

# **Procedure for Creating an Archive for Magazines in PDF Format in the Internet Archive (Archive.org) by Gardner Patton**

First format all the issues of your magazine in .PDF format with one issue per file. Rename the files properly for upload. Each file should have the name of the magazine (separate words with underlines) and the year and month of publication. If you don't have a year and month use the Vol and No.. If you don't have that, number them sequentially. Thus the file name would be "magazine\_name\_yearmonth.pdf" or "magazine\_name\_volxxserxx.pdf". The yearmonth or volser sequence allows for a date order sort of the files.

Now that the files are organized and named there are 2 ways to upload them. One way is to load each one as a single item with one magazine in each item. The second way is to have a single item (which will become a collection of issues) with each .PDF as a file within that item. Note we are assuming a .PDF file, not individual .jpg images which is often the case when an item represents one issue.

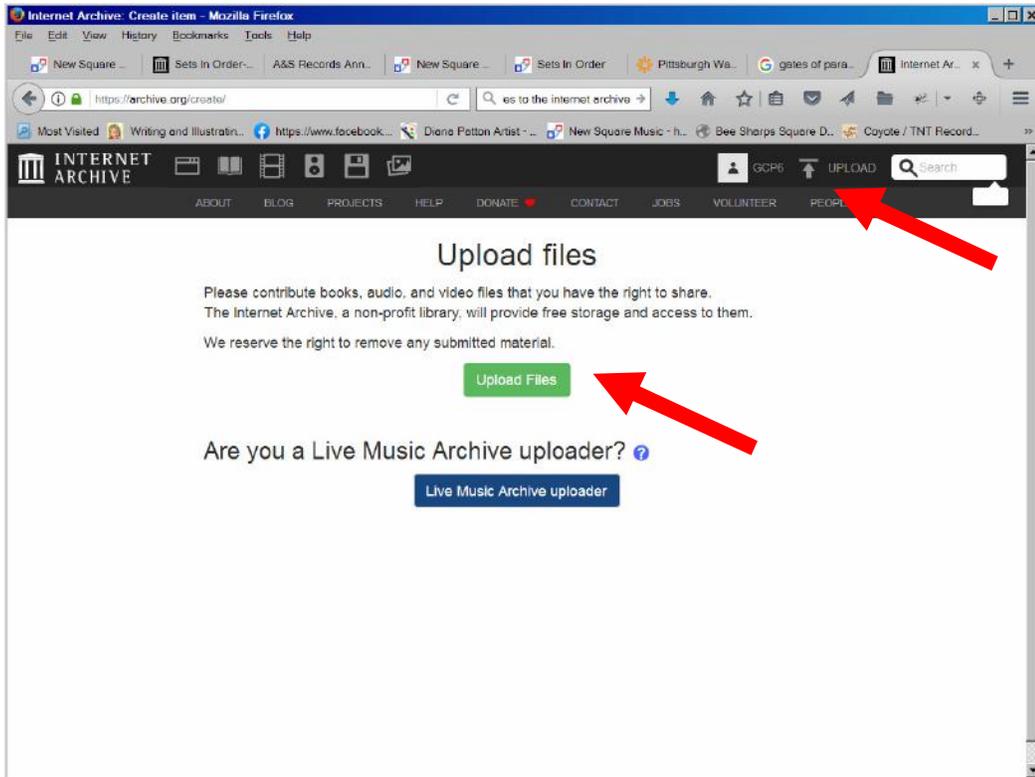
The first method is the more traditional way and there is meta data for each issue which contains among other things the issue date of the magazine and the number of pages. Also with this method the first page of the issue is preserved and shown by the reader. With this method however the meta data must be supplied for each item. This can be speeded up by using parameters in the command used to invoke the uploader page but that information may need to be modified slightly for each issue (date changes, etc.).

Using the second method the meta data is only entered once for all the issues. Thus you lose the opportunity to enter the creation date of the issue and the number of pages. However, if you have named the files with the creation date, that date is easily available to viewers. The number of pages is rarely significant. With the second method you can queue up hundreds of files at a time for upload. The upload times for each file seem about as fast as FTP. Using the procedure to add files to an existing item (described below), an upload of a large number of issues can be broken into manageable sessions. However, if you upload hundreds of large files at one time the Archive may take many days to process them all (create derivative files). In the meantime they are available as normal .PDF files with the permalinks assigned as they are uploaded.

