ECP-2008-DILI-518001

BHL-Europe

File Submission Guidelines

Deliverable number	*None*
Dissemination level	*Public*
Delivery date	*20 January 2012*
Status	*Version 1.0*
Authors	*Melita Birthälmer, Wolfgang Koller*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

0 Document History

0.1 Contributors

Person	Partner
Henning Scholz	MfN
Jiri Frank	NMP
Heimo Rainer	NHMW
Michael Malicky	LANDOE
Michaela Hierschläger	LANDOE
Chris Sleep	NHM

0.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
09 December 2011	Melita Birthälmer	0.1	First draft with open questions
23 December 2011	Wolfgang Koller Heimo Rainer	0.2	Review of first draft and answering open questions
05 January 2012	Melita Birthälmer	0.3	Integrating answers for open questions and highlighting remaining open questions
09 January 2012	Jiri Frank	0.4	Adding comments
12 January 2012	Wolfgang Koller	0.5	Second review with regards to remained open questions
20 January 2012	Melita Birthälmer	1.0	Finalisation

0.3 Reviewers

This document requires no reviews and approvals.

0.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
BHL-Europe Content Provider	23 January 2012	1.0

1 Table of Contents

2 Purpose

This document aims to give guidance how to provide and upload content to BHL-Europe. The file submission guidelines allow BHL-Europe to facilitate the automatic processing of submitted content and thus archive and finally display content in the BHL-Europe portal. The target readers of this document are content providers.

3 Background

File submission guidelines were already available beginning of the year 2011 and therefore content providers followed the previous guidelines in communication with the BHL-Europe ingest team. BHL-Europe decided to complement the already available rules in order to allow an easier automatic ingest of the provided content and to clarify open questions from the content providers. These guidelines are based on experiences made during the last year and open questions might still come up.
The guidelines presented within this document will be valid from date of publication. Content that already has been uploaded to the BHL-Europe server is not affected by these rules. BHL-Europe will try to handle the content that has been already uploaded prior to this and do not follow the new guidelines. However the ingest team still might get back to individual content providers and ask to make changes e.g. adding additional metadata files in case we are not possible to handle the content and make changes independently.

There are five basic elements to take into consideration when providing content to BHL-Europe:

- The metadata that describes your content and makes it retrievable.
- The digital object that is an electronic representation of your physical item or a born digital.
- There might be optional an Optical Character Recognition (OCR) text available for your digital object that you would like to provide.
- The data upload and creation of a directory structure which allows BHL-Europe to understand the structure of your physical item and is used for the portal view.
- The file naming that allows BHL-Europe to understand the file sequence and is used for the portal view.

These five basic elements will be described within the next chapters in more detail.

4 Metadata

This section aims to give an overview of the metadata that is required for BHL-Europe in order to ensure that content is retrievable within BHL-Europe. 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 in addition to the scanned images or born digitals. The higher the metadata quality is, the easier the content can be retrieved.

BHL-Europe is using its own metadata schema for processing the content, which is based on MODS, and called the **O**pen **L**iterature **E**xchange **F**ormat - OLEF[2]. Therefore BHL-Europe developed additionally an open source Schema Mapping Tool and provided metadata will be mapped to the BHL-Europe schema using this Schema Mapping Tool. Provided metadata needs to be mapped to the BHL-Europe schema before we are able to ingest your content to the BHL-Europe system. The mapping needs only to be done once per content provider as long as you don't change the format, fieldnames and structure of your metadata within your uploaded data. During the project time BHL-Europe will provide the mapping as a service for all content providers. Each content provider needs to be mapped individually depending on the used metadata standard within the own institution. For this process, BHL-Europe will create a configuration file for the Schema Mapping Tool for each content provider in cooperation with each content provider individually. 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 information to the content provider. The Schema Mapping Tool is an open source product and is available for download at Github[3]. The tool is also deployed under http://bhl.nhm-wien.ac.at/smt/launch.html.

4.1 Metadata format

Metadata can be provided in following formats to BHL-Europe:

- BHL-Europe internal metadata format OLEF (Open Literature Exchange Format): Providing your metadata using the OLEF schema is an optimum.

- Standard metadata format:
  - MARC21
  - MARCXML
  - MODS
  - Dublin Core
  - RefNum

[2] http://www.bhl-europe.eu/bhl-schema/v0.3/
[3] https://github.com/bhle/bhle/tree/master/pre-ingest/schema-mapping-tool

- Proprietary metadata format
  Custom metadata formats should be avoided. In case you are using non-standard bibliographic format for recording your metadata, please get in contact with us in order to find a solution how to map your metadata to our internal OLEF schema.

4.2 Metadata fields

4.2.1 Descriptive metadata fields for BHL-Europe

4.2.1.1 Serials

Series level
In order to identify a folder as a serial the metadata must have the type set to 'serial' (e.g. MARC21 LEADER 07 set to's')[4].

