



ECP-2008-DILI-518001

BHL-Europe

Technical Note

Pre-Ingest File Submission Guidelines

Deliverable number	<i>TN-SPRINT03-314</i>
Dissemination level	<i>Confidential</i>
Delivery date	<i>October 2010</i>
Status	<i>Draft</i>
Author(s)	<i>Bernd Sproger</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.

Table of contents

1	DOCUMENT HISTORY	3
1.1	CONTRIBUTORS.....	3
1.2	REVISION HISTORY	3
1.3	DISTRIBUTION.....	3
2	FILE SUBMISSION GUIDLINES.....	4
2.1	GENERAL RULES FOR DIRECTORY AND FILE NAMES	4
2.2	DIRECTORY STRUCTURE AND FILENAME PATTERN.....	4
2.3	EXAMPLES	5
3	FAQ.....	6
4	OPTIONAL: PACKAGE INFORMATION METADATA.....	7

1 Document History

1.1 Contributors

A discussion about the specifications was initiated and the following persons provided input that was used for the present document.

Person	Partner
Bernd Sproger	AIT

1.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
2010-09-28	Bernd Sproger	0.1	1. Draft

1.3 Distribution

This document has been distributed to:

Group	Date of issue	Version
BHL-E content providers (bhl-e.cp@lists.hu-berlin.de)	2010-10-19	0.1
BHL tech group (bhle-tech@googlegroups.com)	2010-10-19	0.1
BHL-E WP3	2010-10-19	0.1
BHL WIKI (https://bhl.wikispaces.com/BHL-E_WP3_PREINGEST)	2010-10-19	0.1

2 File Submission Guidelines

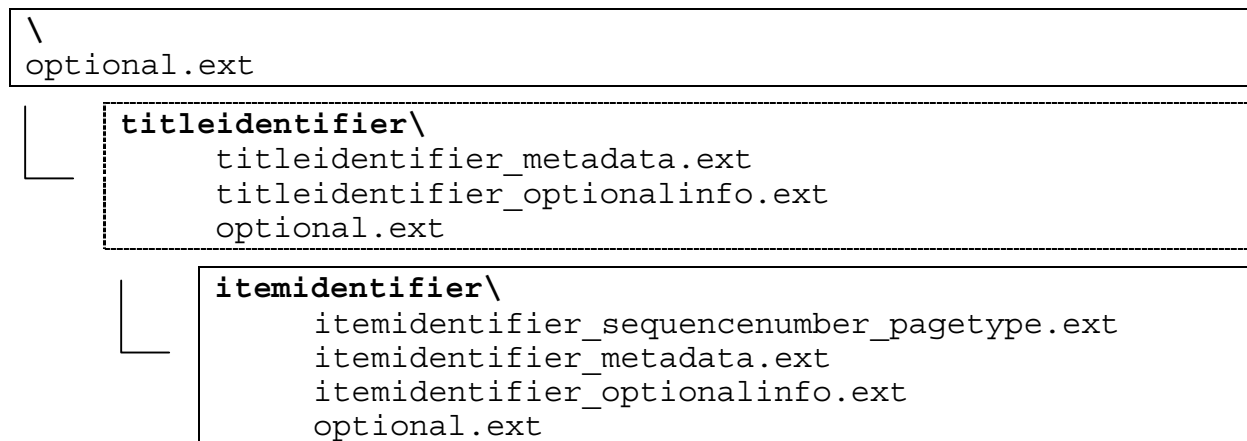
The file submission guidelines are necessary to allow automatic processing of submitted content and metadata. If you can't adhere to the file submission guidelines, please send an email to kochg@ait.co.at.

2.1 General Rules for Directory and File Names

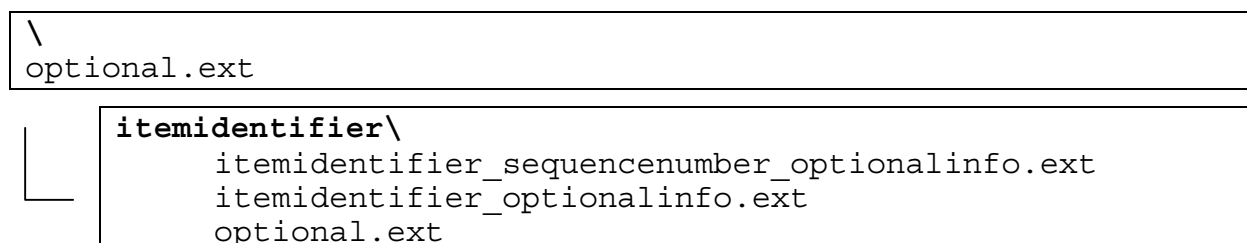
- Preferably use only ASCII characters and Western/Arabic numbers (0-9)
- Don't use: <, >, ", /, |, ?, *
- Avoid to use blank spaces
- Lower case file names are preferred ("abc" instead of "Abc")
- Use three-letter extensions for file names (.tif for TIFF images, .jpg for JPEGs, etc.)

