Data Upload to Internet Archive

AIT Forschungsgesellschaft mbH

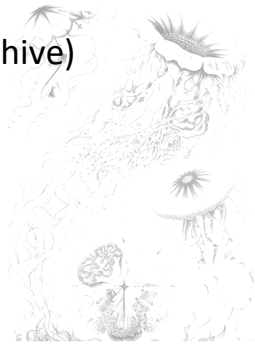
Biodiversity Heritage Library Europe

co-funded by the Community programme eContent*plus*

europeana project group member

Outline:

- Prepration to Upload to IA (Internet Archive)
- Upload
- Page Numbering



co-funded by the Community programme eContent*plus*

europeana project group member

Preparation (1)

- Before you upload an item to IA you should give your item an unique identifier (alphanumeric) which will be used for naming conventions. One format is
  [title:8-16][vol:2][author:4][scanninglocation:0-4]

- The naming of the files depends on the type of files you want to upload.

co-funded by the Community programme eContentplus

Preparation (2)

- For our test we had the single pages in TIFF format and the MARC record.

Name	Type
{IDENTIFIER}_tif.zip	Single Page Processed TIFF ZIP
{IDENTIFIER}_meta.mrc	MARC Binary Description
{IDENTIFIER}_marc.xml	MARC XML Description

- The structure of the {IDENTIFIER}_tif.zip file must be: {IDENTIFIER}_tif/

  {IDENTIFIER}_0001.tif
  {IDENTIFIER}_0002.tif
  …
  {IDENTIFIER}_0273.tif

co-funded by the Community programme eContentplus

Uploading (1)

- You need an account at http://www.archive.org by clicking on "join us" and fill out the form.
  - After you're logged in you press "Upload"
  - In the first form you fill in your {IDENTIFIER}
  - The system checks if the {IDENTIFIER} is free
    - If the {IDENTIFIER} is not free you'll have to take another one and rename all files to match the new {IDENTIFIER}
  - If the {IDENTIFIER} is free you will get a new form in which you just fill out the required fields, add the files for upload and define the item type (e.g. "texts")
  - After that you press "Upload files"
  - Wait until the upload are finished and you'll get the URL of your item.
  - Depending of the files you uploaded the system will try to derive some other formats, see http://www.archive.org/help/derivatives.php

Uploading (2)

- You can check which task are running with the "Item Manager". The "Item Manager" can be accessed over the "Edit Item"
  - http://www.archive.org/details/{IDENTIFIER}
  - Click on "Edit Item" (you must be logged in and the uploader of the item)
  - Click on "change the information"
  - Click on "Item Manager"
  - Click on "show history for item" in the "Miscellaneous Operations" section
  - Now you'll see all the task of the item which are running/pending/finished/…
  - By clicking on a "task id" you will see the log of the task

Uploading (3)

- Following files could be generated by just uploading the single TIFFs and the MARC record:

Name	Type	Description
{IDENTIFIER}_dc.xml	DublinCore	derived from MARC
{IDENTIFIER}.gif	Animated GIF	derived from "Single Page Processed TIFF ZIP"
{IDENTIFIER}_flippy.zip	Flippy ZIP	derived from "Single Page Processed TIFF ZIP"
{IDENTIFIER}_abbyy.gz	Abbyy GZ	derived from "Single Page Processed TIFF ZIP"
{IDENTIFIER}_djvu.xml	Djvu XML	derived from "Abbyy GZ"
{IDENTIFIER}_scandata.xml	Scandata	derived from "Djvu XML"
{IDENTIFIER}.djvu	DjVu	derived from "Djvu XML"
{IDENTIFIER}.pdf	Text PDF	derived from "Djvu XML"
{IDENTIFIER}_djvu.txt	DjVuTXT	derived from "DjVu"

Page Numbering (1)

- After letting IA finishing the derive process we updated the {IDENTIFIER}_scandata.xml to get correct pagenumbers with the online reader.
- Just download the file and edited it with a normal texteditor.

Page Numbering (2)

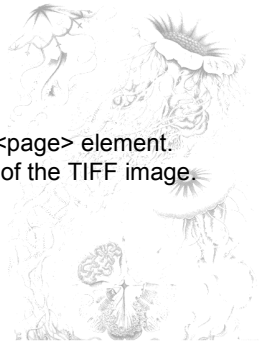
```
<book>
 <bookData>
  <bookId>{IDENTIFIER}</bookId>
  <leafCount>273</leafCount>
 </bookData>
 <pageData>
  <page leafNum="1">
   [...]
  </page>
  <page leafNum="2">
   [...]
  </page>

  [...]

  <page leafNum="273">
   [...]
  </page>
 </pageData>
</book>
```

The xml format is quite simple.

We are just interested in the <page> element.
The "leafNum" is the number of the TIFF image.

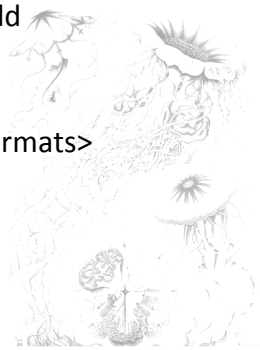


Page Numbering (3)

As the first image was the **spine of the book** we don't want it to be displayed with the online reader, we set/add

```
<pageType>Delete</pageType>
<addToAccessFormats>false</addToAccessFormats>
```

to the element <page leafNum="1">.

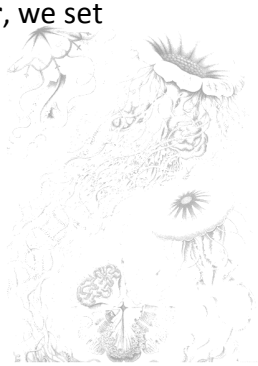


Page Numbering (4)

The second and the last image are the **bookcover**, we set

<pageType>Cover</pageType>

to the element <page leafNum="2"> and <page leafNum="273">.

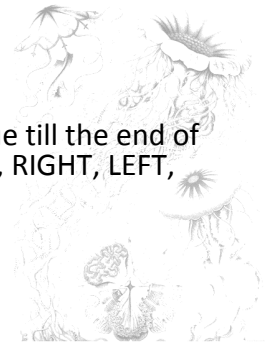


Page Numbering (5)

The second image is also the **start of the book**, we add

<handSide>RIGHT</handSide>

to the element <page leafNum="2"> and continue till the end of the book with alternating values (RIGHT, LEFT, RIGHT, LEFT, ...).



Page Numbering (6)

The fourth image is the **first page** of the book, we add

<pageNumber>1</pageNumber>

to the element <page leafNum="4"> and continue till the end of the book with ascending values (1, 2, 3, 4, ...).

co-funded by the Community programme eContentplus

Page Numbering (7)

The fifth image is the **title page** of the book, we set

<pageType>Title</pageType>

to the element <page leafNum="5">. This will be the start page the online reader will show.

co-funded by the Community programme eContentplus