On series level the following metadata has to be provided:

- Title
- Publisher
- Place of publication
- ISSN (if available)
- Start of publication
- Language of publication (iso639-2b[5])

Section level
In order to identify a folder as section the metadata must have the type set to 'serial' (e.g. MARC21 LEADER 07 set to's'). All folders marked as 'serial' within another folder marked as 'serial' as well will be treated as section (except for volumes, see below).

On section level the same metadata as for serials have to be provided.

Volume level
In order to identify a folder as volume the metadata must have the type set to 'serial' (e.g. MARC21 LEADER 07 set to's'). In addition the volume information of the serial must be set (e.g. MARC21 490$v).

The minimum requirement on metadata for BHL-Europe is metadata on the volume level. Following fields should be contained within the metadata and are mandatory:

- Title of series
- Title of volume
- Authors of volume

[4] http://www.loc.gov/marc/bibliographic/bdleader.html
[5] http://www.loc.gov/standards/iso639-2/php/code_list.php

- Year(s) of publication
- Language of publication (ISO 639-2b[6])

Article level
In order to identify a folder as an article the metadata must have the type set to 'component part' (e.g. MARC21 LEADER 07 set to 'a').

Following fields should be provided if the article level is used:

- Title of article
- Author(s)
- Language of publication (ISO 639-2b[7])
- Date of publication

4.2.1.2 Monographs

Monographs level
In order to identify a folder as a monograph the metadata must have the type set to 'monograph' (e.g. MARC21 LEADER 07 set to 'm').

The following metadata fields should be provided:

- Title of monograph
- Authors of monograph
- Place of publication
- Date of publication
- ISBN if available
- Language of publication (ISO 639-2b[8])

Chapter level
In order to identify a folder as a chapter the metadata must have the type set to 'component part' (e.g. MARC21 LEADER 07 set to 'a').

4.2.1.3 Pages

Page level information is encoded into the filenames. For details see chapter 7.

Page level metadata is recommended for the exchange of BHL-Europe data to the BHL-US portal (via Internet Archive), which at the same time serves also as a data backup for the BHL-Europe content. It is possible to upload content to the Internet Archive without page level metadata, but the Table of Contents view of the item would then give no meaningful results. Therefore, at some stage we need to aim to generate page level metadata for our items.

[6] http://www.loc.gov/standards/iso639-2/php/code_list.php
[7] http://www.loc.gov/standards/iso639-2/php/code_list.php
[8] http://www.loc.gov/standards/iso639-2/php/code_list.php

Additional fields

Following fields may be provided additionally and helps improving the retrieval of your content:

- Keywords
- Abstract text

4.2.2 Technical metadata fields for BHL-Europe

Technical metadata is not relevant for BHL-Europe.

4.2.3 Intellectual Property Rights metadata field for Europeana

The Europeana Data Exchange Agreement (DEA) requires that content providers apply a statement about the rights status of the digital objects described in the metadata submitted to Europeana. The *europeana:rights element* is part of ESE v3.4 and is mandatory[9]. The value of this element should be the URL of the appropriate Creative Commons or Europeana rights statement.

In order being able to provide your content to Europeana you need to let us know under which rights statement you would like to publish your digital object. Applying a rights statement will be possible for each item within the BHL-Europe Pre-ingest tool.

The following values can be selected for the *europeana:rights* metadata field according to the 'Guidelines for the europeana:rights metadata element' that can be found on the DEA pages[10]. Further information about Europeana and legal aspects can be found in the Europeana Licensing Framework document[11] and the legal pages on the Europeana Professional knowledge-sharing platform[12].

[9] http://www.pro.europeana.eu/technical-requirements
[10] http://www.pro.europeana.eu/data-exchange-agreement
[11] http://www.pro.europeana.eu/c/document_library/get_file?uuid=b16bdaf6-4e53-4f58-968a-9d4943a5d297&groupId=858566
[12] http://www.pro.europeana.eu

CC Licences	URL
Public Domain Mark CC – Zero (universal)	http://creativecommons.org/publicdomain/mark/1.0/ http://creativecommons.org/publicdomain/zero/1.0/
CC BY CC BY-SA	http://creativecommons.org/licenses/by/3.0/ http://creativecommons.org/licenses/by-sa/3.0/
CC BY-NC CC BY-NC-SA	http://creativecommons.org/licenses/by-nc/3.0 http://creativecommons.org/licenses/by-nc-sa/3.0/
CC BY-ND CC BY-NC-ND	http://creativecommons.org/licenses/by-nd/3.0/ http://creativecommons.org/licenses/by-nc-nd/3.0/

Table 1: Selectable CC Licences for the Europeana:rights metadata field

Europeana rights statements	URL
Rights Reserved - Free Access	http://www.europeana.eu/rights/rr-f/
Rights Reserved - Paid Access	http://www.europeana.eu/rights/rr-p/
Rights Reserved - Restricted Access	http://www.europeana.eu/rights/rr-r/
Unknown	http://www.europeana.eu/rights/unknown/

Table 2: Selectable Europeana rights statements for the Europeana:rights metadata field

5 Digital object

In addition to the metadata of an object a content provider needs also to provide an electronic representation of a physical item - such as the scanned images of a printed monograph or journal - or if the item already has been created digitally, than you need to provide the born digital. To clarify, in this section we only concentrate on the scanned images of an item and accordingly the born digital. BHL-Europe aims to archive the highest quality available and prefers therefore submissions in a lossless format. If possible, BHL-Europe asks to provide Master Files. For multi-page image formats we also ask to only provide one physical page per file.