2.2 Directory Structure and Filename Pattern

The directory structure aims to reliably identify titles and items and to reliably identify what files belong to which title or which item. You can create multiple folders for your items and multiple folders for your titles. Item folders are nested inside a title folder. Scanned images always belong to items:



If an item has no title, the titleidentifier directory must be omitted:



2.3 Examples

An item consists of a directory with multiple image files and one metadata file:

```
301519\  
  301519_0001.tif  
  301519_0002.tif  
  301519_metadata.xml
```

An item consisting of a directory with multiple image files, metadata files, and others:

```
301519\  
  301519_0001_cover.tif  
  301519_0002_blank.tif  
  301519_0003_title.tif  
  [...]  
  301519_0899.tif  
  301519_metadata.mrc  
  301519_article_01_metadata.mrc  
  301519_article_02_metadata.mrc  
  301519_articles_start_end_fileinfo.xml  
  301519_scanning_info.xml  
  301519_document_ocr.pdf  
  general_info.txt  
  thumbs.db  
  bag-info.txt
```

A title may consist of multiple items. Items could contain other items such as articles:

```
gartenkal\  
gartenkal_title.xml  
  821\  
    821_0001_cover.tif  
    821_0002_blank.tif  
    821_0003_title.tif  
    [...]  
    821_1002_page_999.tif  
    821_metadata.mrc  
    821_scanning_info.xml  
  831\  
    [...]  
    831_01\ [...]  
    831_02\ [...]  
  834\  
    [...]  
    834_01\ [...]  
    834_02\ [...]
```

Files that don't belong to titles or items need to be part of the top-level directory:

```

\
general_info.txt
all_metadata_database.sql
  gartenkal\
  gartenkal.mrc
    821\
      [...]
    831\
      [...]
    834\
      [...]
  magnaturkdehelvet\
  magnaturkdehelvet.mrc
    871\
      [...]
    881\
      [...]
    891\
      [...]

```

Example of a title where there are no separate item metadata files:

```

gartenkal\
gartenkal_title_and_all_items_metadata.xml
  821\
    821_0001_cover.tif
    821_0002_blank.tif
    821_0003_title.tif
    [...]
    821_1002_page_999.tif
  831\
    [...]

```

3 FAQ

- My local identifiers (URNs) are using special characters which can't be used in filenames or directory names. What should I do?

We understand that URNs are a popular way to organise identifiers because they add namespaces to make identifiers more globally unique. URNs contain “:” which can't be used on Windows filesystems. Please exclude the URN prefix containing your namespace, e.g.: urn:nbn:de:gbv:089-3321752945 becomes 089-3321752945. Please always inform us about such steps, so we can correctly process your submissions afterwards.

4 Optional: Package Information Metadata

Attention: Following information is optional only and the intended audience are content providers with a strong technical background.

Submissions are going to be packaged once they are uploaded, and package information metadata will be generated and included. This prevents unwanted modification of the package payload (submitted content and metadata) and enables further processing of data and enrichment of metadata. We are using a software library developed by the Library of Congress called BagIt¹. If you or your digitisation provider are able to include package information metadata with your submission, please add a file called “bag-info.txt” to your items directory containing:

<Element>: <Content>

```
Source-Organization: AIT Angewandte Informationstechnik
Organization-Address: Klosterwiesgasse 33, Graz
Contact-Name: Dr. Walter Koch
Contact-Phone: +43 316-835359-74
Contact-Email: admin@ait.co.at
External-Description: old german book in Gothic print
Bagging-Date: 2010-10-01
External-Identifier: cat00098
Bag-Size: 2 GB
```

For your convenience we append the section 4.2 about BagIt metadata extracted from the BagIt v0.96 specification paper:

4.2. Other bag metadata: bag-info.txt

The "bag-info.txt" file is a tag file that contains metadata elements describing the bag and the payload. The metadata elements contained in the "bag-info.txt" file are intended primarily for human readability. All metadata elements are optional. A metadata element consists of a label, a colon, and a value. Whitespace after the first non-whitespace in the value is considered part of the value. Long values may be folded (continued) onto the next line by inserting a newline (LF), a carriage return (CR), or carriage return plus newline (CRLF) and indenting the next line (any combination of spaces and tabs). It is recommended that lines not exceed 79 characters in length. Reserved metadata element names are case-insensitive and defined as follows.

Source-Organization

Organization transferring the content.

Organization-Address

Mailing address of the organization.

Contact-Name

Person at the source organization who is responsible for the content transfer.

Contact-Phone

¹ <https://confluence.ucop.edu/display/Curation/BagIt>

International format telephone number of person or position responsible.

Contact-Email

Fully qualified email address of person or position responsible.

External-Description

A brief explanation of the contents and provenance.

Bagging-Date

Date (YYYY-MM-DD) that the content was prepared for delivery.

External-Identifier

A sender-supplied identifier for the bag.

Bag-Size

Size or approximate size of the bag being transferred, followed by an abbreviation such as MB (megabytes), GB, or TB; for example, 42600 MB, 42.6 GB, or .043 TB. Compared to Payload-Oxum (described next), Bag-Size is intended for human consumption.

Payload-Oxum

The "octetstream sum" of the payload, namely, a two-part number of the form "OctetCount.StreamCount", where OctetCount is the total number of octets (8-bit bytes) across all payload file content and StreamCount is the total number of payload files. Payload-Oxum is easy to compute (e.g., on Unix "wc -lc `find data/ -type f`") and should be included in "bag-info.txt" if at all possible. Compared to Bag-Size (above), Payload-Oxum is intended for machine consumption.

Bag-Group-Identifier

A sender-supplied identifier for the set, if any, of bags to which it logically belongs. This identifier must be unique across the sender's content, and if recognizable as belonging to a globally unique scheme, the receiver should make an effort to honor reference to it.

Bag-Count

Two numbers separated by "of", in particular, "N of T", where T is the total number of bags in a group of bags and N is the ordinal number within the group; if T is not known, specify it as "?" (question mark). Examples: 1 of 2, 4 of 4, 3 of ?, 89 of 145.

Internal-Sender-Identifier

An alternate sender-specific identifier for the content and/or bag.

Internal-Sender-Description

A sender-local prose description of the contents of the bag.

In addition to these metadata elements, other arbitrary metadata elements may also be present.