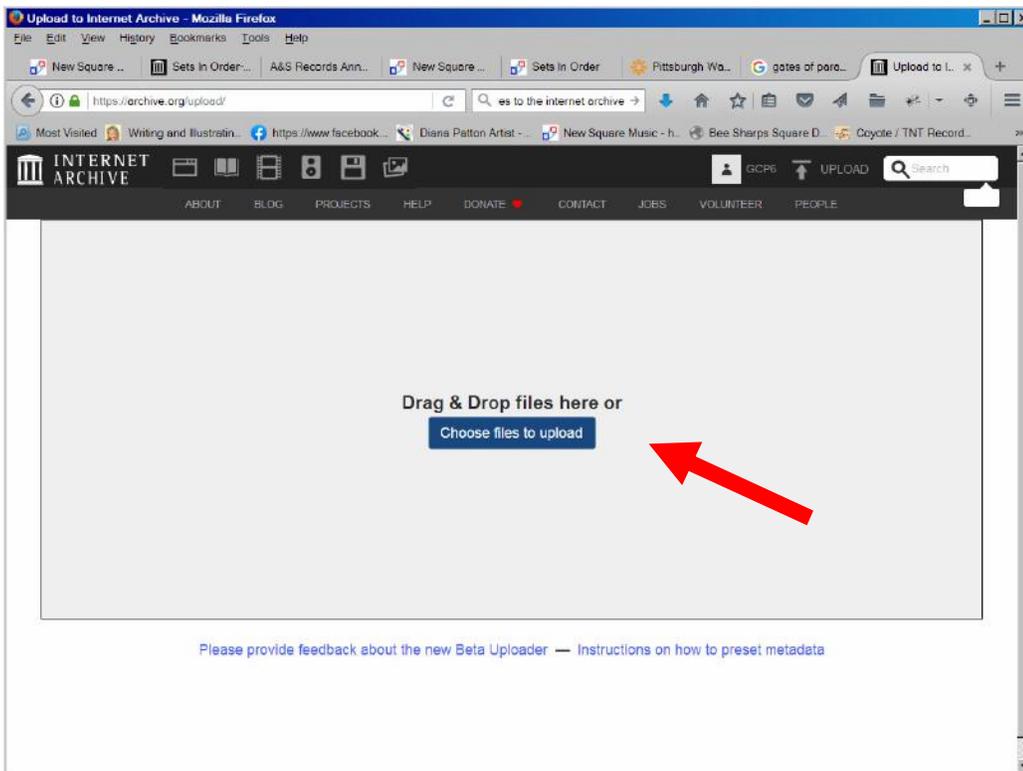
For these reasons I prefer to use the second method for uploading multiple .PDF files each containing one issue of a magazine, a whole collection of the same magazine in one or several uploads, but not one upload for each issue.

1) To do this log in the Archive and click the *Upload* button at the top of the page. See image on next page.

## Upload PDF files to Internet Archive



2) Click *Upload Files*.



3. Drag 1 or more files to the pink box that you want to upload.

## Upload PDF files to Internet Archive

The screenshot shows the 'Upload to Internet Archive' page in Mozilla Firefox. The browser address bar shows 'https://archive.org/upload/'. The page title is 'INTERNET ARCHIVE'. The navigation menu includes ABOUT, BLOG, PROJECTS, HELP, DONATE, CONTACT, JOBS, VOLUNTEER, and PEOPLE. The main content area is divided into two sections: 'Click on any field below to edit it' and 'Drag and Drop More Files Here or Select files to add'. The 'Click on any field below to edit it' section contains the following fields:

- Page Title \***: Sets In Order 194901 (checked)
- Page URL \***: https://archive.org/details/sets-in-order-194901 (checked)
- Description \***: Add a description of the item page (checked)
- Subject Tags \***: Add keywords, separated by commas (checked)
- Creator**: Creator of the content (checked)
- Date**: Date work was created/published (checked)
- Collection \***: Community texts (checked)
- Test Item**: No (checked)
- Language**: Language of the work (checked)
- License**: No license selected (checked)
- More Options**: Add additional metadata...

The 'Drag and Drop More Files Here or Select files to add' section shows a table of files:

Name	Size	x
Sets_in_Order_194811.pdf	5.8 MB	x
Sets_in_Order_194901.pdf	6.1 MB	x

A red arrow points to the 'Page Title' field.

Meta data window appears. Modify page title as needed in page title box. This is the title for the whole collection of issues so it should have magazine name and date range of issues.

The screenshot shows the 'Upload to Internet Archive' page in Mozilla Firefox. The browser address bar shows 'https://archive.org/upload/'. The page title is 'INTERNET ARCHIVE'. The navigation menu includes ABOUT, BLOG, PROJECTS, HELP, DONATE, CONTACT, JOBS, VOLUNTEER, and PEOPLE. The main content area is divided into two sections: 'Click on any field below to edit it' and 'Drag and Drop More Files Here or Select files to add'. The 'Click on any field below to edit it' section contains the following fields:

- Page Title \***: Sets In Order/Square Dancing (checked)
- Page URL \***: https://archive.org/details/bobsogood (checked)
- Description \***: Add a description of the item page (checked)
- Subject Tags \***: Add keywords, separated by commas (checked)
- Creator**: Creator of the content (checked)
- Date**: Date work was created/published (checked)
- Collection \***: Community texts (checked)
- Test Item**: No (checked)
- Language**: Language of the work (checked)
- License**: No license selected (checked)
- More Options**: Add additional metadata...

The 'Drag and Drop More Files Here or Select files to add' section shows a table of files:

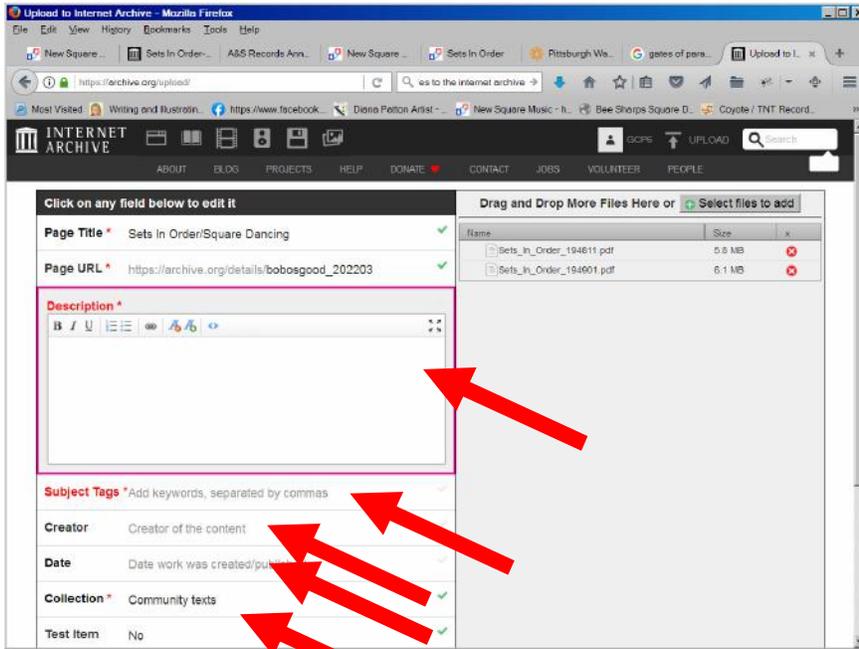
Name	Size	x
Sets_in_Order_194811.pdf	5.8 MB	x
Sets_in_Order_194901.pdf	6.1 MB	x