5.1 Resolution of the digital object

Type	Resolution (minimum)	Bit-depth	File format
Bitonal	300 ppi	1-bit	TIFF
Greyscale	300 ppi	8-bit	uncompressed TIFF or lossless compressed image
Colour	300 ppi	24-bit	uncompressed TIFF, or lossless compressed images

Table 3: Resolution of digital objects

Please 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.

5.2 File format of the digital object

Digital items (scan, born digital) can be submitted to BHL-Europe in following formats:

- TIFF
  This is our preferred format as our archive is working with this.

- PDF
  PDF files will be internally converted to TIFF files during the ingest process as the BHL-Europe archive is working with TIFF files.

- Any other image formats
  Any other image formats will be converted to TIFF using ImageMagick[13] in order to process them.

Use three-letter extensions for file names (e. g. tif for TIFF images)

5.3 Optical Character Recognition

BHL-Europe will use free OCR software (Tesseract)[14] for electronic translation of scanned images of typewritten or printed text into machine-encoded text. However, BHL-Europe content providers are asked to provide their literature with OCR if possible. If not, BHL-Europe will automatically run the freeware OCR engine for content provided without an OCR text which in most cases will be of poorer quality, as most of the OCR text provided by service providers and content providers are generated by using commercial software (e.g. ABBYY FineReader) and manual quality assurance.

In case you would like to provide us additionally OCR text you can choose to provide following files, depending on the availability of file formats for your data:

a) ***PDF in high quality resolution with included OCR***
   This is the optimum scenario. BHL-Europe will use the PDF file for the preservation and portal view as well as for extracting the OCR text for further portal functionalities (e.g. search using TaxonFinder[15]).

b) ***Image files (TIFF preferred) in original quality with additional text file containing the OCR text***
   Text files must be named exactly like the image it contains the text for, expect for a txt extension.

[13] http://www.imagemagick.org/
[14] http://code.google.com/p/tesseract-ocr/
[15] http://www.ubio.org/index.php?pagename=soap_methods/taxonFinder

Example: NHMW12345_0001.tif and NHMW12345_0001.tif.txt

c) ***Image files in original quality with additional PDF files including OCR text***
For this scenario BHL-Europe will use the high quality image files for the preservation and portal view and will extract the OCR text from the PDF files.

5.3.1 Page orientation

To improve OCR, we kindly ask you to rotate pages so that the text is readable in horizontal direction.

6 Data upload

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.

Example:
Folder-Name: "9184JDUK".
The identifier passed to the OIA-PMH service will be "9184JDUK".

6.1 Directory structure

BHL-Europe asks content provider to follow a hierarchical directory structure in order to reliably indentify which files belong to which item. A hierarchical structure allows representing information using parent-child relationships: each parent can have many children, but each child has only one parent. In other words, a journal title can have many volumes, but each volume belongs only to the one journal.
Your directory structure should follow the structure of your physical item, as BHL-Europe will use the directory structure to represent the hierarchy within the portal view.

General rules:

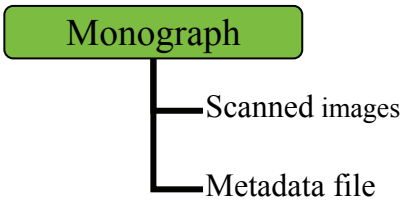
- Multiple subfolders can be created.
- Item folders are nested inside a title folder, e.g. a volume folder is nested inside the serial folder.
- Metadata files must be included at each level.

- The digital object (scanned image or born digital) are placed within the lowermost hierarchical folder. Except for ToC (Table of Contents), Index, etc.

- Preferably use only ASCII characters and Western/Arabic numbers (0-9)
- Don't use: <, >, ", /, |, *
- Avoid to use blank spaces

