PROJECT DOCUMENTATION

INTERNAL INTERIM PROGRESS REPORT

Project: BHL-Europe

Release: Final

Date: 2 March 2010

Author: Henning Scholz

Table of Contents

Internal Interim Progress Report

1 Purpose

To report on a quarterly base the status of work for each member of the consortium. This report will not duplicate the content of the offcial progress and annual reports requested by the EC every 6 and 12 months, respectively.

2 Follow-ups from previous reports

For further information on past activities I refer to D1.1 - Progress Report 1, provided in November 2009 (https://bhl.wikispaces.com/file/view/BHL-E_D1pt1_091107.pdf).

3 Activities

Last deliverables: We submitted the last deliverables due at M6 of the project on 27 November 2009. They arrived in Luxembourg safely. All deliverables we submitted so far were accepted by the EC.

Project Server: NHM moved its hardware equipment in December to a different building. This delayed some necessary work on the project server environment. In the first days of 2010 these problems were solved and all webVPN data were provided for the partners. Jana Hoffmann spent four days in London to further adapt the functionalities of the project server. As the workplans of the five WP are not finished yet, the project server environment is not really useful for the BHL-Europe staff members.

Memorandum of Understanding: During the last project meeting in Prague we discussed and finalised the wording of the BHL-Europe MoU. It was sent of for filling the underlying content appendix and signature. To date, I have received 11 out of 16, properly filled and signed.

Google Groups: To facilitate communication on specific topics and to initiate working groups, we established so far three Google Groups. The first one was created before the meeting in Prague in early November 2009 and is focused on BHL-Europe technology. It currently has 16 members and to date 95 contributions were collected. A second group was established on 3 December 2009. This group is discussing BHL-Europe metadata details and to date 18 team members provided 42 contributions. A third group has 9 members and is discussing the work towards D2.2. To date 33 contributions were collected. Altogether 18 BHL-Europe team members are very active users of these groups and involved in various discussion threads.

BHL-Europe central index: Although the DoW of BHL-Europe gives a clear indication of the requirements of a central bibliographic database (Task 2.1.3), the technical concept and the plan to realise this was not worked out in detail. A meeting was held in London from 8-9 December to prepare that concept. BHL-Europe was represented by Henning Scholz, Adrian Smales, Graham Higley, and Dennis Zielke. As we also discussed the possible interaction with the ViTaL activity of the EDIT project (see below), Boris Jacob was representing EDIT during that meeting. The central bibliographic database of BHL-Europe will be called GRIB, i.e. Global References Index to Biodiversity. Further details are available in D2.2.

Evaluation of possible GRIB platforms (BHL serial list, GVK): Between December and January we evaluated the available platforms for GRIB. Henning Scholz, Dennis Zielke, Wolfgang

Koller, Michael Malicky, Chris Sleep, and Boris Jacob are the team members involved in the evaluation and testing process. The first option to work with is the BHL serial list currently running under http://www.nhm.ac.uk/library/bhlseriallist/index.php/. This is now transformed in a BHL-Europe bidlist containing data of European partners (Vienna, Linz). The lessons learned and details are provided with D2.2. As a summary it can be stated that it still needs a lot of development work to move this system into a useful tool for librarians and users. As an alternative, we discussed and evaluated a solution provided by the GBV, the Common Library Network of more than 400 German libraries. The GBV has more than 20 years of experience in building library catalogues. The GBV is very interested in cooperatin with us and building our union catalogue with the full functionality we need (deduplication, bidding, user interaction, link to digital page images in BHL). As the technology is all ready and just needs to be recombined for us, it would save BHL-Europe lots of resources that we can use to improve the BHL portal functionality and data quality. The GBV is a non-profit organisation that is well funded and sustainable in its own to host our library data and work as a service provider for us even beyond the end of the BHL-Europe project in 2012. It is a very cost efficient solution. It is also low risk for us using their mature and high performance system. Therefore, the PMG of BHL-Europe is very much in favor chosing GBV to build our union catalogue (GRIB, i.e. D2.2, D2.3, D2.5). If we experience no objections from members of the BHL-Europe consortium, we would contracting the GBV as a service provider. More details of the evaluation process are provided with D2.2 within the next few weeks. By then we are also able to present a first GBV pilot based on real library catalogue data of European partners. This is currently work in progress, but will be ready hopefully by mid of February. We are using the partner feedback and our experiences so far to decide during our Berlin meeting in February how to proceed with the GRIB development and how to integrate GRIB in the BHL-Europe system.