A red arrow points to the 'Page URL' field.

Put unique id in URL box. It will become the Identifier for all the issues and become the next text in the created URLs. The last field of the URL will be the full file name.

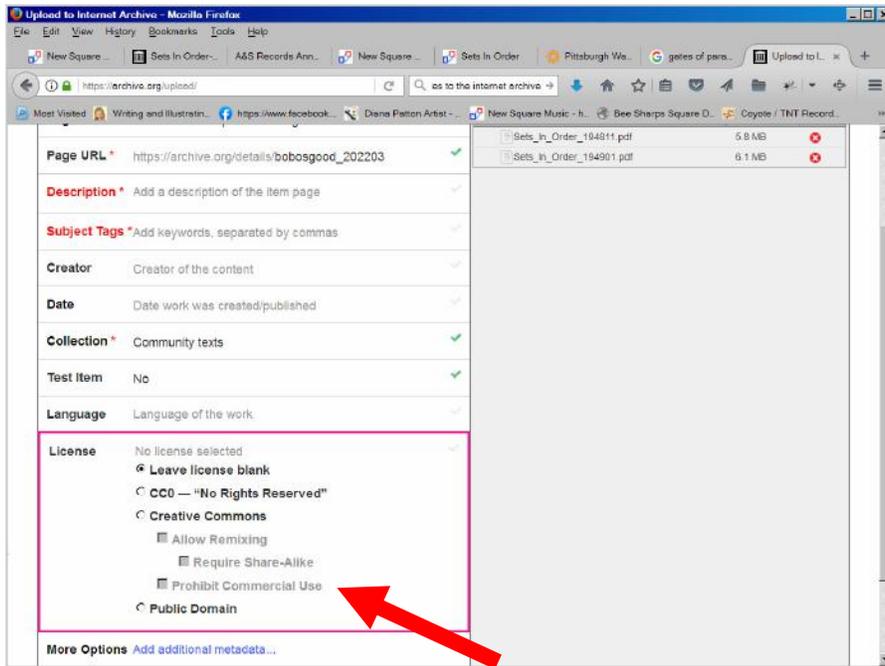
## Upload PDF files to Internet Archive

Since the URL field had bobosgood, the first .PDF file's final full permalink URL will be: [https://arhive.org/download/bobosgood/Sets\\_in\\_Order194811.pdf](https://arhive.org/download/bobosgood/Sets_in_Order194811.pdf).

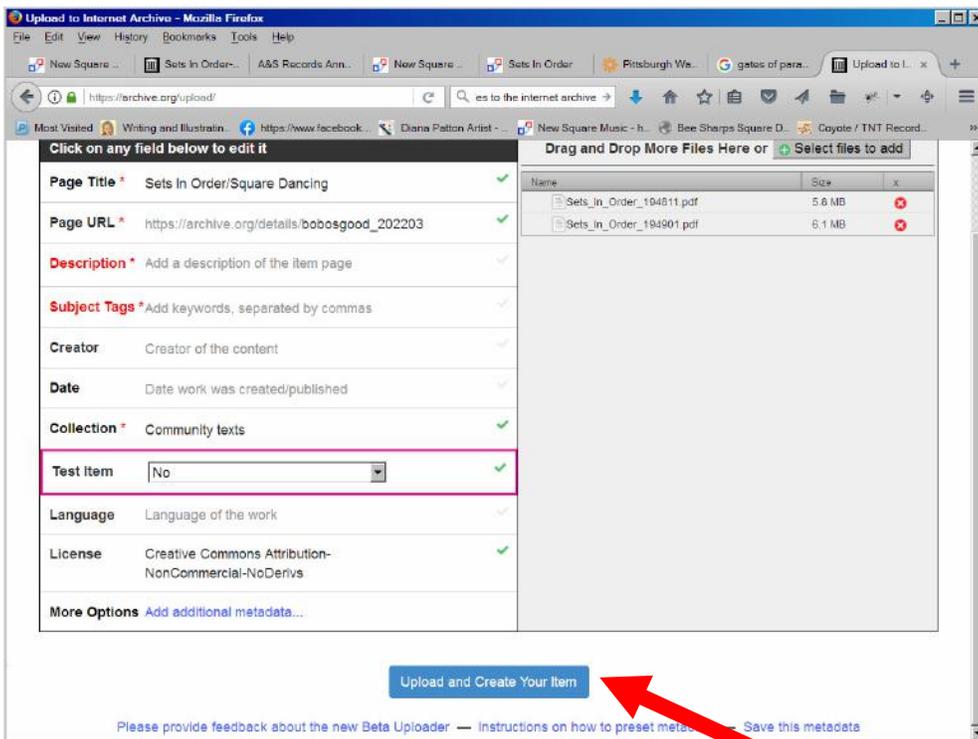


Add the description by clicking on the line to open the box. Next add *Subject Tags*. Carefully consider what to put in these two fields since they will be particularly useful to people discovering items you've uploaded if they're doing a general search but don't already know the title of the magazine. Then put in the person or organization who created the magazines. Put in the date of the first magazine or leave blank. Leave the default collection since initially they offer you a limited selection. After your issues are loaded, however, contact an administrator at [info@archive.org](mailto:info@archive.org) and request your item be added to the collection you think is best or a new collection you suggest. Select the appropriate language from the language box.