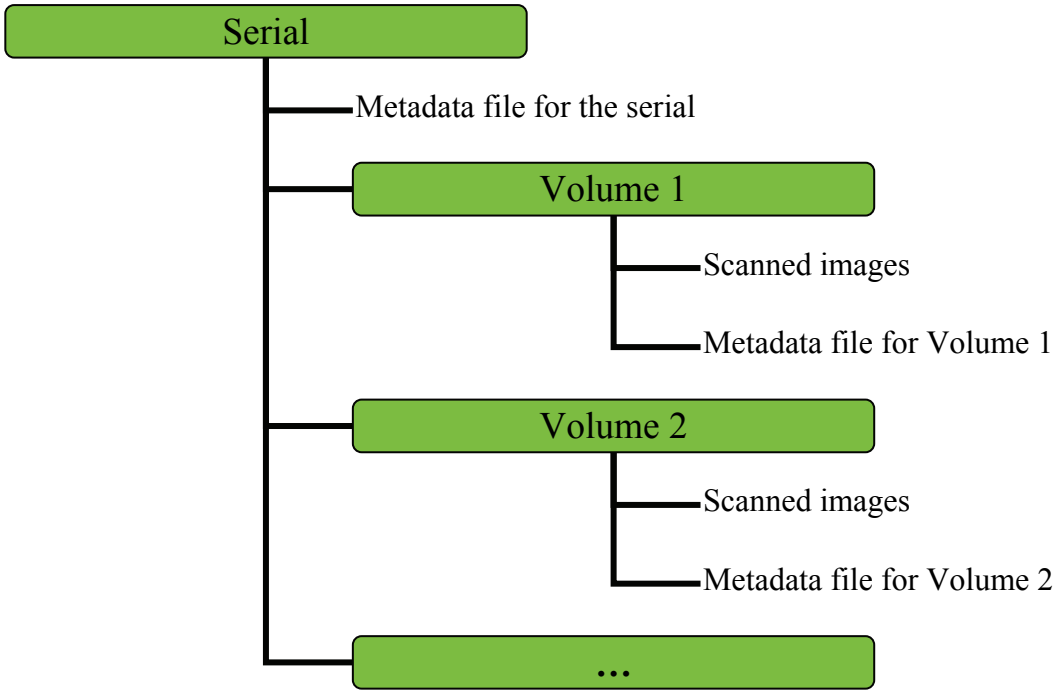
6.1.1 Monograph - directory structure

An exemplary diagram of a hierarchical structure is shown in the following figure for a monograph with minimum required files. In case you provide OCR text it either can be included already in the PDF file (scanned images) or it can be uploaded additionally as a text file (.txt) in the same directory with the image files.



