Global BHL Quarterly Report (March to June 2012[1])

– The Biodiversity Heritage Library for Europe –

Henning Scholz
Museum für Naturkunde Berlin

Introduction

The BHL-Europe project has come to an end on 30 April 2012. By that time all deliverables for the project closure were in progress to be finalised on time for the final review meeting on 6 June 2012. The most important work in the last quarter was the start of the ingest process after the Pre-Ingest tool was delivered for production. Other highlights of the reporting period were the BHL-Europe final events between 2 and 6 June 2012 and the Global BHL meeting from 7 to 8 June, all hosted by Museum für Naturkunde Berlin. For a report on these meetings and events, please see the BHL-Europe[2] and BHL-US blogs[3].

Content statistics

It is still difficult to analyse the full corpus of content available on the BHL-Europer servers. However, based on the Memorandum of Understanding signed by our 28 content providers, BHL-Europe will make about 5.3 million pages of biodiversity literature from European content providers available to BHL. More content is coming as most of our current providers have ongoing digitisation programmes. Some of our partners recently started their digitisation work and are now also active content providers of BHL-Europe. Among them is the Museum of Zoology of the Polish Academy of Sciences and the Museum für Naturkunde Berlin.

New software and functional development

Best Practice guidelines and standards:

The Approved Best Practice Guidelines and Standards provide the first standardised guidelines for processes used by BHL-Europe for the digitisation of biodiversity content. The Best Practice Guide is designed to be easily understood by all persons using them; in particular, it is designed to guide prospective and current content providers simply and clearly through a digitisation workflow from either a print or digital version of an original publication to the final digitised form of that publication in BHL-Europe.

The Best Practice Guide is of special interest to technology users – in particular libraries, digitisation centres and digital library networks. Both, existing and new content providers will find the Best Practice Guide useful to simplify and speed up the whole process of digitisation. Hence an efficient digitisation workflow will be ensured and a direct connection with Europeana and BHL-US will be enabled.

The Best Practice Guide is also a means of developing the case for long-term sustainability of BHL-Europe. BHL-Europe content providers will provide more then 25 million pages of biodiversity literature by the end of the project in April 2012. However, there are many more pages of potential content that could be included in BHL-Europe but which cannot be, due to the limited timeframe of the project. Therefore, it is anticipated that new partners will join the digitisation process and in the future provide digitised biodiversity literature to BHL-Europe and thus to the European citizens. The Best Practice Guide recommends the most efficient way to do so and aims not only to assist BHL-

[1] This is not really a quarter. Based on the agreement during the 3rd gBHL meeting the global reports should cover the calender year. Therefore, I compiled the report covering the time period between the last report and the full quarter. The next report will then cover the quarter properly.
[2] http://bhleurope.blogspot.de/2012/06/bhl-europe-final-meeting-4-june-2012-6.html
[3] http://blog.biodiversitylibrary.org/2012/06/international-outlook-celebrating-bhl.html

Europe content providers, but also prospective content providers after the current BHL-Europe project has ended.

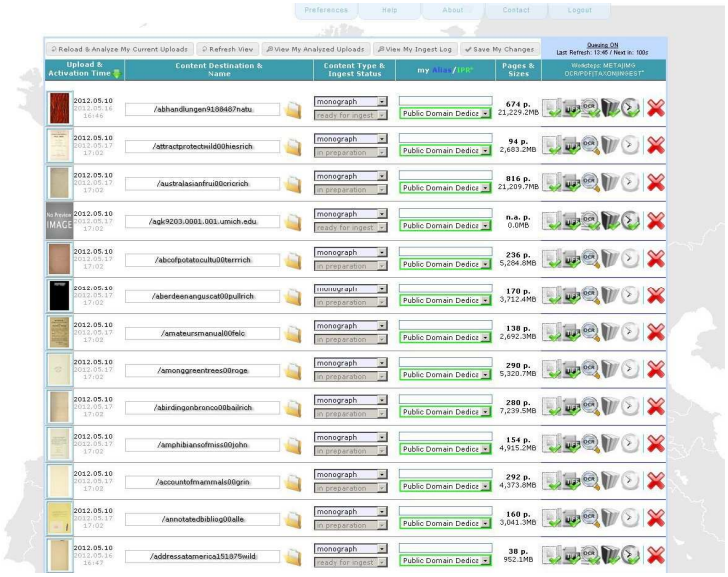
To summarise, this document unifies and simplifies the process of making literature available online. It also provides advice on IPR issues and how to tackle them. The Best Practice Guide is now available as a formal deliverable (D2.9: http://bit.ly/KCyYVB) but also as an illustrated book (D5.11: http://bit.ly/KCx9rN).

Pre-Ingest Tool

The Pre-Ingest component is a set of processing steps and systems which are orchestrated to facilitate data submission, harmonisation and enrichment. This also involves communication and feedback loops with content providers and developers.

This component is the interface to the archives and acts as an adapter for the Ingest module. As external partners store metadata in various formats, the native formats need to be converted, harmonised, enriched and prepared for ingestion. This step is needed for the ingestion, multilingual search, data harmonisation, indexation and search requirement.