Cooperation of BHL-Europe and ViTaL: The Virtual Taxonomic Library (ViTaL) is facilitating the discovery and accessibility of taxonomically relevant literature through the provision of four main services, two of them are important for the content selection of BHL-Europe. (1) ViTaL will include a bibliographic reference aggregator, which will harvest references from a range of sources and to provide a centralised bibliographic resource. (2) ViTaL will provide a facility that will enable users of taxonomic literature to nominate suitable materials for digitisation. ViTaL is being developed in consultation with the EDIT applications development team based in Berlin at the Botanical Gardens (FUB-BGBM). It was agreed recently (27 January) that ViTaL will use the GBV as a service provider and cooperate very intensively with BHL-Europe. This cooperation is also documented by two new staff members currently hired by MfN to work closely with the BHL-Europe team at MfN. The goal is, to build one central bibliographic index system for ViTaL and BHL-Europe using the technology of the GBV. Thus, GRIB is a joint development of two European projects that helps us covering most of the European natural history institutions in providing a sustainable union catalogue for the taxonomic community.

Evaluation of Vifabio as a possible partner: Funded by the DFG (German Research Foundation), a virtual library of biology literature is establish at the university library in Frankfurt/Main: Vifabio (http://www.vifabio.de/?lang=en). During a meeting in Frankfurt (16 December) Henning Scholz and Boris Jacob discussed options for cooperation with BHL-Europe. We also compared advantages of disadvantages of metasearch engines versus index systems, discussed technologies and future directions of both projects. As BHL-Europe is building not a metasearch engine, the potential overlap is relatively low. However, we still might support each other if we work together with the same partners. One example is NHMW that is a partner of Vifabio and BHL-Europe. Obviously both projects already benfits from each other by working together with identical partners.

Analysis of domain content: Our partner libraries are active in analysing relevant literature for scanning operations. Some examples are from:

UH-Viikki - http://www.refworks.com/refshare/?site=014941135929600000/RWWS4A1009351/074281243853638000,

RMCA - http://193.190.223.46/wiki_ext/index.php/List_RMCA),
NBGB -http://193.190.223.46/wiki_ext/index.php/List_NBGB.

Extension of partner network: One important aspect of WP2 is to attract new content providers (Task 2.3.2). BHL-Europe has to enlarge its network of content provider from 16 to 30 at the end of the project. To date we negotiating with the university library Bielefeld. This institutions is running a project focused on the digitisation of German language journals from the 18th and 19th century (http://www.ub.uni-bielefeld.de/diglib/aufklaerung/index.htm), including biodiversity content (http://www.ub.uni-bielefeld.de/diglib/aufkl/naturforscher/naturforscher.htm). The Humboldt University Berlin, one of our project partner but no content provider so far, is also running a repository with biodiversity content, e.g. http://edoc.hu-berlin.de/ebind/mfn/keller1-2005-Mn01674331/XML/index.xml. We have to work out how to get this content. Eventually, the colleagues from Madrid (CSIC) are approaching their partner institutions to provide content for BHL-Europe. The Real Jardin Botanico has interest in sharing their biodiversity content with BHL-Europe. Further details of this sharing are still in discussion.