## Upload PDF files to Internet Archive



Select a license. The Creative Commons, Attribution, NonCommercial, No derivatives seems to be the most protective. Be careful because you can give away your copyright rights here. Also make sure you have the rights to give away if the magazines are not your creation.

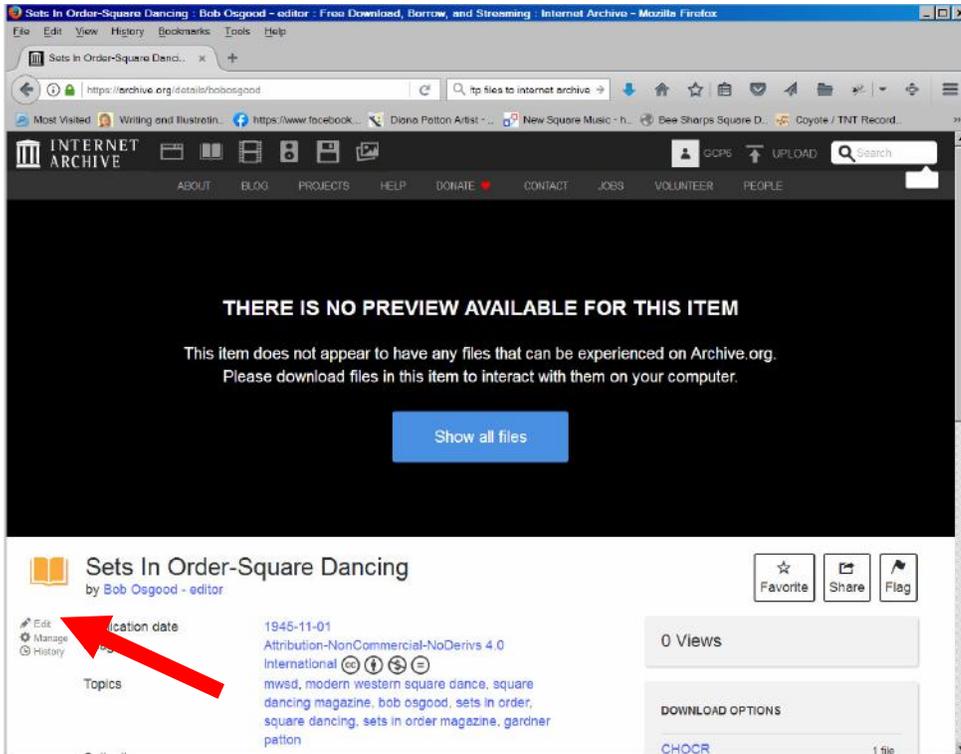


Click the *Upload and Create Your Item* button.