The Pre-Ingest workflow describes the necessary interaction steps to prepare the content for ingest (see also Fig. 5 below). We have carefully investigated other best-practice approaches on how to setup a Pre-Ingest workflow (Archivematica, California Digital Library). Both approaches work with micro-services for several functionalities needed during Pre-Ingest. We compared both approaches and isolated important micro-services which are relevant for Pre-Ingest. We amended the workflow by specific requirements of our dataset, e.g. implementing the TaxonFinder to extract scientific names of animals and plants from the OCR text of the scans. The resulting workflow model can be seen as a best-practice based approach of how a Pre-Ingest process generally could work. For more details on the Pre-Ingest workflow please refer to D3.7[4]. The tool is accessible under http://www.bhl-europe.eu/preingest (Fig. 4) and the code is available under the Modified BSD License via GitHub (see also below). The Pre-Ingest tool now is delivered and tested extensively. It is now working in production.



Screenshot of the Pre-Ingest tool showing content in progress to be processed through the workflow.

[4] http://www.bhle.eu/en/outcomes/documents/key-components-documented-for-output-of-d35-eg-bhl-europe-portal-ocr-demonstrator

Further Development of the Portal System

Over the past couple of months our developers have made continuous improvements to the BHL-Europe system as well as implementing new features for the Portal. Many of these tasks consist of configuration changes and maintenance of the overall performance of the system. This involved the following:

- Bug Fixes: our team of developers have been busy correcting any bugs that were found by the testing team and reported in GitHub. We also corrected bugs found during the public user survey end of April[5].
- Implementation of the Gemini feedback functionality in Drupal for the BHL-Europe Portal; see https://bhl.wikispaces.com/Gemini for the documentation.
- More work was done to manage and display the content hierarchy of serials properly.
- In the final period of the project, we have worked with Species 2000 to implement a reverse look-up in the Catalogue of Life (CoL) and thus establish a better connection between BHL-Europe and CoL. After some preliminary investigations, Species 2000 is now connecting with the Solr access point that is also used by the BHL-Europe Portal. This way, an URL is given in case a title listed in the CoL is available by BHL-Europe.
- Finalisation of the documentation for the BHL-Europe system, including a tutorial[6]. BHL-Europe also provides tutorial videos via its YouTube channel[7].

Collaboration on OCR improvements (IMPACT) – Experiments

After production of the ground truth files in the last reporting period, we have now started experimenting in this period. The first experiments on the data showed an interesting problem for the tools. The BHL test set contained a large number of images and plates. The image pre-processing tools of IMPACT are, however, made for texts and the large number of plates caused serious problems for the algorithm of those tools. These tools are looking for straight lines to optimise the geometric properties of the pages. As they were unable to find such straight lines on a number of pages, the results of the evaluation were unacceptable. The separation of plate and text pages was also not successful. Therefore, the experiments were continued without the image enhancement tools.

Our experiments have shown that Tesseract 3.00 as implemented now is giving very good OCR result. The average quality is only 5% lower compared to the new IMPACT version of the ABBYY Fine Reader 10. In addition to the language indicator in the metadata of an item, a font type indicator is important to add, to facilitate a better transcription of different font types. We decided to publish the ground truth dataset to facilitate the uptake of the data by Tesseract, for example, in order to improve the OCR engine for future releases. A more detailed overview of the outcomes of the collaboration with IMPACT is provided in a separate report[8].

Website statistics

BHL-Europe is currently not monitoring usage of content through BHL-Europe as the portal is not online. However, BHL-Europe drives usage of BHL content through two products, the Biodiversity Library Exhibition (BLE) and Europeana.

- Statistics for BLE itself between 1 March and 18 June 2012: 5,719 visits, 14,245 pageviews, 82.5% new visitors, 58.4% bounce rate; 31.8% visits from CZ, 10% from US, 6.2% from DE.

[5] See here: http://www.bhl-europe.eu/de/publikationen/dokumente/second-user-evaluation-report
[6] http://www.bhl-europe.eu/de/publikationen/dokumente/live-bhl-europe-system-with-distributed-storage-and-management-and-appropria
[7] http://www.youtube.com/user/bhleurope?feature=results_main
[8] http://www.bhl-europe.eu/de/publikationen/dokumente/report-on-the-improvement-and-implementation-of-ocr-techniques

europeana project group member

- Referral from BLE (Spices) to BHL-US/UK: 494 visits, 2,016 pageviews, 68% new visits, 63% bounce rate.
- Referral from BLE (Expeditions) to BHL-US/UK: 206 visits, 962 pageviews, 62% new visits, 59% bounce rate.
- Referral from Europeana to BHL-US/UK: 2,641 visits, 10,089 pageviews, 48% new visits, 52% bounce rate.

Social media statistics

The most important publication stream for BHL-Europe is the Blog[9]: Since the start of the publication activities in December 2011, we have published 52 blog posts until 18 June 2012. 32 of the posts are from the reporting period. We are particularly promoting BLE with the spice of the week and expedition spotlight articles. Altogether we now count 5,703 all time visits of our blog. This makes an average of about 1,000 visits per month in the reporting period.

BHL-Europe also has a Facebook and Twitter account. On 18 June 2012, we have 229 'likers' of the BHL-Europe Facebook page[10] and 182 Twitter[11] followers.

[9] http://bhleurope.blogspot.com/
[10] http://www.facebook.com/pages/BHL-Europe/151086001600041
[11] http://twitter.com/BHLEurope