

ECP-2008-DILI-518001

BHL-Europe

Delivery of the final revised best practice guidelines and standards

Deliverable number	<i>D2.9</i>
Dissemination level	<i>public</i>
Delivery date	<i>30 April 2012</i>
Status	<i>In progress</i>
Authors	<i>Birhölmer Melita Hierschläger Michaela</i>



eContentplus

This project is funded under the *eContentplus* programme¹,
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
Michael Malicky	LANDOE
Fritz Gusenleitner	LANDOE
Tom Gilissen	NAT
Dennis Zielke	UBER
Jane Smith	NHM
Constance Rinaldo	MCZ
Laszlo Peregovits	HNHM
Antonio Valdecasas	CSIC
Marian Ramos	CSIC
Manuel Sanchez Ruiz	CSIC
Francisco Welter-Schultes	UGOE
Lee Namba	ATOS
Wolfgang Koller	NHMW
Graham Hardy	RBGE

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
15.02.2011	M. Hierschläger	1	First compliment of chapters sent to Melita BIRTHÄLMER for first corrections
22.02.2011	Melita BIRTHÄLMER	2	Proofread of M. BIRTHÄLMER back to M. Hierschläger
25.02.2011	Jane Smith	3	Linguistic proofread by Jane Smith
01.03.2011	M. BIRTHÄLMER	4	BPG sent to Consortium plus selected people from BHL-US
21.03.2011	M. Hierschläger	5	Feedback incorporation
07.04.2011	M. Hierschläger	6	BPG revision
11.04.2011	M. BIRTHÄLMER	7	BPG revision
12.04.2011	M. Hierschläger	8	BPG review incorporation

0.3 Reviewers

This document requires the following reviews and approvals.

Name	Position	Date	Version
Youssef Mikhail Bassily	Head of Software and System Development Section, ICT Department, Bibliotheca Alexandrina	11.04.2011	4

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
BHL-Europe consortium plus selected people from BHL-US	01.03.2011	4

1 Table of Contents

0	DOCUMENT HISTORY	2
0.1	CONTRIBUTORS.....	2
0.2	REVISION HISTORY.....	2
0.3	REVIEWERS.....	3
0.4	DISTRIBUTION.....	3
1	TABLE OF CONTENTS	4
2	PURPOSE	6
3	BACKGROUND	7
4	PRE-DIGITISATION SETUP	11
4.1	IMPORTING OF ALREADY EXISTING METADATA TO THE GLOBAL REFERENCES INDEX TO BIODIVERSITY 11	
4.2	WHAT DIGITISE?.....	12
4.2.1	<i>Collection Policy</i>	12
4.2.2	<i>Collection Analysis</i>	12
4.2.3	<i>IPR issues and how to tackle them</i>	16
4.3	CHOOSING BOOK/JOURNAL.....	21
4.4	CALCULATE THE TOTAL COSTS OF THE WHOLE PROCESS.....	22
4.5	CHOOSE THE FINAL BOOK/JOURNAL.....	27
4.6	METADATA.....	27
4.6.1	<i>Recording Metadata in database</i>	30
4.6.2	<i>Metadata Requirements – Metadata Bucket</i>	32
4.6.2.1	Granularity.....	32
4.6.2.2	Intellectual Property Rights/Metadata.....	33
4.6.2.3	Volume level.....	37
4.6.2.4	Article level.....	37
4.6.2.5	Page level.....	37
4.6.2.6	Creator.....	37
4.7	PUT METADATA ON HOMEPAGE.....	37
4.8	FUNDING OF THE DIGITISATION PROCESS.....	39
5	DIGITISATION WORKFLOW	40
5.1	PRELIMINARY STEPS.....	40
5.1.1	<i>Analysing and selecting the material</i>	40
5.1.2	<i>Preparing the scanning order</i>	42
5.1.3	<i>Checklist</i>	44
5.2	IMAGING REQUIREMENTS.....	45
5.2.1	<i>Basic principles</i>	45
5.2.2	<i>Resolution recommended, Bit depth</i>	48
5.2.3	<i>Post-processing</i>	48
5.3	SCANNING.....	49
5.4	OCR.....	49
5.5	ARTICLE LEVEL ACCESS.....	51
5.6	QUALITY CONTROL.....	60
5.7	FURTHER REFERENCES.....	61
6	PREPARATION OF DATA FOR BHL-EUROPE	61
6.1	SET UP TECHNICAL INTERFACE/PROVIDER SOFTWARE INTERFACE.....	61

6.2	DATA HARMONIZATION	61
6.2.1	<i>File submission Guidelines</i>	61
6.2.1.1	Files to be supplied.....	62
6.2.1.2	Example.....	62
6.2.1.3	Delivery method.....	62
6.2.1.3.1	Implementing an OAI-PMH web service.....	62
6.3	SCHEMA MAPPING TOOL	64
6.4	PRE-INGEST	64
6.5	INGEST.....	65
6.6	POST OR RE-INGEST.....	65
6.7	INGEST TO EUROPEANA	65
7	GENERAL INFORMATION.....	65
7.1	WORKFLOW EXAMPLES	65
7.1.1	<i>Example from the Georg-August Universität Göttingen Stiftung Öffentlichen Rechts, Göttingen (Germany)</i>	65
7.1.2	<i>Example from the Ernst Mayr Library, Cambridge, MA-US (BHL-US)</i>	69
7.1.3	<i>Example from the Natural History Museum Library, London, UK</i>	76
7.2	WORKFLOW AND TOOLS.....	77
7.2.1	<i>Individual steps in the digitisation process</i>	77
7.2.2	<i>Workflow management software Goobi</i>	83
7.2.3	<i>Workflow management software DAF</i>	87
7.3	GENERAL ARCHIVATION – PERMANENT STORAGE.....	89
	ABBREVIATIONS.....	92
	FIGURES.....	96
	TABLES.....	97
	REFERENCES.....	97
	APPENDIX.....	102
	A: EXAMPLE OF A MODEL BILATERAL CONTRACT	102
	B: PRE-INGEST FILE SUBMISSION GUIDELINES	103

In the following the “Delivery of the first version of the approved best practice guidelines and standards” is referred to as “BPG” (Best Practice Guide) or “Approved Best Practice Guidelines and Standards”.

2 Purpose

These Approved Best Practice Guidelines and Standards aim to provide the first standardised guidelines for processes used by the Biodiversity Heritage Library for Europe for the digitisation of biodiversity texts. The BPG is designed to be easily understood by all persons using it; 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 BPG will always describe the minimum required standard and where indicated the ideal. Existing digitised items are not covered and may not be able to strictly follow the standards described within this document; however the recommendations should be followed for new items to be digitised.

Why is there a need for Approved Best Practice Guidelines and Standards?

Firstly, the Approved Best Practice Guidelines and Standards are of special interest to technology users – in particular libraries, digitisation centres and digital library networks. Both, existing and new content providers will find the Approved Best Practice Guidelines and Standards useful as they will simplify and speed up the whole process of digitisation. Hence an efficient digitisation workflow will be ensured and their connection with EUROPEANA and BHL will be enabled.²

Secondly, the BPG is also a means of developing the case for long-term sustainability of BHL-Europe. BHL-Europe content providers will provide 25,000,000 pages of biodiversity literature by the end of the project (30.04.2012). However there are many more pages of potential content that could be included in BHL-Europe but cannot be, because of the limited timeframe of the project. Therefore it is anticipated that new partners will join the digitisation process and in future provide digitised biodiversity literature to BHL-Europe and thus to the European citizens. The BPG recommends the most efficient way to do so and do not aim only to assist BHL-Europe content providers during the remaining project time, 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. The BPG is similar to cookery recipes which users follow step-by-step to reach a definite end product. In the case of BHL-Europe the final product is a downloadable document. The analogy between digitisation and cooking is also used by BHL-US who calls their Approved Best Practice Guidelines and Standards, the “BHL Cookbook.”

Therefore the BPG is a summary of the processes and standards developed through the whole BHL-Europe project. Through the BPG, this knowledge will be no longer limited to a small number of people but available for everyone who has access to the Internet.

² See Annex 1 Description of Work (Best Practice Networks) [ECP 518001] Biodiversity Heritage Library for Europe BHL-Europe

Nonetheless, best practices are recommendations based on the knowledge of experts. So be aware that the “best” refers to great efforts making the BPG as good as possible with only the noblest intentions. Hence the BPG will hopefully simplify the process in order to encourage libraries and other institutions to join and make scientific literature available for everyone.

3 Background

Summary of the Approved Best Practice Guidelines and Standards:

The introductory chapters of the BPG provide background information and arguments for the usefulness of such a guide.

The BPG is designed so that a user can work systematically through each chapter, to clearly, and fairly easily, achieve the outcome of making biodiversity literature available for online reading and downloading from BHL-Europe and Europeana.

The important steps which should be carried out before the real digitisation work commences can be found in Chapter 4 Pre-Digitisation Setup.

Chapter 5 covers the next step which is the digitisation workflow and primarily refers to how digitisation can be done in practice. After production of a digital version of the respective literature it has to be ingested into BHL-Europe and Europeana. This is described step by step in Chapter 6. A real highlight here is the fact that you no longer need to be an IT-specialist to be able to perform this task which in turn can save resources.

After chapter 6 the main work is finished. General Information and helpful tips and examples are included in the Chapter 7.

Last but not least helpful references can be found in the Appendix and an alphabetical list of abbreviations contains useful explanations of the technical terms used. This contributes to an easier reading of the Approved Best Practice Guidelines and Standards.

The BPG includes links which at the present time are only accessible to current project members. Non-BHL partners will gain access to these sites when they become a content provider of the project. This concerns mainly links to the internal BHL-Europe Wiki³.

Background, related projects and similar documents:

Of course there have been other digitisation projects before BHL-Europe. Some of them have also produced “Best Practice Guides”.

Nevertheless BHL-Europe is unique in many ways. In the BPG, chapters are written by several experts and also content providers for BHL-Europe, from all over Europe and the USA. All relevant sources will be cited in a consistent fashion.

A short review of other “Best Practice Guides”, with comparison to BHL-Europe’s Approved Best Practice Guidelines and Standards is given below. After this, an overview of the unique features of BHL-Europe is presented.

³ <https://bhl.wikispaces.com/BHL-Europe> Non-BHL-Europe partners will gain access to these sites when they become a content provider of the project.

Review of other “Best Practice Guides” with comparison to BHL-Europe’s “Approved Best Practice Guidelines and Standards”:

➤ “DFG Practical Guidelines on Digitisation for programmes funding Scientific Library Services and Information Systems”⁴

The guideline by the DFG (Deutsche Forschungsgemeinschaft) is a practical guide for digitisation and the implementation of digitisation projects. It refers to standards of the DFG and DFG funded projects. It contains important practical information on the digitisation process and shows possible ways how digitisation can be done in detail. It refers to physical items and not to born-digital items.

The aim of “The Practical Guidelines on Digitisation” is to make it easier for DFG applicants to plan digitisation projects and for reviewers to compare proposals. The Guidelines are meant to formulate standards in order to ensure that projects funded by them will be sustainable and viable over the long term. To summarise, it is recommended to work through this document before starting a digitisation project or digitising as it contains practical information for both areas.

➤ Minerva Knowledge Base Digitising Content Together: “Good Practice Handbook”⁵

This document aims to give an overview of how to write good practices on digitisation. A practical handbook to the establishment, execution and management of digitisation projects with particular focus on the cultural area, including libraries, museums and archives, is presented.

To compare the BHL-Europe Approved Best Practice Guidelines and Standards with the Minerva “Good Practice Handbook”⁵, their coverage is completely different. The Minerva “Good Practice Handbook”⁵ describes how to write documents for digitisation projects but does not cover how to organise the whole digitisation workflow, which is what the BHL-Europe BPG does.

➤ ATHENA Access to cultural heritage networks across Europe

“Recommendation and best practice report regarding the application of standards, including recommendations for a harvesting format and fact sheets for dissemination.”⁶

This document is related to the Minerva project. Its aim is to meet the needs of cultural organisations when starting digitisation projects.

⁴ Deutsche Forschungsgemeinschaft (2009): Scientific Library Services and Information Systems (LIS): DFG Practical Guidelines on Digitisation for programmes funding Scientific Library Services and Information Systems http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_en.pdf

⁵ Drake K-M., B. Justrell & A.M. Tamaro (2003): Good Practice Handbook. Version 1.2. edited by the Minerva Working Group 6 Identification of good practices and competence centres. Minerva Knowledge Base Digitising Content Together. http://www.minervaeurope.org/structure/workinggroups/goodpract/document/bestpracticehandbook1_2.pdf

⁶ Mckenna G., Collection Trust (UK), C. DeLoof & Royal Museums of Art and History (BE) (2009): Recommendation and best practice report regarding the application of standards, including recommendations for a harvesting format and fact sheets for dissemination. D3.2. Athena Access to cultural heritage networks across Europe. <http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents>

Athena has another document which is similar to a best practice guide: “Digitisation: Standards Landscape for European museums, archives, libraries.”⁷ This document gives an overview of the different standards which are in use among the partners in the ATHENA project. Its aim is to increase the interoperability within the cultural heritage sector.

These Athena project documents, like the Minerva project documents already described, have some similarities to, but not the same key objectives as the BHL-Europe BPG. The BPG of BHL-Europe focuses on how the digitisation process is done which is described neither by the “Recommendation and best practice report...”⁶ nor by the “Digitisation: Standards Landscape...”⁷.

➤ **Europeana: ‘Europeana Aggregators’ Handbook’**⁸

The purpose of the Europeana Aggregators’ Handbook “is to provide sufficient information to aggregators that wish to submit data to Europeana and become sustainable partners of the service.”⁸

Compared to the BHL-Europe’s BPG the overall aim is different, as BHL-Europe publishes guidelines on how to best proceed in digitising literature.

➤ **‘Internet Archive Book Digitization Process’**⁹

The Internet Archive (IA) is archiving digitised texts of BHL-US and is also scanning books for BHL-US. The “Internet Archive Book Digitization Process”⁹ document describes how the in-house digitising process is done. It notes that the workflow is reviewed frequently in order to incorporate the discovery of new “best practices”.

This document is similar to the BPG of BHL-Europe except that it describes only Internet Archive digitisation processes and does not describe any other approaches. Furthermore, the IA guide does not include sections on “collection analysis” and selecting books to be digitised as the IA mostly scans items selected by other institutions.

➤ **‘Benchmark for Faithful Digital Reproductions of Monographs and Serials’**¹⁰

This benchmark documents the minimum characteristics of digital reproductions to ensure usability, persistence and interoperability – regardless of whether or not they are registered in the Digital Library Federation (DLF) or other registries. The minimisation or elimination of the need to digitise a work more than once is one important objective, ensured by defining baseline levels of quality.

⁷ Mckenna G., Collection Trust (UK), C. DeLoof & Royal Museums of Art and History (Belgium) (2009): Digitisation: Standards Landscape for European museums, archives, libraries. Athena Access to cultural heritage networks across Europe. La Tipografia di Umbero Frisardi, Roma, Italy. <http://www.emuseum.cz/download/athena-soubory/27-imp-athenalibrettinoStandardCopertinaLastINCIANO.pdf>

⁸ Europeana Team (2010): Europeana Aggregators’ Handbook Edition 1. Europeana think culture. http://version1.europeana.eu/c/document_library/get_file?uuid=94bcddb3-3625-4e6d-8135-c7375d6bbc62&groupId=10602

⁹ Internet Archive Operations Team (2010): Internet Archive Book Digitization Process. Ed. Robert Miller, Director of Books, Internet Archive. <http://www.archive.org/details/ProcessDocument&reCache=1>

¹⁰ The Digital Library Federation Benchmark Working Group (2002): Benchmark for Faithful Digital Reproductions of Monographs and Serials. The Digital Library Federation. <http://www.diglib.org/standards/bmarkfin.pdf>

Although this document by the DLF¹⁰ is very similar to the BPG of BHL-Europe, it is still a lot shorter in length. Moreover, the Approved Best Practice Guidelines and Standards of BHL-Europe are tested guidelines presented in a step-by-step way, unlike the DLF document.

To sum up, comparing the documents of the different projects reviewed above with the Approved Best Practice Guidelines and Standards of BHL-Europe, it becomes clear that the BHL-Europe BPG is very unique in its aim to give a sustainable guideline on how to best proceed in the process of digitising literature. Only, the documents produced by the Internet Archive and the Digital Library Federation are similar in content and intention to the BPG of BHL-Europe. However one should not forget that these two fairly similar documents have been produced in the USA and reflect only the North-American perspective which mainly means large scale scanning projects. In addition the Internet Archive is also a partner of BHL-US. Therefore some overlap is inevitably. Moreover the BPG of BHL-Europe complements the “Internet Archive Book Digitization Process”⁹ and “Benchmark for Faithful Digital Reproductions of Monographs and Serials”¹⁰ of the DLF.

Unique Features of the BHL-Europe project:

Europeana has a large collection of information on digitised literature but only links are provided to content holder homepages where the literature can be downloaded. Here BHL-Europe complements by providing online readable and downloadable documents.

Also of note is the Gallica project of the Bibliothèque National de France. Gallica provides online readable literature, however most of the literature is in French. BHL-Europe sets value on a multilingual aspect, not only in the search process but also in the literature provided to read online and complements the literature provided by Gallica with literature from all over Europe.

BHL-Europe has many more aspects which are new in the landscape of literature digitisation projects in Europe.

The Global References Index to Biodiversity (GRIB) and the Scan List represent compilations of all bibliographic information on biodiversity literature held in the major European libraries which should be digitised in future or are already digitised in order to avoid duplication.

In addition, BHL-Europe will provide Optical Character Recognition (OCR) for all the literature available. This is very convenient as it opens up the opportunity to perform a more specific and elaborated search. The TaxonFinder (<http://www.ubio.org/>) enables searches for species names in the digitised literature. For example, it is possible to perform a search for “*Lobaria pulmonaria*” (L.) HOFFM. (lichen species) and all literature that contains this species name or related names is being displayed. This useful tool will be of special interest to scientists but it is also of use for non experts as a lot of knowledge on one species will be instantly available by one simple click.

Moreover, BHL-Europe is a multilingual project which honours the many languages and different cultures within Europe. The multilingual search functionality enables a term, for example “dog”, to be inserted in the search field and retrieval and display of every digitised literature item where “dog” is mentioned whether the language is in Italian, Slovenian or French, or any other European language.

Furthermore, historic literature is preserved in a sustainable way. This will facilitate the work of taxonomists. Rather than spending precious time travelling to a library which holds the book they need for their research or sending a request for a loan, the BHL-Europe search

provides a direct and instant access to the literature needed. By this the search for historic and more recently published literature is more efficient for all kind of users, whether university professors, students or the interested public.

BHL-Europe is the largest project which provides digitised literature in the field of natural sciences for Europe. Major achievement is clearly the gathering together of content of all available biodiversity literature in Europe.

In a world facing nature destruction, climate change and consequently the destruction of the basis of life it is one of the most important tasks to raise awareness for the protection of our nature. One attempt is making the knowledge about our planet's biodiversity better accessible and sustainably available in order to also allow our children and the children's children to live in such a beautiful and still biodiversity rich world as the one we know.

Furthermore BHL-Europe is obligated to provide all digitised literature at no costs for the users and free for use and reuse (open access). Thus, also scientific work in developing countries is supported as they hold the highest biodiversity but have the least access to biodiversity publications due to various reasons. Offering the information for free is a kind of repatriation of knowledge.

BHL-Europe is to date the only European project offering directly online readable and downloadable literature solely for biodiversity related sciences. Hence BHL-Europe is in its concept and execution unique to Europe and of great benefit for Europe's citizen.

4 Pre-Digitisation setup

The Approved Best Practice Guidelines and Standards start here with important steps which should be done prior to the real digitisation workflow, which is described in chapter 5 "Digitisation Workflow".

4.1 *Importing of already existing Metadata to the Global References Index to Biodiversity*

Short Description of the GRIB

The Global References Index to Biodiversity (GRIB) is a collaborative resource developed by EDIT¹¹ and BHL-Europe, and created by the Head Office of the Common Library Network GBV¹² (VZG¹³). GRIB is designed to allow users to browse and search records for deduplicated titles held in the libraries of EDIT and BHL-Europe partners. As well as providing this bibliographic information GRIB is designed to indicate at which stage a text is in a digitisation workflow and to link to fully digitised texts.

Detailed information and documentation on the GRIB prototype can be found using the links below:

- The GRIB prototype: <http://grib.gv.de>
- Technical specifications: <http://bhleurope.gbv.de/>

¹¹ <http://www.e-taxonomy.eu/>

¹² <http://www.gbv.de/vgm/index?lang=en>

¹³ <http://www.gbv.de/vgm/info/biblio/01VZG/>

- Further specifications for BHL-Europe partners on the internal wiki:
<https://bhl.wikispaces.com/Global+References+Index+to+Biodiversity+%28GRIB%29>¹⁹

Importing existing metadata to the GRIB is a very important thing to do as it avoids literature being digitised twice. It ensures the proper management of literature digitisation. An explanation how this is done can currently not be given because the GRIB was still in the test phase at the point of writing.

Therefore, only a brief review of the Scan List (predecessor of the GRIB) is given here: Adding additional library data from a relational database to the Scan List on the table/view level was demanding a rather high knowledge in the IT domain. The tables and views had to be examined by hand to gain knowledge of the semantics of each data field. In the second step building a view from the own data to match at least the minimum requirements was necessary. As the process requires IT knowledge it is not practicable for users without that expertise. GRIB hopefully will facilitate librarians to import their metadata to a global database of deduplicated titles.

Additional information about the GRIB will be included in the second and final revised version of the BPG in April 2012. However till then, more detailed information on the GRIB will be available from May 2011 in the BHL-Europe deliverable “Final and enhanced Web-database for content management and collection analysis” **D2.5**.

4.2 What digitise?

4.2.1 Collection Policy

4.2.2 Collection Analysis

This chapter provides an answer to the main question: “What shall be digitised?” in the wide scope of biodiversity; and its surrounding themes.

- **Collection analysis from the point of view of content: some preliminary considerations**

There are, at least three possible views of collection analysis:

- a) the technical view
- b) thematic via content analysis (the researchers approach)
- c) mixed

However, content analysis has different meanings in different activities, depending mainly on applications. We have found the following (but there are many more):

Industry: Strategies, processes, skills and technologies providing the means for an efficient management of digital content (Brocke & Simons, 2008)¹⁴.

¹⁴ Brocke v. J. & A. Simons (2008): Towards a Process Model for Digital Content Analysis – The Case of Hilti. Available at: [http://ecom.fov.uni-mb.si/proceedings.nsf/0/4c1351e08e8d0f87c1257482003c3852/\\$FILE/44Brocke.pdf](http://ecom.fov.uni-mb.si/proceedings.nsf/0/4c1351e08e8d0f87c1257482003c3852/$FILE/44Brocke.pdf)

Wikipedia: Content analysis is a methodology in the social sciences for studying the content of communication. Earl Babbie defines it as “the study of recorded human communication, such as books, websites, painting and laws.” Content analysis is also considered a scholarly methodology in the humanities by which texts are studied as to authorship, authenticity or meaning.

Academic: Content analysis is a research tool used to determine the presence of certain words or concepts within text or sets of texts.

Our own operational definition: Research content which is relevant for BHL-Europe objectives.

➤ **How do we decide what relevant for BHL-Europe is?**

This is not a very easy question to answer: simple subject matter terms frequently are not the best way to find works of relevance.

For example: words like ‘*evolution*’ sent to a generalist database may retrieve diverse material as ‘*evolution of custom*’, ‘*evolution of dress*’, etc. This may be one of the reasons for the existence of ‘thematic’ libraries, just to avoid the implicit multi meaning of many terms. However, thematic libraries may have their role in a restriction of holdings not by matter but ‘level’. The most well known example is ‘Google Scholar’. Not all pages addressing ‘biological evolution’, for example, are allowed in this search engine, only those that have – either with merit or not – academic recognition.

‘Galapagos’ is another good example. In Spanish this term may refer to some islands and to a special kind of turtle. Unless context is provided, no one will know to what entity being referred to.

➤ **General answer**

There are several strategies to decide what is relevant for BHL-Europe.

- A) One possibility is to elaborate ‘*trajectories*’ of individual researchers whose studies are focussed on general biodiversity topics. This will allow the identification of ‘main’ and ‘peripheral’ pertinent works to include in the database. ‘Main’ works are easily understood. Linnaeus *Systema Naturae* is a clear example. ‘Peripheral’ is difficult to define: relevant but not necessarily mainstream. An example: our partner CSIC (Consejo Superior de Investigaciones Cientificas) built a taxonomist data base in the 1990s and published a *Directory of Taxonomists* (listing all organisms, and the professional and amateur taxonomists researching them). Anyone interested in the ‘health’ of taxonomic studies in Spain during those years will be able to get very useful information from that work.
- B) Another approach to delimit the ‘relevant content’ for BHL-Europe comes from the resources used and needed by historians to recover the activity of such a discipline in a certain period. These may include:
 - 1) The materials inherited from a previous period.
Due to the increasing specialisation in the scientific research, present day publications may not be the examples for deciding which past publications should be digitised. For

example, in Linnaeus time, botany and medicine (including pharmacology) were intimately related. Although it may be relevant to include those works by Linnaeus in BHL-Europe, it does not seem sensible to include present day ‘pharmacological’ treatises. In the same sense, almost all of the literature of the Renaissance, with its recovering of Greek and Roman legacy, that intermix agriculture, geology, *sui generis* ‘naturalia’, zoology and botany, are appropriate for the content of BHL-Europe. To conclude this point, that could be argued longer, ‘*legacy is a criterion for inclusion*’.

2) The items produced.

Each historical period has its own ‘standard’ of disciplines, practices and contents. In each period there are master works, associated literature and additional and peripheral materials. These works are not necessarily important for one single domain, but rather are applicable across various domains. The works of Louis Pasteur on the impossibility of spontaneous generation is not only important for the domain of basic biology, but also very important for any understanding of biodiversity and should therefore be included in BHL-Europe. Something similar could be done with the works of Claude Bernal along with the works of the great naturalists of the nineteenth century, like Charles Darwin, Alfred Russell Wallace, and others researching large biodiversity themes across a number of areas.

3) The meanings associated to terms.

Dictionaries and encyclopaedias are extremely useful tools to know the proper meaning of the words and terms in a certain period. The meaning of key words changes with time and usage, and proper interpretation of content in a certain period may be a compromise if we lack the availability of contemporary tools like these.

4) Additional tools.

A great part of the biodiversity literature is not written in English. On-line translators of Latin and modern languages may be an invaluable help to those interested in the content of BHL-Europe. For additional relevant sources of material that serve a biodiversity purpose, we suggest the reading of Brian W. Ogilvie: “*The Science of Describing*” a scholarly analysis of the biodiversity studies of the 15th-16th centuries and may perfectly illustrate the kind of content required.

C) Within the zoological domain, another approach to select content is the use of the Zoological Record (ZR) general database. Criteria used to include/exclude published literature in the Zoological Record are:

- Inclusion: all zoological publications including from pure taxonomic work to more applied and experimental studies.
- Exclusion: fringe areas dealing with animals solely from an agricultural, pharmacological, pathological, toxicological or immunological point of view or when the animal is mentioned because of medical or veterinary research on another species.

As a conclusion, we advice as a thumb of rule - regardless of which strategy or approach you decide to use for your collection analysis - to rather include collections than maybe exclude collections from BHL-Europe.

➤ **Are there any limitations of content?**

Assuming that size of content is not a limit, and that the priorities of digitisations have been already established, the limitations may come from what has been mentioned at the beginning of this chapter: *quality of a search to retrieve a desired item*.

The search result may depend on the search system used and is illustrated in the following Figure (Fig. 1).

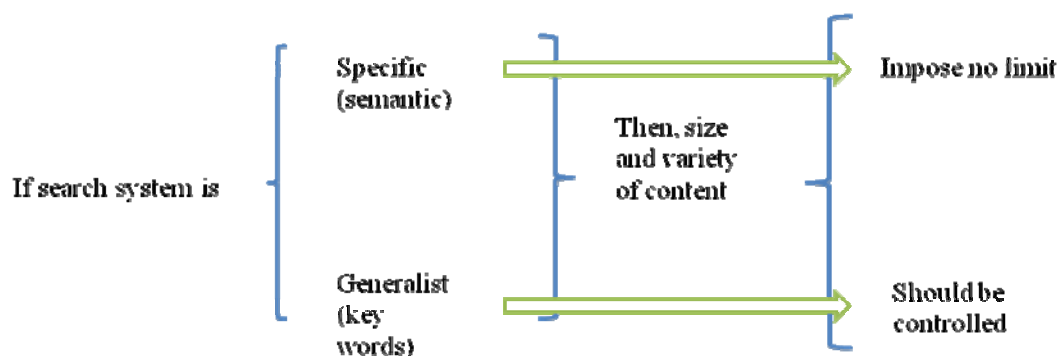


Fig. 1: Limits of database in relation to search criteria

Retrieving information can be done, either by using a specific (semantic) or a generalist (key words) system.