BHL content storage: As a first step to move the BHL content from the US to Europe, BHL-Europe provided a hardware kit to BHL. This kit was set up in MBL Woods Hole by Adrian Smales during a 5 days workshop before Christmas 2009. To date, almost 4Tb of BHL content are downloaded from Internet Archive to this system. Once the download is finished, the BHL content is ready to be transferred to Europe. A big storage solution will be purchased for BHL-Europe to be set up in NHM London. After a lot of configuration and budget discussions (including the EC) the system is now ready to be ordered. Once the system is set up, BHL-Europe is ready to receive and store the BHL data. From that time onward we can also start harvesting our European partner for the German prototype of BHL-Europe in Autumn 2010.

Technical Director for the German prototype of BHL-Europe: It was discussed in Prague to have technical director for the German prototype of BHL-Europe. We appointed Lee Namba (Atos) to fill that position. Lee also was the deliverable responsible for D3.4, which was provided as a draft on 16 January 2010.

Europeana ingest plan: A very important task of the last weeks with several rounds of discussion was the planning of the ingest of BHL-Europe content by Europeana. In a first step, a dump of BHL metadata and thumbnails will be provided to AIT to do the mapping to ESE before end of March 2010. Currently, AIT is testing the mapping using the Europeana Content Checker. We will then add 1-2 BHL-Europe content provider for a re-harvesting in June 2010. This content will then be available for the Europeana Rhine release this summer. The following re-harvesting steps were also specified. EDLF, AIT, Henning Scholz Adrian Smales, and Chris Freeland were involved in this process.

BHL user survey: BHL-US is currently preparing a questionnaire for a user survey. Henning Scholz and Francisco Welter-Schultes are aligning with BHL. Currently it is planned to pre-date our survey and launch the BHL-US and BHL-Europe survey at the same time. The work on the questionnaire is still in progress and some more details are available on the wiki: https://bhl.wikispaces.com/BHL-US+Survey.

Dissemination activities: BHL-Europe was active in participating in various events. Some examples are the Biologentag (Berlin, 21 November 2009), the EDIT General Meeting (Portugal, 15-18 December 2009), the official launch of the Year of Biodiversity (Berlin, 11 January 2010).

Improvement of website: The colleagues at NMP are continuously improving the BHL-Europe website at http://www.bhl-europe.eu. This work will continue for the rest of the project.

Workshop participation: Tom Gilissen (NAT) was attending DISH 2009 on 10 December 2009 to learn about user participation in Europeana (http://www.dish2009.nl/node/130). Digital Strategies for Heritage (DISH) is a new biannual international conference on digital heritage and the opportunities it offers to cultural organisations.

KeyToNature: Since 13 December 2009, BHL-Europe is Associated Member of KeyToNature (http://www.keytonature.eu/wiki/Associated_Members#International).

ICT PSP proposal: The ICT Policy Support Programme (or ICT PSP) aims at stimulating innovation and competitiveness through the wider uptake and best use of ICT by citizens, governments and businesses. The 2010 Work Programme has been formally approved, the 4th call is open as of 21 January until 1 June 2010 (see http://ec.europa.eu/information_society/activities/ict_psp/index_en.htm). It is currently discussed in the BHL-Europe consortium to prepare a proposal. RBINS, RMCA, MSN, MfN, NHM are among the institutions that are interested to participate in the bid.

4 Deliverables

12	D2.2 Prototypes of deduplication tool and bibliographic database system for monographs and serials
13	D3.4 Implement plans for all components in WP3, incl. data models, technology standards etc.
14	D4.1 Delivery of IPR working documents, including best practice guide, due diligence guide, pro forma agreements and process for formally agreeing rights management with rights holders. Complete agreement with EUROPEANA and BHL for reciprocal access and Rights metadata.

5 Schedule status / problems / issues / deviations

The leadership of WP2 is the most important issue we currently have. Bernard Scaife had this position during the first three months before Kai Stalmann took over as planned in August. Bernard then left NHM and is not working anymore for BHL or BHL-Europe. After 3.5 months on the project, Kai stopped working for BHL-Europe and officially left the project end of December 2009. The result of these changes are threefold: First, it required lots of re-organisation of actual work. Second, the engagement of BHL-Europe content provider and the distribution of work among them was not pushed the way it was envisaged. Third, as both Bernard and Kai were dealing with software development, we have lost code and the particular expertise for some specific products. MfN is currently in progress to hire a replacement for Kai as NHM is in progress to hire a replacement for Bernard. We hope this will be successful in spring 2010 to have more resources driving the work of BHL-Europe WP2. At the moment, BHL-Europe benefits from the ViTaL activity of the EDIT project both with resources and expertise, to fill the gap left by Bernard and Kai.