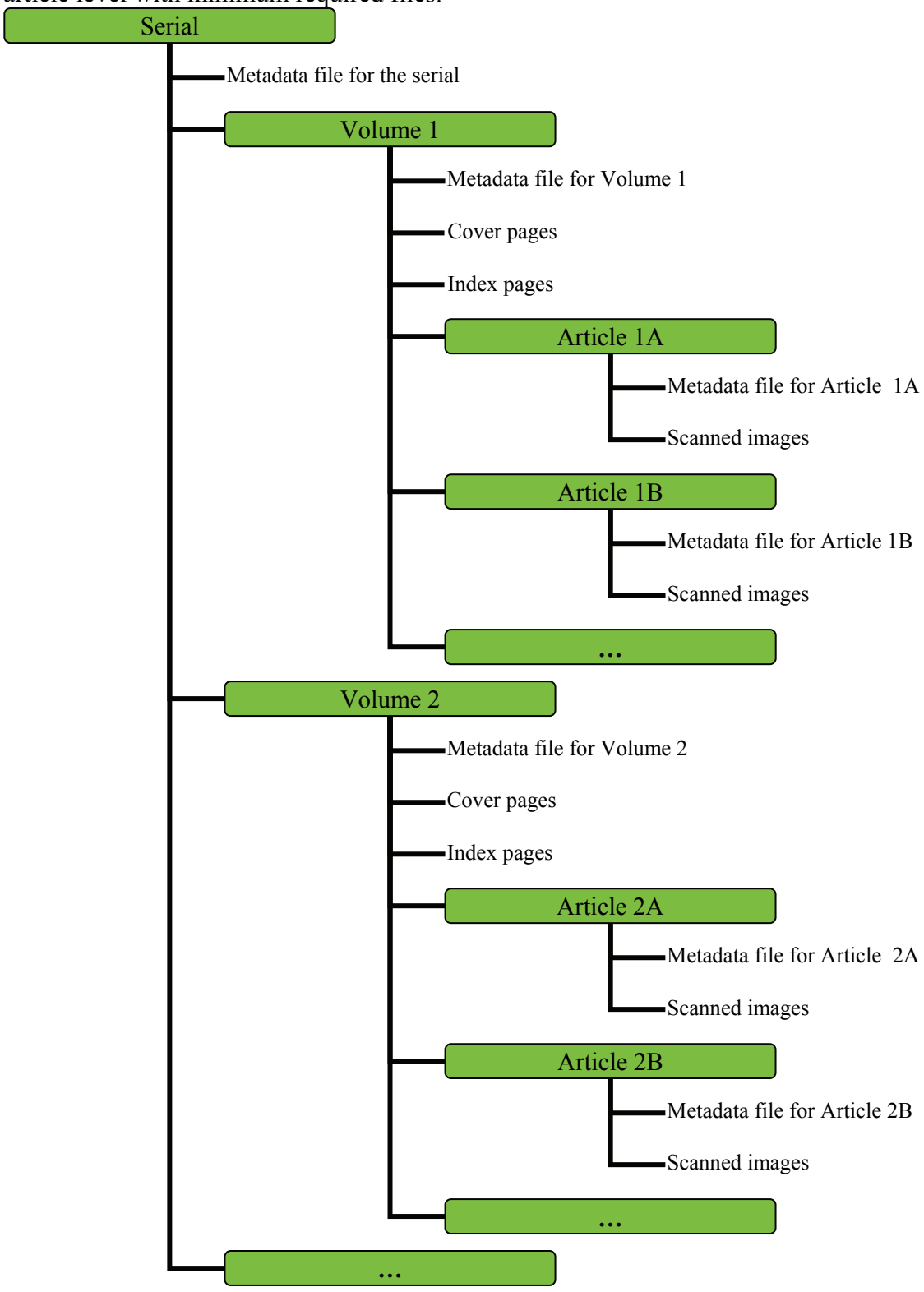
6.1.2 Serial - directory structure

Following example shows an exemplary diagram of hierarchical structure for a serial on volume level with minimum required files.

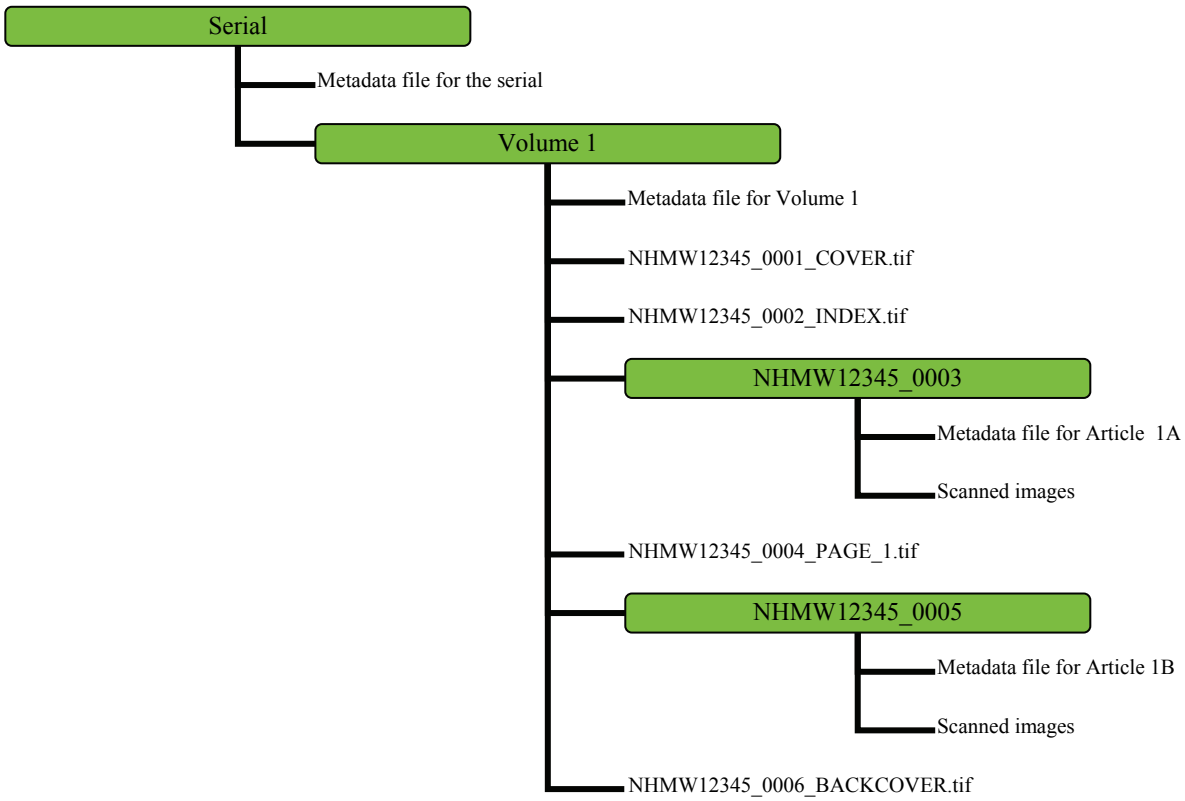


6.1.3 Serial on article level - directory structure

Following example shows an exemplary diagram of hierarchical structure for a serial on article level with minimum required files.



In order to maintain the sequence of articles and pages outside an article (like Index, TOC, etc.) the naming of the article folders has to follow the file naming conventions (see chapter 0).
An example is outlined below:



7 File naming

How you name your files is important for BHL-Europe. The names of your files allow us to understand the file sequence and will also have an influence how your content is shown in the book viewer of the portal.