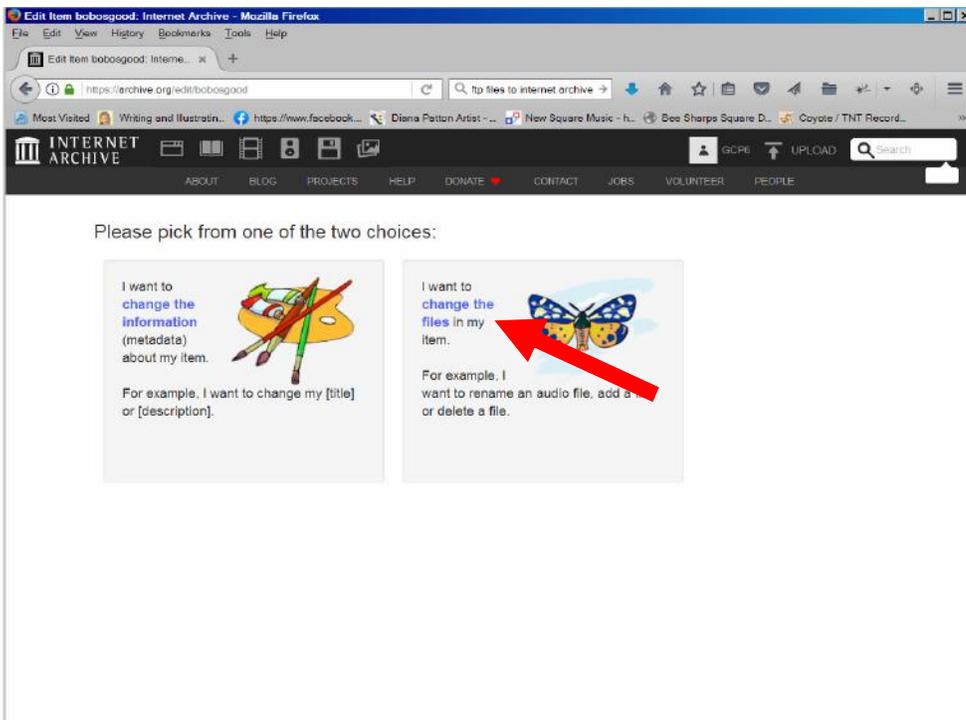
## Upload PDF files to Internet Archive

### Procedure to add files to an existing item

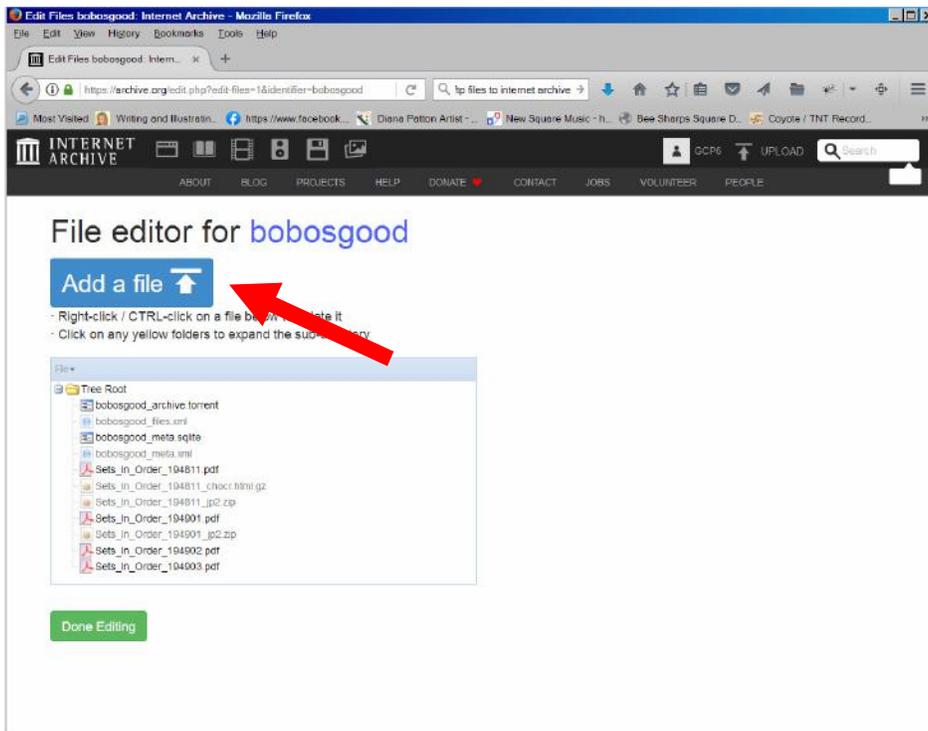
First set up your files so each .pdf file has a name similar to the ones already in the item. Go to the page for the item.



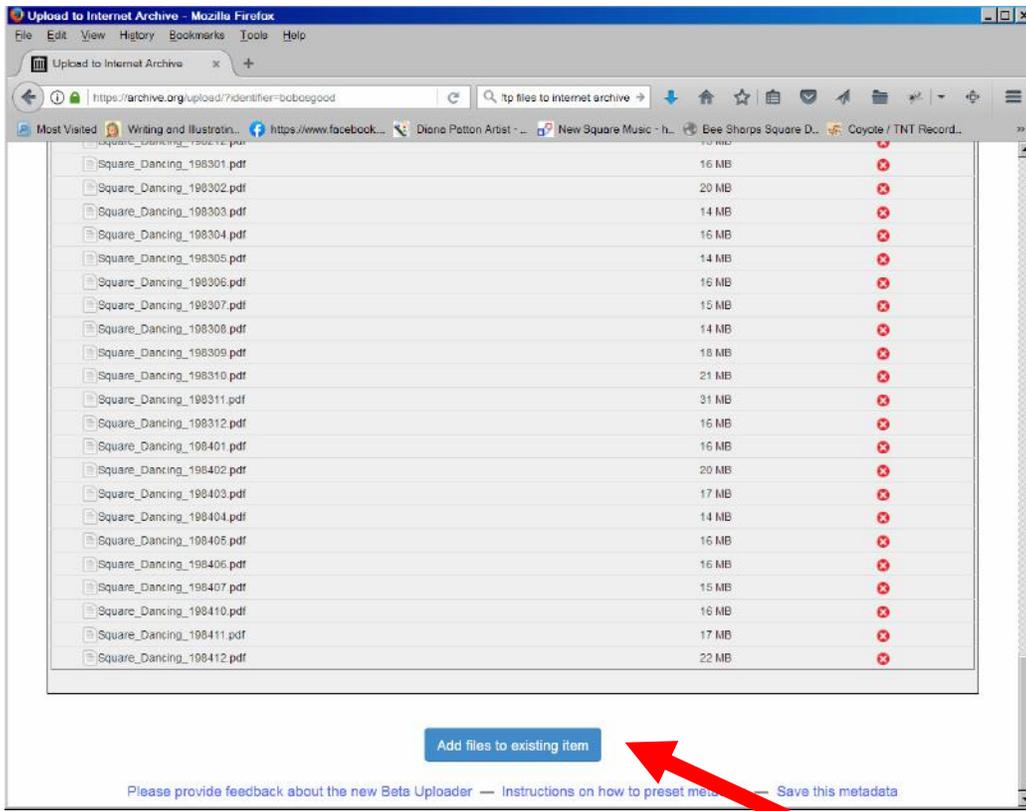
Click on *Edit*.



Upload PDF files to Internet Archive  
Click on *change the files*.