The specific system is “intelligent” and understands alternative or potential meanings of terms from the context by using semantics. In that case there is no need to impose limit in the collection of content. However this implies a very efficient semantic search system, which may not be fully available in a near future.

The second alternative – the generalist system – is based on key words. In that case the search results should be controlled as not all retrieved items might be desired.

Besides the search system, Fig. 2 shows further criteria for choosing content to be included in BHL- Europe.

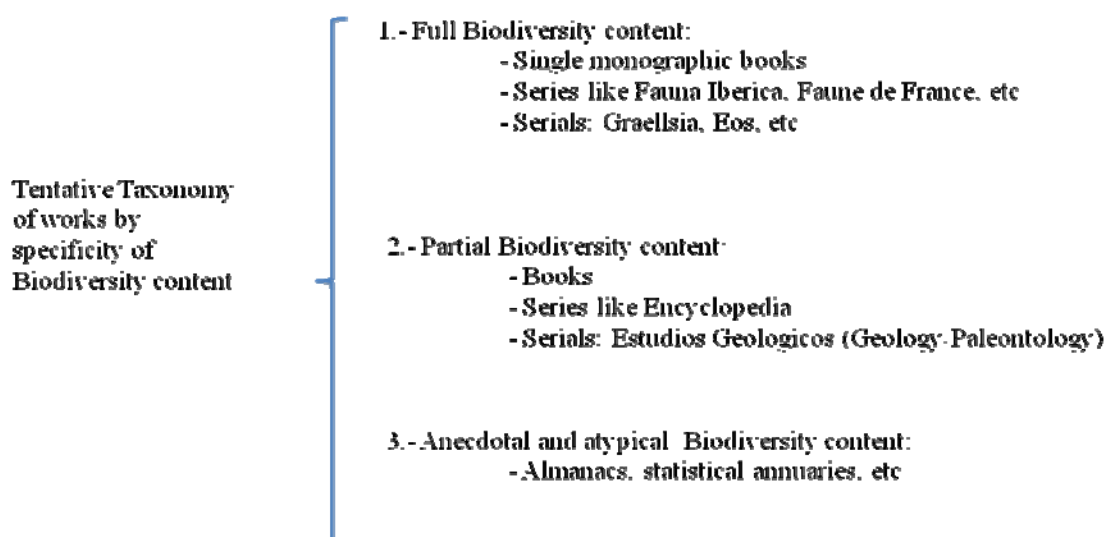


Fig. 2: Criteria for choosing content to be included in BHL-Europe

4.2.3 IPR issues and how to tackle them

Intellectual Property Rights (IPR) questions may quickly arise after content for digitisation has been chosen.

This chapter aims to identify and to clarify the IPR factors that need to be considered when selecting and digitising titles for BHL-Europe.

BHL-Europe aims to provide scanned legacy literature and make it available with some rights reserved under a Scientific Commons licence on the internet. As it is an international project, its activities cross many legal jurisdictions in terms of the partners involved, origin of the material to be scanned and IPR legislation which applies to them. Participating institutions to BHL-Europe need to take into consideration the Copyright and other Intellectual Property legislation applicable within each member state.

More detailed information and guidance relating to Intellectual Property Rights can be found in the Delivery of IPR working documents D4.1¹⁵ which includes guidance on due diligence, pro forma license agreements and metadata agreement for BHL-Europe and Europeana contributors and users. The IPR Best Practice Guide aims to outline IPR factors that need to be considered in relation to the scanning and online display of library biodiversity collections of the BHL project partners. Scanning copyright works and making them publicly available without due consideration of rights ownership and clearance presents serious risks. The guide describes risk management strategies and suggested rights clearance processes for BHL-Europe project partners to minimise those risks. Relevant sections from the Delivery 4.1 relating to the selection of titles for digitisation are below.

Key rights issues and considerations

Many of the works being digitised and made available for BHL-Europe will be out of Copyright. For those works, no permission is needed to scan and post them online. Works that are still protected by copyright will require permission from the rights owner.

Rights complexities

Copyright is a complex system; ascertaining the copyright status of a work can be problematic; works protected by copyright may have multiple rights owners; rights owners may be unknown or traceable. Reasonable efforts must be made to secure permission to use works believed to be protected by copyright. However it is acknowledged that it may not be possible in all cases to identify all rights owners or secure all necessary permissions to scan and make copyright works available online for the project. Nonetheless, partners should follow risk management strategies and due diligence guidance in order to mitigate the risk to themselves, to the project and to fellow project partners.

For titles still in copyright where the rights owner cannot be identified following a process of due diligence to ensure reasonable steps have been undertaken to identify and seek permission from right holders will minimise the risks involved.

General considerations in Selecting a Title for Digitisation

The following extract from Delivery 4.1¹⁵ summarises the main IPR issues to be considered when selecting a title for digitisation:

¹⁵ <http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

BHL-Europe partners should:

- Assess material for scanning to determine its likely copyright status.
- Follow due diligence guidelines as detailed under the 'Rights clearance' section of this document.
- Use the risk category bands described in this document to determine risk.
- Identify the date of publication of material for scanning by checking either the physical item, library catalogue or both. The risk category will be assigned on the basis of this information.
- Where possible, base copyright duration upon the death date of the author, or in the case of multiple authors, the last of those to die plus 70 years, regardless of who owns the rights.
- When identifying and preparing material for scanning, assess the likelihood of embedded third-party content within the work and whether or not this renders the work too risky to scan.
- Limit any liability should rights information that was provided in good faith by the purported rights owner is subsequently found to be incorrect, e.g. where the person granting the permission is not the rights owner, by ensuring any agreement made with a publisher contains relevant warranties and indemnities to protect it and the Project partners against any losses and damages.
- Only supply content to the project that does not infringe third party rights, or that is considered acceptable risk having followed risk management and due diligence guidelines.
- Provide accurate and comprehensive rights and licensing metadata with the content they supply to the project.

Difficulties in tracing rights holders and orphan works:

The rights for each serial title may or may not be owned by a publisher. If it is a historical serial title and the rights were originally owned by a publisher, the publisher may have gone into liquidation or gone bankrupt or merged with another company. In the UK, the assets (including the IP) may be Bona Vacantia (i.e. passed to the Crown) and administered by the Treasury Solicitor.

The rights holder for serials material is not always clear and the following should be considered:

- If the original contributors to an article assigned their copyright to the publisher, then the publisher is the rights holder, so long as the initial period of assignment (if stated) has not expired.
- The rights holder may be the contributor's employer.
- Where possible, the last known publisher who was a rights owner should be contacted to ensure that they are in a position to determine the rights of the serial title(s) in question.
- If the contributor is the owner of the rights, the expiry of Copyright after the death of the author may be different.

The country of origin is significant. This is the country in which the publication was *first published*. If this is in multiple jurisdictions, any of those can apply (provided that they were published in those jurisdictions within thirty days of each other). In Europe copyright duration upon death date is usually life plus 70 years of all authors. Other countries may have their own terms.

Risk Evaluation when decision made to digitise titles

The purpose of risk evaluation is to measure the probability and impact of the various identified risks and enable a judgement to be made about how this is to be handled. The risk continuum ranges from deciding that the risk is too high and the material should not be scanned, through to low risk, which may allow scanning to be conducted, accompanied by a procedures for due diligence and the presence of a robust “notice and takedown” procedure in the event that the risk was inaccurately judged in a particular case, but in good faith. The greater the time period that has elapsed since the material in question was published, the more that the risk diminishes.

In order to alleviate risks of copyright infringement, material that is intended for scanning can be classified into one of three bands. Each BHL-Europe partner needs to decide which of them to adopt, including the option to reject material in particular bands which are presents too great a risk. BHL-Europe partners may scan and supply low and medium-risk items as defined below, giving a maximum cut-off year of 1910 for material whose rights are not owned by the BHL-Europe partner or if the author died before 1939. The risk bands are (Tab. 1):

Level of Risk	Date of publication creation	
Low Risk	Published before 1860	Material published before 1860 is very likely to be out of copyright but there is a minimal risk that this may not always be the case.
Medium Risk	Published between 1860 and 1907	A proportion of this material will be out of copyright, but it is impossible to say how much. This type of material presents an acceptable level of risk for scanning if rights are not cleared, as long as suitable “due diligence” procedures are followed and recorded
High Risk	Published after 1910	The use of this material without rights clearance is considered high risk. It is therefore recommended that it should not be scanned without obtaining express permission from the publisher or current right holders.

Tab. 1: Risk bands

Risk band and licensing guide

The flow chart shows where rights clearance is necessary and where there are licensing implications (Fig. 3).

Extract from D4.1¹⁶

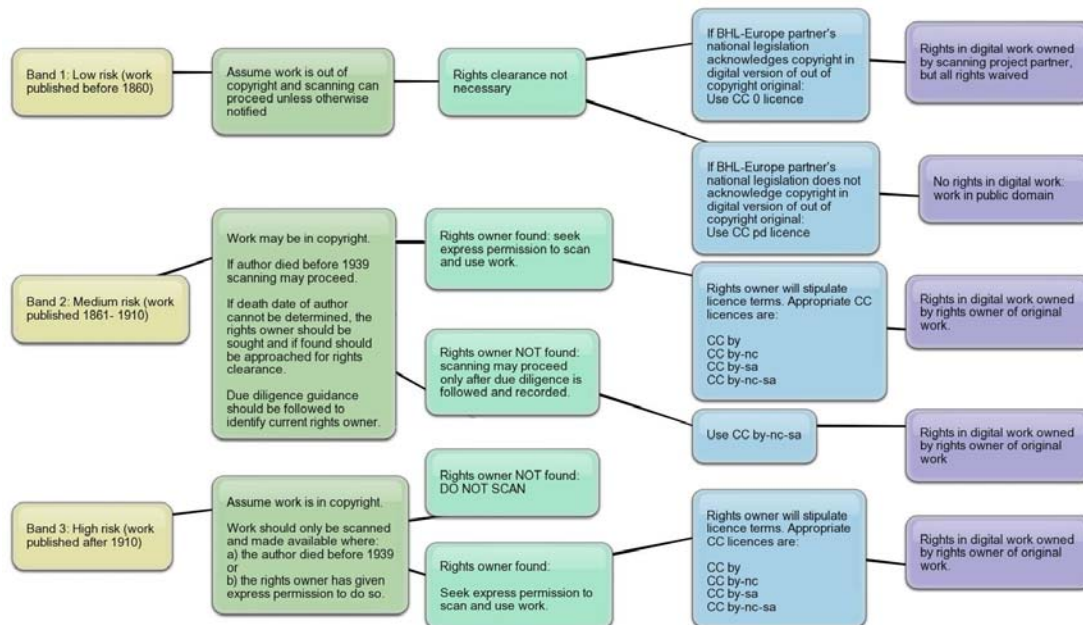


Fig. 3: Risk band and licensing guide

Rights clearance and due diligence

Detailed guidance can be found in the Delivery of IPR working documents D4.1¹⁶

Most original works being scanned for BHL-Europe will not be protected by copyright as the protection term will have lapsed. Works published after 1861 *may* still be protected by copyright; rights clearance is necessary for these works in accordance with risk management strategies described above. The BHL-Europe project has provided a template permission request letter and licence for project partners to assist in this endeavour.

Due diligence is the term used to describe the types of efforts that BHL-Europe partners should undertake to trace rights owners in order to fulfil the requirements of a “reasonable” search.

Third Party content

Where a work contains embedded third party content (e.g. drawings or photographs supplied from external sources), these may be subject to their own copyright or licensing restrictions. For example, a journal publisher may have had to obtain a licence from a commercial Picture Library to print such material prior to the original print run (and paid specifically for such copyright clearance), but this will not automatically mean that this can be legally reused as part of the digital work being created.

¹⁶ <http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

Where the material has an insubstantial amount of third party content, it will be considered that no further action to clear this is required and that the notice and take-down procedure will suffice.

All scanning decisions for Band 3 works must be subject to an assessment of embedded third party content. Where a substantial proportion of work includes embedded third party content, the item will not be scanned or supplied to the project.

Check list for finding rights owners

When seeking rights owners, sources which should be searched include:

- The library catalogue record should be checked for rights owner details
- Physical examination of a sample of volumes from the publication to seek any credit attributions, signatures, marks or signs relating to authorship and/or rights owners details. Additionally, where there appears to be third-party content, this should be noted.
- Google scholar and Google (web) should be checked for presence of the work in electronic form and evidence of the status of the publisher and current rights owner.
- Copyright specific search engines such as WATCH File¹⁷ should be searched.
- Checking with collecting societies and other known rights owners, such as the publishing and authors associations and societies.
- Partner institution accession registers (where they exist) should be checked for rights owner details.

Multiple publishers owning parts of serial runs

The names of other identified publisher(s) and date ranges will be noted. A date restriction may be imposed to ensure that only relevant volumes are noted as under consideration (unless all the other publishers are known to have been through the due diligence procedure and a successful outcome reached).

Multiple authors (monographs)

The rights of all authors should be considered when calculating the duration of copyright and seeking permissions.

Applying for permission

BHL-Europe project partners should apply to suspected rights owners formally in writing for permission to scan and use copyright works.

The preferred method of permission receipt is upon signature from the rights owner on a formal licence including confirmation from purported rights owner that they do indeed own the necessary rights to grant the permission.

A pro forma introductory letter and agreement are provided for this purpose in the Delivery of IPR working documents D4.1¹⁸.

The BHL-US Web site also provides guidance and permissions pro forma. The direct link is: <http://biodivlib.wikispaces.com/Licensing+and+Copyright>

Or can be access via the BHL web site home page by clicking on the **Copyright** option. <http://www.biodiversitylibrary.org/>

¹⁷ <http://tyler.hrc.utexas.edu/>

¹⁸ <http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

Keeping records

The information below will need to be sought and recorded where possible and may be used in the metadata (see chapter 4.6.2.2).

Checklist of information to retain:

- When the work was first published
- Whether the publisher can be identified
- The name of the publisher who published the material
- The country of origin (where the work was first published). Where a publication is published in more than one location at a similar time, the most favourable country of origin will be noted.
- The term of copyright duration given the country of origin.
- Whether the publisher still exists.
- Whether the identified owner or the publisher owns the rights to the work.
- Whether there is evidence of embedded third party material (photographs etc.) which may have separate rights.
- If scanning serials, is there evidence of more than one publisher during the proposed date range.
- Whether the publisher has been contacted to ask them for the right to scan and use the work.
- Whether the publisher/individual confirms that they are the rights owner
- Whether a legal agreement is in place
- What is the status of this and what level of scanning it allows or otherwise
- Any works excluded from the agreement.
- The date of an agreement and how long it lasts.
- BHL-Europe partners should retain paper copies of any agreements with rights owners.

What are bilateral contracts and what are they for? Answers are provided in the following paragraph:

Why are bilateral contracts entered?

Periodicals finance themselves with sales; the exchange of works facilitates the development of one's own library. Hence the editors categorise the issue of digitisation and the making available of PDFs as a sensitive matter. A written agreement, including stipulations about all conditions and the at-will terminability which is signed by the editor and the website provider is of utmost importance in order to clarify why, in which way, and based on which agreement, PDFs may be presented on the Internet. This makes the editors feel secure and encourages them to grant the rights to make their periodicals publicly available. An example of a model contract can be found in the Appendix on page 102.

4.3 Choosing Book/Journal

By checking either the Scan List / serials bid list (chapter **Fehler! Verweisquelle konnte nicht gefunden werden.**), or the GRIB (chapter **Fehler! Verweisquelle konnte nicht gefunden werden.**) you have now compiled a list of journal and/or book items you can start digitising. From all the items on this list you can safely assume that they have not yet been digitised by any BHL partner (not only BHL-Europe, but of all global partners including BHL-US) nor that they are in the process of being digitised. This will not only save you the

effort of doing something that has already been done before, but more importantly, also save you the expenses which you can now use for items that still need digitising.

Normally you will not digitise items that have already been digitised by others, but there may of course be exceptions. There may be several reasons for that. To name just a few examples: you may have an item that contains annotations by the original author; your item has hand coloured plates, while the digitised copy has not; your copy contains dated library stamps that are not in the BHL-US copy. There will of course be many other kinds of exceptions. You will have to decide for yourself how unique the copy you are planning to digitise is, and what the added value of this extra digital copy to the end user will be.

In addition, it is wise to include the cost calculation in your decision process on whether or not an item is to be digitised. Hence if you have considered the Scan List, GRIB, IPR issues and costs you may assign your chosen item for digitisation in the Scan List and the GRIB. Chapters 4.7 and 4.8 respectively provide further information in this regard which may help in the whole process of the Pre-Digitisation Setup.

[Further Description of how to check the GRIB needed!](#)

Check GRIB

After importing existing bibliographic metadata to GRIB decisions can then be taken about which texts to digitise. See Chapter **Fehler! Verweisquelle konnte nicht gefunden werden.** for further assistance with text selection issues. Key amongst these selection issues is the question of IPR.

After deciding on possible texts for digitising a further check of GRIB is required in order to see that no other institution has or is in the process of digitising the selected texts, and if no one has taken hand on your chosen item, from the GRIB point of view the way is free for digitising the chosen literature.

How you practically do check and sign in literature to the GRIB or the Scan List and what they really are is explained in the following two subchapters.

4.4 Calculate the total costs of the whole process

This chapter demonstrates the calculation of costs for the whole digitisation process by referring to the experiences of two BHL-Europe partners.

The first example is that of BHL-E consortium partner Upper Austrian State Museums, Biology Centre (LANDOE):

Almost all steps of the digitisation workflow are done in house at LANDOE, except the OCR process. Therefore, the example hereafter demonstrates the calculation of costs for an in-house workflow, only the OCR process is calculated directly.

Assumption: The calculation is based on a series of 100 books each 300 pages with a total of 1500 articles.

Steps for calculation of costs:

1) Prepare metadata in Text based form, including additional data on authors (address, full name):

Estimation of costs: 50 hours work by senior academic staff required at 30 Euro/hour = 1500 Euro

2) Import metadata to database:

Estimation of costs: 50 hours work by senior typist staff required at 20 Euro/hour = 1000 Euro

3a) OCR of books (destroyable) for automated scanning:

Estimation of costs: 30.000 pages, 12 Cent/page = 3600 Euro

3b) Scan of books in house and submitting of files:

Assumption:

- 400 pages can be scanned per hour
- 30000 pages in totally requires 75 hours for a scan assistant with a wage of 10 Euro/hour
- Charge for OCR: 10 Cent/page

Estimation of costs:

10 Euro/hour*75 hours =750 Euro
+ 10Cent/page for OCR = 3000 Euro
 =3750 Euro

4) Cutting full book PDFs to article level:

Estimation of costs: 50 hours work by senior typist are required with a total cost of 1000 Euro (20 Euro per hour)

5) Quality control (matching articles to metadata + putting article PDFs online):

Estimation of costs: 50 hours work by senior assistant with a wage of 20 Euro/hour are required = 1000 Euro

The summed total costs for 100 books, 1500 articles and 30000 pages are 8100 Euro (following step 3b) or 8250 Euro (following step 3a).

Total for 100 books = 1500 articles = 30000 pages: 8100-8250 Euro

Second example contributed by our BHL-US partner (Museum of Comparative Zoology - Ernst Mayr Library) (Tab. 2)

The second example provides a spreadsheet to detail cost of participating in BHL-US for the Smithsonian Institution. This table (Tab. 2) is based on their digitisation workflow.



Special Collections Scanning Project											
Institutions/Vendors	Harvard Imaging	NYBG	MOBOT	Smithsonian	Smithsonian w/IA	ANS, Phil.	Internet Archive	Parrot	Kirtas	Robot	Other
Costs											
Personnel											
(specify cost unit)			page	item	Item						
Administrative			\$0.01	\$1.00	\$1.00						
Technical				\$1.61	\$1.61						
Pre/Post Prep		\$1.06		\$10.56	\$10.56						
Cataloguing		\$0.08	\$0.08	\$2.11	\$2.11						
Other (explain):											
QA					\$8.64						
Scanning Costs											
(specify cost unit)						per page					
Structural metadata	\$25.2 per page	\$0.13	\$0.02	\$2.00							
Scanning, regular	\$0.75 per page	\$0.32	\$0.28	\$30.00		\$0.83					
Scanning, oversize	\$25 per page	\$0.32	-	\$30.00		\$35		\$73-133 per page			
Stitching	\$15 per page	n/a									
OCR	\$0.2 per page	\$0.08									
Infrastructure costs											
DVD duplication	\$10 per dvd	\$5 per dvd									



Special Collections Scanning Project											
Institutions/Vendors	Harvard Imaging	NYBG	MOBOT	Smithsonian	Smithsonian w/IA	ANS, Phil.	Internet Archive	Parrot	Kirtas	Robot	Other
Other (explain)	\$5 deposit fee per vol.				\$0.10						
Overhead											
(specify cost unit)											
Hardware											
Software											
Space											
Recurring costs											
Other (explain)											
Personnel											
Vehicle/fuel											
Insurance											
Other(explain)											
Other costs (explain)							Can only do foldouts, not text				
(specify cost unit)											
Total per book	\$518.	\$608.		\$9,016.00	\$53.29						
Total cost per page	\$0.61	\$1.67	\$0.39	\$30.05	\$0.18			\$0.394			
Average pages per item					303			\$0.394			



Special Collections Scanning Project											
Institutions/Vendors	Harvard Imaging	NYBG	MOBOT	Smithsonian	Smithsonian w/IA	ANS, Phil.	Internet Archive	Parrot	Kirtas	Robot	Other
Insurance											
Appraisal											
Digital Storage costs											
Art handlers											
Usage stats											

Tab. 2: Cost spreadsheet example BHL-US

The Ernst Mayr Library of the Harvard University tracks in a simpler way by adding the following costs:

A) Full-time staff costs:

- Transportation costs (for sending material off-site for scanning)
- Conservation and repair of items before scanning
- Travel costs
- Equipment costs
- Software development
- Other costs (digital storage, fee for service imaging)

B) Imaging services costs:

- Image capture per standard page
- Image capture per oversized page
- Uncorrected OCR price per page
- Structural metadata generation and deposit (fixed per volume)
- Digital archive deposit (fixed price per volume)
- DVD delivery (price per disc). The DVD delivery cost is incurred because the digital archive of the institution cannot directly deposit to the Internet Archive so we have to deliver the content by disk.

4.5 Choose the final Book/Journal

Strategic explanation of working steps

After cost calculation you can reconsider whether you want to digitise your chosen item or not. The working steps that follow now are very work intensive and at this point of the procedure you should mark your chosen item in the GRIB as being scanned to avoid duplication. “Description how signing literature to the GRIB should be inserted here.” – Can be taken out of **D2.5**

When you have done the previous points you should start to think about Metadata on the technical side but also how they can help you in terms of IPR which is explained in chapter 4.7 Put metadata on homepage. This can be a strategic move to gain the favour of rightsholders.

4.6 Metadata

The metadata chapter provides some general aspects as well as an example and requirements for recording metadata.

A permanent revolution is going on nowadays in the digital domain. The evolution of information technologies makes this revolution possible. The maturation of technological developments on the one hand and the unprecedented wide access to IT equipment on the other hand are the driver behind this process. As a result enormous amounts of digital data/assets have been produced and consequently new problems have arisen about how to

navigate/find one's way in this rather chaotic mass. The metadata provide the compass for setting course in the digital ocean.

The most widely used definition of metadata is simple: data about data. However, this definition is hard to interpret for most people. The etymology of metadata draws from the „Greek prefix meta which means “after, behind or higher” and the Latin word data, the plural form for datum means “a piece of information or something given”. There are sources using the term “information” instead of data but it can also be used as a synonym for “keywords”. Fortunately, irrespective of the term used, they all share two important characteristics namely they try to provide information consistently structured (that makes it possible to use automated processes on), and in an extremely shortened/epitomised way. Furthermore usually metadata is created once the object exists or rarely simultaneously with the object creation itself (e.g. the technical metadata of digital photos is recorded simultaneously during the creation process). The concept of metadata has not recently been discovered, but has already been used for fifty years in library card indexes and nowadays in digital bibliographic catalogues. One natural intention was of course to share data. This was the key driving force for the standardisation efforts which created automatic and machine readable computer processes.

An obvious question is: What does “metadata” mean in the real world? Have most of us dealt with it before at all. The answer is, yes, which might be astonishing. However, nowadays everybody is using digital equipment day-to-day and as a result creates metadata by working with digital technologies, even if most users are not aware of this.

Whenever we take a digital photo with our digital camera or we modify a digital file (e.g., word processor or spreadsheet) these metadata are automatically created. Just get an insight into your digital photograph (practically all image viewer/manipulation programs provide this option) you will find at least technical information, the so-called EXIF data (data of the camera associated with the given photo like camera model, exposure, aperture, date and time, and many optional things like photographer's name, etc.) integrated/incorporated into the file. Nowadays metadata is not only created while taking a picture, but also while using plenty of programs, such as word processors. Often only the metadata fields are automatically created and therefore usually plenty of metadata fields are unfortunately empty as they are rarely filled in.

Now we have learned that metadata surrounds us. Yet it may not be clear why metadata, do we really need them? The explosion of digital born/available data and the development of the World Wide Web, the universal access to distributed sources enforced the necessity of navigation. Metadata itself are the firm basis of it without it resource discovery would be extremely difficult if not impossible at all.

There are different types of metadata¹⁹:

- Bibliographic metadata, also called descriptive metadata – is the information used to search and locate an object such as title, author, subjects, keywords or publisher.
- Structural metadata – gives a description of how the components of an object are organised, for instance how pages are ordered to form chapters.
- Administrative metadata – provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can

¹⁹ <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

access it. Two sub-types of administrative metadata are rights management metadata and preservation metadata.

Metadata standards

As in all aspects of life the wide spread use of non-isolated things lead to standardisation. There is no single standard that suits all purposes. Therefore many formal metadata standards are being developed based on user functionality and data management needs. Often they are not isolated and overlap to an extent. There are simple ones (e.g. Dublin Core) and loosely structured ones (e.g. Dublin Core) while others are very rich and highly structured (ABCD). There are domain or subject specific (e.g. DwC) standards and some are universally applicable like Dublin Core. Metadata standards have been formulated in response to needs of resource types, domains or subjects.

Below, a list of some selected BHL-Europe relevant standards is provided. They can be roughly assigned to content, bibliographic or technical metadata standards. Within this domain you might often note the terms “schema” and “standard”. A schema is a formal description of a standard and a standard is an agreement of things verified by an authority.

Content metadata standards

- **Dublin Core (DC):** <http://dublincore.org/> - One of most general purpose and simple (15 core elements), tries to cover “all” domains, resource types, and subjects (there is a more detailed one, the Qualified DC).
- **qDC:** Qualified Dublin Core, more detailed Dublin Core
- **IPTC:** Information Interchange Model is a file structure and set of metadata attributes that can be applied to text, images and other media types. It was developed in the early 1990s by the International Press Telecommunications council (IPTC) to expedite the international exchange of news among newspapers and news agencies.

Bibliographic metadata standards

- **Encoded Archival Description (EAD):** <http://www.loc.gov/ead/> - Mostly used by archives.
- **MARC** <http://www.loc.gov/marc/> - **MA**chine-**R**eadable **C**ataloging, standards for the representation and communication of bibliographic and related information in machine-readable form. Very widespread, mostly used by libraries describing bibliographic records.
- **MARC21** - is a standard bibliographic metadata format, see also MARC; a formatting, record structure, and encoding standard for electronic bibliographic cataloguing records developed by the Library of Congress. The “21” refers to the version of MARC issued in 1998 that integrated the U.S. and Canadian versions of MARC.
- **MODS** - Metadata Object Description Schema; is an XML-based bibliographic description schema and a derivate of MARC 21; a metadata scheme for rich description of electronic resources. <http://www.loc.gov/standards/mods/>
- **RSLP Collection-level description** <http://www.ukoln.ac.uk/metadata/cld/> - General, covers collections of all subjects, domains and types.
- **SPECTRUM** - Standard for describing museum objects.