7.1 Characters within the file name

- Preferably use only ASCII characters and Western/Arabic numbers (0-9)
- Don't use: <, >, ", /, |, *
- Avoid to use blank spaces

7.2 Pagination within the filename

Pagination within the filename helps us to display the pages in the right sequence in the portal.
Pagination information should be encoded using the following pattern:

InternalIdentifier_**FileSequenceNumber**_**PageType**_**PrintedPageNumber**.tif

An example is shown hereafter:
NBGB013726AIGR1889FLOREELE00_**0007**_**PAGE**_**3**.tif

- **NBGB013726AIGR1889FLOREELE00** is the internal identifier for a specific monograph.
- **0007** represents the image sequence number.
- **PAGE** represents the page type within the monograph. Different page types are available (see 7.2.1).
- **3** represents the printed page number.

7.2.1 Page type

Following page types can be used within BHL-Europe:

Page type	Description
COVER	Cover page of the object
FRONTCOVER	Front cover page of the object
BACKCOVER	Back cover page of the object
IMPRINT	Imprint page of the object
INDEX	Table of contents page
TITLE	Title page
HALFTITLE	Half title page
BLANK	Blank page within the object, e. g. imprint page
PLATE	Plate page within the object
FIGURE	Figure page within the object
FOLDOUT	Foldouts within the object
PAGE	Simple page

Table 4: Page types within BHL-Europe

The names of the file are not case-sensitive.

8 FAQ & Examples

- **Monographs are published in several volumes. Do we need to export the metadata of each individual volume and the metadata of the parent?**

Example:

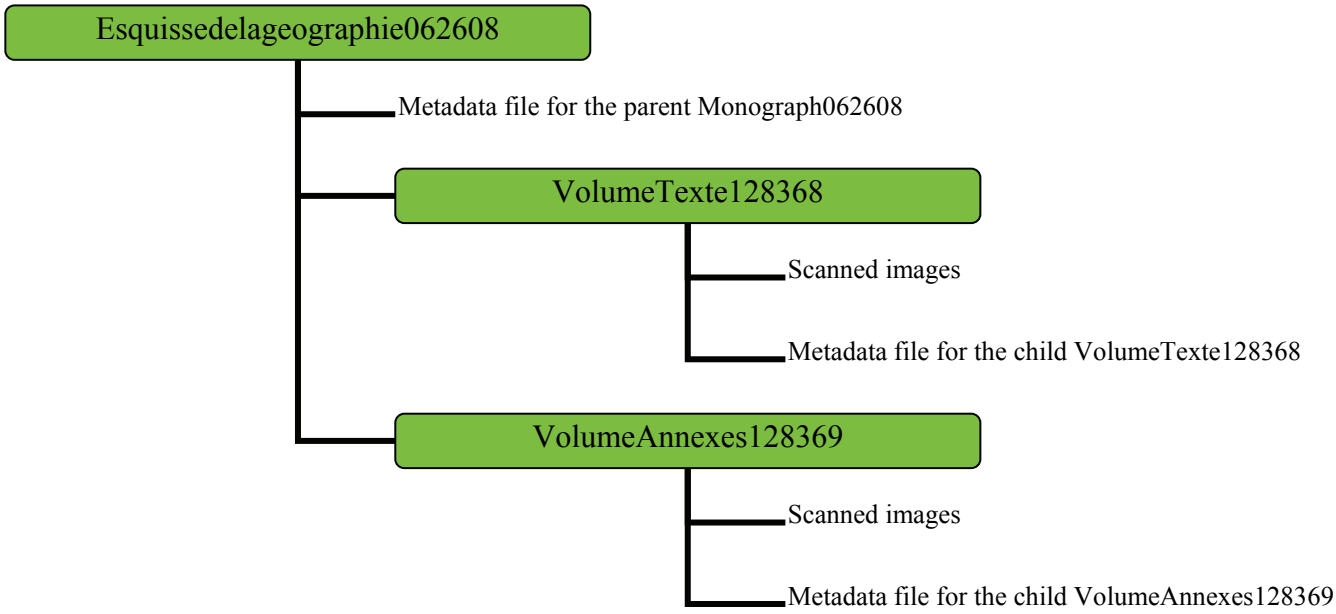
Esquisse de la géographie botanique de la Belgique (parent description = 062608)

Esquisse de la géographie botanique de la Belgique : volume texte (child description = 128368)

Esquisse de la géographie botanique de la Belgique : volume annexes (child description = 128369)

Answer:

Yes, each directory needs a metadata file (see also 6.1.2). The directory structure could look as followed:



- **For naming our directories we use the unique description number from our catalogue. Each volume has a unique number. For the directory containing the different volumes do we use the description number of the first volume or the number of the parent description**?

Example:
The unique description number of the institutional catalogue of the parent directory is 062608.
The number of the first volume (volume text) is 128368.
The number of the second volume (volume annex) is 128369.

