Download Archived Magazines

History

Oh what a tangled web we weave…

If at first you don't succeed…

What a long, strange trip it's been…

And undoubtedly others.

Recently my friend Gary emailed me that issues of Galaxy magazine were available on archive.org. Great! They published a *lot* of classic science fiction stories by many of the top authors of the day such as Asimov and Heinlein. I took a look at the site[1] and decided that there were too many to want to download them by hand. So I'd write a program to do it for me. Shouldn't be a problem. Right?

When I first started to write this program, I had the hope that all the files would have a consistent naming convention. For example,

> https://archive.org/download/GalaxyMagazine-1953-11/GalaxyMagazine-1953-12.pdf
>
> https://archive.org/download/GalaxyMagazine-1953-12/GalaxyMagazine-1953-12.epub

And so on.

I should be so lucky!

Logically, this made sense. There was a directory for an issue (e.g. GalaxyMagazine-1953-12), then several files inside that directory, one for each file type (pdf, epub, etc).

But then the filenames tended to use underscores where the directory name used hyphens.

> GalaxyMagazine-1953-12/GalaxyMagazine_1953_12.pdf

But then some directories put volume and number into the name:

> http://archive.org/download/Galaxy_v22n03_1964-02

And if that weren't enough, in some cases, presumably for copyright reasons, pages seem to have been deleted for some issues. The directory name in that case might be

> http://archive.org/download/Galaxy_v35n08_1974-08_modified

And there are several others.

Also, the articles were available in several formats (pdf, epub, etc). But some of the more obscure formats were in a secondary web page that also had to be downloaded and parsed.

[1] I was familiar with the site, having downloaded many Old Time Radio (OTR) audio files from them, such as Fibber McGee and Molly, The Great Gildersleeve, My Friend Irma, and many, many more.

Finally, publishing schedules were, to say the least, erratic. Sometime the magazine came out monthly. Then later switched to bi-monthly. Then back to monthly for a while. And at times, it was a crap shoot whether an issue was published or not. Now this, while a nuisance, wasn't actually a show-stopper. I could just go through the time span when the magazine was published, and try to get every month's issue. If the web page for an issue wasn't found, fine, assume it wasn't published that month and continue.

But so much for the pipe dream for the pipe dream that it was going to be a simple case of a template URL that I could just plug in the year and month and download that file.

Overview

This document describes a program I wrote to automate downloading "all" (see later) of the Galaxy magazine issues. And after I understood the structure of how archive.org organized the files involved, I realized I could use the same approach to download other magazine collections.

Note: Archive.org has individual issues of magazines available for download, but most magazines aren't in a comprehensive collection. This program downloads only the collections; issues not in a collection must be downloaded manually.

See https://archive.org/details/pulpmagazinearchive for more information.

This program has not been tested on anything other than the collections pre-configured into the program. It ***may*** work on other collections found through the main archive.org page, but you're on your own if you try it.

This document describes some relevant notes about the program. See below for a user's guide.

Please note that the program is hardly polished. Since it's meant to be run essentially once (it's not like new issues are being added every month), I didn't bother making it (or these notes) too pretty.

The other file formats (e.g. epub, mobi, etc) seem to be a result of running the pdf image through an OCR program and then converting the result all the other file formats. Needless to say, the OCR results aren't stellar. I'll have to go back and check that the same OCR mistakes are at the same place in all the file types, but at this point I'm assuming they are. So you might want to just grab the pdf files and ignore the others.

Issues

My program works, but there are a few issues:

- The program is multithreaded to allow more than one file to be downloaded at once, making better use of bandwidth. So messages may be (are usually!) out of chronological order. I have no plans to fix this.
    - This occurs since we necessarily process many interim web pages before we can download actual issue contents. And factors such as file size, responsiveness of

the archive.org web site, etc, can (and does) lead to something other than simple linearity.

- A variant of the above concerns the number of issues of a magazine in the status messages. As we'll see below, magazine collections involve getting information about up to 75 magazines at a time, each with a web request. So a message might say that it had parsed 10 out of 75 issues, but the very next message might talk about 20 out of 150. I could have arranged things so that I got as much information as possible as early as possible so the messages made a bit more sense. But that would have slowed down the program somewhat. And again, for a one-shot (per magazine collection), it wasn't worth it.