- **Text Encoding Initiative (TEI)** <http://www.tei-c.org/> - Standard for the representation of texts in digital form.
- **Darwin Core (DwC)** <http://rs.tdwg.org/dwc/index.htm> - An extension of DC for biodiversity information.
- **Access to Biological Collections Data (ABCD)** <http://rs.tdwg.org/abcd/> - A data and metadata specification for the exchange of biological collections and observations data.
- **International metadata standard for geographic information** <http://marinemetadata.org/references/iso19115/> - Defines the conceptual model required for describing geographic information and services
- **Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (CSDGM)** <http://www.nbi.gov/> - National Biological Infrastructure Initiative metadata standard including geographical metadata.
- **Ecological Metadata Language** <http://knb.ecoinformatics.org/software/eml/> - for ecological data.

Technical metadata standards

- **XML:** Extensible Markup Language; a set of rules for encoding documents in machine-readable form <http://www.w3.org/XML/>
- **XMP:** Extensible Metadata Platform is a standard, created by Adobe Systems Inc. for processing and storing standardised and proprietary information relating to the contents of a file.

4.6.1 Recording Metadata in database

Recording metadata is a very important task and should not be underestimated. Always keep in mind that content is only retrievable within any database if you provide the corresponding metadata. The higher your metadata quality is, the easier you retrieve your content.

This chapter provides an example from our partner LANDOE about how to record metadata in a database for their particular library usage. For each database field the significance is indicated.

Indicators: 1 must have, 2 recommended, 3 nice to have, 4 obsolete

General Note: BHL-Europe is using its own schema, which is based on MODS (<http://www.loc.gov/standards/mods/>). Metadata provided needs to be mapped to the BHL-Europe schema. The mapping to the BHL-Europe schema will be done by using an open source Schema Mapping Tool in collaboration with each content provider during the project time. Standard mappings, for instance, from MARC21, DC will be available. At time of writing the technical development was in progress and therefore more detailed information will follow in BHL-Europe Deliverable 2.9 (final revised BPG due 30.04.2012) which is due to 30.04.2012.

The <bibliographicInformation> Field of the BHL-Europe Schema is a full MODS record, which means bibliographic metadata provided as MODS can be used without any modifications.

Bibliographic/Descriptive Metadata

1) Volume level metadata (1)

<bhlElement> with <bhlLevel>volume</bhlLevel>

1.1) Title (1)

<bibliographicInformation><titleInfo>

1.2) Number of volume within series (1)

<bibliographicInformation><part><detail type="volume">

1.3) Year(s) of publication (1)

<bibliographicInformation><originInfo><dateIssued>

1.4) ISBN (2)

<bibliographicInformation><identifier type="isbn">

1.5) Title of Series including ISSN (if available) (1)

<bhlElement> with <bhlLevel>serial</bhlLevel>

<bibliographicInformation><titleInfo>

1.6) Authors(s) (1)

<bibliographicInformation><name type="personal">

with <role><roleTerm type="text">creator</roleTerm>

2) Article level metadata - handy piece of information, matches the classical paper, from LANDOE's point of view the best compromise between volume and page level data (2)

<bhlElement> with <bhlLevel>article</bhlLevel>

2.1) Title (2)

Same as 1.1

2.2) Author(s) including additional info like scientific biography, portraits.

LANDOE got an overall of very positive echo on that, as they can link the literature data to this data now, and a lot of researchers are very happy with this option on LANDOE's web pages (2)

Same as 1.6

2.3) Keywords - will be dropped in the future within LANDOE as they do full text OCR and want to build a good search engine on the top of the OCR text (3)

<bibliographicInformation><subject><topic>

2.4) abstract text - has been done in the past before OCR (4)

<bibliographicInformation><abstract>

2.5) Start page (2)

<bibliographicInformation><part><extent unit="pages"><start>

2.6) Stop page (2)

<bibliographicInformation><part><extent unit="pages"><end>

At LANDOE, the data is stored in a relational database, giving them the opportunity to map the data against almost all target schemas by building views and database functions (Example see chapter 6.7 Ingest to Europeana)

Administrative Metadata

In the case of LANDOE, Administrative Metadata is stored in a simple Word-Table. All employees working on the project have read and write access. In the table all digitisation steps are recorded mapped including information on which person has to do what and when. Once a task is finished by a person, the follow up task is marked yellow, thus allowing the next worker to simply identify all open tasks concerning him/her. In detail the following information is stored in the table:

- Name of the Serial/Monograph
- Name of the worker responsible for putting descriptive metadata into the database initially and once a new volume of a serial is released
- Finalised descriptive Metadata – up to which volume
- Information when and who has signed the agreement allowing to put the objects on the homepage
- Volumes currently in OCR process
- Finished volumes in PDF form
- Volumes cut to article level
- Article level PDFs digitally signed
- Article level PDFs online
- Online Quality control with worker signature
- Prepared Volumes for BHL/Europeana
- Extracted articles concerning biographical data on natural history scientists

4.6.2 Metadata Requirements – Metadata Bucket

Where are metadata stored?

Data can be stored either externally or internally (internally means into the files themselves) as well as externally and internally. There is no fit for all solutions and one has to carefully consider that the IT background is rapidly changing; therefore data security/long-term maintenance is essential. Both solutions have a particular advantage and probably the best practice is a combination of two, however, that increases the IT administration investments.

4.6.2.1 Granularity

It is an important issue to decide what level of granularity will be used. Metadata creation for different items should not necessarily be on the same level of granularity. Here we suggest considering carefully your demands. Different types of works need different approaches; the decision may also be influenced by the technical implementation and feasibility. For books, monographs, journals and other items different level of metadata granularity may be practical. In the everyday practice the minimal level in general should be the bibliographic description on article level (e.g., MARC) complemented with the Intellectual Property Information. The optimal solution would be to add keywords and abstracts (if available) to the elements of the minimal level data because in most cases they contain important information even if we use

imprecise free-text search. The ideal model for metadata would be one which includes information on taxon and geographical location. However, in most cases this is not achievable because of resource costs. To reach this level of granularity the use of reference structures (controlled vocabularies) is highly recommended. Nevertheless, a relatively low cost way of solving part of this problem is provided by BHL-Europe automatically enriching the metadata with taxonomic terms.

Special attention should be paid to the non-sequential pages/items like illustrations bound at the given volumes which are not incorporated into the page series of a given article/item.

4.6.2.2 Intellectual Property Rights/Metadata

Metadata relating to Intellectual Property issues pertaining to a title or an item is important. The metadata attached to a digital image may include information about the author or creator, date of creation, copyright holder and/ or license holder. Most images cannot be made publicly available on the Internet without the presence of this information (metadata). If this metadata does not appear with the image, the image user may be liable for copyright infringement.

Including Copyright and other IPR related information in the metadata is essential as Copyright or Intellectual Property law is not standardised across Europe so image holders and image users face a variety of rules depending on where they are situated.

There are often many layers of copyright within an image, all of whose rights need to be considered. This may include the photographer, the creator of the original work, the location which holds the original work and/ or other rights holders. Determining all the rights holders within an image requires an understanding of the source of the image, the content portrayed and the creation of the image. Where the rights owner is known this information should be included in the metadata.

In some cases, copyright information about the rights holder(s) of an image is not known. These images are referred to as 'orphan works'. This information must also be included.

Metadata will include the Copyright status of a work (whether it is subject to Copyright protection or in the public domain). Where a work is protected by Copyright and permission has been sought from the right holder(s), contributors should supply details of the license terms under which it is being made available through the BHL-Europe site. Where possible licence terms should be selected from a range of Open Access licensing solutions, most likely, Creative Commons licences. There may be exceptions whereby rights holders prefer not to license a work under such terms and in these cases partners will need to supply accurate licensing information in the metadata of the object.

Guidance on Intellectual Property Rights is also covered in Chapter 4.2.3 above and in greater detail in the BHL-Europe Delivery of IPR working documents D4.1²⁰. BHL Europe follows guidance developed by *Europeana Specification for the Europeana Semantic Elements V 3.1.1, 24/01/2011*²¹. The Europeana Semantic Elements (ESE) consists of the Dublin Core (DC) metadata elements, a subset of the DC terms (defined at Dublin Core Metadata Initiative web site at: <http://dublincore.org>). The specific page covering Semantic Elements is <http://dublincore.org/documents/dcmi-terms/> and a set of thirteen additional which were

²⁰<http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

²¹http://version1.europeana.eu/c/document_library/get_file?uuid=a830cb84-9e71-41d6-9ca3-cc36415d16f8&groupId=10602



created to meet Europeana's (and BHL-Europe's) needs. The elements reflect metadata required for both digitised titles and born-digital content.

The elements identified by Europeana can be found below (Tab. 3).

Extract from Specification for the European Semantic Elements V3.1.1 24/01/2011 page 3²²

Source	Element	Element Refinement(s)
DC	<u>title</u>	<u>alternative</u>
DC	<u>creator</u>	
DC	<u>subject</u>	
DC	<u>description</u>	<u>TableOfContents</u>
DC	<u>publisher</u>	
DC	<u>contributor</u>	
DC	<u>date</u>	<u>created</u> ; <u>issued</u>
DC	<u>type*</u>	
DC	<u>format</u>	<u>extent</u> ; <u>medium</u>
DC	<u>identifier</u>	
DC	<u>source</u>	
DC	<u>language*</u>	
DC Europeana	<u>relation</u>	<u>isVersionOf</u> ; <u>hasVersion</u> ; <u>isReplacedBy</u> ; <u>replaces</u> ; <u>isRequiredBy</u> ; <u>requires</u> ; <u>isPartOf</u> ; <u>hasPart</u> ; <u>isReferencedBy</u> ; <u>references</u> ; <u>isFormatOf</u> ; <u>hasFormat</u> ; <u>conformsTo</u> <u>isShownBy</u> ; <u>isShownAt</u>
DC	<u>coverage</u>	<u>spatial</u> ; <u>temporal</u>
DC	<u>rights*</u>	
DC terms	<u>provenance</u>	
Europeana	<u>country</u>	
Europeana	<u>dataProvider</u>	
Europeana	<u>hasObject</u>	
Europeana	<u>language*</u>	
Europeana	<u>object</u>	
Europeana	<u>provider</u>	
Europeana	<u>rights*</u>	
Europeana	<u>type*</u>	
Europeana	<u>unstored</u>	
Europeana	<u>uri</u>	
Europeana	<u>userTag</u>	
Europeana	<u>year</u>	

Tab. 3: elements identified by Europeana

The specific elements under Dublin Core and Europeana specific Requirements relating to Rights are below (Tab. 4, Tab. 5)

²² Europeana think culture: Europeana Semantic Elements Specification version 3.3.1 24/02/2011. Europeana v1.0. <http://www.version1.europeana.eu/web/guest/technical-requirements/>

Extract from Specification for the European Semantic Elements V3.1.1 24/01/2011 page 22-23²³

Element name: rights	
Namespace	Dublin core
URI	http://purl.org/dc/elements/1.1/rights
Label	Rights
Definition	Information about rights held in and over the resource.
Europeana note	Information about intellectual Property Rights, access rights or license arrangements for the digital object (digitised or born digital).
Obligation &	Optional (Minimum: 0, Maximum: unbounded)

Tab. 4 elements, requirements, rights

Element name: rights	
Namespace	Europeana
URI	http://www.europeana.eu/schemas/ese/rights
Label	Europeana Rights
Definition	Information about copyright of the digital object as specified by isShownBy and isShownAt
Europeana note	The value is a URL constructed according to the specifications in the “Rights Guidelines” document on the Technical Requirements page at: http://version1.europeana.eu/web/guest/technical-requirements/ The URLs are constructed by adding a code indicating the copyright status of an object to the domain name where that status is defined. For users of Europeana.eu this copyright information also applies to the preview specified in <code>europeana:object</code> . The rights statement will be represented as a badge on the object page and as a text string in the metadata display. The provision of this element has an obligation level of “recommended” in this version of ESE. It will be changed to “Mandatory” in a later version.
Obligation & Occurrence	Recommended (Minimum: 0, Maximum: 1)
Europeana search and display features	Simple search Facet Timeline Advanced search Full search result display X
Example	<code><europeana:rights>http://www.europeana.eu/rights/rr-f/</europeana:rights></code>

Tab. 5: elements, requirements, rights

²³ Europeana think culture: Europeana Semantic Elements Specification version 3.3.1 24/02/2011. Europeana v1.0. <http://www.version1.europeana.eu/web/guest/technical-requirements/>

4.6.2.3 Volume level

The volume level is an absolute must for BHL-Europe as the information is shown on this level. The Metadata should contain the following fields:

- 1) Title of book
- 2) Title of Series (if within a series)
- 3) ISBN if available
- 4) ISSN if available
- 5) Authors of book
- 6) Year(s) of publication

4.6.2.4 Article level

A huge volume is split in handy pieces matching the classic "paper". If the article level is used, it should at least contain the following information:

- 1) Title of article
- 2) Author(s)
- 3) Start page
- 4) End page

BHL-Europe recommends to record bibliographic description on article level.

4.6.2.5 Page level

Recording on this level is a huge effort and one should weigh the cost-benefit ratio. The page level metadata allows access to the targeted page for a scientific research directly, but in most situations this can be done also very quickly on article level. On this level the following information is required:

- 1) Page number
- Etc.

4.6.2.6 Creator

Is Metadata on the creator(s) and their roles in the digitisation process: gives a better feeling for the quality of the data?

- 1) Metadata recorder(s)
- 2) Metadata enhancer(s)
- 3) PDF Splitter(s)
- 4) Quality controller(s)

4.7 Put Metadata on homepage

In this chapter it is explained why it will be helpful if metadata are provided on the institutional homepage with regard to receive funding and in the respect of building up a well structured offer of digitised literature.

Editors consider their periodicals and other types of publications private projects. Assigning (partial) rights to a third party is an issue that has to be dealt with circumspectly. Therefore, a third party's request to digitise periodicals which they do not publish themselves has to be set up carefully. The first step towards this is to draw up a bibliography of the tables of content for a literature portal. By doing this, respective editors are shown that the portal provider is

also investing money, energy and time in order to logically prepare a later presentation of the PDFs. A short summary gives basic information about the respective periodical as well (Fig. 4). This advance effort can be seen as a vital part in order to draw up a mutual agreement in writing concerning the integration of digital contents. Experience has shown that such an agreement could only be reached after supplying a bibliography combined with basic information linked to the respective authors (Fig. 5). In-house it has proved to be an advantage not to include the collected data in the database immediately. Instead our content provider LANDOE record it by means of a text editor first to ensure that articles can be clearly assigned to an author when included in the database. In this way it is possible to include any missing information, e.g. full names for authors, and also standardise authority terms, e.g. Linné for Linnaeus. Furthermore it is possible to unmistakably assign information to existing data set numbers and prevent redundant data sets (Fig. 6). Experienced typists are to collect the textual data. The correct implementation in the database as well as the correction of the aforementioned data, however, requires experienced scientists' know-how whenever the document does not provide the needed specifications.



ÖKO L Zeitschrift für Ökologie, Natur- und Umweltschutz

Gegründet: 1979

Ehemalige Benennung: Apollo ISSN: 0003-6528

Herausgeber: Magistrat der Landeshauptstadt Linz, Stadtgärten Linz, Abt. Botanischer Garten und Naturkundliche Station, Url: http://www.linz.at/Umwelt/umwelt_10159.asp,

E-Mail: gerold.laister@mag.linz.at

Redakteur: Dr. Fritz Schwarz, Ing. Gerold Laister

Erscheinungsrhythmus: meist viermal jährlich

Inhaltliche Ausrichtung: populärwissenschaftliche Bearbeitung zoologischer, botanischer und umweltrelevanter

Themen meist das Bundesland Oberösterreich betreffend

Sonstige Angaben: Abokosten jährlich 13 Euro (Stand 2006)

Bisher erschienene Bände:

1979 1: (1979): ÖKO L 1979 (1)

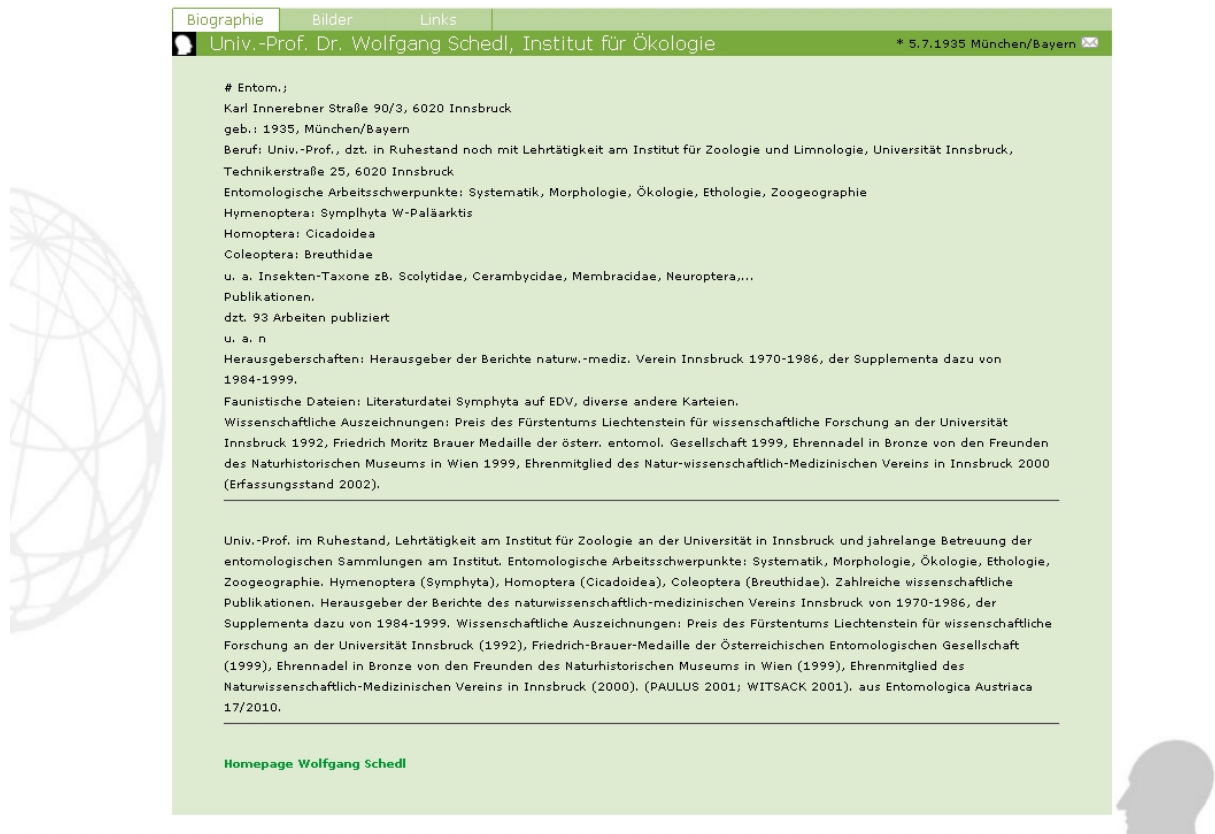
1979 2: (1979): ÖKO L 1979 (2)

1979 3: (1979): ÖKO L 1979 (3)

1979 4: (1979): ÖKO L 1979 (4)

1980 1: (1980): ÖKO L 1980 (1)

Fig. 4: short summary of the journal

Biographie Bilder Links

Univ.-Prof. Dr. Wolfgang Schedl, Institut für Ökologie * 5.7.1935 München/Bayern

Entom.:

Karl Innerebner Straße 90/3, 6020 Innsbruck
geb.: 1935, München/Bayern
Beruf: Univ.-Prof., dzt. in Ruhestand noch mit Lehrtätigkeit am Institut für Zoologie und Limnologie, Universität Innsbruck,
Technikerstraße 25, 6020 Innsbruck

Entomologische Arbeitsschwerpunkte: Systematik, Morphologie, Ökologie, Ethologie, Zoogeographie
Hymenoptera: Symphyta W-Paläarktis
Homoptera: Cicadoidea
Coleoptera: Breuthidae
u. a. Insekten-Taxone zB. Scolytidae, Cerambycidae, Membracidae, Neuroptera,...

Publikationen.
dzt. 93 Arbeiten publiziert
u. a. n
Herausgeberschaften: Herausgeber der Berichte naturw.-mediz. Verein Innsbruck 1970-1986, der Supplementa dazu von 1984-1999.
Faunistische Dateien: Literaturliste Symphyta auf EDV, diverse andere Karteien.

Wissenschaftliche Auszeichnungen: Preis des Fürstentums Liechtenstein für wissenschaftliche Forschung an der Universität Innsbruck 1992, Friedrich Moritz Brauer Medaille der österr. entomol. Gesellschaft 1999, Ehrennadel in Bronze von den Freunden des Naturhistorischen Museums in Wien 1999, Ehrenmitglied des Naturwissenschaftlich-Medizinischen Vereins in Innsbruck 2000 (Erfassungsstand 2002).

Univ.-Prof. im Ruhestand, Lehrtätigkeit am Institut für Zoologie an der Universität in Innsbruck und jahrelange Betreuung der entomologischen Sammlungen am Institut. Entomologische Arbeitsschwerpunkte: Systematik, Morphologie, Ökologie, Ethologie, Zoogeographie. Hymenoptera (Symphyta), Homoptera (Cicadoidea), Coleoptera (Breuthidae). Zahlreiche wissenschaftliche Publikationen. Herausgeber der Berichte des naturwissenschaftlich-medizinischen Vereins Innsbruck von 1970-1986, der Supplementa dazu von 1984-1999. Wissenschaftliche Auszeichnungen: Preis des Fürstentums Liechtenstein für wissenschaftliche Forschung an der Universität Innsbruck (1992), Friedrich-Brauer-Medaille der Österreichischen Entomologischen Gesellschaft (1999), Ehrennadel in Bronze von den Freunden des Naturhistorischen Museums in Wien (1999), Ehrenmitglied des Naturwissenschaftlich-Medizinischen Vereins in Innsbruck (2000). (PAULLUS 2001; WITSACK 2001). aus Entomologica Austriaca 17/2010.

[Homepage Wolfgang Schedl](#)

Fig. 5: Screenshot database www.zobodat.at

Wiener Entomologische Monatsschrift (1857-)

- ANONYMUS (1857): Vorwort. — Wiener Entomologische Monatsschrift 1: 1.
- LOEW Hermann Dr., Meseritz (1857): Dipterologische Notizen. — Wiener Entomologische Monatsschrift 1: 1-10.
- MILLER Ludwig. *21.8.1820 Laibach (Krain) +4.4.1897 Wien (1857): Die Apionen der Wiener Gegend. — Wiener Entomologische Monatsschrift 1: 10-25.
- LEDERER Julius. *24.6.1821 Wien +30.4.1870 Wien (1857): Lycaeniden-Gattungen der europäischen Fauna. — Wiener Entomologische Monatsschrift 1: 25-32.
- LOEW Hermann Dr., Meseritz (1857): Dipterologische Mittheilungen. — Wiener Entomologische Monatsschrift 1: 33-56.
- MILLER Ludwig. D-Nr. 13312 *21.8.1820 Laibach (Krain) +4.4.1897 Wien (1857): Die Apionen der Wiener Gegend. — Wiener Entomologische Monatsschrift 1: 56-59.

Fig. 6: Example

4.8 Funding of the digitisation process

Several possibilities and examples with better or less chances are given by our partner LANDOE for the funding of scanning hereafter:

- 1) Overheads from other projects – As our Technical/Scientific staff do the administrative work for the projects themselves, we get the overheads as additional free budget. E.g.: A project with worth 120.000 Euros with 20 percent overhead included raises our funds by 20.000 Euro that is approx. 150.000 pages with OCR.

- 2) Internal budget – if your department head is asking for some ideas at the end of a fiscal year, raise your hands!
- 3) National funds from public bodies. In Austria Museums can ask for additional budgets (10.000-20.000 Euros) at the Ministry of Culture for digitisation processes.
- 4) Very rare international projects with funds for such work.

5 Digitisation workflow

In this chapter we will try and describe the basic steps of a digitisation project for books and / or journals. How you will actually organise each of the steps will depend on your own circumstances. It might be that you will be scanning just one back-list of a journal, or that you have acquired a grant for digitising your entire library. In each case the basics underlying the process will be generally the same. It is of course the scale that differs. This will certainly reflect the number of people working on the project and how you need to manage them. That however lies outside the scope of this document and will not be covered here. There is plenty of literature available on how to successfully manage a project.

Another difference between institutions will be that some do scanning of the books / journals themselves; others will make use of a specialised (commercial) company for parts of the digitisation project. Whether you will do everything yourself, or outsource (part of) the project, there are certain elements that need to be taken care of in both cases. In the next few paragraphs we will focus on the steps that need to be taken to digitise your collection so that it can be included into the Biodiversity Heritage Library for Europe.

5.1 Preliminary steps

5.1.1 Analysing and selecting the material

After parts of your collection have been identified as possible candidates for digitisation, and collection analysis and deduplication matters have been investigated thoroughly, the identified books and journals need to be carefully assessed.

A) Structure

First, you should ask yourself how your collection is structured. Is it all monographs you are digitising, or all journal volumes, or both? In the case of journal volumes, how many issues are there per volume? Does this number stay the same or does it change over time? While you are doing this, check also for completeness. This is especially true for journal issues that have not been bound in volumes, and for all kinds of loose leaf publications. Verify also that the sequence of pages and/or issues is correct. It might be relevant to make notes about this and write instructions for the scanning company on how to deal with inconsistencies that you have noticed. An easy way to do this is to record all these remarks in a spreadsheet, give each one a number, and put a (coloured) sheet containing this number at the location it concerns.

This procedure should help you analyse the information you have, so that you can decide what information digitising this collection should bring you.

B) Access

Is the material easy accessible, or not? Is it all bound in volumes, or is (part of) it all loose issues, or perhaps loose leaf pages stored in folders, or boxes? And, how well sorted is it?

You should check that every item is actually available. It might be that items are borrowed by users, or even have been lost or stolen. You will have to decide, how to act on this. Are you going to replace any missing material? If not, you will need to delete its record from the scan list. When you decide to replace it, you can either buy a replacement copy, or try to obtain a (high quality) photo copy. Do not forget to mark each title as “temporarily not available” to prevent disappointments with your users.

C) Physical condition/State of the material (quality of paper, binding etc.)

While you are busy checking the books and journals you will also start to get an overview of the condition of your collection. There may be volumes or issues which are that fragile, that you might have to treat them different from the rest. Some documents may be damaged: spines may be broken; pages may be loose or torn. Sometimes the quality of the paper itself will be that poor that the document cannot be handled without causing damage. You may sometimes even have to decide for restoration, before these items can be sent for scanning. It is good practice to restore the documents before sending them away for scanning. Often repairs will be fairly simple and cheap, and can be done in house. Sometimes however, they will be difficult and extensive and you will need to call in help from professional paper restoration services.

D) Range of sizes

The range of sizes (smallest, largest) and the degree to which they vary should be considered. This will influence the price of scanning. Also, very large and very small sizes may be difficult to scan, and specialised equipment might be needed. Other things you may wish to look for are images, maps, fold outs, loose leaf pages containing a table of contents, indexes or errata. Make note of this all, because you will need to decide on how to deal with this later on in the process.

At the end of this phase, it should be clear what the exact volume of the collection you are going to digitise is, how it is structured, and what its physical state is.

Book scanning decisions factors²⁴

The following elements are part of the evaluation process to determine if a volume is suitable for scanning at a high-volume centre or if they must be handled differently (Tab. 6, Tab. 7).