A second issue is our Project Server. Although the system is fully working and functional right now and all project members should have full access to the server, the detailed project plans are not ready and online yet to allow everybody seeing the work ahead of us in detail. We are planning to solve this by end of February after we've met in Berlin for the next project meeting.

6 Work planned

The work planned for the next phase is explained in more detail in Annex 1. On the IT development side, we have to move towards the 2nd generation of GRIB by end of April (D2.3). We also have to provide the final and detailed plan how to implement the German prototype (D3.5). The preparation of the survey of BHL Portal users is a third important work that needs to be done (D5.7). In the following I will to present some more details of the BHL-Europe roadmap for the next three months (February to April) and further in the future.

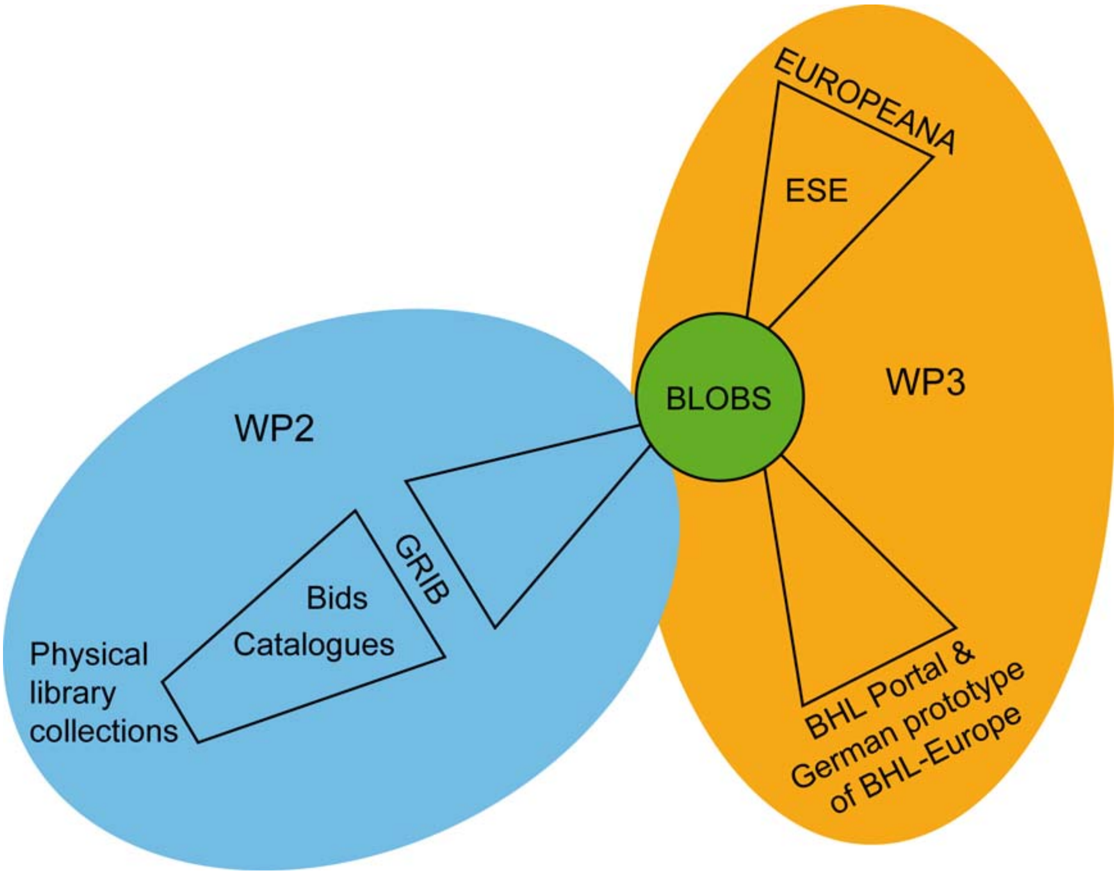
EUROPEANA
ESE
WP3
BLOBS
WP2
GRIB
Bids
Catalogues
Physical library collections
BHL Portal & German prototype of BHL-Europe