Following example demonstrates the naming of the directory according to the institutional unique description number:

NBGB**062608**MASS1910ESQUISSE_02 (parent directory)
NBGB**128368**MASS1910ESQUISSE01 (volume text)
NBGB**128369**MASS1910ESQUISSE02 (volume annex)

Following examples demonstrates the naming of the directory according to the unique description number of the first volume:

NBGB**128368**MASS1910ESQUISSE_02 (parent directory)
NBGB**128368**MASS1910ESQUISSE01 (volume texte)
NBGB**128369**MASS1910ESQUISSE02 (volume annexe)

Answer:
Both examples can be used, as the naming of the directory will not be used within BHL-Europe (except for articles and pages which belong to the volume, see chapter 6.1.3). Only the logical structure matters for BHL-Europe and the naming of the files within the directory.

- **We receive books from different libraries and act as an aggregator for BHL-Europe. How can we mention the names of the provider of an item in addition to our institutional name as a content provider?**

  This can be done using different logins.

- **We have two physical pages scanned in one TIFF file. Can you handle this? How does the book viewer display the two pages? Will the pages be shrunk?**

  This should be strictly avoided, if not try to split the pages before uploading. Otherwise you can maintain the correct printed page by leaving out numbers.

  Example:

  NHMW12345_0001_PAGE_1.tif
  NHMW12345_0002_PAGE_3.tif
  NHMW12345_0003_PAGE_5.tif
  NHMW12345_0004_PAGE_6.tif

- **We provide content on article level. How do we provide pages that do not belong to articles but to the volume?**

  See chapter 6.1.3.

Appendix

A: How to connect to the BHL-Europe server

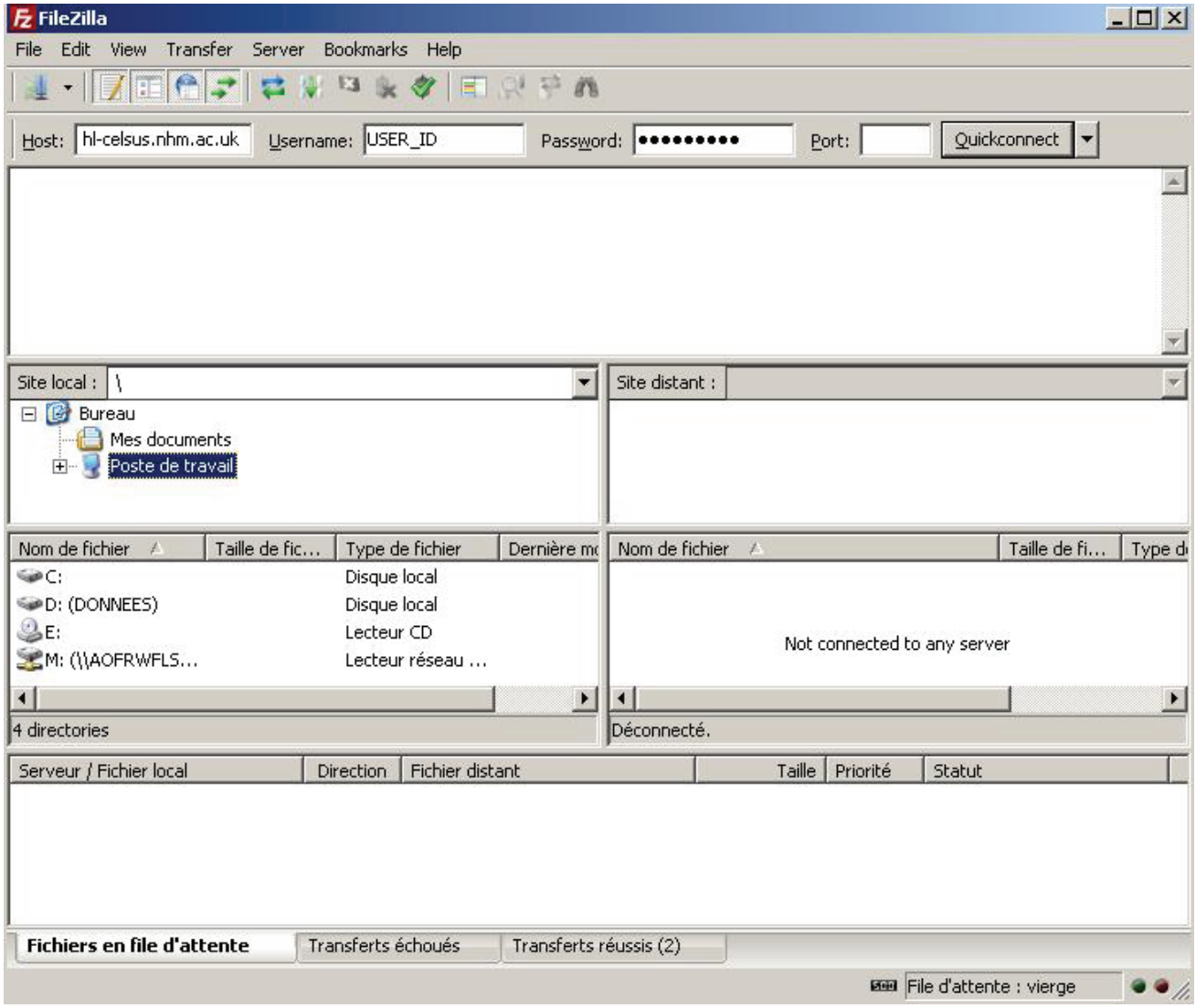
How to connect to the BHL-Europe server

Requirements:

BHL-Europe provides FTPS access to the upload server. Each content provider receives logon details for uploading data to the BHL-Europe server. For this purpose this section will demonstrate exemplary how to use the FTPS client Filezilla. Filezilla is used as an example and you are free to use any client that supports a FTPS connection.
You can get Filezilla at this address http://filezilla-project.org/download.php. Once downloaded and installed, you can connect to the BHL-Europe server to upload files.