General condition factors

too fragile to send
 too large to send
 too small to send
 too valuable to send

Tab. 6: General condition factors

²⁴ Hanken J. (2010): Retooling Special collections Digitization in the age of Mass Scanning (LG-50-08-0058-08. URL: <http://biodivlib.wikispaces.com/file/detail/Retooling+Special+Collections+Digitization+in+the+Age+of+Mass+Scanning.odt>

Detailed condition factors	
<i>Size</i> Oversize = exceeds scanning bed capabilities (i.e. over 46 cm or 18") Undersize = too small for scanning bed (i.e. under 10 cm. or 4") Odd Other	<i>Binding</i> Broken binding Leather binding has red rot Broken spine Unbound Water damage (smelly, mouldy) Hollow or flat Latches Other
<i>Text block/paper</i> Broken text block Insufficient gutters (bound too tight) Faint text See-through paper Uneven sizes Tip ins Torn pages Water damage (smelly, mouldy) Deckle-edged pages (feather edges) Bolted (uncut) pages Other	<i>Fold outs</i> Fold outs that exceed scanning capabilities Fragile Detached Torn Stressed Brittle Other
<i>Value</i> Rare Unique Manuscripts Age Insurance issues Handling issues Other	<i>Added Value</i> Provenance Marginalia Inserts Research Other

Tab. 7: Detailed condition factors

5.1.2 Preparing the scanning order

Digitising books or journals is useless if you don't provide metadata to go with them. So, apart from the physical objects, you will also need to analyse and select the available metadata. In the scanning process the scans (of the pages) will be grouped according to the available metadata. In digitisation projects this is sometimes called an *object*. The *object* then stands for the unit of scans that represents the digital copy of a physical item. It can consist of just one scan (e.g. the scan of the front page of a print), or of multiple scans (e.g. the scans of all the pages of an entire book). Because each scan belongs to a specific object, there has to be a metadata record for each object. How many objects there are, will depend on the level of detail your metadata contains.

In the BHL three levels of metadata are recognised²⁵: title, item and page level metadata. This is also described as three levels of granularity. On the title level, the “title” is described as the bibliographic authority for a single or multiple scanned item(s). In other words, at this level all the bibliographic information of the book or journal is recorded, like title, authors(s), publisher, etc. On the item level, an “item” is described as the physical bound object scanned within a title. This is the level of the physical items, e.g. the copy of the book your library holds, or the volumes of a journal sitting on the shelves. Barcode, Local Library Identifier, Start and End volume are typical metadata entries that are recorded at this level. Finally, on the page level, a “page” is described as a single image file contained within a scanned item. This is the level where the file name of the image scan will be stored. Here you will also find the page type, which holds a description of the contents of the page (Like: title page, text, map, table of contents, etc.).

In the beginning of the digitisation process you should start with selecting the metadata you have available. Next, you should consult the “Catalogue of content holder requirements” https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf²⁶ which is only accessible for current partners of BHL-Europe and analyse whether the selection you have meets the requirements. If not, it is important that you extend (enrich) your metadata until it does. You can do this yourself, but – if you have chosen to outsource the scanning process – you can also ask the scanning company to do this for you. Please note, that this is as important an aspect of the digitisation process as the scanning of the pages. No matter how high the quality of scanning, without sufficient metadata, the entire project will be rendered useless.

To be able to link the metadata to the file names of the scans, you need to decide on a unique ID that you can use. This can be for instance the record number from your local catalogue, or the PPN (Pica Production Number) from a union catalogue, or any other ID that is unique to your object. Furthermore, you will also need to know the data format of the metadata you have and check whether you can export this from your local catalogue or database. For more information on both of these issues you can consult Chapter 6 of the BPG: Preparation of data for BHL-Europe. More specifically paragraph 6.2.1 File submission guidelines.

Additional general information can be found using the following link: Moving theory into practice: digital imaging tutorial - 5. Metadata (Cornell University Library, 2000-2003): <http://www.library.cornell.edu/preservation/tutorial/metadata/metadata-01.html>

It is wise to write a status report on the documents you send and compile a shipment list.

If you are not scanning the material yourself, there are two things you have to keep in mind. First, you will have to be able to answer questions by the scanning company about the documents. Since the documents are not available for checking by you after you have shipped them, you will have to depend on the shipment list and the status report you have created. Second, no matter how careful the documents are being handled, there will always be the chance of something going wrong and items getting damaged. It is in these situations that a shipment list and status report will be indispensable.

²⁵ Scaife B. (2009): Catalogue of content holder requirements (quality, quantity, accessibility, standards, specifications of content and metadata), (ECP-2008-DILI-518001) URL: https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf

²⁶ Non-BHL partners will gain access to these sites when they become a content provider of the project.

In case you need to a quotation for the scan job, you will need to get the best estimate of the total number of master scans that you expect to be created. This of course depends on the decision whether empty or blank pages will be scanned or not. Often it is decided that empty pages are scanned as well, because that way the digitised copy is a true representation of the physical copy. For the total price however, not only the total number of pages counts, but also the number of scans per format, and per sort of material (in case there might be differences).

If you are preparing a first set to be shipped to a scanning company, you may want to select a set of a few volumes you will be using as a test sample. When you are scanning a long running journal, you will see that its printing quality changes over the years. The printing technique used may have changed, as well as the paper used, but also the print font, the layout and other aspects may have changed as well. Choose a few volumes that represent the major changes you have observed. You can use this set later for (fine) tuning the scanning process (See: paragraph 5.3 Scanning).

5.1.3 Checklist

- 1) Retrieve the documents from the stacks or - when they are borrowed – from the user(s)
- 2) Make sure the documents cannot be borrowed for the duration of the project, and communicate this to your users (e.g. in your web catalogue)
- 3) Record the physical condition of the documents (status report)
- 4) Collate and identify missing pages and damage
- 5) Repair and replace any missing or illegible pages
- 6) Prepare intermediates (e.g. photocopies, transparencies)
- 7) Disbind originals (when required)
- 8) Create documentation for bibliographic control, indexing, tagging and encoding information (when required)

Selection of materials

- Identify materials (Determine legal restrictions; Investigate the availability of digital and other versions.)
- Eliminate items which are in poor condition or incomplete (Determine appropriate conversion process e.g. film, then scan, disbind originals etc.)

Determine the size of the collection

- Count number of titles, volumes and pages to be imaged, from bound or unbound documents.
- Count number of frames, fiche or reels of micro-images to be converted.
- Count number of finding aids required.

Prepare documents

- Retrieve documents from storage.
- Remove documents from circulation.
- Record the physical condition of documents.
- Collate and identify missing pages and damage.
- Repair and replace missing or illegible pages.
- Prepare intermediates (e.g. photocopies, transparencies).
- Disbind originals (when required).

- Create documentation for bibliographic control, indexing, tagging and encoding information (when required).

Determine imaging requirements (benchmarking)

- Assess essential document attributes to determine scanning requirements (resolution, bit depth, enhancements, file format, compression).
- Confirm results by scanning a sample.
- Perform inspection of sample on screen and in print.

Determine requirements for and create metadata

- Create catalogue entries for digital resources.
- Determine file naming and structuring strategies (e.g. individual images cf. Groups of images).
- Create additional indexes (e.g. index at article level for journal literature) or revise/enhance existing finding aids.

Source: AHDS (2008), "Information Papers",

URL: <http://www.ahds.ac.uk/creating/information-papers/checklist/index.htm> [07.10.2010]

5.2 Imaging requirements

5.2.1 Basic principles

When you get involved in digitisation, you will learn a lot of new technical terms and jargon. In this paragraph we will discuss some of the basic concepts. We do not aim to give you a full introduction to the subject. We hope however to provide you some basic knowledge, that should help you to get familiar with the more technical side of the subject.

1	1	1	1	1	1	1	1	1
1	0	0	0	1	1	0	0	0
1	1	0	1	1	1	1	0	1
1	1	0	1	1	1	1	0	1
1	1	0	1	1	1	1	0	1
1	1	0	0	0	0	0	0	1
1	1	0	1	1	1	1	0	1
1	1	0	1	1	1	1	0	1
1	1	0	1	1	1	1	0	1
1	0	0	0	1	1	0	0	0
1	1	1	1	1	1	1	1	1

Fig. 7: Bitmap

There are two types of graphic images: vector graphics and bitmap or raster graphics. Vector graphics are created by specific graphical programs and exist of vector defined shapes. These images have relative small file sizes and are scalable.

A raster graphic or bitmap is a digital image composed of a matrix consisting of a (large) amount of small units, called pixels. Scanning software produces bitmap.

In a bitmap each pixel can be assigned a value. A bitmap that just has the values black and white (or the binary values 0 and 1) is said to be bitonal. Fig. 7 is an example of how a character might look like in a bitmap. Greyscale images are bitmaps that show different shades of grey. Usually they store 8 bits or 256 (28) of different tones per pixel. For colour images up to 24 bits (224) = 16.7 million tones are being stored in the bitmap. This is also referred to as bit depth: the greater the bit depth of an image, the more tones (in colour or shades of grey) it can represent. Tab. 8 sums up the bit rates that are most commonly used.

Bits per pixel	Number of colours available	Common name
1	2	Monochrome
8	256	VGA
24	16777216	SVGA / True Colour

Tab. 8: Commonly used bit depth rates

http://en.wikipedia.org/wiki/Color_depth

While bit depth is a good measure for the level of colour detail stored in an image; resolution is a measure for the level of spatial detail an image can hold. As mentioned earlier, a digital image is in fact a matrix of pixels. In a scan or digital copy, each square centimetre is represented by a certain amount of pixels. The more pixels per area are being used, the more detail can be stored and hence a higher resolution is achieved. The level of resolution is expressed as the number of *pixels per inch*, abbreviated as *PPI*. So, when you read that the resolution of a scan is 300 PPI, this means that each inch of the original paper is represented by 300 pixels in its digital image (or approximately 118 pixels per centimetre).

Apart from PPI there is often talked about *DPI* or *dots per inch*. Most of the time, this is used as a synonym of PPI. When used strictly however, DPI means something different. DPI or dots per inch refers to the amount of ink droplets that a printing device uses per inch of paper. The amount of DPI has to be much higher to gain the same level of detail as the number of PPI. When you for instance want to make a print of an image with a resolution of 300 PPI, you will need a printer that prints 1200 dots per inch to produce a similar amount of detail.

Most of the time however, users will view images not on print but on their computer screens. It's therefore very useful to realise how a digital image is represented on such a device, because this is often misunderstood. Take for instance a computer screen that measures 15 x 12 inch (or 20 inch in diameter) and that has pixel dimensions of 1280 x 1024 pixels. Its screen resolution equals 85 pixels per inch (1280/20 or 1024/12 = 85 ppi). This means that if you have a (full screen) picture of 1280 x 1024 pixels its maximum resolution will be 85 ppi. Offering this picture at a resolution of 300 ppi is then useless, because this level of detail can't be shown on this screen (without zooming that is). Because of the extra detail, the 300 ppi picture has a larger file size. And this not only takes extra storage space, it will also take extra time to load when viewed on a website. You may take this into account when you decide on what resolution you need for you digital images (see also calculation example below).

Starting with the resolution you have chosen for the scanning and the size of the object to be scanned, you can calculate the file size per image in advance. You do this by taking the surface area of the page to be scanned (height x width) and multiply this by the bit depth and the PPI²:

file size=height× width × bit depth × ppi/8 bytes

(Note: height and width are in inches, and all is divided by 8 because file size is expressed in bytes, which are made up of 8 bits)

Example: one page of size A4 (8.27 × 11.69 inch):

scanned in greyscale (8 bits) at 300 PPI

file size= $8.27 \times 11.69 \times 8 \times 300/8 = 8700867$ bytes= 8497 Kilobyte KB= 8.3 Megabyte (MB)

scanned in greyscale (8 bits) at 100 PPI

file size= $8.27 \times 11.69 \times 8 \times 100/8 = 966763$ bytes= 944 Kilobyte KB= 0.92 Megabyte (MB)

For both examples we have assumed that *all* information and about *each* pixel will be stored in the image file. In practice however, file sizes will be smaller because various compression techniques are being used to store pixel data much more efficiently. There are two types of compression: *lossless* and *lossy* compression. When *lossless* compression is being applied, data that is being repeated is stored in an abbreviated form (E.g. instead of storing *100 times* in a row *this pixel is white*, you could also store just *1 time* the *next 100 pixels are white*). After de-compression such an image is exactly the same as the original (non-compressed) image. This kind of compression is especially effective when applied to line drawings and texts.

When however *lossy* compression is being applied, some data will be lost. This is because the algorithm that takes care of the compression, replaces ranges of data with their average value. This means that the result after de-compression is not as good as the original image file. It can save you quite some storage space though and may therefore still be suitable for your purpose. Often it will be used for images that contain detailed grey tones or colour images. The efficiency of each compression technique will vary, depending on the frequency of repeated details in the digital image.

A file format that uses lossless compression is DjVu (.djvu, .djv). Note however that also TIFF image files can be compressed lossless. So, if you require uncompressed TIFF files, you have to specify that explicitly. A well know example of lossy compression is the JPEG (.jpeg, .jpg) image file format. For a few years now there is also file format with a very sophisticated compression technique called JPEG2000 (.jp2). This format applies lossless as well as lossy compression. For master scans BHL accepts both uncompressed TIFF, as well as lossless compressed JPEG2000 images.

Normally, when a document is being scanned, some extra space surrounding the page will be scanned as well. This is done to prevent loosing parts of a page in case it has been misplaced on the scanning machine. This first image, often stored uncompressed or at least lossless, is referred to as the raw image or the *master image* file. This master image will then be used for further processing, the products of which are then called *derivative images*.

One common manipulation of the master image is cutting off the unwanted edges of the scan. This is called *cropping*. The derivative image is then called a *cropped* image, while the master image is *uncropped*. Other possible manipulations include de-skewing images, or levelling an image that is not straight, removing noise from the pages like dots and bloats, applying some changes to the text to prepare it for OCR like filling the cuts, and OCR (optical character recognition. Read more on this in paragraph 5.4). But it may also concern converting images to other file formats. For instance converting to JPEG images is very common because these,

due to their smaller file size (lossy compression), are often much more suitable for presentation in websites.

5.2.2 Resolution recommended, Bit depth

The recommended minimum and preferred values for the resolution, bit depth and file format for the supply of digital images are as followed (Tab. 9):

Type	Resolution (minimum)	Resolution (preferred)	Bit-depth	File format
Bitonal	300 ppi	600 ppi	1-bit	TIFF
Greyscale	300 ppi	600 ppi	8-bit	uncompressed TIFF or lossless compressed image (e.g. LZW, JPEG2000 [* .jp2])
Colour	300 ppi	400 ppi	24-bit	uncompressed TIFF, or lossless compressed images (e.g. LZW, JPEG2000 [* .jp2])

Tab. 9: Resolution and Bit depth

NOTE: BHL Europe prefers submissions in a non-lossy format, either Tiff or JPEG2000. Legacy scans with a maximum of 15% compression are accepted, maintaining 85% quality. General aim is that the archive preserves the highest quality available.

Do keep in mind that the above quoted values reflect the minimum requirements. This means that you can, and should, if that is possible or desirable for your purpose, use qualifications that exceed this minimum. The given preferred values may serve as an extra guideline for when you are deciding on the specification for your project.

BHL-Europe partners can access more detailed information on the requirements in the document “Catalogue of content holder requirements” on the BHL wiki.²⁷

Scaife B. (2009): Catalogue of content holder requirements (quality, quantity, accessibility, standards, specifications of content and metadata), BHL-Europe Deliverable 2.1.²⁸

URL wiki: https://bhl.wikispaces.com/BHLE_WP2

URL PDF: https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf

5.2.3 Post-processing

According to the “Catalogue of content holder requirements” https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf²⁹ “Master images [should] be supplied without post-processing.” This means that you should provide, when possible, the unprocessed (e.g. uncropped) master images. When possible the colour profile used during

²⁷ Non-BHL partners will gain access to these sites when they become a content provider of the project.

²⁸ The Catalogue of content holder requirements is largely based on the Digital Library Federation’s (DLF) “Benchmark for Faithful Digital Reproductions of Monographs and Serials”: <http://www.diglib.org/standards/bmarkfin.htm> (2002)

²⁹ Non-BHL partners will gain access to these sites when they become a content provider of the project.

the scanning workflow should be embedded. Every device that captures scanned images can have its own colour profile. BHL-Europe will store these images for preservation purposes. Next however, you are also asked to provide cropped derivatives of the master images. These will then be used within the page turner software of the BHL-Europe portal.

5.3 Scanning

Before starting to scan a large volume of work (by a scanning company), you should start with a small set of volumes that you have collected earlier (See paragraph 5.1 Preliminary steps). This set of volumes should represent most of the differences this scanning job will encounter; differences in the physical quality of volumes, but also the structural differences. Starting with this set enables you to verify if you and the scanning company both understand the scanning instructions you have compiled in the same way.

The scanning company scans the set and delivers the results to you. You will then need to check these results in detail and check whether they meet your scanning instructions. This includes the technical instructions (resolution, bit depth, file formats etc.) but you will also have to check if this is a true copy of the original work. Have all the pages been scanned? Including the blank ones, for instance? What has happened with fold outs? Or loose leaf errata or addenda?

After this first check, you may have discovered some mistakes. You may also have discovered that some of your instructions have been interpreted differently than you expected. It may also be that you discover that some of your original instructions need to be adjusted. Create a clear and detailed report on everything you have found and send this report back to the scanning company. Together with their comments on this report, you can now create an updated version of the scanning instructions, which can then be used for a second test (when needed of course).

The test phase cannot continue for any length of time of course. Part of the success of your scanning enterprise however will depend on how well you use this testing period. Be prepared to invest time in it. This will save you a lot of time (and trouble) later on.

For quality control see paragraph 5.6 Quality control and for detailed information on Quality assurance BHL-Europe partners have additional access to the following link: <https://bhl.wikispaces.com/QA+Procedures>³⁰

5.4 OCR

When a page has been scanned, the result is an image file or, to put it differently, a *picture* of that page. When you open this file, you can perfectly read the text, but you cannot select the text with your mouse to, for instance, copy and paste it. Nor can the file be used for full text search, simply because nothing in this *picture* will be recognised as text by the search software.

There is however specialised software that is capable of recognising text characters in picture files. This software can add as it were an extra layer to the picture containing the text of that page. That way, you can select the text at the page, and the file can be used for text search.

³⁰ Non-BHL partners will gain access to these sites when they become a content provider of the project.

The technique that makes this possible is called Optical Character Recognition, commonly abbreviated as OCR.

For BHL-Europe, no OCR is needed. Content partners are asked only to provide raw images in TIFF or JPEG2000 format. After the ingest, the OCR software of BHL-Europe will make these pages machine readable for use in search engines and the BHL-Europe portal. This however does not mean that content providers should not aim for OCR in their digitisation efforts. There are many reasons to do this anyway. The most important of which will be that the content will be provided at the own institutional repository or digital library.

OCR of image files does not have to be done at the same time of the scanning, although this will often be the case. Most scanning companies offer this as a service and it may be convenient to let the same company scan and OCR your material.

OCR software is fairly good at recognising text nowadays, but errors will occur. Accuracy rates are between 95% and 100% for documents using the Latin alphabet. It might be difficult to obtain this level of accuracy for other languages like Arabic for instance. The level of accuracy depends on many factors, among which are the scanning quality, the type face and font size, the language and the OCR software itself. Of major influence is also the use of a (specialised) lexicon. This can greatly increase the accuracy of the OCR process. This may be just one reason, why in BHL-Europe the OCR is done afresh: with OCR you can build a specialised biodiversity literature lexicon, which enables you to improve the OCR accuracy rates.

As mentioned earlier, if the quality of the scanned text is not that good: e.g. the text contains cut, toothed or jagged letters, some processing might be useful to increase the OCR accuracy. Processing will help in filling the cuts, levelling the teeth through a process of filing and so on. This kind of processing can be done automatically by software, but it requires a human operator to inspect the document to decide which processing technique is required.

If OCR accuracy rates are low and we still need to obtain a high level of text recognition, then the OCR process can be further enhanced through a learning phase. The OCR tool can be trained to recognise the font of this particular volume being scanned. If this doesn't work as well, then the generated text might need manual correction or rekeying in the worst case.

When you commission an OCR job you can ask for a specific accuracy rate. You can however also specify specific accuracy rates for specific blocks of text. You might for instance want an accuracy rate of 99.50% for titles and abstracts, and settle for an accuracy rate of 95% for the body of texts. If you have a (specialised) lexicon available, this can be used to further increase the accuracy rate. A lexicon of taxonomic names might for instance be very useful in this respect.

For a list of common OCR software, see for instance:

http://en.wikipedia.org/wiki/List_of_optical_character_recognition_software

Chapter 7.2 will further evaluate tools for the digitisation workflow.

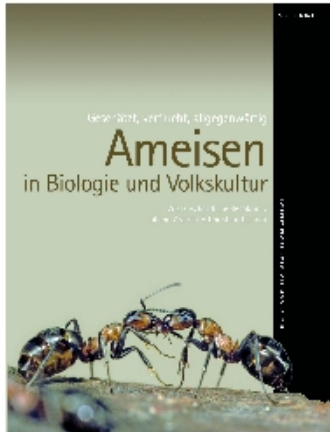
5.5 Article level access

This chapter is based on the experienced BHL-Europe partner LANDOE and aims to provide help, new ideas and work patterns about developing access at article level.

Division at article level

Our content provider LANDOE lists all publications at article level. Since this practice involves substantial additional effort compared to the listing of volumes and the automatic division of the whole files is only possible to a limited extent, questions concerning the purpose and the advantage of such an approach might be raised. The following paragraph outlines the reasons for this approach.

Periodicals usually contain a number of articles dealing with specific scientific questions and observations. Publishers of electronic journals currently display and allow searching at article level. Compared with downloading a complete journal volume or part downloading individual articles requires a smaller data volume. Furthermore full text indexing of stored data only includes relevant parts and is not adulterated by articles that the researcher does not require, and this allows more efficient searching. Being able to definitely link articles to an author is another advantage of listing individual articles. This not only provides an opportunity to generate lists of publications for a specific author but it also makes it possible to include additional information about the author such as CVs, obituaries, pictures, travel information etc. The inclusion of individual articles in the database also allows for a directed search in the metadata without OCR-errors. Therefore the supply of individual articles logically results from analogous collection of special prints, which merely include individual contributions of various authors too. Additionally, the online presentation of the individual articles resembles a printed table of contents (Fig. 8): A PDF-download button (free or fee-based) can be found on the right and the author's name on the left, as shown in the illustration. The author's name is underlined to mark it as a link where additional information about the author can be obtained.



Denisia 0025:diverse (2009): Geschätzt, verflucht, allgegenwärtig
Ameisen in Biologie und Volkskultur 188 pp.



Inhaltsverzeichnis












Christian O. Dietrich, Erich Steiner Das Leben unserer Ameisen - ein Überblick	7- 36 Details	 4.0MB
Johann Ambach Zur Geschichte und Entwicklung der Myrmekologie in Österreich	37- 52 Details	 1.5MB
Hans Christian Luschützky Myrmekonymia Europaea Ein wortgeschichtlicher Streifzug auf den Spuren der Ameisen	53- 66 Details	 2.7MB
Jens Dauber Ameisen (Hymenoptera, Formicidae) in Kulturlandschaften	67- 78 Details	 1.3MB
Florian Glaser Ameisen (Hymenoptera, Formicidae) im Brennpunkt des Naturschutzes Eine Analyse für die Ostalpen und Österreich	79- 92 Details	 627.0kB
Johann Ambach Hügelbauende Waldameisen (Formica rufa-Gruppe) Ein aktueller Überblick	93- 106 Details	 2.3MB
Veronika Mayer Tragedienste gegen Nahrung: Ameisen als Frucht- und Samenverbreiter	107- 118 Details	 5.3MB
Wolfgang Rabitsch Gebietsfremde Ameisen: Eine Übersicht (Hymenoptera: Formicidae)	119- 140 Details	 1.4MB
Thomas Schmickl Schwarmintelligenz am Beispiel der Ameisenstraßen	141- 155 Details	 1.1MB
Elisabeth Schuster Ame(i)s als Bestandteil von Orts- und Flurnamen in Nieder- und Oberösterreich	157- 163 Details	 2.3MB
Franz Groß Ameise und Volkskultur	165- 188 Details	 914.7kB

Fig. 8: <http://www.landesmuseum.at/biologiezentrum/> online presentation of individual articles belonging to a volume of a journal

How can articles be extracted from the entire volumes?

PDF-files of the entire volumes provide the basis for the articles that need to be extracted. To extract the individual parts Adobe Acrobat Professional, more precisely Version 7 of this program is used (the process is similar in newer versions). After opening the entire volume divide the window into two horizontal halves (i.e. Menu – divide window) (Fig. 9). On the top half the first page of the document that needs to be separated remains open while on the bottom half you have to navigate to the last page of the document (Fig. 10). When the cursor points on the top half a page number is shown in the status bar. Make a mental note of the page number and move the cursor to the bottom half. Open document-extract pages in the menu bar next. In the input window that pops up the number next to "from" is already highlighted (Fig. 11). Type the number that you kept in mind in and as a result you will get the extracted article. You can see that the separation went according to plan since the first page of the separated document has to be shown on the screen. When pressing the save-button you will be directed to a folder where the documents that were separated before are saved as well. In this folder you click on the file that was saved last and update the page number in the name of the document under file name (Fig. 12). Depending on the outline of the volume it might be the case that illustrations that relate to particular articles are found in other parts of the journal volume or part. These illustrations need to be extracted as described above and saved under a temporary name. To include the illustrations open the article and use the insert page command in the menu bar. Then select the illustration file that you just saved and in the next pop up window choose where you want to insert it (Fig. 13).

You will possibly have to make a few adjustments when opening the entire volume. Click on View, Page Layout in the menu bar and select single page (Fig. 14), then go to Edit, Preferences in the menu bar and remove the check mark next to "use logical page numbers" when setting the page numbers (Fig. 15).

Instead of the printed document's pagination the actual page numbers as shown in the status bar have to be taken into account (Fig. 16). Therefore it is possible that the printed page is cited as page 1 while the PDF-file is already at page 5. It may take some time to get used to this.

Linzer biol. Beitr.	39/2	727-730	18.12.2007
---------------------	------	---------	------------

Hieracia nova Alpium IV

G. GOTTSCHLICH

A b s t r a c t: *Hieracium rottii*, a new species of the Southern Alps (Italy, Piemonte, Prov. Vercelli) is described and illustrated.

K e y w o r d s: *Hieracium*, Compositae, new species, Alps, Italy

Linzer biol. Beitr.	39/2	727-730	18.12.2007
---------------------	------	---------	------------

Hieracia nova Alpium IV

G. GOTTSCHLICH

A b s t r a c t: *Hieracium rottii*, a new species of the Southern Alps (Italy, Piemonte, Prov. Vercelli) is described and illustrated.

K e y w o r d s: *Hieracium*, Compositae, new species, Alps, Italy

Denominatio alterum Botanices fundamentum.

LINNAEUS

Einleitung

In einer Revisionssendung mit Hieracien aus den Penninischen Alpen (Provinz Vercelli, Piemont), die mir von Adriano Soldano, Vercelli, vermittelt wurde, war eine schon auf den ersten Blick als neu postulierte Art enthalten. Eine Überprüfung am Wuchsort, zu-

Fig. 9: window divided into two horizontal halves (first step)

Linzer biol. Beitr.	39/2	727-730	18.12.2007
---------------------	------	---------	------------

Hieracia nova Alpium IV

G. GOTTSCHLICH

A b s t r a c t : *Hieracium rottii*, a new species of the Southern Alps (Italy, Piemonte, Prov. Vercelli) is described and illustrated.

K e y w o r d s : *Hieracium*, Compositae, new species, Alps, Italy

Denominatio alterum Botanices fundamentum.

LINNAEUS

730



Fig. 10: second step

Linzer biol. Beitr.	39/2	727-730	18.12.2007
---------------------	------	---------	------------

Hieracia nova Alpium IV

G. GOTTSCHLICH

Abstract: *Hieracium rottii*, a new species of the Southern Alps (Italy, Piemonte, Prov. Vercelli) is described and illustrated.

Key

Seiten entnehmen

Von: Bis: von 616

Seiten nach Entnahme löschen

Seiten als einzelne Dateien entnehmen



Abb. 1: *Hieracium rottii* GOTTSCHL.