Figure 1. The Binary Large Objects (BLOBS = scanned page images) are in the centre of the approach. The Global References Index to Biodiversity is composed on the catalogue records of the physical library collections, the index and the link to the BLOBS. It is an important component of WP2 of BHL-Europe and provides one access route for the digital images files in the repository. It is currently thought to be the access route of the librarians managing the content hosted by BHL and BHL-Europe. The actual content integration of the BLOBS will be done by WP3. The BHL-Europe users (taxonomists, general public) will access the BLOBS either through the BHL Portal or Europeana. For them, WP3 also improves the functionality of the BHL Portal and connects the BHL-Europe content to Europeana.

BHL-Europe is expected to deliver **three products** (prototype versions first) in the following months (Figure 1). The first product is the access of BHL content via **Europeana** for the Rhine release (see http://version1.europeana.eu/web/europeana-project/home for more details of Europeana v1.0). This work is currently in progress mainly by AIT in cooperation EDLF. As we don’t have established yet our BHL-Europe repository, the first step is a provisional one. A dump of metadata will be hosted by AIT together with the thumbnails. AIT will map the existing metadata to ESE to be interoperable with Europeana. This is currently in progress for the BHL data. With LANDOE and NAT two of our European partners are identified to provide their metadata for the Rhine release. The tests with their data is also in progress. Once we have established our own repository and uploaded content from our European partners to this system, we can provide more content to Europeana. Several re-harvesting periods are identified to update the BHL-Europe content in Europeana. After the establishment of the German prototype of BHL-Europe in fall 2010, we are able for the first time to provide content to Europeana from a central European source. From that time onwards, all content that is available in our prototype system is also available through Europeana.

The second product we have to build is the **German language prototype** mentioned above (D3.6). This is expected to be finished end of October 2010. The main development work will be done by AIT and ATOS with support from other partners which still needs to be specified. The series of documents prepared by WP3 (D3.2 & D3.3 in November 2009 coordinated by AIT, D3.4 in January 2010 coordinated by ATOS, D3.5 in April 2010) are necessary to specify this system step by step and identify the steps to develop and implement the prototype system. The German language prototype from a user perspective can be envisaged as the existing BHL Portal with some new features and two operating languages, i.e. English and German. This is the first step to make the BHL Portal a multilingual portal.

For the foundation of the prototype development we need a lot of different input from various partners and groups of the BHL-Europe consortium and beyond. We have had a first **user survey** last year to collect experiences with user surveys, but of course also to collect first feedback on user requirements. These user requirements are a first input the the IT development work as well. In a second step, we now adapt our questionnaire for the first large user survey this spring. Currently we identify our strategic goals for that survey and work on the questionnaire. This survey will be a review of the existing BHL Portal to identify user needs for a new BHL Portal in the future. The results of the survey will feed the IT development of the prototype to a large amount. The survey is currently prepared by UGOE, MfN, NMP, and BHL.

We also need to identify **use cases** to enhance the functionality of the portal. For example, it might be required to have a specific image search functionality that is helpful when working with species identification keys. A workgroup headed by Heimo Rainer (NHMW) was created to go into more detail here and the results will also help the IT development team in building the prototype. The user survey will be also used to test the necessity of new functions.