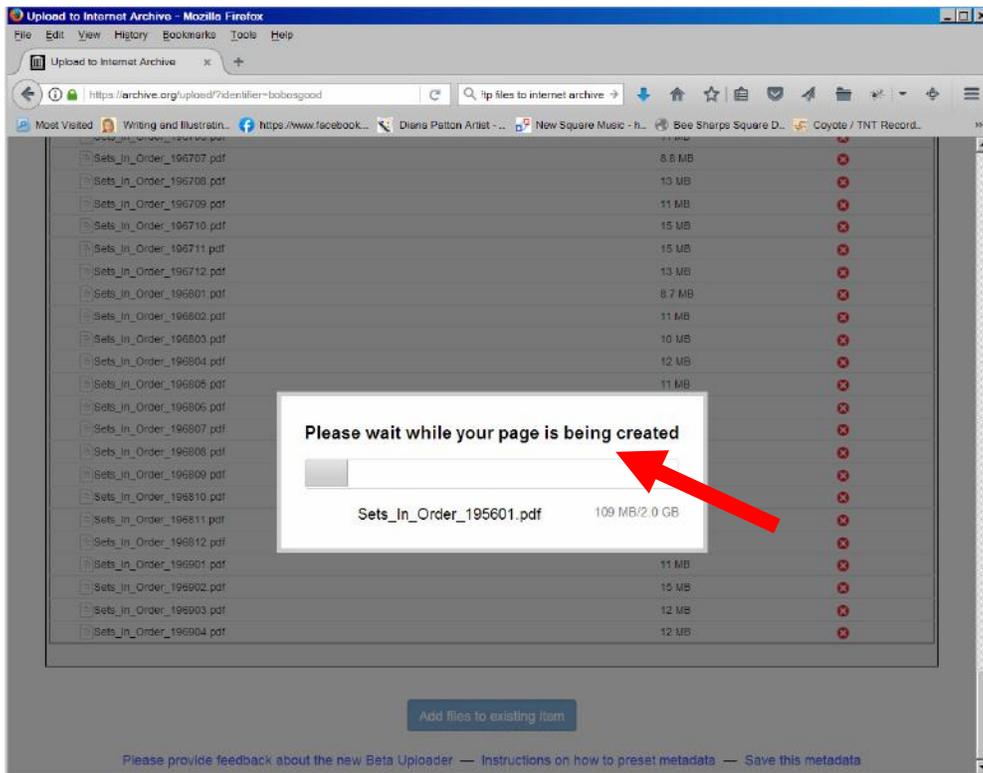


Click on *Add a File*.



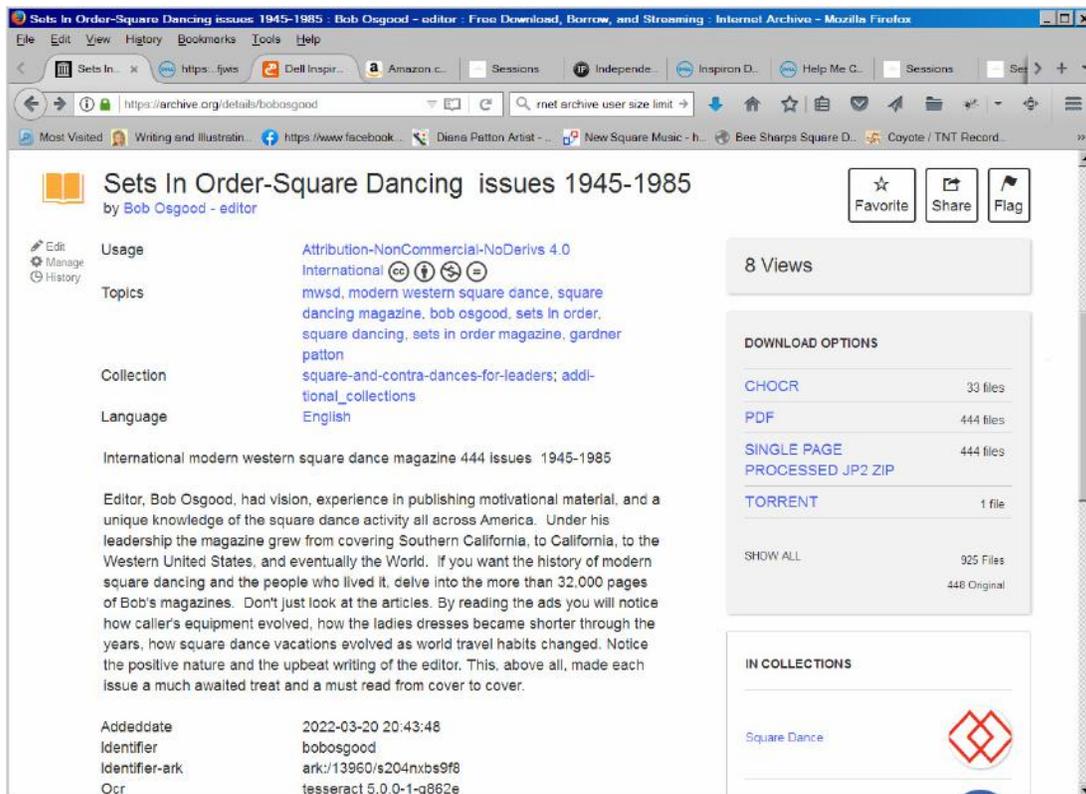
## Upload PDF files to Internet Archive

Drag over the new files to add. Click *Add Files To Existing Item*.



Wait for files to be uploaded.