Fig. 11: third step

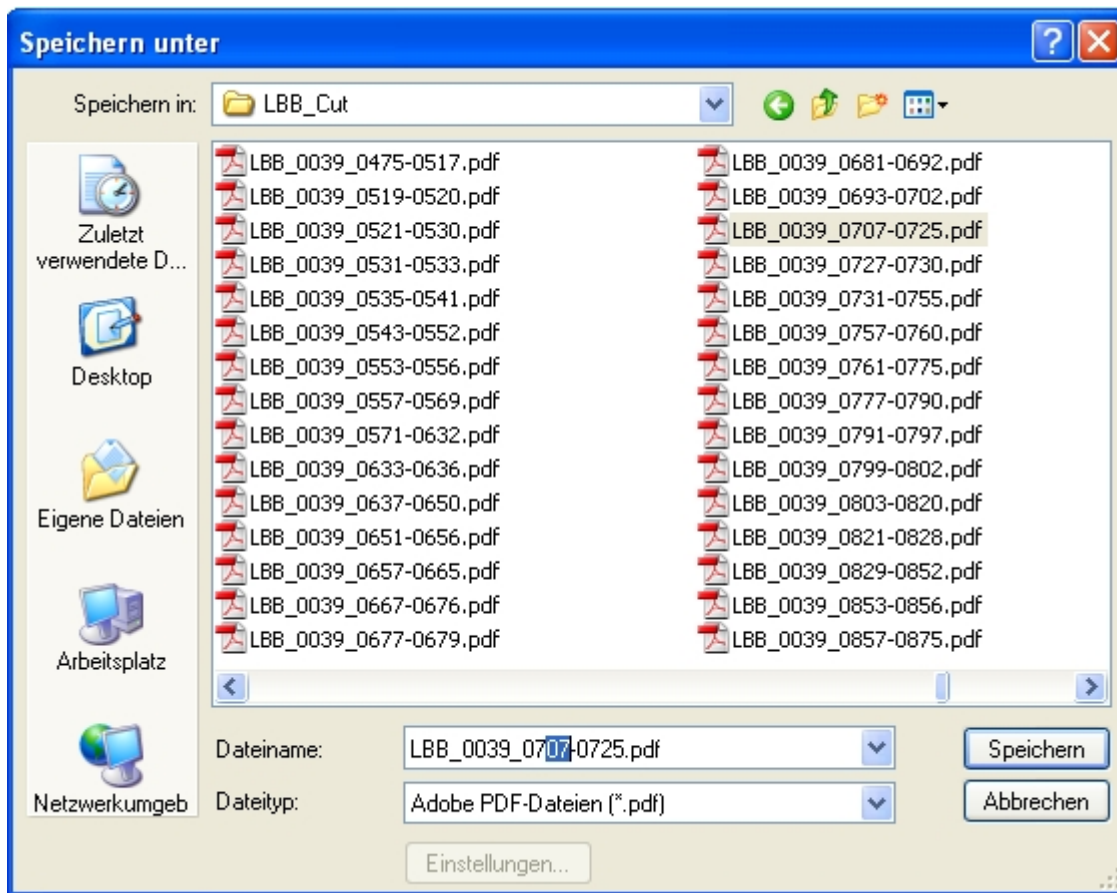
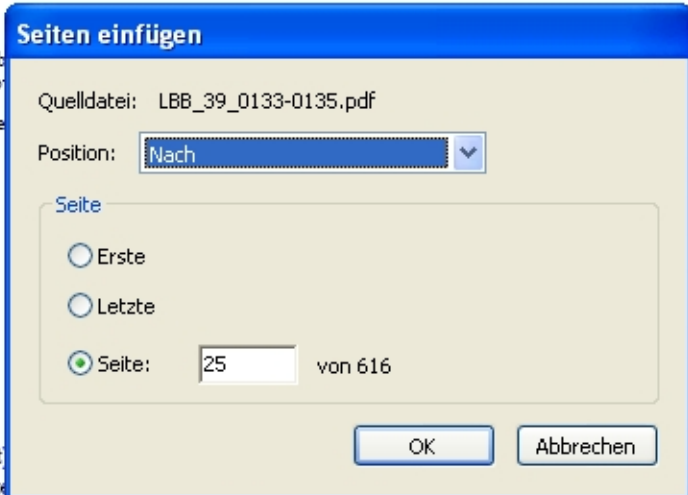


Fig. 12: fourth step

Linzer biol. Beitr.	39/2	727-730	18.12.2007
---------------------	------	---------	------------

Hieracia nova Alpium IV

G. GOTTSCHLICH



A b
Pro
K e

Quelle, fonte,

Seiten einfügen

Quelldatei: LBB_39_0133-0135.pdf

Position: Nach

Seite

Erste

Letzte

Seite: 25 von 616

OK Abbrechen

mentum.
LINNAEUS

In einer
Piemont
den erste

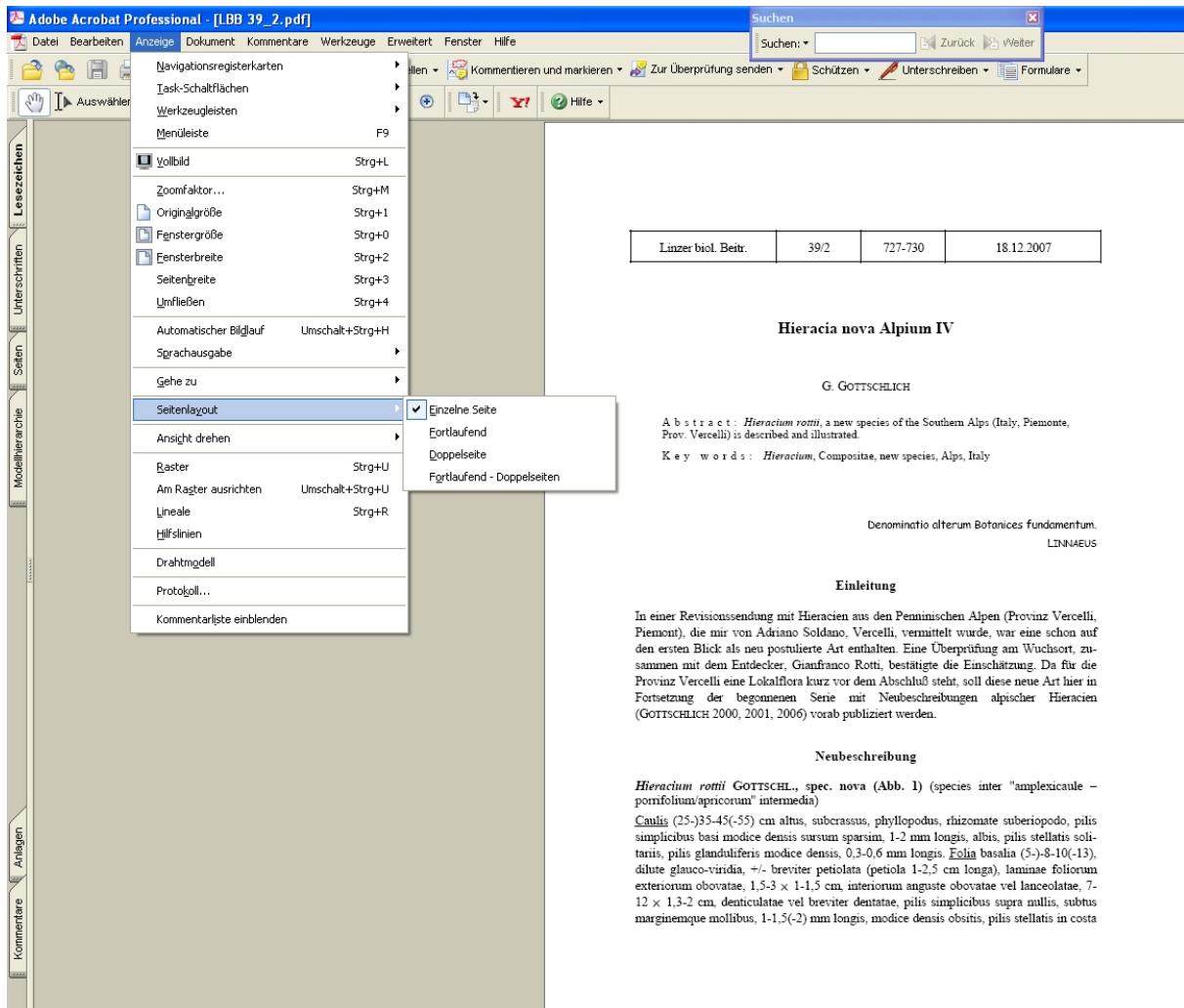
z Vercelli,
schon auf
hsort, zu-
sammen mit dem Entdecker, Gianfranco Rotti, bestätigte die Einschätzung. Da für die Provinz Vercelli eine Lokalflorenkurze kurz vor dem Abschluß steht, soll diese neue Art hier in Fortsetzung der begonnenen Serie mit Neubeschreibungen alpischer Hieracien (GOTTSCHLICH 2000, 2001, 2006) vorab publiziert werden.

Neubeschreibung

Hieracium rottii GOTTSCHL., spec. nova (Abb. 1) (species inter "amplexicaule – pomifolium/apricorum" intermedia)

Caulis (25-)35-45(-55) cm altus, subcrassus, phyllopodus, rhizomate suberipodo, pilis simplicibus basi modice densis sursum sparsim, 1-2 mm longis, albis, pilis stellatis solitariis, pilis glanduliferis modice densis, 0,3-0,6 mm longis. **Folia** basalia (5-)-8-10(-13), dilute glauco-viridia, +/- breviter petiolata (petiola 1-2,5 cm longa), laminae foliorum exteriorum obovatae, 1,5-3 × 1-1,5 cm, interiorum anguste obovatae vel lanceolatae, 7-12 × 1,3-2 cm, denticulatae vel breviter dentatae, pilis simplicibus supra nullis, subtus marginemque mollibus, 1-1,5(-2) mm longis, modice densis obsitis, pilis stellatis in costa

Fig. 13: fifth step



Adobe Acrobat Professional - [LBB 39_2.pdf]

Suchen

Suchen: Zurück Weiter

Menü: Kommentieren und markieren Zur Überprüfung senden Schützen Unterschreiben Formulare

Navigation: Navigationsregisterkarten, Task-Schaltflächen, Werkzeugleisten, Menüleiste (F9)

Leseeigenschaften: Vollbild (Strg+L), Zoomfaktor... (Strg+M), Originalgröße (Strg+1), Fenstergröße (Strg+0), Fensterbreite (Strg+2), Seitenbreite (Strg+3), Umfließen (Strg+4), Automatischer Bildlauf (Umschalt+Strg+H), Sprachausgabe, Gehe zu

Seitenlayout (ausgewählt): Einzelne Seite (ausgewählt), Fortlaufend, Doppelseite, Fortlaufend - Doppelseiten

Ansicht drehen, Raster, Am Raster ausrichten (Umschalt+Strg+U), Lineale (Strg+R), Hilfslinien, Drahtgmodell, Protokoll..., Kommentarliste einblenden

Linzer biol. Beitr.	39/2	727-730	18.12.2007
---------------------	------	---------	------------

Hieracia nova Alpium IV

G. GOTTSCHLICH

Abstract: *Hieracium rottii*, a new species of the Southern Alps (Italy, Piemonte, Prov. Vercelli) is described and illustrated.

Key words: *Hieracium*, Compositae, new species, Alps, Italy

Denominatio alterum Botanices fundamentum.
LINNAEUS

Einleitung

In einer Revisionsendung mit Hieracien aus den Penninischen Alpen (Provinz Vercelli, Piemont), die mir von Adriano Soldano, Vercelli, vermittelt wurde, war eine schon auf den ersten Blick als neu postulierte Art enthalten. Eine Überprüfung am Wuchsort, zusammen mit dem Entdecker, Gianfranco Rotti, bestätigte die Einschätzung. Da für die Provinz Vercelli eine Lokalflora kurz vor dem Abschluß steht, soll diese neue Art hier in Fortsetzung der begonnenen Serie mit Neubeschreibungen alpischer Hieracien (GOTTSCHLICH 2000, 2001, 2006) vorab publiziert werden.

Neubeschreibung

Hieracium rottii GOTTSCHL., spec. nova (Abb. 1) (species inter "amplexicaule - porrifolium/apricorum" intermedia)

Caulis (25-)35-45(-55) cm altus, subcrassus, phyllopodus, rhizomate suberipodo, pilis simplicibus basi modice densis sursum sparsim, 1-2 mm longis, albis, pilis stellatis solitariis, pilis glanduliferis modice densis, 0,3-0,6 mm longis. Folia basalia (5-)8-10(-13), dilute glauco-viridia, +/- breviter petiolata (petiola 1-2,5 cm longa), laminae foliorum exteriorum obovatae, 1,5-3 x 1-1,5 cm, interiorum anguste obovatae vel lanceolatae, 7-12 x 1,3-2 cm, denticulatae vel breviter dentatae, pilis simplicibus supra nullis, subtus marginemque mollibus, 1-1,5(-2) mm longis, modice densis obsitis, pilis stellatis in costa

Fig. 14: sixth step

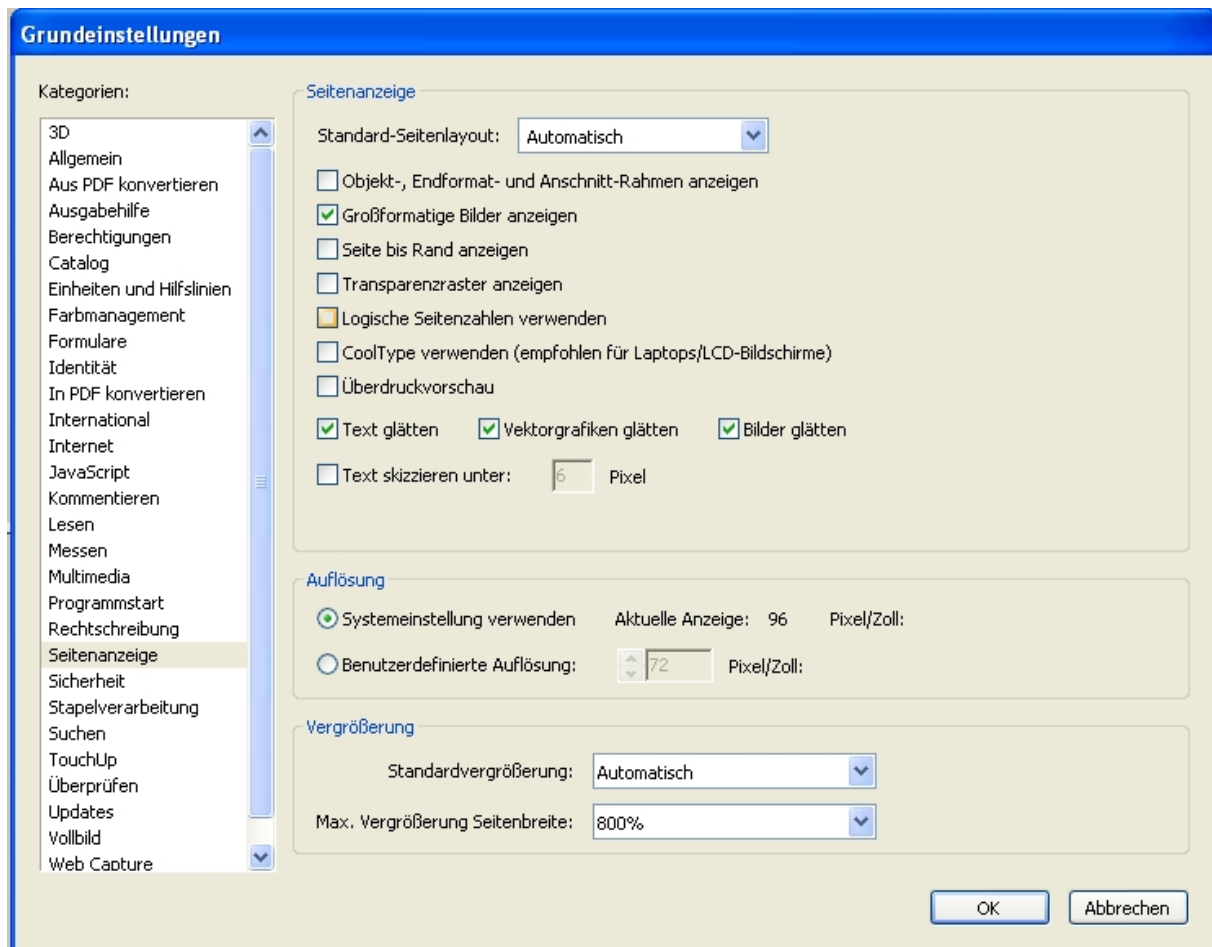


Fig. 15: seventh step

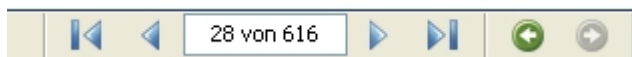


Fig. 16: eighth step

5.6 Quality control

This chapter draws on the experience of LANDOE to describe a three step quality control workflow.

First quality step: enhance the Article level data of the Author(s) with biographical data, so one can more easily distinguish and match author names (full names, address).

Second quality step: Names of article level PDFs have a schema that has to match the metadata; otherwise the article level PDF will not be shown in the web:

Series code, number of volume, start page and end page are cross checked here.

Third quality step: The quality control person has to read through the web table of contents and check that the volume or part is complete.

5.7 Further References

This chapter gives a selection of useful further references to assist in further understanding of the various parts of a digitisation workflow.

There are many useful links in the World Wide Web and the following is only a representative selection to assist you in your digitisation project planning:

- Moving theory into practice: digital imaging tutorial (Cornell University Library, 2000-2003). The Digital Imaging Tutorial helps the user to learn about the use of digital imaging to convert and make accessible cultural heritage materials.
<http://www.library.cornell.edu/preservation/tutorial/>
- Bit depth tutorial: <http://www.cambridgeincolour.com/tutorials/bit-depth.htm>
- Benchmark for Faithful Digital Reproductions of Monographs and Serials: <http://www.diglib.org/standards/bmarkfin.htm>
- The *Make It Digital* guides: <http://makeit.digitalnz.org/guidelines>

Further information is available for BHL-Europe partners on our wiki:³¹

BPG: https://bhl.wikispaces.com/BHLE_WP2_BPG

Cook Book: <https://bhl.wikispaces.com/BHL+Cookbook>

WP2: https://bhl.wikispaces.com/BHLE_WP2

6 Preparation of data for BHL-Europe

At the date of writing, the development of the BHL-Europe portal and the associated modules for ingesting data were in progress and therefore it is not possible to give a detailed guideline to the designated themes within this version of the “Approved Best Practice Guidelines and Standards”. However, in this chapter we provide a review of the current information. More details will be provided in the BHL-Europe Deliverable 2.9 final revised BPG which is due in April 2012. Thank you for your understanding!

6.1 Set up technical interface/provider software interface

6.2 Data harmonization

6.2.1 File submission Guidelines

BHL-Europe provides files submission guidelines for its content providers. The file submission guidelines are necessary to allow automatic processing of submitted content and metadata. At the time of preparation of the BPG the latest version of these guidelines was version 3, which is attached as an Appendix (page 102) to this document.

For project partners the up-to-date file submission guidelines are available on the internal wiki navigating from the WP2 main page within the section “Submission Guidelines for Content Providers”.

³¹ Non-BHL partners will gain access to these sites when they become a content provider of the project.

6.2.1.1 Files to be supplied

BHL-Europe asks content provider to provide digital content (scanned images, born digitals) together with the corresponding metadata. Content providers are asked to follow the file submission guidelines mentioned in the chapter 6.2.1.

6.2.1.2 Example

Examples of how to provide your data according to the file submission guidelines can be found in the attached file submission guidelines. As the file submission guidelines are a working document we recommend that our content providers regularly check the current version of “Submission Guidelines for Content Providers” on the WP2 main page of the BHL-Europe wiki, when they wish to check for worked examples for content submission.

6.2.1.3 Delivery method

A BHL-Europe content provider receives an FTP account for uploading data (metadata and scanned images or born digitals) to the BHL-Europe server.

BHL-Europe also offers to harvest metadata via an OAI-PMH interface. Nonetheless, if metadata is provided over OAI-PMH, the content provider still needs to upload the corresponding scans to the BHL-Europe server. In that case it is required to name the uploaded folders using the same identifiers that are used by the content providers OAI provider.

An example of how to implement an OAI-PMH web service is given in the next chapter by using the Europeana Semantic Elements Schema.

6.2.1.3.1 Implementing an OAI-PMH web service

Our content provider LANDOE implemented an OAI-PMH web service for their institution as it was more useful for them. The implementation of this service took place before the BHL-Europe Schema was defined and therefore this example uses the Europeana Semantic Elements Schema. After the BHL-Europe Schema has been finalised and BHL-Europe data workflow is well proven, we will update this example for the final revised version of the BPG at the end of the project (30.04.2012).

LANDOE has lots of items digitised and plans to do even more and on a regular basis in the future. Therefore they have built a flexible interface for the delivery of the metadata information.

A general starting point for implementation of an OAI-PMH web service is the homepage of the Open Archives Initiative (www.openarchives.org). Under the bullet point Tools one can find lots of different implementations for OAI PMH web services, in doing so LANDOE decided to take the PHP OAI Data Provider from the University of Oldenburg (Germany). The software comes in one TGZ package (compressed tar file format) and includes several PHP scripts and configuration files.

The archive extracts to the following directory structure:

OAI (base with php basic scripts)

- doc (documentation)

- oai2 (configuration (oaid-config.php) and php sub scripts. These two config files were relevant.

First, file to be configured

oaidp-config.php

Parameters that should be filled out are shown here using LANDOE as an example:

\$repositoryName: BHL Land Oberoesterreich, Oberoesterreichische Landesmuseen, Biologiezentrum

\$baseURL: <http://www.biologiezentrum.at/oai/oai2.php>

\$adminEmail: array('mailto:m.malicky@landesmuseum.at')

\$repository Identifier: biologiezentrum.at

\$MAXRECORDS: 300 (maximum records delivered by query)

\$MAXIDS 300

\$SETS array(array('setSpec'=>'bhllandoe', 'setName'=>'BHL Land Oberoesterreich', 'setDescription'=>''))

\$METADATAFORMATS

```
Array ('ese' => array('metadataPrefix'=>'ese',
    'schema'=>'http://www.europeana.eu/schemas/ESE-V3.3.xsd',
    'metadataNamespace'=>'http://www.europeana.eu/schemas/ese/',
    'myhandler'=>'record_ese.php',
    'record_prefix'=>'dc',
    'record_namespace'=>'http://purl.org/dc/elements/1.1/'));
```

\$DB_HOST='//IP address db server//'

\$DB_USER='...' database username

\$DB_PASSWD='...' database password

\$DB_NAME='zobo' database instance

\$DSN = „pgsql://...“ Postgresql connect string

\$charset = „utf-8“ database instance encoding

\$SQL['identifier'] = 'oai_identifier' (unique identifier within the table oai_records)

Second, file to be configured

record_ese.php

As a copy of the original file **record_dc.php**. The **record_ese.php** is the actual mapping file between the database table(s) – oai_records - and the ESE schema. Our partner LANDOE copied the original file (record_dc.php) and renamed it to record_ese.php. Afterwards, the record_ese.php has been modified according to their needs.

The modified lines are stated hereafter:

```
$output .= xmlrecord($record['dc_title'], 'dc:title', "", $indent);
    Title of Volume e.g “Phyton, Annales Rei Botanicae, Horn 35_2 (1995)”
$output .= xmlrecord($record['dc_creator'],'dc:creator', "", $indent);
    Author(s) of the Volume
$output .= xmlrecord($record['dc_description'], 'dc:description', "", $indent);
    Table of contents of the volume
$output .= xmlrecord($record['dc_date'], 'dc:date', "", $indent);
    Publishing date of the volume
$output .= xmlrecord($record['ese_isshownby'], 'ese:isShownBy', "", $indent);
    Link to the digital volume (object in PDF)
```

```

$output .= xmlrecord($record['ese_isshownat'], 'ese:isShownAt', "", $indent);
    Link to the generated table of contents of the volume on our homepage
$output .= xmlrecord($record['ese_object'], 'ese:object', "", $indent);
    Link to a picture of the cover page
$output .= xmlrecord($record['ese_provider'], 'ese:dataProvider', "", $indent);
    "Oberoesterreichische Landesmuseen, Biologiezentrum"
$output .= xmlrecord($record['ese_type'], 'ese:type', "", $indent);
    "TEXT"
  
```

A further requirement was a machine running an apache2 webserver with the PHP module (Version 4) included. The oai directory has to be included into the www structure of the apache webserver. Therefore LANDOE simply copied it directly as a sub directory under the apache base directory /var/www/ -> /var/www/oai to the machine.

6.3 Schema Mapping Tool

BHL-Europe is using its own schema - called BHL-Europe schema. Therefore BHL-Europe will provide an open source Schema Mapping Tool. The content provider's metadata will be mapped to the BHL-Europe schema using the Schema Mapping Tool. During the project time BHL-Europe will provide the mapping as a service for all content providers. However, this should not disengage content providers from their responsibility to maintain the quality of their metadata. A content provider first uploads test content to the BHL-Europe server including the scanned images and the corresponding metadata according to the file submission guidelines (see chapter 6.2.1). The test content is then checked by our technical team and BHL-Europe will start with the data mapping to the BHL-Europe schema creating a configuration file for every content provider to accommodate for the characteristics of the content provider data. The mapping configuration file is defined for every content provider individually once and will remain the same as long as a content provider does not change the structure of the provided metadata. However, standard mappings (for e.g. MARC21, DC) will be available and only partners with non-standard data will need more work to map the data properly. The mapping process is accompanied by a quality assurance process to negotiate and verify the result of the mappings in collaboration with the content providers. After the test content is approved, the content provider can upload the bulk of the material to be ingested into the BHL-Europe system.

In case a content provider has the expertise and resources to work on the mapping configuration file himself/herself, BHL-Europe can offer this as an option and will provide all necessary tools to the content provider. The Schema Mapping Tool will be open source and documentation will be available before the end of the project.

6.4 Pre-Ingest

A content provider is asked to log into the Pre-Ingest module to trigger the conversion of all metadata into the BHL-Europe schema based on the previously defined configuration file (see chapter 6.3 for more details) by selecting the folder(s) for procession via the Pre-Ingest interface. The Pre-Ingest module will also provide up to date information on the status of the processing and the results.

More detailed information on the Pre-Ingest module can only be given in the next BHL-Europe Deliverable 2.9 (final revised BPG due 30.04.2012), as to date of writing the module was in progress.

6.5 Ingest

Due to the current development status, information about the ingest module can only be given in the next BHL-Europe Deliverable 2.9 (final revised BPG due 30.04.2012).

6.6 Post or Re-Ingest

This chapter will be elaborated in the next BHL-Europe Deliverable 2.9 (final revised BPG due 30.04.2012), due to the current development status.

6.7 Ingest to Europeana

BHL-Europe acts as a content aggregator for Europeana and will therefore additionally ingest metadata to Europeana after a content provider has provided content (metadata and scanned images/born digitals).

A detailed description of the workflow is not available at the time of writing and will be provided in the next BHL-Europe Deliverable 2.9 (final revised BPG due 30.04.2012).

7 General Information

7.1 Workflow examples

This chapter will show workflow examples from BHL-Europe partners.

7.1.1 Example from the Georg-August Universität Göttingen Stiftung Öffentlichen Rechts, Göttingen (Germany)

The first example explains the collection analysis for scanning of books from the point of view of our BHL-Europe partner in Göttingen, Germany.

In Göttingen we had two scanning projects in the past 10 years; both were funded by the German Science Foundation DFG. The first project (EZOOLO 1, this means Early Zoological Literature Online) ran from 2003 to 2005, the second project (EZOOLO 2) from 2008 to 2011.

In both projects we scanned zoological works exclusively, particularly those which were important for taxonomic studies. In this way all books selected were biodiversity literature in the strict sense.

The persons who selected the books were exclusively zoologists with relatively deep experience in taxonomy.

The selection methods were slightly different in both projects.

A) The selection method for the EZOOLO 1 (2003-2005) project:

a) Book selection

We selected some 350 early zoological works. Our aim was to digitise the very old (pre-1771) zoological literature which is important for the taxonomic studies, especially because many original descriptions of animal taxa had been described in these very old works, and many zoologists had no access to the rare books (particularly outside Europe). We knew from our zoological experience that the content of the old books was not outdated because in modern taxonomic research the original description where a name was made available, must always be consulted, for example to verify the spelling of a name, or the original genus-species combination. Those who selected the books were zoologists with only marginal experience in library issues.

We arbitrarily set our limit at 1770 and decided to digitise the basic zoological works by Linnaeus and some other authors published until 1770, together with those works to which these authors had referred when they established the taxonomic names for the species. The cut at 1770 was set due to financial reasons; we feared that if our application was too expensive it would not be approved. We calculated a possible financial margin prior to setting this limit at 1770.

We analysed a printed version of Linnaeus 1758 (*Systema Naturae*) page by page and extracted the bibliographical references, to get to know the books we would need to digitise. Linnaeus 1758 gave some 10,000 references to 400 different works. Of these, 97 % of the references referred to 130 works published between 1550 and 1758. We knew that our library (SUB Göttingen) held about 95 % of these old works. We compiled a list of these 130 works, added some more (about 100) works published until 1770, and some other important works corresponding to the remaining 3 % of the Linnean 1758 and 1766/1767 references, together 400 works.