It is very important to define **data standards and data quality** for the BHL-Europe metadata. A metadata workgroup was established by Wolfgang Koller (NHMW) that is currently discussing using Google Group. Parallel to these discussions, AIT will analyse the test datasets provided by our partners last summer to understand differences. This needs to be discussed with the libarians as well to get more background information. A **content provider meeting** is foreseen in March to bring librarians from all partners together for proper alignment. All these discussions and analyses are very important to identify the ideal and best quality level we can achieve with the resources we have in our project. This eventually will define the work effort required by every partner to fit their data to the standards and quality we agree on. Once this is finished every content provider can start working on their data. Depending the work effort this may also define the schedule for harvesting data by BHL-Europe.

Based on the information we have received with the first library questionnaire and the BHL-Europe MoU we (AIT, MfN) have developed a first idea of the **content ingestion plan for the portal**. For the German prototype (October 2010) we have selected the following partners to contribute their content: UGOE, LANDOE, NAT, RMCA, BnF, MNHN, UH-Viikki. By January 2011 we should have finished the ingestion of NMP, RBGE, UCPH, NBGB, CSIC. By April 2011 we should have finished the ingestion of RBINS, HNHM, and hopefully Wiley. We keep those partners with now content to date for the last ingestion phase in winter 2011. This ingestion planning is still in discussion and subject to change. The content provider meeting this spring is also a good opportunity to align with each other to find the best way forward. The ingestion planning will be coordinated by Gerda Koch (AIT) as she has lots of experience with such tasks from other projects. It is very important to have in mind at this point, that the coordination of content ingestion is a WP3 task, whereas the actual work on the data (mapping, quality control, data enhancement) in the local libraries needs to be charged to WP2.

The third product we have to build and deliver is a system to manage content contributions and the digitisation progress in all our projects (D2.2, D2.3, D2.5). The **Global References Index to Biodiversity (GRIB)** is designed to fulfill this function. GRIB actually is a database of biodiversity literature that indicates:

a) monographs and serials that are relevant for the biodiversity community (i.e. the library catalogue records of the BHL-Europe partners)
b) the distribution of this relevant literature in the partner libraries
c) the portion that is already available in digital form (BHL Portal)
d) the portion that is in the process of being digitised (BHL-Europe partners)
e) the portion for which plans have been created for digitisation
f) responsibilities for content contribution (bidding process).

The further evaluation of options and building of demonstrators and prototypes will be coordinated in the next weeks from Berlin (MfN) by Boris Jacob. MfN is about to hire two more staff members to support the IT development of GRIB and to fill the WP2 lead of BHL-Europe again. Based on the previous working relationship in evaluating and enhancing the BHL serial bidlist, MfN will closely cooperate with FUB-BGBM, UGOE, NHMW, LANDOE, and UBER for the further development of GRIB.

As GRIB is a sort of union catalogue of our library holdings, we need to bring the library catalogues of all our partner libraries with biodiversity content in GRIB. That is a process we have to do in addition to the ingestion of metadata records associated with BLOBS and it is therefore also interesting for libraries that don't have BLOBS yet, but still very valuable libarary catalogues. We need to establish procedures how to do that and develop an ingest plan for this. As a first step, MfN will send another library questionnaire in March latest to update the information from the library questionnaire from last year and get some additional information. A first data ingest is expected not before May this year. As we have test data already available, we can do the first prototyping without additional data to deliver the D2.2 and D2.3.

At one point, GRIB needs to integrate the current work of partners in building list of publications to be scanned. From that date onwards, all commitments to scan can be managed by GRIB and every partner can use GRIB to bid on items to scan. Following the current workplan and assuming the next steps run smoothly, it is expected from May onwards.

An additional work of BHL-Europe WP2 is working on a best practice guide for scanning operations. For this purpose a wiki page was created that have not received a lot of attention: https://bhl.wikispaces.com/BHLE_BP-guide. We first need some more input before we can put this in a formal document. Following our contract, we need to deliver a first version by April 2011. It may be worth establishing a group to develop draft guidelines that can be extended over the next months. I encourage experienced BHL-Europe team members to indicate if they are ready to join such a working group.

