Europe aims towards a digital global library of life by bringing together existing digital collections of biodiversity literature from all over Europe into one freely accessible online portal.

- **Why do we need BHL-Europe?**

The libraries of the European 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. As yet this wealth of knowledge is only currently available to those few people who can gain direct access to these collections. The body of biodiversity knowledge is thus effectively withheld from use for a multiplicity of potential users.
Much of the early published literature is rare or has limited global distribution and is available in only a very few libraries. From a research perspective, these collections are of exceptional value because the domain of systematic biology depends – more than any other natural science – upon historic literature. Once the collections of biodiversity literature are freely available on the Internet, this will be of great value to scientists, and also to a much wider public.

- **What is the focus of BHL-Europe?**

BHL-Europe focuses on the interoperability of existing European digital libraries and repositories with the goal of providing open access to the general public and scientists via one unique Web portal and Europeana. There is no single natural history museum or botanical garden library which holds the complete corpus of legacy literature. Therefore, BHL-Europe needs to be a federation of numerous institutions in which your institution can also be a part.

- **What are the benefits of becoming a partner of BHL-Europe?**

  - BHL-Europe increases the visibility of your content, thus increases the visibility of your institution.

  - BHL-Europe establishes a sustainable preservation and archive system to store, curate, manage and migrate your data. Thus BHL-Europe will have strategies and processes in place for long-term preservation of the data produced by biodiversity digitisation programmes.

  - BHL-Europe provides multilingual access to your content, thus saving you the costs for creating a multilingual portal.

  - BHL-Europe enriches your metadata so that every content provider can retrieve and reuse the enhanced data.

  - BHL-Europe provides you access to an important up-to-date and cost-effective pool of information and network of collaboration to help you use best practice approaches.

  - BHL-Europe provides tools and technologies to cost-effectively present the digital content and manage digitisation projects in the biodiversity domain.

  - BHL-Europe helps you to create good quality OCR text from the scanned page images for further data analysis and metadata enrichment.

  - BHL-Europe provides you access to Taxonomic Intelligence tools to facilitate the search for taxon specific information.

  - BHL-Europe dissemination activities will reach a large group of target users to increase the use of your data and content. Thus, BHL-Europe gives a cost-effective way to multiply the impact of your editorial activities and build a presence with users, the professional community, and other organisations including national governments.

  - BHL-Europe makes all the content available through Europeana – a search platform to a collection of European digital libraries. This will increase the visibility of BHL-Europe content, and it will also increase the visibility of your data. As Europeana enriches your metadata, this new metadata can be retrieved through BHL-Europe too. BHL-Europe benefits from the network and knowledge building capacity of Europeana, thus increasing the access to state-of- the-art research and technologies. A sustainable Europeana will also support the sustainability of BHL-Europe.

europeana
project group member

Intellectual Property Rights

The main aim of the BHL-Europe project is to make biodiversity knowledge accessible on an open access basis to a wide spectrum of end-users. The value the project delivers is not merely in making works available online, but in the open access terms under which the works are available. Therefore understanding Intellectual Property considerations is fundamental in achieving this objective.

The purpose of this summary is to outline Intellectual Property Rights (IPR) factors that need to be considered in relation to the scanning and online display of library biodiversity collections of BHL-Europe project partners. Scanning copyright works and making them publicly available without due consideration of rights ownership and clearance presents serious risks.

Adherence to BHL-Europe's Intellectual Property principles and requirements is essential to the success and sustainability of the project. All BHL-Europe partners must therefore understand the following:

- BHL-Europe partners are responsible for providing content and data to the project that does not infringe the intellectual property rights of third parties.
- The activities of BHL-Europe cross many legal jurisdictions in terms of the partners involved, origin of material to be scanned and applicable IPR legislation. It is therefore necessary for project partners to know the relevant legislation of their own country.
- The project shares its data (and some content) with other projects, including BHL, EOL and Europeana. It is essential that project partners ensure that the works and data they make available to BHL-Europe are also licensed for sharing with each other and these other projects.
- BHL-Europe partners must ensure that digital content provided to the project are licensed under Creative Commons open access terms as stipulated in the project's Description of Work to the European Commission.
- Neither the project BHL-Europe nor the data/content providers will seek to assert any intellectual property rights over digital copies of public domain original work. Original public domain material can be reused or exploited by anyone who wishes to use it including educational, non-commercial, and commercial users.

europeana
project group member

- The use of technical protection measures such as visible digital watermarking or copy prevention are not compatible with the Open Access principles of BHL-Europe. Such enforcement technologies cannot respect or acknowledge the subtle and subjective concepts of fair dealing or fair use or other exceptions to the monopoly rights copyright confers.

- Neither data nor content provided to BHL-Europe may be restricted by subscription fees or other access controls that require payment.

New BHL-Europe partners will be required to sign a memorandum of understanding which further details the terms upon which content and data are made available to the project. BHL-Europe will provide project partners with an IPR best practice guide that includes further information about IPR risk management, rights clearance, due diligence and Creative Commons licensing.

B: Calculation table for estimation of number of pages within Europeana

	Quantity according to MoU [pages]			Quantity [volumes]			Mean value [pages per volume]	Items in Europeana	Pages [BHL-Europe indicator]
European Content Providers	**April 2010**	**April 2011**	**April 2012**	**April 2010**	**April 2011**	**April 2012**			
NAT	86,500	88,500	90,500	3,153	3,220	3,290	27	3,516	96,604