In the EZOOLO 1 project application to the funders we finally added a complete list containing the titles and page numbers (black and white or colour) of all works. This long and detailed list must have looked very impressive to the reviewers (in fact we had invested one entire year of unpaid work for compiling this detailed list) and so they approved our application and granted the financial funds.

At that time, the costs for bitonal digitising were 1/10 lower than for greyscale scans and 1/100 lower than for colour scans. This was why we had to check all the original outprints of the books prior to setting them on the list, and to count the bitonal, greyscale and colour figure pages. 5 years later this differentiation was not necessary any more because colour scans were practically the same cost as bitonal scans.

b) Deduplication

At this time almost no digitised zoological works were available online, so we had no need to search for the presence of online content somewhere else.

c) Obtaining books from other libraries

Several books were not present in SUB Göttingen. We compiled a list of these and looked them up in the GVK³² interlibrary compound catalogue system, to find out if we could obtain these books from other libraries in loan.

³² <http://gso.gbv.de>

In 2003 it was not easy to obtain books in loan, and the head of SUB Göttingen asked the heads of the other libraries personally, explaining the aim of our project, so that they would give us the books for scanning. Every library reacted individually, some were more collaborative and friendly and some others not. Erlangen decided not to give us the books on loan, but most other libraries finally did. We obtained about 50 books from various different libraries in a relatively easy procedure, not much more bureaucratic than a regular interlibrary loan request.

B) The selection method for the EZOOLO 2 (2008-2011) project:

a) Book selection

Book selection in the second project was generally much more chaotic than in the first project. Selecting the books began generally only after the funds had been granted. Initially we screened Sherborn's Index Animalium, which is a very broad compilation of 1902/1922 of all zoological works that contained original descriptions of animal species and genera, between 1758 and 1850. It contains some 440,000 taxonomic names of species and genera, listing for every name the original source where this name was established. It was relatively easy to decode the abbreviations of the corresponding literature references, and then to look up the titles in the following workflow:

- (1) Consulting BHL if the title is already online (if yes then check for quality, if quality was acceptable then link this title in AnimalBase³³ and go on with the next title, if not then go to step (2).
- (2) Consulting the SUB Göttingen library OPAC catalogue to see if we had the title. If yes then list this title in AnimalBase³³ and enter it in the list to give it to the digitisation department. If no, then go to step (3).
- (3) Looking for the title in the GVK compound catalogue, to see if the Staatsbibliothek zu Berlin - Preussischer Kulturbesitz (SPBK) has it. If yes, then put the title on the list for Berlin. If the title is in GVK but Berlin does not have it, then go on with the next title (and send a scanning request to BHL). If the title is not contained in GBV, then go to step (4).
- (4) Looking for the title in Karlsruhe Virtual Catalogue (KVK) to see if any other catalogue in the world has this book. If yes, go on with the next title (and send a scanning request to BHL). If not, then something must have been wrong with the search request, incorrect title citation, for example, and it is necessary to look more systematically for the item using author, year and components parts of a book title.

We proceeded chronologically with years of publication. So we started at 1770 and screened Sherborn for all books published in 1770; then after that, 1771, then 1772, and so on, year by year. When we arrived at 1813 the proportion of "unsuccessful" searches was so high that it took a very long time to find a book that was (1) not already online at BHL-US online and (2) in our library stock.

Then we screened our rare book holdings list. The workflow was the following:

- (1) Checking the title in the list, and taking a decision if this proved to be an important work for the project or not. This was usually not difficult, but it was necessary to have some experience in the background, for example to know the names of the most important

³³ <http://www.animalbase.uni-goettingen.de/>

zoological authors. If the title did not prove to contain important zoological content, then go on with the next title, if yes then go to step (2).

(2) Consulting BHL-US if the title is already online (if yes then check for quality, if quality was acceptable then link this title in AnimalBase and go on with the next title, if not then put the title on the scanning list).

We did not check the original works prior to digitisation, because these works were in the rare works stacks of the Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB Göttingen) (RARA) department and obtaining them for the loan department means that they have to be transported, bearing a potential danger to damage the books. Also, we would have had to check hundreds of books, for which we did not have sufficient man power. So we decided to digitise them without checking them previously, and to hope that everything went well.

b) Obtaining books from other libraries

When analysing the library holdings with the literature list in Sherborn's Index Animalium, we quickly saw that with ongoing time from the 1780s onwards, the proportion of works held by SUB Göttingen decreased continuously. This had to do with continuously decreasing funds, a process that is still ongoing today (SUB Göttingen is still unsubscribing journals that have been subscribed for many decades).

In 2008 most libraries we asked again (the same libraries we had asked in 2003), had their own scanning machines and for this reason did not like to give us any book (they usually offered to scan the books themselves, against payment and in a different - much lower than our - quality standard). The negotiations were much more complicate, this was a very bureaucratic procedure and the costs needed for the energy and man power to be invested for obtaining such books were much higher than in 2003. We finally decided to do this only with one library namely the Staatsbibliothek zu Berlin - Preussischer Kulturbesitz (SBPK Berlin), which covered about 30-40 % of the books that were not present in SUB Göttingen, and to neglect the others. By this method we lost more than 100 very precious and rare works we intended to scan and which we knew were really important for scientists.

c) Deduplication

Deduplication was very important. For every title we found we consulted BHL (for journals, the BHL Scanlist) to see if the book was digitised in the BHL project. If yes, and if the quality was acceptable, we did not include such a book to our list. We did not consider books as "digitised" if they were digitised by Google Books or other private or commercial institutions, outside the public domain. At the end of the project, about 70-80 % of the books we intended to digitise, were already online at BHL, and we had to develop strategies to find methods to find higher proportions of books that were not online at BHL (because it took much energy/man power to look up all these books at BHL and in most cases find that the works were already online). One strategy was, to focus on non-English and non-German works, and instead to select for example more Italian works. Another strategy was to screen the department of rare book holdings of our library. Our library staff compiled a list of zoological works held in the rare book department, and we went through the titles to estimate if these titles (or authors/years) were promising to be useful. When checking these titles at BHL we found that even if they were English they were rarely online at BHL. Obviously there were restrictions in American libraries and in London to scan the most precious and rare books.

C) Comparison between EZOOLO 1 and EZOOLO 2

a) Book selection

In 2003 the books were selected prior to requesting the funds. In 2008 the books were selected after the funds had been approved.

In 2003 we mainly analysed Linnaeus 1758 and extracted the literature references. In addition we knew from our zoological experience most of the zoologically important works that were published until 1770. We looked these up in FishBase³⁴ and some other resources.

In 2008 we initially worked with Sherborn's Index Animalium, later we developed other strategies for book selection, for example screening our rare books department stock for titles that sounded important (title, author, year).

b) Estimating the work volume

In both projects we estimated the financial volume needed for the scanning before sending our request to the funders.

In 2003 we counted every one of the 400 books page by page, and differentiated bitonal, greyscale and colour pages. We finally offered our funders a complete and detailed list of 400 works to be scanned.

In 2008 we just estimated roughly the average page number of an old work, and simply asked to scan 644.000 pages. We did not give a list, and although the funders expressly requested such a list, we did not see much sense in compiling such a list prior to the actual start of the project. If we had compiled such a list we would have digitised many works for which digitised versions already existed (perhaps 80 % of the works). Due to the removal of this obligation to strictly follow such a list, we were able to remove those works we had already found digitised somewhere else.

c) Obtaining books from other libraries

In 2003 most libraries had not even heard about what book scanning was. So, after our library head had talked with them, they trusted us and we obtained many precious and rare books on loan for the scanning. In 2008 most libraries we asked again (the same libraries we had asked in 2003), had their own scanning machines and for this reason did not like to give us any book (they usually offered to scan the books themselves, against payment and in a different - much lower than our - quality standard). The negotiations were much more complex, this was a very bureaucratic procedure and the costs needed for the energy and man power to be invested for obtaining such books were much higher than in 2003. We finally decided to do this only with one library (SBPK Berlin), which covered about 30-40 % of the books that were not present in SUB Göttingen, and to neglect the others. By this method we lost more than 100 very precious and rare works we intended to scan and which we knew were really important for scientists.

7.1.2 Example from the Ernst Mayr Library, Cambridge, MA-US (BHL-US)

This example shows the initial workflow of the Ernst Mayr Library which contributes content to BHL-US.

³⁴ <http://www.fishbase.org/>

Reports are generated from the Ex Libris Aleph integrated library system³⁵. These reports are used to create ‘picklists’ of items to pull for scanning. Picklists are the lists of items that we are allowed to scan, based on date criteria (before 1923 for US publications and before 1908 for non-US publications). Items that have been scanned or will be scanned by another partner are removed from the picklists. These lists are in shelf-list order and this simplifies the process of pulling the correct items from the shelves. Items on the list are barcoded and cross-checked for suitability for scanning. Items that are too large or too fragile or contain foldouts that are too large are removed from the lists. Fold-outs are identified, measured and flagged because they must be scanned in a separate workflow. Then a final packing list is generated, items checked out of the library and the selected items are shipped to the scanning centre. When the items are scanned and returned, the packing list is checked and items are checked back in. Quality assurance process is detailed below.

We maintain a database as an inventory of all items scanned for the Ernst Mayr Library (MCZ). Items rejected, either by the Library or the scanning centre, are tracked via the integrated library system and other scanning solutions are sought.

It takes a minimum of 2.5 FTE devoted full time to BHL work to keep the project moving at a pace of about 200 books every other week. The Harvard University Library systems staff worked with the Internet Archive to ensure a smooth connection over the Z39.50 protocol. The Z39.50 connection enables the Internet Archive to pull metadata from the Harvard University catalogue to populate the Internet Archive and BHL portal records. We have not yet accomplished the final step of providing access to the BHL digitised material through Harvard’s portal.

The following steps demonstrate the workflow within the Ernst Mayr Library:

- 1) Reports are generated from the Ex Libris Aleph integrated library system.
- 2) From report, create picklist of items that can be pulled for scanning (First pass is by date for us—pre 1923 US published and pre-1908 non-US published)
- 3) Items that have been scanned or will be scanned by another partner are removed from the picklists. (in shelf-list order)
- 4) Pull items from the shelves.
- 5) Barcode and cross-check for suitability for mass scanning.
- 6) Items that are too large or too fragile or contain foldouts that are too large are removed from the lists. These items are tracked in a database and go through a separate workflow.
- 7) Fold-outs are identified, measured and flagged because they must be scanned in a separate workflow.
- 8) Final packing list is generated and printed to be included with shipment.
- 9) The selected items are shipped to the scanning centre.
- 10) When the items are scanned and returned, the packing list is checked.
- 11) Items are checked back in.
- 12) Quality assurance (QA)
- 13) Return items to shelves.
- 14) Add URL to integrated library system.
- 15) Update or correct metadata in BHL portal if necessary. (This includes identifying illustrations, pagination, and isolating articles in serial volumes.)

³⁵ <http://www.exlibrisgroup.com/category/Aleph>

A more detailed current workflow within BHL-US is shown for the creation of a packing list and for the process after the shipment has been returned:

A) Creating Internet Archive Packing List/WonderFetch Spreadsheet

Before starting with the actual steps we will give you a brief background to WonderFetch. WonderFetch is the term used for pre populating the Internet Archives Biblio metadata forms, so named because it is more wonderful than regular Z39.50 fetching. Using WonderFetch, BHL Partner libraries can populate fields with item-level (rather than title level) data that would not normally be populated as part of the standard IA process, and then store those values in the foo_meta.xml file alongside each scanned item in the IA repository. WonderFetch is a procedure developed by the Internet Archive, in conjunction with BHL-US staff, that uses URL parameter/value pairs to automatically fetch the title-level MARC record via Z39.50 while also passing item-level data such as enumeration (vol. 2 issue 14), chronology (April 1903), and IP information (Permission to Digitise Granted by Rights Holder) directly to the Biblio software before scanning.

The steps for creating the Internet Archive Packing List/Wonder Fetch spreadsheet are as followed:

- 1) When a shipment (2 Meyer carts) is ready for transfer to IA, all volumes are checked out to the BHL pseudopatron via Aleph Circulation.
- 2) An Aleph report is generated listing all items checked out. This forms the basis of the packing list, which is created in Excel.
- 3) Items on the packing list spreadsheet are ordered according to the order of the items on the carts (per IA requirement).
- 4) Combined serial volumes and “bound-widths” are divided on the spreadsheet with separate lines created to describe each volume/bibliographic item.
- 5) Final Excel spreadsheet is copied into OpenOffice WonderFetch template.
- 6) Excel packing list is printed and hard copy accompanies shipment to IA.
- 7) WonderFetch spreadsheet is emailed to IA at time of shipment.
- 8) Carts are shipped to IA scanning centre at Boston Public Library via W.B. Meyer.
- 9) A separate spreadsheet listing monographs is extracted from the packing list. This list is added to the monographic de-duping tool

B) Return Shipment

- 1) Return shipments are unpacked and checked against the packing list to ensure that all volumes have been returned.
- 2) Based on number of items per cart, a QA sample is selected and removed from each cart.
- 3) Remainder of items are checked in and reshelfed.
- 4) Quality Assurance is performed on the selected items. Problems requiring rescanning are noted in the internal note field of the Aleph item record. Such items are included with the next shipment for correction. Quality Assurance volumes with no problems are checked in and reshelfed.
- 5) Book IDs (provided by Internet Archive on returned WonderFetch spreadsheet), are added to the original Excel packing list. This list is then imported to a FileMaker database of all BHL items scanned to date

- 6) Items rejected by the scanning centre are noted in the internal note field of the Aleph item records.
- 7) BHL Scan List bids are updated to exclude rejected serial volumes.
- 8) Monographic de-duping tool is updated to exclude rejected monographs.
- 9) Gemini issues are updated to “Complete” status.
- 10) An 830 field with the text: “Museum of Comparative Zoology—Biodiversity Heritage Library digitization project” is added to all bibliographic records.

BHL-US uses the Gemini Project Issue Tracking software to give users the possibility to give feedback and place scanning requests. The workflow for responding to a scanning request by a user is shown below:

C) Processing Gemini Requests (Fig. 17)

- 1) Gemini requests are received via email.
- 2) If item to be scanned is in Special Collections:
 - a) Component is set to “Special Handling” on Gemini issue page.
 - b) Request is recorded on spreadsheet. Such issues will be addressed as time allows and processed in the same manner as main collection items.
- 3) If item to be scanned is in the main collection, item is pulled from shelf and checked to see if it may be scanned.
- 4) Items bound for Widener Imaging Services (oversized, items with large foldouts, etc.) are separated and placed on a cart thus designated.
- 5) For items that can be scanned, the Gemini issue number is recorded in the internal note field of the Aleph item record (e.g., “issue 4376”) for tracking purposes. If requested item consists of several volumes (or all volumes) of a serial, the issue number is recorded in the Internal note field for all requested volumes.
- 6) Response is made to Gemini request indicating whether we can scan the item(s). For serials, volumes we will/won’t be able to scan are listed.
- 7) For serials, a bid is placed upon volumes to be scanned in the BHL Scan List.
- 8) Items to be scanned are placed on carts bound for either Internet Archive or Widener Imaging Services.

Gemini Commercial License registered to CounterSoft Hosted Service Search Logout crinaldo Help

Home Create Issue ▾ Dashboard ▾ Profile

Projects ▾ BHLFEED: BHL Feedback ▾ Create Issue ▾ Issues ▾ Email Alerts ▾

<< Issue 68 of 802 Issue(s) >> <http://biodiversitylibrary.countersoft.net/Default.aspx?p=7&i=4645>

BHLFEED-4645 - Archiv fur Physiologie Created: 11/1/2010 4:24:59 PM Revised: 11/9/2010 5:28:15 PM

<p>Type: Task</p> <p>Priority: Minor</p> <p>Severity: Trivial</p> <p>Status: Assigned</p> <p>Resolution: Unresolved</p> <p>Assigned To: Matt Person, Joe deVeer, Katie Mullen</p> <p>Reported By: Grace Costantino</p> <p>Component: Collections, Special Handling</p> <p>Fixed In Version: None</p> <p>Versions Affected: None</p> <p>Risk Level: No Risk</p> <p>Votes: 0</p> <p>Visibility: Everyone</p> <p>Start Date: ?</p> <p>Due Date: ?</p> <p>Percent Complete: </p>	<p>Description</p> <p>Title from priority list. Has preceding and succeeding titles.</p> <ol style="list-style-type: none"> Archiv für Anatomie, Physiologie und wissenschaftliche Medicin <ul style="list-style-type: none"> OCLC 1481913 in bhl full bid by NHM, but have scanned through 1860, which is actually their cut-off date Missing 1835, 1850, 1857-58, 1861-76 Archiv für Physiologie <ul style="list-style-type: none"> OCLC 6028347 no in BHL no bids covers 1877-1919, 43 volumes Pflügers Archiv für die gesamte Physiologie des Menschen und der Tiere <ul style="list-style-type: none"> OCLC 1640487 not in BHL no bids covers 1910-1968 (bd. 132-301)
--	--

Comments (10) Attachments (0) Issue Watchers (3) Sub Issues (0) Issue Links (0) Time (0) History

[Assign Issue to me](#)
[Edit \(+\)](#)
[Comment \(+\)](#)
[Move](#)
[Copy](#)
[Vote](#)
[Watch this issue](#)
[Update Issue Description](#)
[Print](#)

Comment (+)

[Expand All](#) [Collapse All](#)

Katie Mullen
11/9/2010 5:28:15 PM [\[Edit\]](#)

For Title:

Archiv für Anatomie, Physiologie und wissenschaftliche Medicin.
Leipzig [etc.] Veit & Comp. [etc.]
42 v. : ill. ; 23 cm.
HOLLIS: 000128237

1835 and 1850 are in our Special Collections. I've added them to the list and will advise as to scanability as soon as possible.

1863 and 1865 have obscured text.
1867 appears to have missing pages.
1870 and 1875 have narrow margins.
1874 has missing text.

Otherwise we can scan requested volumes.
We will send 1873 to Widener for scanning due to its condition.

Grace Costantino
11/9/2010 4:03:04 PM [\[Edit\]](#)

Thanks for the clarification. Katie I have updated what I asked MRI. In scan based on what you just said

Fig. 17: Processing Gemini Requests (screenshot)

For more information about the workflow one can have a look at: http://si-pddr.si.edu/dspace/bitstream/10088/7182/1/Gwinn_Rinaldo_IFLA_2009.pdf

Partners of BHL-Europe have further access to:
<https://bhl.wikispaces.com/Workflow+documents>³⁶

³⁶ Non-BHL partners will gain access to these sites when they become a content provider of the project.

BHL and Quality Assurance (QA) (extracted from BHL Cookbook):

<https://bhl.wikispaces.com/BHL+Cookbook>³⁶

GOAL: To assure that the digital manifestations of the items scanned and added to the Biodiversity Heritage Library match the quality of their physical counterparts.

Steps:

In order to ensure that scans of those items that are sent for scanning by the Biodiversity Heritage Library match the quality of the physical items that are sent, BHL staff developed quality assurance policy and procedures. This involves QA'ing 100% of a sample of books from a given shipment cart.

The QA process begins once a cart of books is returned from scanning. Upon receipt of book trucks, the invoice sent along with the returned cart is matched against the cart itself. If items are missing, the scanning centre is contacted and provided with a list of missing (or excess) items. Staff then document missing/excess items (documentation of missing or excess items is maintained by each individual institution - there is no BHL-wide documentation of this). All invoices/manifests of items sent and returned are stored by each individual institution for future audit purposes.

QA is done on a *statistical sampling* of books in the returned shipment, using either the PDF or flip book on the IA site. The number of books chosen to QA is based on the number of books in the returned cart. BHL staff use the following chart that details how many books in a given shipment should be QA'd (Tab. 10).

Items in set ('bin')	9-15	16-25	26-50	51-90	91-150	151-280	281-500
Number of items to QA	3	5	8	13	20	32	50
# of errors to Fail the set	1	1	1	2	3	4	5
# of major errors = acceptable	0	0	0	1	1	2	3
# of minor errors = fail	1	1	2	2	3	3	4
# of minor errors = acceptable	0	0	1	1	2	2	3

Tab. 10: Chart how many books in a given shipment should be QA'd

This sampling procedure is also used by IA during their QA process. There is no documented method for choosing which books are part of the sample. Staff often attempt to identify and prioritise for QA those books on the returned cart that look like they may be problematic (foreign language, bound-with monograph, etc.).

In general, a shipment will fail QA if more than 2% error rate is found. For QA purposes, BHL staff differentiates between "major" and "minor" fails of items QA'd (See <https://bhl.wikispaces.com/QA+Policy>³⁷). If a shipment has a 2% error rate, the entire shipment (minus any books that passed QA) is sent back to the scanning centre, after which time the scanning centre can decide to either QA the entire cart or rescan the entire cart.

³⁷ Non-BHL partners will gain access to these sites when they become a content provider of the project.

The procedures for checking a physical item in-hand against the digital manifestation of the item on IA involves both a metadata check and a scan images check.

For metadata checks, staff makes sure that the metadata for the item on IA correctly reflects the item in hand. If the metadata is attached to the wrong item, staff notify the scanning centre, which should in turn be able to fix the problem. If QA is being done on material that was scanned more than a month ago, metadata should be corrected by hand in the portal via the editing tool.

For scan image checks, staff simply check each page of the physical book while clicking along with the flip-book or PDF on IA (i.e. turn a page, click the scan; repeat for entire volume). Staff are particularly looking for missing pages, unreadable text, text that has been cut off, the scanner's hand turning the page in the shot, anything that the camera might have caught that obscures the text, etc. If any problems such as these are discovered, the books are put aside and sent back to IA on the next shipment for correction. If the problem involves only a few pages, IA will simply rescan the pages in question and insert them in place of the offending pages in the scan file. If the problem is with the entire scan file (all pages are too light/too blurry), the entire book is rescanned. The corrected books are then returned to the BHL partner institution.

Results:

By performing consistent QA on returned carts, BHL staff have seen a decrease in the number of QA errors. The policy of returning entire carts with more than a 2% error rate prompted our scanning partner to perform extensive QA on the materials as they're scanned - a policy they were lacking. As a result, many of the errors are found before they are returned back to BHL partners and the number of returned carts for QA fails has significantly decreased.

Lessons learned: When performing QA, it is important to have the physical item in-hand. Page numbers often lie or hide plates within otherwise numerically correct spans.

For any pages that seem too light, off-colour, etc. download and check those pages in the PDF. Rarely, staff have found that pages missing (or seeming too light to read on the flip-book) will, in fact, be readable and present in the PDF. Furthermore, ensure that the OCR'd text in the PDF for suspect pages is there and more or less correct (it is OCR...).

Bottom line: if there seems to be something wrong with the flip-book double check the PDF against the book in hand page by page.

It is important to have a methodology in place that allows for certain QA corrections to be made without having to rescan the entire item. Often, the problem with a book involves only a few pages (a few pages missing, text obscured on a few select pages, etc.). Rather than having to rescan an entire book for only a few pages, develop a procedure that allows a few pages to be rescanned and inserted into the existing scan file in place of the "bad" page scans.

It is important to rely on user feedback for QA. It is impossible for staff to find all QA problems with all materials in BHL, but users will find these problems. However, it is important that digital libraries have a means of dealing with QA problems that may be discovered by staff or users long after the item has been scanned. This involves working out an agreement with scanning partners that allows books for which errors are found long after scanning is complete to be sent back for correction.

Documentation & further references are available for BHL-Europe partners on the wiki³⁸:

QA in general: <https://bhl.wikispaces.com/Cookbook+Quality>)

QA procedures: <https://bhl.wikispaces.com/QA+Procedures>

QA policies: <https://bhl.wikispaces.com/QA+Policy>

NISO chart: <https://bhl.wikispaces.com/QA+sampling+chart>

7.1.3 Example from the Natural History Museum Library, London, UK

The Natural History Museum, London (NHML) is a partner of BHL-Europe and BHL-US and has an Internet Archive Scribe operating in their institution. Therefore the NHM follows the BHL-US and Internet Archive (IA) protocols and workflow for scanning serials and delivering associated metadata. The workflow applied at the NHM Library is similar to the version described from the Ernst Mayr Library Cambridge. The equivalent of one full time staff member is assigned to the pre and post Internet Archive processes. All items are scanned on site.

Simplified overview of the NHML workflow is below

- 1) The NHM Library creates “Picklists” for selected titles from reports generated from the Library Catalogue. The NHM Library works with the SirsiDynix Library Management System Unicorn (from June 2011 this will be upgraded to Symphony).
- 2) The items are selected for scanning based on the following criteria:
 - Subject content or response to direct request from Library or BHL user
 - Unless filling gaps identified via the Gemini system, serial titles where most of the proposed date range can be scanned are selected.
 - No existing scanned copies or “bids” to scan by other institutions
 - Date of publication (before 1860, post 1860 following receipt of Copyright permissions or completion of due diligence checks). Copyright records are entered.
 - Suitability and condition of item to meet IA criteria and NHML conservation criteria , for example frailty, size (no items larger than 12.5 X 10.5X 4 inches, bindings, Pages must open at 75 degrees with ¼ inch gutter to withstand scanning process
 - Foldouts, which need to be tagged separately for separate and simultaneous workflow
- 3) A packing lists is created for out of Copyright or Copyright “cleared titles” only
- 4) The final excel spreadsheet is copied into OpenOffice Wonderfetch template
- 5) Completed packing lists are printed, items are retrieved from the shelves and copy of packing slips inserted
- 6) Trolleys of retrieved titles are delivered to the Internet Archive unit with packing lists and
- 7) Wonderfetch copy available on password protected area of NHM Network
- 8) delivery and temporary location of items is recorded
- 9) Monitor status of packing list. When “Scanned by Internet Archive”, collect trolley and check all items accounted for as per packing list.
- 10) Quality Assurance checks and return for rescan or reshelve.
- 11) URL and catalogue amendments are added to Library Catalogue

³⁸ Non-BHL partners will gain access to these sites when they become a content provider of the project.

12) Metadata updates or corrections added to the BHL portal as required.

The workflow for scanning monographs is currently different and NHML is working towards bringing its workflow in line with the workflow applied by the other BHL institutions. NHML started scanning monographs in 2010 via the Internet Archive and are following the same approach as MBLWHOI Library at Woods Hole. A spreadsheet with required metadata fields is populated manually from reports for the selected titles generated from the Library Management System Unicorn. IA manually creates a matching IA record for each title scanned.

The NHM also operates a Special Digitisation Unit to scan artwork, large volumes and the rare, delicate, more difficult to scan material not appropriate for the IA process. Monographs and serials are scanned for BHL through the Special Digitisation Unit, and the images and metadata are sent on a hard drive to MBLWHOI for processing and uploading to BHL. The Specialist Digitisation Unit follows BHL standards and protocols.

7.2 Workflow and Tools

7.2.1 Individual steps in the digitisation process

This chapter provides an overview of software tools and solutions that may be used as automated workflow systems to control and document each step in a digitisation process.

The workflow starting with the printed book and ending with the provision of a digital copy within a digital library contains a variety of processing steps. Different organisations have different demands of these steps regarding the quality of the final product in terms of resolution or metadata on the one hand and the software tool for the realisation of the different workflow steps on the other hand.

Difficulties arise when a high amount of data is produced and when differently skilled staff are involved in the workflow. Therefore, there is an urgent need for a system which allows for the scheduling of processes and resources and documents the work done step by step.

This challenge is supported by workflow systems such as Goobi^{39,40} and DAF⁴¹.

The following paragraphs show different tools, organised per category according to the steps in the digitisation process.

A) Bibliographic recording or bibliographic import

B) Scanning

Name of tool: Scan-n-Stitch Deluxe

Description: Scan-n-Stitch Deluxe allows users to digitise their large-scale objects, such as fonts, pages from scrapbooks, drawings, magazines and magazine pages.⁴²

³⁹ <http://www.carpet-project.net/en/catalogue/carpet/production-level-goobi-1/>

⁴⁰ <http://www.carpet-project.net/en/catalogue/carpet/presentation-level-goobi-1/>

⁴¹ Digital Assets Factory by Bibliotheca Alexandrina <http://wiki.bibalex.org/DAFWiki/>

⁴² http://www.arcsoft.com/de-de/software_title.asp?productCode=SNSD

Language: English/ German/ French/ Italian/ Spanish/ Russian

Price: US \$ 39.99 or 39.99 €

Conditions:

- Windows XP, Vista, 7
- Pentium III 500 MHz or higher
- 256MB
- 16-Bit or higher colour display with an resolution from 800x600
- A4 Flatbed-Scanner

Name of tool: ACDSSee Photo Manager 12

Description: ACDSSee Photo Manager 12 allows users to digitise objects quickly and intuitively with easy-to-use editing tools.