I don't want to forget to mention that we still have to finish our consortium agreement. It is only one section that needs particular attention and we have received advice by experienced lawyers to finish this. The recent deadlines of WP4 prevent us from implement the advice and finish the document. This hopefully will be done after the Berlin meeting.

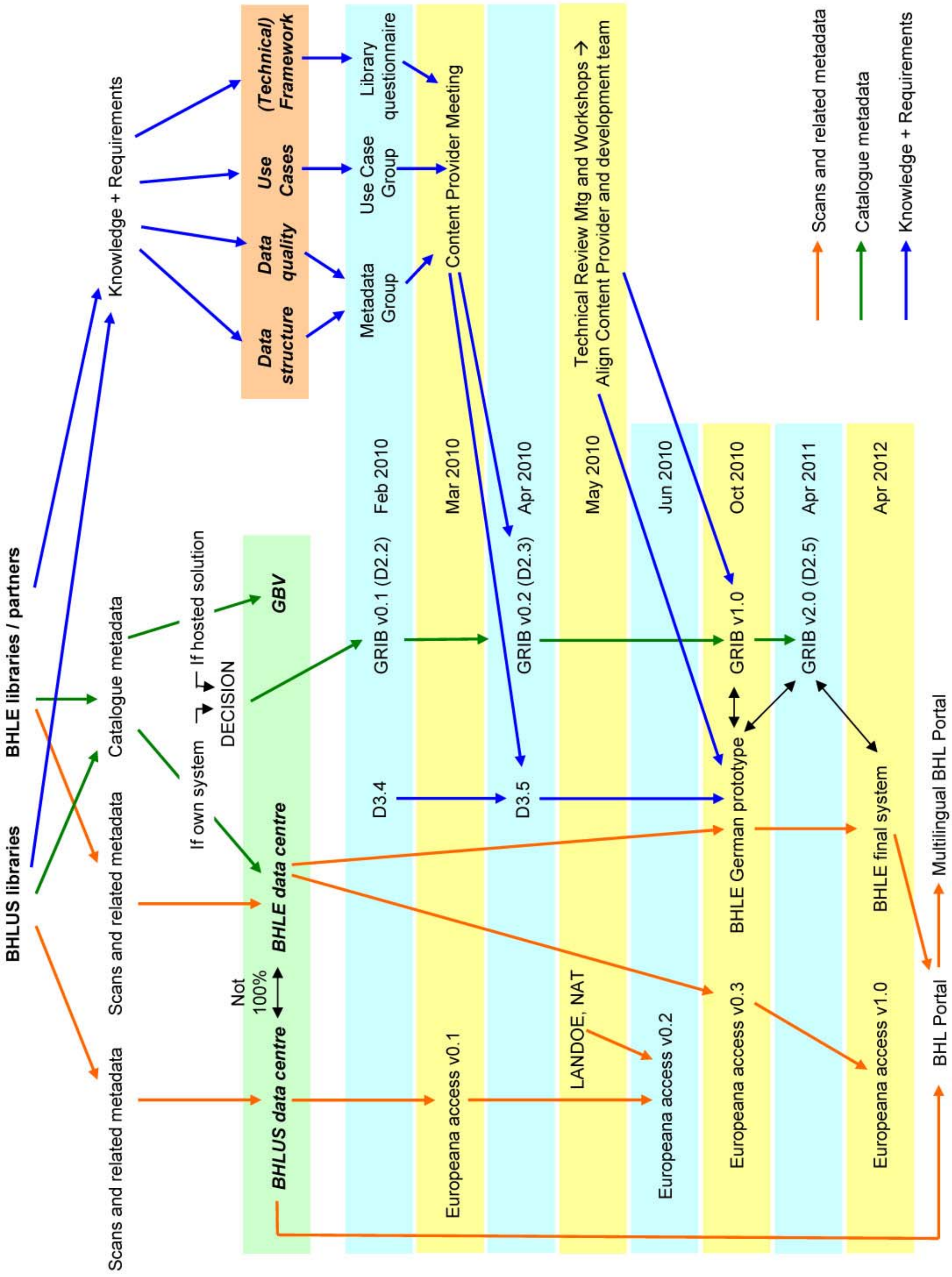

Figure 2. High level relationship of institutions, data and products in BHL-Europe. For the various GRIB versions, the ingestion plan is not set and depends very much on the results of the content provider meeting in March. It is anticipated to have all our library catalogues in GRIB v1.0. For the start, we have chosen the catalogues of MfN, NHM, FUB-BGBM, and NAT (GRIB v0.1). The ingestion of scans is explained above. For the German prototype we anticipate UGOE, LANDOE, NAT, RMCA, BnF, MNHN, and UH-Viikki to