Configuration File

The program comes with a configuration file, *DownloadMags.xml*, that currently specifies three things:

- The target directory, into which files will be saved.
- A <Mags> section that currently has URLs for several magazine collections.
- A <Formats> section that defines the filetypes (e.g. .pdf, .epub, .mobi, etc) that the program supports.

Target Directory: Where the files are placed. Note: On my system it points to my G: drive. You probably want to change this!

If you find some other magazine collections on the archive.org site, you can add them to the <Mags> section. This is simply the name of the collection (which will show up on the UI) and the URL of the collection main page.

You shouldn't have to modify the <Formats> section unless you want to delete some entries. Adding an entry will have no effect, unless there are additional file formats on collections I haven't looked at.

Output

Files are stored in the Target Directory from the configuration file. Each magazine will have its own subdirectory, and individual issues are themselves in a subdirectory defined by the year or Volume. For example,

```
 Directory of G:\LRS\#####Magazines\Computer Language\1984

08/31/2017  10:58 PM    <DIR>          .
08/31/2017  10:58 PM    <DIR>          ..
08/31/2017  11:00 PM        48,393,850 Computer_Language_Issue_01_1984-09_CL_Publications_US.pdf
08/31/2017  11:00 PM        50,381,445 Computer_Language_Issue_02_1984-10_CL_Publications_US.pdf
08/31/2017  10:59 PM        50,470,331 Computer_Language_Issue_03_1984-11_CL_Publications_US.pdf
08/31/2017  10:59 PM        49,214,863 Computer_Language_Issue_04_1984-12_CL_Publications_US.pdf
               4 File(s)    198,460,489 bytes

 Directory of G:\LRS\#####Magazines\Computer Language\1985

08/31/2017  10:58 PM    <DIR>          .
08/31/2017  10:58 PM    <DIR>          ..
08/31/2017  10:59 PM        47,756,391 Computer_Language_Issue_05_1985-01_CL_Publications_US.pdf
08/31/2017  11:01 PM        60,770,257 Computer_Language_Issue_06_1985-02_CL_Publications_US.pdf
08/31/2017  10:59 PM        55,184,800 Computer_Language_Issue_08_1985-04_CL_Publications_US.pdf
08/31/2017  10:59 PM        55,491,500 Computer_Language_Issue_09_1985-05_CL_Publications_US.pdf
08/31/2017  10:59 PM        56,006,480 Computer_Language_Issue_10_1985-06_CL_Publications_US.pdf
08/31/2017  11:00 PM        66,012,816 Computer_Language_Issue_11_1985-07_CL_Publications_US.pdf
               6 File(s)    341,222,244 bytes
```

Output File Formats

This is preliminary, but it looks like the following. The .pdf files are image scans of the pages. The fact that there are letters within those images is irrelevant. So the file is 100% graphics and each file is thus 10s of megabytes each.

As this program is really just meant for me and my friends, I haven't gone to the bother of putting in a UI to let the user the file types. Just edit the .xml file directly using Notepad or whatever.

Note that by no means does archive.org provide *collection* pages for all magazines. This program supports only batch downloading of issues in a collection. Other magazines must be downloaded manually.

Structure of Archive.org Magazine Collections

It's a bit embarrassing to admit how long it took me to work this out, but then I've only had time to work on this in dribs and drabs…

I'll use the science fiction magazine *Worlds of IF* as an example. It has 176 issues in its collection. It's main collection page is https://archive.org/details/ifmagazine.

This page (you might think) would have links to all 176 individual issues. However it's organized as multiple pages with a maximum of 75 links per page. So in this case we'll need to read 3 pages to get all the links.

We'll then have to read 176 pages to find the Table of Contents (TOC) of each issue. This in turn has links to several main formats for the issue (.pdf, .epub, .mobi and others). In addition, there is a link to a web page with secondary (often obscure) formats. So that's another 176 web pages to read.

Then comes the articles themselves. Normally you might want only the .pdf files. That would be another 176 web pages read (and these, being graphics, large, usually 10s of megabytes in size). And if you wanted, say, .epub and .mobi, this would involve another 176 + 176 = 352 web pages.

So to download the .pdf, .epub and .mobi files for *Worlds of IF*, you'd have to download 3 + 176 + 176 * 3 = 707 pages! And even if you wanted only just the .pdf files, that would add up to 355 web pages.