Language: English/ German/ French

Price: US \$ 69.99

Conditions:

- Windows XP, Vista, 7
- Pentium III 500 MHz or higher
- 512 MB
- 16-Bit or higher colour display; recommended resolution 1280x1024
- A4 flatbed-Scanner

C) Quality control

After digitisation the scans have to be processed before they can be published. There are many ways to optimise the raw data from the scanning process. The image files can be edited manually or automatically and OCR⁴³ software makes text in an image file machine readable and therefore enables searching within the full text of a scanned page.

➤ *Image optimisation*

Digital images can be optimised in many ways e.g. with the scanner software itself or with image editing tools. The decision whether or not optimisation methods are used depends on the particular scan.

Name of tool: GIMP

Description: GIMP is an acronym for GNU Image Manipulation Program⁴⁴. The focus of the software is the intensive processing of individual images, for which a variety of effects ("filters") are available.

Documentation: <http://www.gimp.org/docs/>

Download-link: <http://www.gimp.org/downloads/>

Licence: GNU General Public License⁴⁵ (GPL)

Operating systems: GNU/Linux (i386, PPC), Microsoft Windows (XP, Vista), Mac OS X, Sun OpenSolaris, FreeBSD

Features and Capabilities: see here: <http://www.gimp.org/about/introduction.html>

⁴³ Optical Character Recognition

⁴⁴ <http://www.gimp.org/about/introduction.html>

⁴⁵ <http://www.gnu.de/documents/gpl-3.0.en.html>

Name of tool: IrfanView (for Batch-conversion)

Description: IrfanView is a fast, small and compact graphic viewer for Windows 9x, XP 2008, Vista, Windows 7.⁴⁶

Language: English/ German/ French/ Italian/ Spanish/ Russian and more

Licence: free for non-commercial use (means at home)

Software-Agreement under <http://www.irfanview.com/eula.htm>

Some features:

- Many supported file formats e.g. TIF, JPG, BMP, PNG and others⁴⁷
- Thumbnail/preview option
- Paint option - to draw lines, circles, arrows, straighten image etc.
- Support for Adobe Photoshop Filters
- Fast directory view (moving through directory)
- Batch conversion (with image processing)
- Multipage TIF editing
- Support for embedded colour profiles in JPG/TIF
- Change colour depth
- Scan (batch scan) support
- Cut/crop
- Effects (Sharpen, Blur, Adobe 8BF, Filter Factory, Filters Unlimited, etc.)
- Extract icons from EXE/DLL/ICLs
- Lossless JPG rotation
- Unicode support
- Many command line options⁴⁸

Download-link: <http://www.software.com/irfanview-9337-1>

Producer/author/developer: Irfan Skiljan

➤ *Full text digitisation*

Name of tool: ABBYY Finereader 10.0

Description: Intuitive text recognition with good OCR engine, can be used to convert printed documents, digital photographs and PDFs into editable and searchable formats, according to the manufacturer, with maximum accuracy and productivity

Language: German English

Operating system: Windows XP, 2003, Vista, 7

Download-size: 50,4 MByte

Licence: commercial (demo-version available)

Price: 139 € (Professional, Download), 999 € (Corporate)

Name of tool: gocr

Description: Font-independent character recognition (omnifont-OCR omnifont) as a command line version, includes a bar code recognition⁴⁹

Language: English

Operating system: Windows XP, 7, Linux

Conditions: 133 MHz CPU, 4 MByte RAM, 4 MByte HD

⁴⁶ <http://www.irfanview.net/>

⁴⁷ http://www.irfanview.net/main_formats.htm

⁴⁸ <http://www.irfanview.net/>

⁴⁹ <http://www.heise.de/software/download/gocr/7147>

Download-Link: <http://jocr.sourceforge.net/download.html>

Licence: Open Source

Price: for free

Restrictions: no

Producer/author/developer: Joerg Schulenburg

D) Metadata recording

There are different types of metadata:

- Bibliographic metadata, also called descriptive metadata – is the information used to search and locate an object such as title, author, subjects, keywords or publisher.
- Structural metadata – gives a description of how the components of an object are organised for instance how pages are ordered to form chapters.
- Administrative metadata – provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.⁵⁰ Two sub-types of administrative metadata are rights management metadata and preservation metadata.

It is very important to record descriptive, structural and administrative metadata (for further information please see chapter 4.6, as these types of metadata make your document searchable and understandable by a system.

➤ *Bibliographic metadata*

Name of tool: DSpace

Description: DSpace is open source software for operating a document server. It provides tools for capturing, storing and dissemination of digital resources available and is used mostly by universities, libraries and research institutions for the "Institutional Repository".

Licence: BSD⁵¹-License

Download-Link: <http://www.dspace.org/latest-release>

Documentation: <http://www.dspace.org/training-summary/training-materials.html>

Operating system(s): Linux / Unix

Use of external software systems (libraries) or services:

- Unix operating system (Tested on HP-UX, AIX, Solaris)
- Java 1.4 (Standard SDK is sufficient)
- Apache Ant 1.6.2 or higher version
- Database: PostgreSQL 7.3 (or higher), or Oracle 9
- Servlet Container: Jakarta Tomcat 4.x/5.x or similar e.g. Jetty or Caucho Resin

Hosting-Services for DSpace:

Service provider:

- OpenRepositories.org (info@openrepository.com)

More information: <http://openrepository.com/products/open-repository-dspace>

⁵⁰ <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

⁵¹ Berkeley Software Distribution

- @mire (info@atmire.com) More information:
<http://atmire.com/consultancy.php>
 Source: Repositorienhosting
http://www.carpet-project.net/fileadmin/user_upload/documents/2010-11-26-Repositorienhosting) by Anja Oberländer, Stand: 26.11.2010)

➤ *Structural metadata*

Name of tool: Phaidra

Description: PHAIDRA is the acronym for Permanent Hosting, Archiving and Indexing of Digital Resources and Assets⁵², which represents a comprehensive university digital asset management system with long-term archiving functions⁵³. The persistent identification allows the exact location and retrieval of prepared digital objects.

Phaidra can currently only be used by all employees and students of the University of Vienna (or mailbox-u: net-account). The objects can be seen worldwide.⁵⁴

Licence: still unclear

Contact: Project Management, Dr. Paolo Budroni;

paolo.budroni@univie.ac.at

T: +43-1-4277-151 71

Name of tool: METS-Builder

Description: METSbuilder is a digitisation tool. It extracts technical metadata from master image files and produces METS metadata packages for the item, containing structural, bibliographic (Dublin Core), technical image (MIX) and preservation (PREMIS) metadata.⁵⁵

Download-Link: <http://sourceforge.net/projects/metsbuilder/>

Other possibility: One can build a simple METS editor by using an XML editor (e.g. XML Spy or similar) together with the METS schema⁵⁶ and the metadata format used (e.g. MARC 21⁵⁷ or MODS⁵⁸).

E) Validation

In order to add technical metadata (from e.g. file format information) automatically to an existing metadata structure, or to provide manually generated metadata, a validation tool may be used.

Name of tool: JHOVE

Description: JHOVE (JSTOR/Harvard Object Validation Environment) is an extensible framework for format validation. JHOVE is a format-specific digital object validation API written in Java.⁵⁹ It provides functions to perform format-specific

⁵² <https://phaidra.univie.ac.at/>

⁵³ <https://phaidra.univie.ac.at/>

⁵⁴ <https://fedora.phaidra.univie.ac.at/fedora/get/o:29162/bdef:Content/get>

⁵⁵ <http://sourceforge.net/projects/metsbuilder/>

⁵⁶ <http://www.loc.gov/standards/mets/version18/mets.xsd>

⁵⁷ <http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd>

⁵⁸ <http://www.loc.gov/standards/mods/mods.xsd>

⁵⁹ <http://en.wikipedia.org/wiki/JHOVE>

identification, validation, and characterisation of digital objects.⁶⁰ Format identification is the process of determining the format to which a digital object conforms; in other words, it answers the question: "I have a digital object; what format is it?" Format validation is the process of determining the level of compliance of a digital object to the specification for its purported format, e.g.: "I have an object purportedly of format F; is it?" Format characterisation is the process of determining the format-specific significant properties of an object of a given format, e.g.: "I have an object of format F; what are its salient properties?"

Download-Link: <http://sourceforge.net/projects/jhove/>

Documentation: <http://hul.harvard.edu/jhove/documentation.html>

Licence: GNU Lesser General Public License (LGPL)

Supported language: English

Operating system(s): Linux / Unix, Windows 2000, Windows 7, Windows 95/98, Windows Server 2003, Windows Server 2008, Windows VISTA, Windows XP, MAC OS

Requirements: For the use of JHOVE a Java Virtual Machine⁶¹ is required.

F) Archiving (Repository software)

Name of tool: Fedora repository 3.x (METS Compatible)

Description: Fedora is the acronym for Flexible Extensible Digital Object Repository Architecture. It is a modular repository system for the management and sharing of digital contents. It is especially suitable for digital libraries and archives to provide access to and preservation of data.⁶²

It is also used to deliver subject-related access to very large and complex digital collections of historical and cultural material and scientific data.

Fedoras flexibility allows the integration of its functionalities in a range of institutional- and web based systems and affords scalability (e.g. millions of objects) as well as longevity (e.g. all information is stored in datasets without software dependence, so that the complete data can be rebuilt at any time).⁶³

It also offers the possibility of relations between the digital resources as sets represent and query the repository using the Semantic Web SPARQL⁶⁴ language.

Fedora is used globally by academic and heritage organisations, universities, research institutions, university libraries, national libraries and government agencies.

Key features:

- Stores all types of content and its metadata
 - Digital content of any type can be managed and maintained
 - Metadata about content in any format can be managed and maintained
- Scale to millions of objects
- Access data via Web APIs (REST⁶⁵/SOAP⁶⁶)

⁶⁰ <http://www.carpet-project.net/en/catalogue/carpet/JHOVE/>

⁶¹ http://www.java.com/en/download/inc/windows_upgrade_xpi.jsp

⁶² <http://www.carpet-project.net/en/catalogue/carpet/Fedora%20Commons/>

⁶³ <http://www.carpet-project.net/en/catalogue/carpet/Fedora%20Commons/>

⁶⁴ Is the recursive acronym for SPARQL Protocol and RDF Query Language

⁶⁵ Representational State Transfer

⁶⁶ Simple Object Access Protocol, is a simple XML-based protocol to let applications exchange information over HTTP, <http://www.w3schools.com/soap/default.asp>

- Providers RDF search (SPARQL)
- Rebuilder Utility (for disaster recovery and data migration)
- The entire repository can be rebuilt from the digital object and content files.
- Content Model Architecture (define "types" of objects by their content)
- Many storage options (database and file systems)
- JMS messaging (your applications can "listen" to repository events)
- Web-based Administrator GUI (low-level object editing)
- OAI-PMH Provider Service
- GSearch (fulltext) Search Service
- Multiple, customer driven front-ends⁶⁷

Licence: Apache Licence (2.0)

Documentation:

<https://wiki.duraspace.org/display/FCR30/Getting+Started+with+Fedora>

Use of external software systems (libraries) or services:

- Java SE Development Kit (JDK) 6⁶⁸
- Tomcat 5⁶⁹.x or higher
- Maven2⁷⁰

7.2.2 Workflow management software Goobi

This chapter gives a general overview of the scanning workflow management software Goobi⁷¹.

What is Goobi?

Goobi is an open-source software application for digitisation projects that enables organisations to model, manage and supervise freely configurable production processes.

All working steps, that are necessary to create a digital library, are represented by Goobi.

These include importing data from library catalogues, scanning and content-based indexing and the digital presentation and delivery of results in popular standardised formats.⁷²

It is a web-based application and is designed for simultaneous operation by multiple users. It is available in several languages (e.g. English, German) and is therefore ideal for international digitisation projects.

History of Goobi

In 2004 the development work on Goobi began in libraries. The main focus in the library environment concentrated on large-scale digitisation. From 2004 – 2008 Goobi's development was led by the Göttingen State and University Library (SUB). During this time Goobi's development took place within the framework of international digitisation projects

⁶⁷ <http://www.fedora-commons.org/about/features>

⁶⁸ <http://download.oracle.com/javase/index.html>

⁶⁹ <http://tomcat.apache.org/>

⁷⁰ <http://maven.apache.org/>

⁷¹ <http://www.digiverso.com/en/products/goobi>

⁷² <http://www.digiverso.com/en/products/goobi>

sponsored by the DFG (German Science Foundation). In 2008 intranda GmbH took over the task of developing the software further under the GPL open-source licence.⁷³

Who can use Goobi?

Goobi is the most widely used system for digitisation workflows in Germany. Besides museums and archives Goobi is mainly applied in university libraries. The application is used for digitisation projects of various sizes – from diminutive projects in smaller institutions to mass digitisation in the largest German university libraries. Since about 2008, interest in Goobi outside Germany has been growing.

Because of the flexibility and multilingualism of the entire application, Goobi is now in productive use for coordinating digitisation project workflows at several institutions outside Germany, e.g. in Austria, Great Britain and Spain.

Goobi is open to integrate different software tools into its workflow component.

There are already several user groups, which have their own Goobi installation. The current users demonstrate that the needs and requirements of Goobi are different for each institution.

Goobi users can be roughly categorised as follows:

- Libraries, archives and museums using Goobi for their in-house digitisation projects
- Scan service providers using Goobi for the coordination of their internal work
- Digitisation projects of museums
- Scan service providers for the coordination of their work internally
- Scan service providers using the Goobi installation of their client (e. g. library) in order to manage and coordinate the outsourced scanning work within the clients' Goobi digitisation workflow.
- A shared Goobi installation that allows external partners to deliver information to the collaborating institutions.

Costs for Goobi (Installation, Support)

The costs for installation vary according to the needs of institutions. The following factors can have an impact on the costs:

- If the OCR-Service is used, number of pages (fracture, antigua)
- Possible procurement of necessary hardware
- Desired type and volume of support
- Use of additional modules e.g. universal catalogue interface for Z39.50
- Travel costs for on-site training, client is in Germany or in any other country
- Modules that need to be installed according to the users needs ,e.g. Goobi core, presentation module can be used independently or combined

Technical requirements

- Developed using JavaServer Faces (JSF) technologies
- Supports all current Java application servers (e.g. Apache Tomcat, GlassFish)
- Supports all current SQL Standard-compliant databases (e.g. MySQL)
- Supports index services (e.g. LDAP, Active Directory) based on x. 500 protocol
- Operating system-independent (can run, for example, on UNIX systems such as SUN Solaris)

⁷³ <http://www.digiverso.com/en/products/goobi>

- BSD systems, Mac OSX, different Linux distributions, Microsoft Windows)
- File system-independent using SMB protocol
- Flexible import from different OPACs (e.g. via PICA+, MARC21, MAB2, Dublin Core, MODS)
- Flexible data export (e.g. as METS/MODS, METS/MARC, Dublin Core, OAI, MARC)
- Supports different image formats (e.g. TIF, JPEG, JPEG 2000, PNG)
- ABBYY Finereader support for OCR link
- Search using Apache Lucene or database index
- Editor to capture pagination details, structural data and metadata at whatever level required
- Integration and management of wide range of scanners
- Individual and automatic tasks can be described using the script languages of choice.

Evaluation of Goobi - positive and negative aspects

Positive aspects:

Goobi is highly flexible and generic according to the various workflow requirements

- of institutions,
- of several digitisation projects of the same institution or
- of common cooperative projects where external partners have access to the same data.

In most cases specific technical requirements have to be considered such as LDAP or Active-Directory-arrays, specific storage conditions, different backup and archiving strategies, different optimisation strategies of digitised objects, requirements for quality control or content validation, different types of objects with appropriate structure information and metadata models.

Thus the Goobi software operates with different requirements and covers several methods of integrating external components. It has already enabled mapping and successful digitisation of large-scale projects.

Validation of each stage of the entire workflow is automatically implemented in Goobi. In cases where the result of a working stage has not been validated, Goobi refuses to complete the work step. With this functionality it is possible to control and thus avoid human errors to ensure a high quality at each stage of the process.

The development of Goobi was initially built on the daily digitisation practice of one of the largest German libraries (SUB-Göttingen). Based on this practically oriented design the system is certainly the most efficient and appropriate digitisation software. Its usage is widespread in several large European research libraries which guarantee sustainability of the software and continuity in development. Furthermore Goobi is an Open Source Software licensed under GPL. The consistent use of international standards such as METS/MODS for metadata and structural information storage as well as of software components such as Java, My SQL and Linux assures the continuing readability and software development.

Negative aspects:

A negative issue to be mentioned is the amount of set-up possibilities such as metadata regulations, catalogue interfaces, validation of regular expressions etc. which requires

professional support. On the other hand this applies to most of the library systems and ensures the high flexibility of the software. As well as the high complexity of the systems configuration, documentation about using Goobi has been insufficient. Currently documentation for the various Goobi modules only exists in part, and is only available in German⁷⁴. It is planned to continue the Goobi documentation collaboratively on a multilingual basis by means of the upcoming cooperation with the digitisation platforms digiverso⁷⁵ and the CARPET platform. Accordingly CARPET will provide a growing KnowledgeBase⁷⁶ and a discussion forum⁷⁷ co-ordinated by intranda GmbH. Support will be available both in English and German.

Evaluation of Goobi with regard to BHL-Europe

Goobi has already implemented some digitisation projects in an international context therefore it can be used for a project with many international institutions, such as BHL-Europe. BHL-Europe content providers often outsource the actual scanning process, but all other stages of the digitisation workflow are carried out in-house. Goobi allows the involvement of external service providers, such as scan service providers, within the workflow and enables an unrestricted transfer of the data managed within Goobi to external systems.

An automatic delivery of digitised items along with the structure and metadata to the central BHL- Europe presentation system is therefore applicable to all partners without problems and guarantees compliance to conventions and standards. In addition, a delivery of data to different presentation systems is also possible, for those partners who provide content to other projects as well as BHL-Europe.

A workflow within Goobi can be organised in the following way:

- 1) Bibliographic Recording (e.g. Transfer from a central BHL-system)
- 2) Scanning
- 3) Quality control
- 4) Image optimisation (automatic or manual)
- 5) OCR
- 6) Delivery from structure- and metadata, pagination
- 7) Generation of unique identifiers (e.g. URN⁷⁸ granular, Handle, etc.)
- 8) Content delivery to the BHL-Europe presentation system
- 9) Content delivery to a separate presentation system (e.g. intranda viewer⁷⁹), therefore also OAI etc.
- 10) Archiving

The workflow might be different for each partner and needs to be organised according to the needs of each institution. Therefore the workflow shown above is an example of one use of Goobi and needs to be adjusted individually for each institution.

⁷⁴ <http://goobi-wiki.slub-dresden.de/index.php/Hauptseite>

⁷⁵ <http://www.digiverso.com>

⁷⁶ <http://www.carpet-project.net/en/knowledge-base/>

⁷⁷ <http://www.carpet-project.net/forum/>

⁷⁸ Uniform Resource Name

⁷⁹ <http://www.digiverso.com/en/products/viewer>

Is Goobi capable of doing split processes?

In Goobi workflows can be made in individual steps and can be set up to meet the specific needs of certain partners and staff skills. Thereby a distributed, decentralised operation is possible.

Goobi allows for the integration of external service providers within the digitisation workflow of the institution. This functionality allows for the coordination of the workflow of the external service provider as well as verifying their results as a process within the institutional workflow.

7.2.3 Workflow management software DAF

This chapter gives a rough overview of the workflow management software DAF.

What is DAF?

The Digital Assets Factory (DAF), developed by Bibliotheca Alexandrina (BA), provides a configurable and flexible management tool for any digitisation workflow, integrating with the current tools used for digitisation. A digitisation workflow is defined as a set of phases (Fig. 18), e.g. scanning, processing, quality assurance, encoding ...etc. Administrators can define the sequence of phases required for a digitisation in addition to adding pre-phase and post-phase checks making sure that the process adheres to the digitisation standards in the institution. DAF checks that the correct file types, number of files and naming conventions are in place before and after the current phase. It can even manage several types of workflows for different object types. DAF can integrate automated phases thus lighten the human operators to do tasks in which humans are good: e.g. OCR correction.

It is very important to associate the object with its metadata as early as possible. Therefore, DAF integrates external sources of metadata - e.g. an external ILS, repository or a database - while digital objects are ingested. This is made possible through the development of plugins.

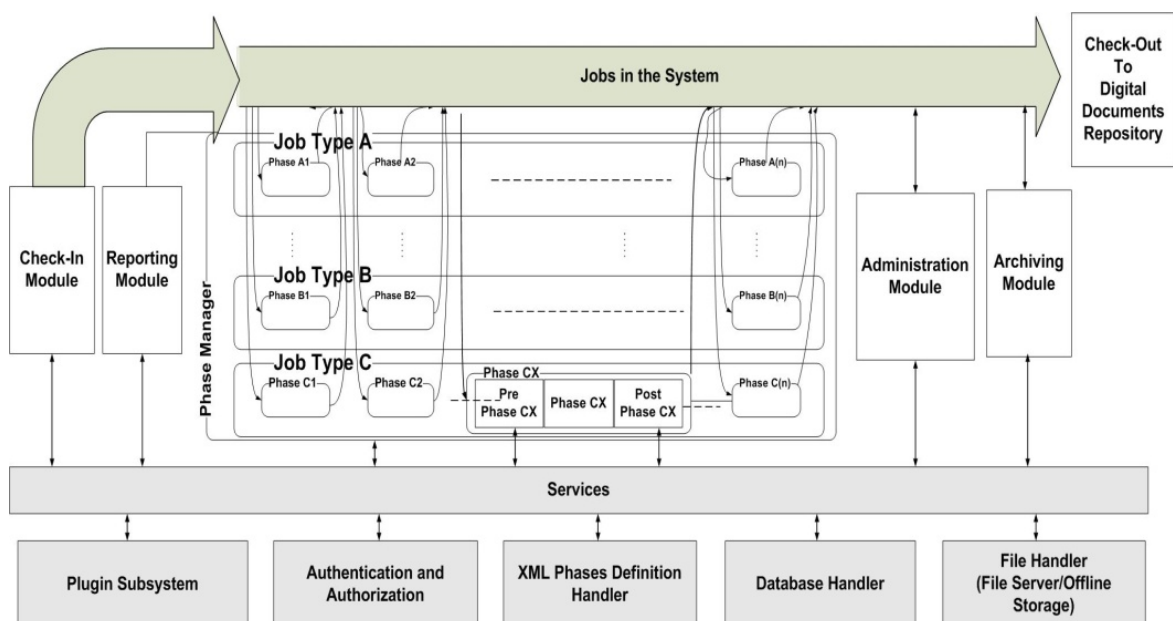


Fig. 18: DAF Architecture

Fig. 18 demonstrates the system architecture of DAF: A *Check-in Module* adds jobs to the system. Operators will be able to work on these jobs through the various phases of the

workflow managed by the *Phase Manager*. Each phase has a pre-phase and a post-phase check to verify the output. After the jobs complete the workflow, they are ingested into the repository through the *Check-out Module*, and an archival copy is sent automatically to the archive through the *Archiving Module*. If a job needs to be retrieved from the archive, DAF is used to fetch it and reload it into the system to perform the necessary actions upon it. DAF allows operators to request a re-do for a previous phase if they discover any problem. The system allows operators to exchange messages indicating problems they face for a particular job so that other operators along the flow would take the necessary actions accordingly. The digitisation supervisor checks the request, and can direct the job to another phase accordingly providing a means to correct the job as soon as the problem is detected.

A typical workflow for digitising a physical book within DAF would be organised as follows:

- 1) Cover scanning
- 2) Scanning
- 3) Processing of scanned images
- 4) Quality assurance
- 5) OCR
- 6) Encoding, e.g. into PDF, DjVu
- 7) Quality Assurance
- 8) Archiving intermediate files
- 9) Ingest into a repository

The quality of work and the performance of the digitisation operators or automated phases can be monitored through the *Reporting Module*. DAF provides timely reports to various levels of management describing the workflow on a daily, weekly or longer basis allowing online queries about the current status of a certain asset during the digitisation workflow. Reports display the number of jobs assigned to different phases of the workflow indicating the pending, running and finished tasks.

DAF is currently being used at Bibliotheca Alexandrina to manage the digitisation process at the library. It is also used in several large-scale digitisation initiatives around Egypt, e.g. the Egyptian Public Record Office currently relies on DAF for mass digitisation of about 2 billion documents scheduled to be done on stages.

BA provides DAF to the community as an open source tool (<http://wiki.bibalex.org/DAFWiki>).

Sources:

<http://www.carpet-project.net/en/>
<http://www.minervaeurope.org/>
<http://www.digitalisierung.ethz.ch/>
<http://www.digiverso.com/en/>
<https://phaidra.univie.ac.at/>
<http://fedora-commons.org/>
<http://www.arcsoft.com/en-us/>
<http://www.irfanview.com/>
<http://www.gimp.org/>

7.3 General archivation – permanent storage

Digital preservation is the active management of digital information over time to ensure its accessibility. Preservation of digital information is widely considered to require more constant and ongoing attention than preservation of other media. This constant input of effort, time, and money to handle rapid technological and organizational advance is considered a major stumbling block for preserving digital information. Indeed, while we are still able to read our written heritage from several thousand years ago, the digital information created merely a decade ago is in serious danger of being lost, creating a digital Dark Age

– Wikipedia.

Introduction

The preservation system used by BHL-Europe is based on the International Organization for Standardization (ISO) Reference Model for an **Open Archival Information System (OAIS)**. An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a **Designated Community**, in our case scientists, students, teachers, and the general public interested in biodiversity literature.

Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely. This could mean hundreds if not thousands of years.

In order to achieve this ambitious goal the preservation system addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. The OAIS itself will also be migrated over time as technologies change. To facilitate this evolution the system is separated into different functional entities or modules.

Functional Model

The OAIS of Fig. 19 below is separated into six functional entities and three information packages. The solid lines connecting entities identify the main communication paths over which information flows through the system.

Submission Information Package (SIP): An Information Package that is delivered by the producer to the OAIS for use in the construction of one or more AIPs.

Archival Information Package (AIP): An Information Package, consisting of the content information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Dissemination Information Package (DIP): The Information Package, derived from one or more AIPs, received by the consumer in response to a request to the OAIS.

In addition to the previously mentioned OAIS model two other components are needed for the BHL-Europe system. As metadata harmonization and augmentation has to be done before the Submission Information Package (SIP) creation to have full control over the created metadata,

a Pre-Ingest module will handle these activities and be the main interface to the content providers. To display the augmented metadata correctly a portal will present a multilingual interface and search functionality to the user.

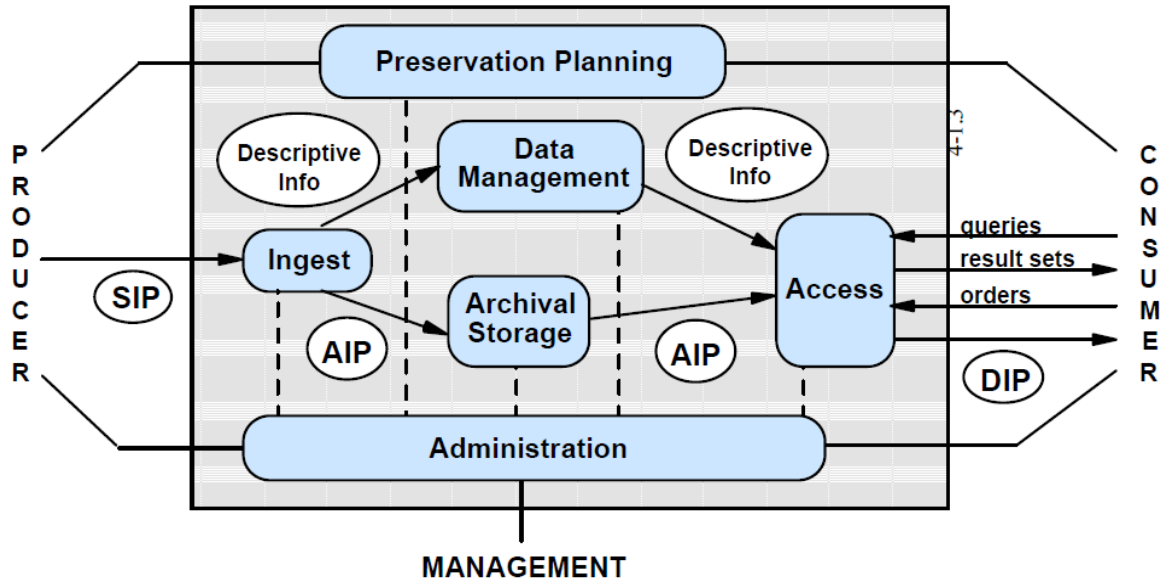


Fig. 19: OAIS

Module Descriptions

The following paragraphs are brief descriptions of the principal modules of the BHL-Europe system.