Once the files are uploaded, you can create your own index to the pages using the pdf files shown if you click the PDF file



## Upload PDF files to Internet Archive

The screenshot shows the Internet Archive interface for the collection 'Sets In Order-Square Dancing issues 1945-1985' by Bob Osgood. The page includes a metadata section on the left, a main description, and a list of PDF files on the right. A red arrow points to the first file in the list, 'Sets\_In\_Order\_194811.pdf', which is 5.9M in size.

File Name	Size
<a href="#">Sets_In_Order_194811.pdf</a>	5.9M
<a href="#">Sets_In_Order_194901.pdf</a>	7.1M
<a href="#">Sets_In_Order_194902.pdf</a>	9.3M
<a href="#">Sets_In_Order_194903.pdf</a>	6.2M
<a href="#">Sets_In_Order_194905.pdf</a>	4.9M
<a href="#">Sets_In_Order_194906.pdf</a>	5.8M
<a href="#">Sets_In_Order_194907.pdf</a>	5.4M
<a href="#">Sets_In_Order_194908.pdf</a>	5.0M
<a href="#">Sets_In_Order_194909.pdf</a>	4.9M
<a href="#">Sets_In_Order_194910.pdf</a>	10.2M
<a href="#">Sets_In_Order_194911.pdf</a>	9.3M
<a href="#">Sets_In_Order_194912.pdf</a>	10.8M

In picture above a list with the .PDF file names is shown. These names are pasted on to the standard Internet Archive download URL to create the link for your index. The full link would be [https://archive.org/download/bobosgood/Sets\\_In\\_Order\\_194811.pdf](https://archive.org/download/bobosgood/Sets_In_Order_194811.pdf). The full link can be copied to the clipboard with ctrl-C while hovering over the link.

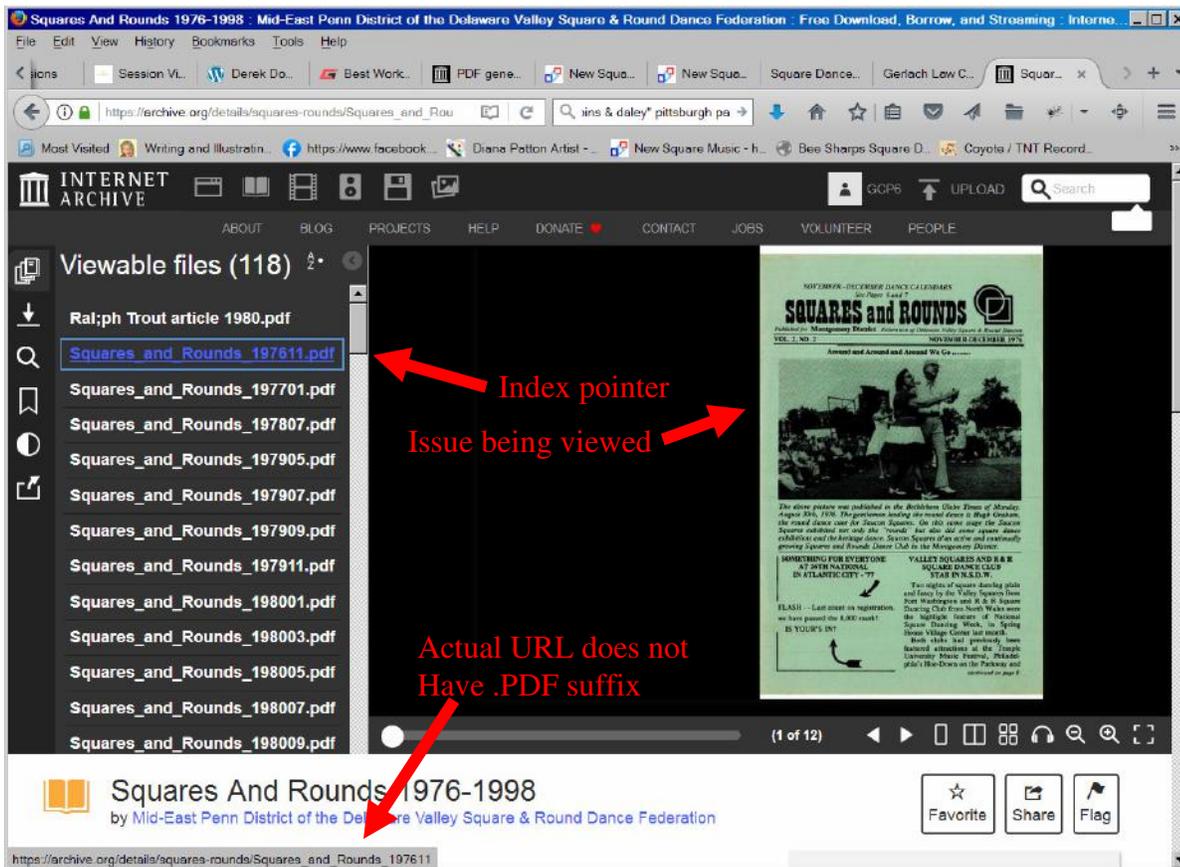
Upload PDF files to Internet Archive

## Internet Archive Derivative Files for PDF's

Once you have uploaded your .PDF files, the Internet Archive will create a number of derivative files. They do this because the files they are given could have been created by a variety of different software packages which may have provided different kinds of data. For example they will create individual .JPG images of each page and do Optical Character Recognition (OCR) to pull off any text that might have been in the .JPG images and they will pull out pictures on a page and compress them for efficient transmission over the internet. In one case, I uploaded 118 files of the Magazine Squares and Rounds and after the derivative files were created there were 1185 files.