be ingested. For the final version all our other content provider plus additional voluntary content provider have to be ingested. For a different view on the planning process and the responsibilities see the Gantt chart below (Figure 3).

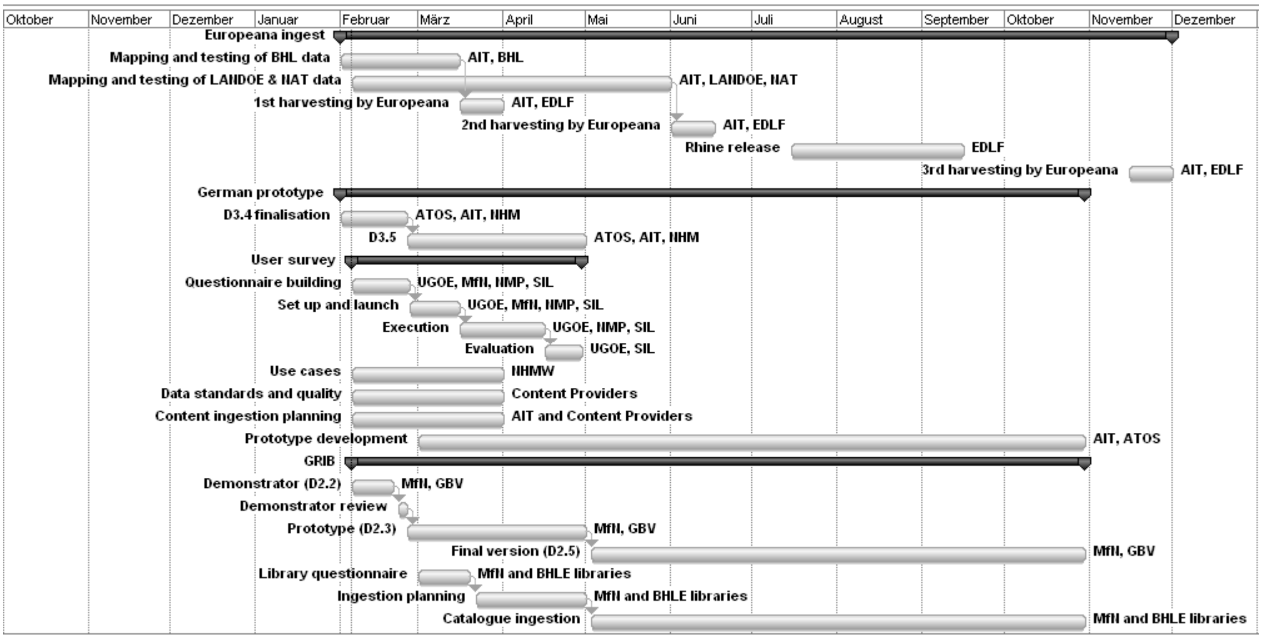

Figure 3. High level overview of the tasks described above with the timeline, resources and dependencies.

7 Products to be completed in the next phase

15	D1.2 Progress Report 2 including pre-financing request
16	D1.3 Annual Report 1 including first ideas for BHL-Europe business plan
17	D2.3 Prototype of Web-database for content management and collection analysis
18	D2.4 Content analysis and management status report 1 (metadata, page numbers, content providers)
19	D3.5 Technical architecture status and progress report with particular focus on the development of the German prototype
20	D5.7 Online questionnaires for user survey