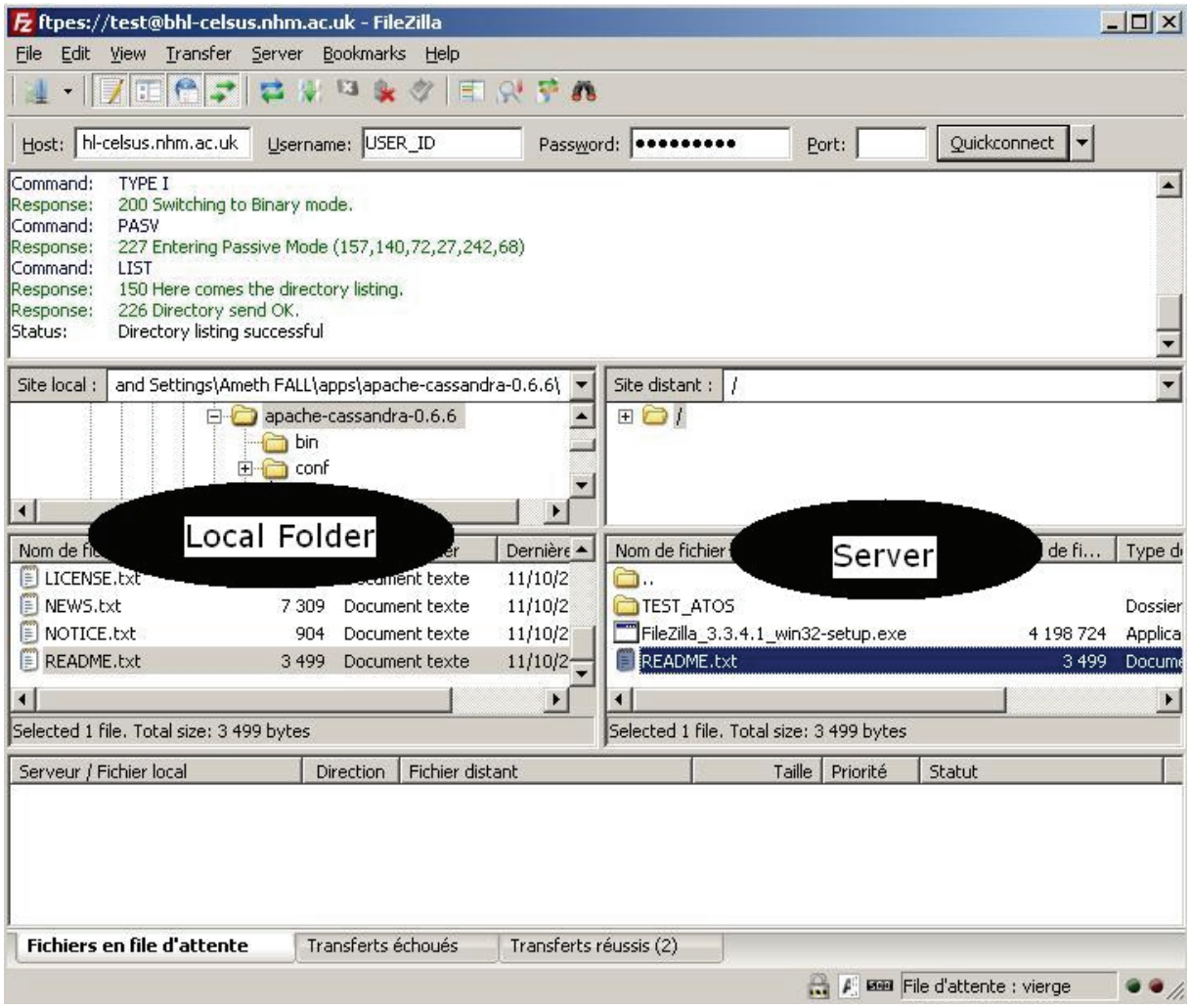
Connection with Filezilla:

Launch Filezilla and fill in the host field with ftpes://bhl-celsus.nhm.ac.uk, the login, and password field like this screen below using your username and password.



Uploading File

Once connected we can start upload files using drag and drop from your local folder to the server.



After the upload you can disconnect to the server using the red cross above the username field or just close the application.