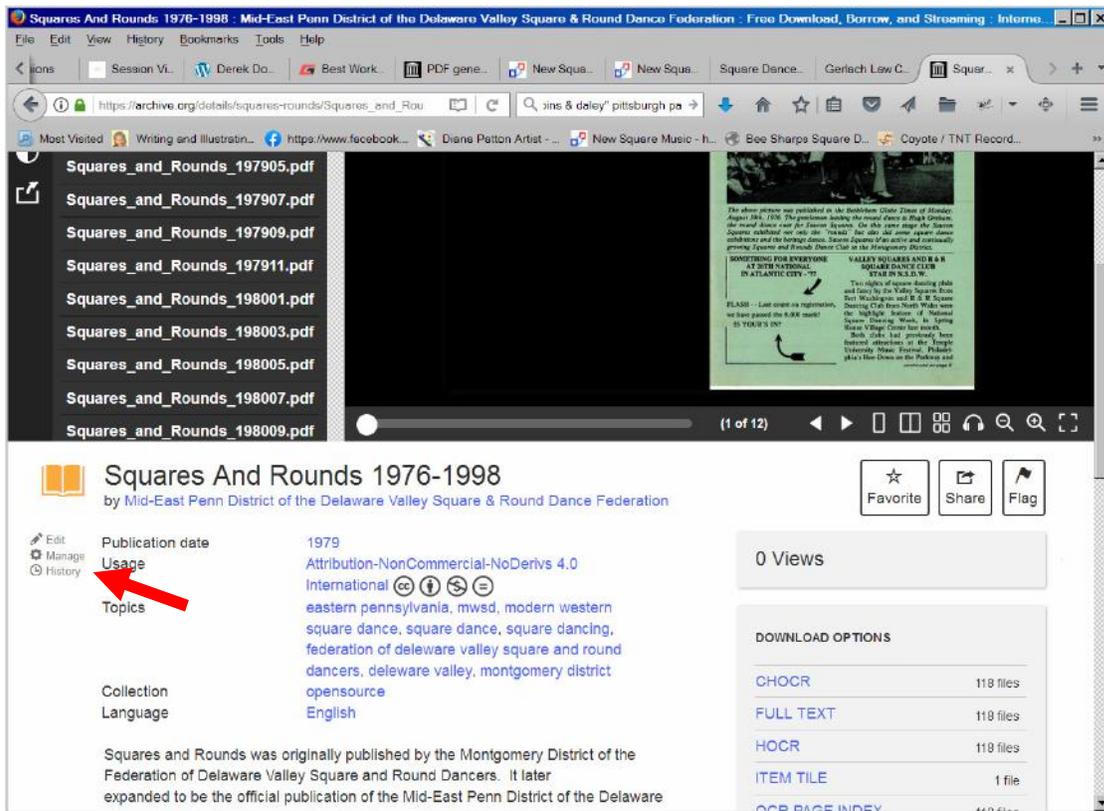
The derivative files are created by a background process which may take hours or days to process a group of input files. When I uploaded 444 issues of a magazine it took about 2 weeks to create the derivative files.

After the derivative files are created, when you view your item, there will be an index pointer to each issue in your item. The pointers are shown in the Internet Archive viewer window. When you click on an issue in the index, that issue will be shown in the viewer.



If you don't see the index it means that the derivative files are still being created or that there is a problem in their creation. You can click on the history button (see diagram below) to see what processing has been done.

## Upload PDF files to Internet Archive



In my case, my the .PDF files I uploaded had been created using PDF Converter Pro and had a text layer created by that program's OCR software. It had also been processed with MRCS so the images had been extracted and separated from the text. Then the files had been optimized for fast internet download. Thus the Internet Archive process duplicated most of what I had done to get my files into their archive required format.

In one Internet Archive post, I found that the Internet Archive does not modify the original .PDF files. If there is a text layer in the .PDF file then the file is searchable. If there is no text layer then they create a derivative file with their OCR to make the file searchable when viewed with their viewer. This would indicate to me that if you are going to pointing to the .PDF file which you uploaded, and your index link will cause it to be downloaded, it should already have been made searchable before it is uploaded. Note, text searching is only as good as the OCR. If the OCR does not read the characters correctly, a search for words you may be able to recognize eye might not may not be found if the OCR recognition is bad.

So the question is, if you create an index to the files you saved in the Archive what link should the index point to?

If you have an issue, for example Squares and Rounds for November 1976 and you point to [https://archive.org/details/squares-rounds/Squares and Rounds 197611](https://archive.org/details/squares-rounds/Squares_and_Rounds_197611) the archive will pop up their viewer with that issue showing in that viewer where you can page through, or search for text in the issue. Your search will use their derived files.

If you point instead to [https://archive.org/download/squares-rounds/Squares and Rounds 197611.pdf](https://archive.org/download/squares-rounds/Squares_and_Rounds_197611.pdf)

Upload PDF files to Internet Archive

(note: the action keyword in the URL is download not details and there is a .PDF suffix is at the end) you will get your original .PDF file downloaded with whatever features (search, MRCS, etc.) you had included in it when it was uploaded.

## **Constraints**

As of June 4, 2021 one Internet Archive page said:

Currently, there is no limit on the size of files nor the number of files. However, from a systems perspective, we do not recommend files larger than 50 GBs to be uploaded or more than 1000 files, per single page.

Gcp 29 March 2022