Pre-Ingest

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

Ingest

This component provides the services and functions to accept Submission Information Packages (SIPs) from the Pre-Ingest module. It prepares the contents for storage and management within the archive. Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP) which complies with the archive's data formatting and documentation standards, extracting Descriptive Information or metadata from the AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management.

Archival Storage

This component provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive

holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfil orders.

Preservation Planning

This component provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the designated user community over the long term. Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the designated community's service requirements and knowledge base. Preservation planning also develops detailed migration plans, software prototypes and test plans to enable implementation of administration migration goals.

Data Management

Data Management provides services and functions for populating, maintaining, and accessing metadata which identifies and documents archive holdings and administrative data used to manage the archive.

Data Management functions include:

- catalogues and inventories on what may be retrieved from archival storage,
- administering the archive database functions (maintaining schema and view definitions, and referential integrity),
- performing database updates (loading new metadata or archive administrative data),
- performing queries on the data management data to generate result sets,

For BHL-Europe, the catalogue schema is derived from the fields and semantics of the BHL Deduplication Subset (BDS), which conforms to core Dublin Core fields.

Access

This component provides the services and functions that support consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing consumers to request and receive information products. Access functions include communicating with consumers to receive requests, applying controls to limit access to specially protected information, coordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering the responses to consumers.

Portal

This component represents the OAIS consumer. The portal is the public face of BHL-Europe. It will allow the end user to search for and access the description, location and availability of items. The portal provides visual aids for browsing the archive as well as multilingual search tools to retrieve items on various levels in various languages.

Abbreviations

ABBYY: is a commercial company, which provides software for OCR

ABCD: Access to Biological data Collections <http://rs.tdwg.org/dwc/index.htm> – A data and metadata specification for the exchange of biological collections and observations data.

AIP (Archival Information Package): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

ANSI: American National Standards Institute

API: Application Programming Interface

ATOS: Atos Origin Integration France (Paris)

BA: Bibliotheca Alexandrina

BDS: BHL Deduplication Subset

BHL: Biodiversity Heritage Library

BHL-E: Biodiversity Heritage Library Europe

BHL-Europe: Biodiversity Heritage Library Europe

BHL Open Access: Images on the portal are available for viewing and reuse with no charges attached. See the Berlin declaration: http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf Link to BHL Copyright and Licensing statement: <http://biodivlib.wikispaces.com/Licensing+and+Copyright>

Bitmap: or raster graphic is a data structure representing a generally rectangular grid of pixels, or points of colour, viewable via a monitor, paper, or other display medium. Raster images are stored in image files with varying formats.

Link: http://en.wikipedia.org/wiki/Bit_map

BPG: Best Practice Guide (shortcut for “Delivery of the first version of the approved best practice guidelines and standards”), also referred to as “Approved best practice guidelines and standards”.

BSD: Berkeley Software Distribution is a version of the operating system Unix.

Colour depth: or bit depth is the number of bits used to represent the colour of a single pixel in a bitmapped image or video frame buffer. This concept is also known as bits per pixel (bpp), particularly when specified along with the number of bits used. Higher colour depth gives a broader range of distinct colours.

CSDGM: Federal Geographic Data Committee Content standard for Digital Geospatial Metadata <http://www.nbi.gov/> - National Biological Infrastructure Initiative metadata standard including geographical metadata.

CSIC: Consejo Superior de Investigaciones Científicas (Madrid), Spain

CVS: Concurrent Versions System or Concurrent Versioning System is a client-server free software revision control system in the field of software development. Version control system software keeps track of all work and all changes in a set of files and allows several developers (potentially widely separated in space and/or time) to collaborate.

DAF: Digital Assets Factory

DC: Dublin Core is a standardised metadata element set which provides small and fundamental groups of txt elements.

DFG: Deutsche Forschungsgemeinschaft (German Science Foundation), Germany

DIP (Dissemination Information Package): The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS.

DLF: Digital Library Federation, USA

DPI (Dots Per Inch): is a measure of spatial printing or video dot density, in particular the number of individual dots that can be placed in a line within the span of 1 inch (2.54 cm). The DPI value tends to correlate with image resolution, but is related only indirectly.

DwC: Darwin Core <http://rs.tdwg.org/dwc/index.htm> – An extension of the DC for biodiversity information.

EAD: Encoded Archival Description <http://www.loc.gov/ead/> – Mostly used by archives.

EDIT: European Distributed Institute of Taxonomy

EOL: Encyclopaedia of Life, a project to create an online reference source and database for every one of the 1.8 million species that are named and known on this planet. <http://www.eol.org/>

EUROPEANA: Europeana is a search platform to a collection of European digital libraries with digitised paintings, books, films and archives. The project was initiated by the European Commission.

EXIF: Exchangeable image file format is a specification for the image file format used by digital cameras and scanners. The specification uses the existing JPEG, TIFF Rev. 6.0, and RIFF WAV file formats, with the addition of specific metadata tags. It is not supported in JPEG 2000, PNG and GIF.

EZOLO: Early Zoological Literature Online

FTE: Full-time equivalent

Gallica: Bibliothèque numérique, Gallica digital library – Over a million books and documents... <http://gallica.bnf.fr/>, France

GBV: Gemeinsamer Bibliotheksverbund (Common Library Network, a public non-profit institution by seven northern German federal states Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen-Anhalt, Schleswig-Holstein, Thüringen and the Stiftung Preußischer Kulturbesitz) <http://www.gbv.de> →VZG

GPL: General Public Licence

GRIB: Global References Index to Biodiversity

GVK: Gemeinsamer Verbundkatalog (GBV Union Catalogue at <http://gso.gbv.de>, a multimaterial bibliographic database, comprises the library holdings of 7 northern German federal states, →GBV)

HNHM: Hungarian Natural History Museum (Budapest), Hungary

IA: Internet Archive (Universal access to all knowledge) <http://www.archive.org>

ILS: Integrated Library System

Image resolution: describes the detail an image holds. The term applies to digital images, film images, and other types of images. Higher resolution means more image detail.

http://en.wikipedia.org/wiki/Image_resolution

IPR: Intellectual Property Rights

IPTC: Information Interchange Model is a file structure and set of metadata attributes that can be applied to text, images and other media types. It was developed in the early 1990s by the International Press Telecommunications council (IPTC) to expedite the international exchange of news among newspapers and news agencies.

ISO: International Organization for Standardization; the primary international standards development organization.

JSF: JavaServer Faces is a Java-based application Framework to building web-based user interfaces

JPEG: Joint Photographic Experts Group, is an image format that used method of compression for digital photography

JPEG2000: is an image format for raster graphics with image compression.

KVK: <http://www.ubka.uni-karlsruhe.de/kvk.html> Karlsruher Virtueller Katalog (Karlsruhe Virtual Catalogue) (KVK) offers a uniform user interface for simultaneous research in union catalogues located around the world. Germany

LANDOE: Land Oberösterreich (Oberösterreichische Landesmuseen, Linz, Österreich), (Upper Austrian State Museums, Biology Centre), Austria

LDAP: Lightweight Directory Access Protocol is an application protocol for reading and editing directories over an IP network.

MAB2: Machine exchange format for metadata used by libraries, especially in the German library system.

MARC: MACHine-Readable Cataloging, standards for the representation and communication of bibliographic and related information in machine-readable form

MARC21: is a standard bibliographic metadata format, see also MARC; a formatting, record structure, and encoding standard for electronic bibliographic cataloguing records developed by the Library of Congress. The "21" refers to the version of MARC issued in 1998 that integrated the U.S. and Canadian versions of MARC.

MBLWHOI Library: Marin Biological Laboratory Woods Hole Oceanographic Institution Library, USA

MCZ: Museum of Comparative Zoology Harvard University, USA

MfN: Museum für Naturkunde (Berlin), Germany

Metadata: structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource.

METS: Metadata Encoding and Transmission Standard; is an XML-based container for metadata; a metadata scheme for complex digital library objects

MODS: Metadata Object Description Schema; is an XML-based bibliographic description schema and a derivative of MARC 21; a metadata scheme for rich description of electronic resources.

MySQL: is the most popular relational database management system.

NAT: Stichting Nationaal Natuurhistorisch Museum Naturalis (Leiden), Netherlands

NHM: Natural History Museum (London), UK

NHML: Natural History Museum Library London, UK

NHMW: Naturhistorisches Museum Wien (Museum of Natural History Vienna), Austria

NISO: National Information Standards Organization (United States of America); a standards development organization, accredited by the American National Standards Institute that develops library and information related standards.

OAI: Open Archives Initiative, develops and promotes Standards for Web Content Interoperability.

OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting

OAIS: Open Archival Information System

OCLC: Not for profit computer service and research organization whose systems help libraries locate, acquire, catalogue, and lend library materials. <http://www.oclc.org>

OCR: Optical Character Recognition; the electronic identification and digital encoding of scanned, printed or handwritten characters by means of an optical scanner and specialized software.

OPAC: Online Public Access Catalogue, is an online catalogue of a library collection that is available to the public

PDI: Preservation Description Information

PHAIIDRA: is the acronym for Permanent Hosting, Archiving and Indexing of Digital Resources and Assets <https://phaidra.univie.ac.at/>

PHP: Hypertext Pre-processor - a scripting language

PICA: The PICA format is the internal data format for cataloguing in “Gemeinsamer Bibliotheksverbund” (GBV)

PNG: Portable Network Graphics, is a raster graphics image format

PPI (Pixels Per Inch): or pixel density is a measurement of the resolution of devices in various contexts; typically computer displays, image scanners, and digital camera image sensors.

http://en.wikipedia.org/wiki/Pixels_per_inch

http://en.wikipedia.org/wiki/Dots_per_inch

PPN: Pica Production number

PREMIS: Preservation Metadata. Implementation Strategies, is an international working group concerned with developing metadata for use in digital preservation.

QA: Quality Assurance

qDC: Qualified Dublin Core, more detailed Dublin Core

RARA: rare works stacks of the SUB Göttingen, Germany

RBGE: Royal Botanic Garden Edinburgh, UK

RDF: Resource Description Framework, is a standard model for data interchange on the Web

REST: Representational State Transfer

RSLP: Collection-level description <http://www.ukoln.ac.uk/metadata/cld/> – General, covers collections of all subjects, domains and types.

SBPK Berlin: Staatsbibliothek zu Berlin - Preussischer Kulturbesitz

SDK: Software Development Kit is a set of development tools that allow an application to be created.

Semantics: The names and meanings of metadata elements.

SIP (Submission Information Package): An Information Package that is delivered by the Producer to the OAIS for use in the construction of one or more AIPs.

SMB: Server Message Block is a network protocol that you can use to provide shared access to files, printers and serial ports.

SOAP: Simple Object Access Protocol, is a simple XML-based protocol to let applications exchange information over HTTP, <http://www.w3schools.com/soap/default.asp>

SPARQL: Is the recursive acronym for SPARQL Protocol and RDF Query Language.

SPECTRUM: <http://www.mda.org.uk/spectrum.htm> -- Describes museum objects.

SQL: Structured Query Language, is the database language of data manipulation (e.g. update, delete) and database creation (e.g. create)

SUB Göttingen: Niedersächsische Staats- und Universitätsbibliothek Göttingen, Germany

TAR: Tape ARchive is a method to unite several single files in one file. No compression takes place.

TEI: Text Encoding Initiative <http://www.tei-c.org/> – Digital texts.

TGZ: A TGZ file is a TAR (Tape ARchive) compressed with GZIP.

TIFF: Tagged image file format

Tip ins: Loose leaf page (or pages) that have been added later; usually containing corrections.

UBER: Humboldt-Universität zu Berlin, Germany

UGOE: Georg-August-Universität Göttingen, Germany

URL: Uniform Resource Locator; a unique address for identifying and locating a resource on the Internet.

URN: Uniform Resource Name, identifies a resource or unit of information independent of its location.

VZG: Verbundzentrale des GBV – Head office of the Common Library Network, Germany

XML: Extensible Markup Language; a set of rules for encoding documents in machine-readable form.

XMP: Extensible Metadata Platform is a standard, created by Adobe Systems Inc. for processing and storing standardised and proprietary information relating to the contents of a file.

Z39.50: is a client–server protocol for searching and retrieving information from remote computer databases; National Information Standards Organization Z39.50 Information Retrieval Protocol (Z39.50/ISO 23950), a computer protocol that can be implemented on any platform, defines a standard way for two computers to communicate for the purpose of information retrieval; a NISO and ISO standard protocol for cross-system search and retrieval. Officially, international standard, ISO 23950, Information Retrieval (Z39.50): Application Service Definition and Protocol Specification, and ANSI/NISO standard Z39.50.

ZR: Zoological Record, general database for zoological literature

Figures

Fig. 1: Limits of database in relation to search criteria.....	15
Fig. 2: Criteria for choosing content to be included in BHL-Europe.....	15
Fig. 3: Risk band and licensing guide	19
Fig. 4: short summary of the journal	38
Fig. 5: Screenshot database www.zobodat.at	39
Fig. 6: Example	39
Fig. 7: Bitmap.....	45
Fig. 8: http://www.landesmuseum.at/biologiezentrum/ online presentation of individual articles belonging to a volume of a journal	52
Fig. 9: window divided into two horizontal halves (first step)	54
Fig. 10: second step.....	55
Fig. 11: third step	56
Fig. 12: fourth step	57
Fig. 13: fifth step	58
Fig. 14: sixth step	59
Fig. 15: seventh step.....	60
Fig. 16: eighth step.....	60
Fig. 17: Processing Gemini Requests (screenshot)	73
Fig. 18: DAF Architecture	87
Fig. 19: OAIS	90

Tables

Tab. 1: Risk bands.....	18
Tab. 2: Cost spreadsheet example BHL-US.....	26
Tab. 3: elements identified by Europeana.....	35
Tab. 4 elements, requirements, rights	36
Tab. 5: elements, requirements, rights	36
Tab. 6: General condition factors.....	41
Tab. 7: Detailed condition factors.....	42
Tab. 8: Commonly used bit depth rates http://en.wikipedia.org/wiki/Color_depth	46
Tab. 9: Resolution and Bit depth.....	48
Tab. 10: Chart how many books in a given shipment should be QA'd	74

References

Non digital References:

- BROCKE v. J. & A. SIMONS (2008): Towards a Process Model for Digital Content Analysis – The Case of Hilti. Available at: [http://ecom.fov.uni-mb.si/proceedings.nsf/0/4c1351e08e8d0f87c1257482003c3852/\\$FILE/44Brocke.pdf](http://ecom.fov.uni-mb.si/proceedings.nsf/0/4c1351e08e8d0f87c1257482003c3852/$FILE/44Brocke.pdf)
- DEUTSCHE FORSCHUNGSGEMEINSCHAFT (2009): Scientific Library Services and Information Systems (LIS): DFG Practical Guidelines on Digitisation for programmes funding Scientific Library Services and Information Systems http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_en.pdf
- DRAKE K-M., B. JUSTRELL & A.M. TAMMARO (2003): Good Practice Handbook. Version 1.2. edited by the Minerva Working Group 6: Identification of good practices and competence centres. Minerva Knowledge Base Digitising Content Together. http://www.minervaeurope.org/structure/workinggroups/goodpract/document/bestpracticehandbook1_2.pdf
- EUROPEANA TEAM (2010): Europeana Aggregators' Handbook Edition 1. Europeana think culture. http://version1.europeana.eu/c/document_library/get_file?uuid=94bcd9bf-3625-4e6d-8135-c7375d6bbc62&groupId=10602
- HANKEN J. (2010): Retooling Special collections Digitization in the age of Mass Scanning (LG-50-08-0058-08. URL: <http://biodivlib.wikispaces.com/file/detail/Retooling+Special+Collections+Digitization+in+the+Age+of+Mass+Scanning.odt>
- INTERNET ARCHIVE OPERATIONS TEAM (2010): Internet Archive Book Digitization Process. Ed. Robert Miller, Director of Books, Internet Archive.
- MCKENNA G., COLLECTION TRUST (UK), C. DELOOF & ROYAL MUSEUMS OF ART AND HISTORY (BE) (2009): Recommendation and best practice report regarding the application of standards, including recommendations for a harvesting format and fact sheets for dissemination. D3.2. Athena Access to cultural heritage networks across Europe. <http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents>
- MCKENNA G., COLLECTION TRUST (UK), C. DELOOF & ROYAL MUSEUMS OF ART AND HISTORY (Belgium) (2009): Digitisation: Standards Landscape for European

museums, archives, libraries. Athena Access to cultural heritage networks across Europe. La Tipografia di Umbero Frisardi, Roma, Italy.
<http://www.emuzeum.cz/download/athena-soubory/27-imp-athenalibrettinoStandardCopertinaLastINCIANO.pdf>

SCAIFE B. (2009): Catalogue of content holder requirements (quality, quantity, accessibility, standards, specifications of content and metadata), (ECP-2008-DILI-518001) URL:
https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf

The Catalogue of content holder requirements is largely based on the Digital Library Federation's (DLF) "Benchmark for Faithful Digital Reproductions of Monographs and Serials": <http://www.diglib.org/standards/bmarkfin.htm> (2002)

THE DIGITAL LIBRARY FEDERATION BENCHMARK WORKING GROUP (2002): Benchmark for Faithful Digital Reproductions of Monographs and Serials. The Digital Library Federation. <http://www.diglib.org/standards/bmarkfin.pdf>

Digital References (Summary of Internet Links):

<https://bhl.wikispaces.com/BHL-Europe>

http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_en.pdf

http://www.minervaeurope.org/structure/workinggroups/goodpract/document/bestpracticehandbook1_2.pdf

<http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents>

<http://www.emuzeum.cz/download/athena-soubory/27-imp-athenalibrettinoStandardCopertinaLastINCIANO.pdf>

http://version1.europeana.eu/c/document_library/get_file?uuid=94bcddbdf-3625-4e6d-8135-c7375d6bbc62&groupId=10602

<http://www.archive.org/details/ProcessDocument&reCache=1>

<http://www.diglib.org/standards/bmarkfin.pdf>

[\[mb.si/proceedings.nsf/0/4c1351e08e8d0f87c1257482003c3852/\\\$FILE/44Brocke.pdf\]\(http://ecom.fov.uni-mb.si/proceedings.nsf/0/4c1351e08e8d0f87c1257482003c3852/\$FILE/44Brocke.pdf\)](http://ecom.fov.uni-</p>
</div>
<div data-bbox=)

<http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

<http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

<http://tyler.hrc.utexas.edu/>

<http://biodivlib.wikispaces.com/Licensing+and+Copyright>

<http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>

<http://grib.gbv.de/>

<http://bhleurope.gbv.de/>

<https://bhl.wikispaces.com/Global+References+Index+to+Biodiversity+%28GRIB%29>

<http://www.e-taxonomy.eu/>

<http://www.gbv.de/vgm/index?lang=en>

<http://www.gbv.de/vgm/info/biblio/01VZG/>

<http://bhl.nhm-wien.ac.at/scanlist/>

<https://bhl.wikispaces.com/Documentation+from+BHLSeriallist>

<http://dublincore.org/>

<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

<http://www.loc.gov/ead/>

<http://www.loc.gov/marc/>

<http://www.ukoln.ac.uk/metadata/cld/>

<http://www.tei-c.org/>

<http://rs.tdwg.org/dwc/index.htm>
<http://rs.tdwg.org/abcd/>
<http://marinemetadata.org/references/iso19115/>
<http://www.nbio.gov/>
<http://knb.ecoinformatics.org/software/eml/>
<http://dublincore.org>
<http://dublincore.org/documents/dcmi-terms/>
<http://www.bhl-europe.eu/en/outcomes/documents/ipr-working-documents>
http://version1.europeana.eu/c/document_library/get_file?uuid=a830cb84-9e71-41d6-9ca3-cc36415d16f8&groupId=10602
<http://www.version1.europeana.eu/web/guest/technical-requirements/>
<http://purl.org/dc/elements/1.1/rights>
<http://version1.europeana.eu/web/guest/technical-requirements/>
<http://www.version1.europeana.eu/web/guest/technical-requirements/>
www.zobodat.at
<http://biodivlib.wikispaces.com/file/detail/Retooling+Special+Collections+Digitization+in+the+Age+of+Mass+Scanning.odt>
https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf
https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf
<http://www.library.cornell.edu/preservation/tutorial/metadata/metadata-01.html>
<http://www.ahds.ac.uk/creating/information-papers/checklist/index.htm>
http://en.wikipedia.org/wiki/Color_depth
https://bhl.wikispaces.com/BHLE_WP2
https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf
https://bhl.wikispaces.com/file/view/BHL-E_2pt1_20090805.pdf
<http://www.diglib.org/standards/bmarkfin.htm>
<https://bhl.wikispaces.com/QA+Procedures>
http://en.wikipedia.org/wiki/List_of_optical_character_recognition_software
<http://www.landesmuseum.at/biologiezentrum/>
<http://www.diglib.org/standards/bmarkfin.htm>
https://bhl.wikispaces.com/BHLE_WP2_BPG
<https://bhl.wikispaces.com/BHL+Cookbook>
https://bhl.wikispaces.com/BHLE_WP2
www.openarchives.org
<http://www.biologiezentrum.at/oai/oai2.php>
<http://gso.gbv.de>
<http://www.animalbase.uni-goettingen.de/>
<http://www.fishbase.org/>
<http://www.exlibrisgroup.com/category/Aleph>
http://si-pddr.si.edu/dspace/bitstream/10088/7182/1/Gwinn_Rinaldo_IFLA_2009.pdf
<https://bhl.wikispaces.com/Workflow+documents>
<https://bhl.wikispaces.com/BHL+Cookbook>
<https://bhl.wikispaces.com/QA+Policy>
<https://bhl.wikispaces.com/Cookbook+Quality>
<https://bhl.wikispaces.com/QA+Procedures>
<https://bhl.wikispaces.com/QA+Policy>
<https://bhl.wikispaces.com/QA+sampling+chart>
<http://www.carpet-project.net/en/catalogue/carpet/production-level-goobi-1/>

http://www.arcsoft.com/de-de/software_title.asp?productCode=SNSD
<http://www.gimp.org/docs/>
<http://www.gimp.org/downloads/>
<http://www.gimp.org/about/introduction.html>
<http://www.irfanview.com/eula.htm>
<http://www.gimp.org/about/introduction.html>
<http://www.gnu.de/documents/gpl-3.0.en.html>
<http://www.irfanview.net/>
http://www.irfanview.net/main_formats.htm
<http://www.software.com/irfanview-9337-1>
<http://www.heise.de/software/download/goocr/7147>
<http://jocr.sourceforge.net/download.html>
<http://www.dspace.org/latest-release>
<http://www.dspace.org/training-summary/training-materials.html>
<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
<http://openrepository.com/products/open-repository-dspace>
<http://atmire.com/consultancy.php>
http://www.carpet-project.net/fileadmin/user_upload/documents/2010-11-26-Repositoryienhosting
<https://phaidra.univie.ac.at/>
<https://fedora.phaidra.univie.ac.at/fedora/get/o:29162/bdef:Content/get>
<http://sourceforge.net/projects/metsbuilder/>
<http://sourceforge.net/projects/jhove/>
<http://hul.harvard.edu/jhove/documentation.html>
<http://www.loc.gov/standards/mets/version18/mets.xsd>
<http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd>
<http://www.loc.gov/standards/mods/mods.xsd>
<http://en.wikipedia.org/wiki/JHOVE>
<http://www.carpet-project.net/en/catalogue/carpet/JHOVE/>
http://www.java.com/en/download/inc/windows_upgrade_xpi.jsp
<http://www.carpet-project.net/en/catalogue/carpet/Fedora%20Commons>
<http://www.w3schools.com/soap/default.asp>
<https://wiki.duraspace.org/display/FCR30/Getting+Started+with+Fedora>
<http://www.fedora-commons.org/about/features>
<http://download.oracle.com/javase/index.html>
<http://tomcat.apache.org/>
<http://maven.apache.org/>
<http://www.digiverso.com/en/products/goobi>
<http://goobi-wiki.slub-dresden.de/index.php/Hauptseite>
<http://www.digiverso.com>
<http://www.carpet-project.net/en/knowledge-base/>
<http://www.carpet-project.net/forum/>
<http://www.digiverso.com/en/products/viewer>
<http://www.carpet-project.net/en/>
<http://www.minervaeurope.org/>
<http://www.digitalisierung.ethz.ch/>
<http://www.digiverso.com/en/>
<https://phaidra.univie.ac.at/>

<http://fedora-commons.org/>
<http://www.arcsoft.com/en-us/>
<http://www.irfanview.com/>
<http://www.gimp.org/>
<http://rs.tdwg.org/dwc/index.htm>
http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf
<http://biodivlib.wikispaces.com/Licensing+and+Copyright>
http://en.wikipedia.org/wiki/Bit_map
<http://www.nbii.gov/>
http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf
<http://biodivlib.wikispaces.com/Licensing+and+Copyright>
http://en.wikipedia.org/wiki/Bit_map
<http://www.nbii.gov/>
<http://rs.tdwg.org/dwc/index.htm>
<http://www.loc.gov/ead/>
<http://www.eol.org/>
<http://gallica.bnf.fr/>
<http://www.gbv.de>
<http://gso.gbv.de>
<http://www.archive.org>
http://en.wikipedia.org/wiki/Image_resolution
<http://www.ubka.uni-karlsruhe.de/kvk.html>
<http://www.oclc.org>
http://en.wikipedia.org/wiki/Pixels_per_inch
http://en.wikipedia.org/wiki/Dots_per_inch
<http://www.ukoln.ac.uk/metadata/cld/>
<http://www.w3schools.com/soap/default.asp>
<http://www.mda.org.uk/spectrum.htm>
<http://www.tei-c.org/>
<http://www.w3.org/XML/>
<http://www.loc.gov/standards/mods/>
<http://www.ubio.org/>
<http://www.biodiversitylibrary.org/>
<http://makeit.digitalnz.org/guidelines>
<http://www.library.cornell.edu/preservation/tutorial/>
<http://www.cambridgeincolour.com/tutorials/bit-depth.htm>
<http://wiki.bibalex.org/DAFWiki>

Appendix

A: Example of a model bilateral contract

City and date

Name of institution

Entitled Person

Address

Website provider

Authorised proxy

Address

Agreement

Herewith **### (Name of institution)** approves of the digital versions of the following periodicals **“Name of periodical”**, **“Name of periodical”**, and **“Name of periodical”** being made available for download on the website of the **provider (www.websiteprovider.xx)**.

The following volumes/issues will be provided for free download:

- 1) #####
- 2) #####
- 3) #####

The remaining volumes/articles will be for sale for **##** Cent per page on **www.websiteprovider.xx**. After the financial service provider’s deduction of expenses the entitled person will receive **##** Cent per downloaded page. If a client downloads 500 pages or more they will be given a discount (as shown in the list below). Such a discount respectively reduces the gains. The financial service provider draws up an annual account including the transfer of the earned money and list of the itemised downloads.

The website provider is obligated to subsequently transfer the money to the account **###** of the entitled person.

It shall be hereby stated that either party may rescind this contract without giving any reasons. At the same time it shall be emphasised that free supply of data can be changed to a fee-based service and vice versa as proposed by the entitled person.

For reasons that are inherent in the system the following discounts are defined for all downloads which the entitled person hereby approves:

Pages	Discount
>=500	10%
>=1000	20%
>=3000	30%
>=5000	40%
>=10000	50%

As of now the entitled person provides the web provider with a PDF-document (printable version) of every volume if available to include in the download platform. Furthermore a written note has to be included about whether the download is subject to a charge and if so for which period the service is fee-based. The conditions as described above apply to future volumes as well. The web provider includes a watermark in the header of every PDF-page reading as follows: © **entitled person, downloaded at www.websiteprovider.##**

Signature of the entitled person

Signature of the authorised website provider

B: Pre-Ingest File Submission Guidelines