And while the .pdf files are large, setting up to read 179 more minor pages (collection plus TOC pages) puts a certain strain on the archive.org site. Sometimes the site can respond quickly and even multi-threading the downloads of these pages doesn't help as much as you might like.

So response time on the collections isn't as fast as you might like/expect. The good news is that normally you'd only want to download a given magazine collection once!

Performance

The original version of this program has undergone several significant modifications as I've understood the URL structure that archive.org uses for their collections et al.

In particular, there may well be hundreds (in some cases over 1,000) of web accesses required to download an entire magazine collection. And since archive.org doesn't have unlimited bandwidth available, downloading a single collection (must less all of them) could take a perhaps surprisingly long time.

Now from one POV, this wasn't a deal breaker. The magazines on archive.org are hardly having new issues added every month. So what if it might take hours to download a collection. Just leave it running overnight.

At least take comfort in knowing that web accesses are multi-threaded with many running at the same time. However, most of the files are small; it's the pdf files that are big, but they're in the minority. And archive.org has limited bandwidth. So even if you have whopping big bandwidth available, don't expect it to immediately shoot up to 1Gb/sec. In fact, you may spend most of your time at a few Mb/sec as we set up things for web page download on the local computer, ask the server for the page (which it will probably have to dynamically construct), download the 2-3K of data, then parse its Html.

User Guide

Note: The program requires the .NET runtime version 4.7 (the latest as of this writing, August 2017) or later that must be installed on the user's computer, for example from https://www.microsoft.com/en-us/download/details.aspx?id=55170.

The app looks like this:

The *Collections* box lists some magazine collections you might be interested in. Click an entry to schedule it to be downloaded. Click again to unselect it. You can click any combination of them, but I'd suggest doing one at a time, since (depending on your download speed and perhaps data limits) each collection can be several (and in some cases, many) gigabytes in size.

As mentioned above, it seems that the magazines were scanned and converted to pdf format. They were then (again, it seems) OCR'd with the ABBYY program (https://www.abbyy.com/en-us/ocr-sdk/) and the result converted into all the other formats. To be honest, I really haven't looked at any of the formats other than pdf in any depth. The PDF option comes pre-set. Click as many (or few) of the others as you please. Note that not all the formats are necessarily available for all collections (e.g. the Comic Book Zip format is usually missing). And I'll leave it to you to decide what (if any) use any of the other formats might have.

During development, I put a lot of "print" statements into the code to help in debugging. Some messages always come out. Others are flagged as *Info*, *Detailed Info* and *Debug.* You can enable the display of these additional messages via the *Message Types* menu item. Each option can be turned on or off (by clicking) independently. But I warn you, the information they provide isn't very useful to the user.

The *Go* button starts the process going. See the Sample Output section below.

Finally, the Copy Messages button puts all the messages (without the colorization) on the clipboard. You can paste it into Notepad or whatever to view at your leisure.

Sample Output

I've used *Computer Language* a lot during my testing, since there are only 10 issues on archive.org. Here's what the default output looks like in that case.

Each message has a sequence number on the left followed by the elapsed time from clicking *Go*.

Note: If a file has already been downloaded (if for some reason you've re-run the program on the same collection), you'll get one or more *Skipping* messages. These can be ignored. Also, I've occasionally seen messages that the remote host (archive.org) has closed the Internet connection. I'm guessing that it's detected that the site is too busy. However if there was an error downloading a file, the download will be retried.

Magazine Collection Sizes

Collection Name	Number of .pdf files	Total File Size
Galaxy	349	7,984,540,227
Worlds of IF	176	3,341,315,237
Amazing Stories	703	5,133,900,402
Byte	211	52,909,159,556
PC Magazine	179	4,236,925,929
Computer Language	10	539,682,733
Scientific American (1845-1909)	2,965	34,050,835,013

Note: In some cases, even though the magazine collection has an entry for an issue, there may be no files of the type you asked for. For example, even though Galaxy has a nominal 355 issues, only 349 of them have a pdf file -- https://archive.org/details/Galaxy_2nd_Series_v01n02_1994-03 has filetypes Comic Book RAR (.cbr) and Torrent (.torrent) and so on, but no .pdf.