Download Archived Magazines

Thoughts – not part of this document

* In the mainline, I put (say) 5 instances of MagCollection into the queue, then call the Run routine.
* Run dequeues the first one, issues a semaphoric Wait, then queues a work item onto a thread pool thread.
  + Eventually the thread proc for MagCollection runs. Ideally when the code there finishes, I should called Release(). Meanwhile the thread code could/should put more items onto the queue.
    - The queue adds should be ones for each MagCollectionPage

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! The published a *lot* of classic science fiction stories by many of the top authors of the day. I took a look at the site[[1]](#footnote-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/GalaxyMagaine-1953-11/GalaxyMagzine-1953-11.pdf>

<https://archive.org/download/GalaxyMagaine-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>

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.

Good luck just motoring

* By URL/filename – nope; names weren’t consistent. And had no real way to conveniently specify the months they were published in.
* Wanted to generalize it to other mags (e.g. Worlds of IF, Astounding Stories, and many others
* Finally found a way to get a list of just the published (and available online) mags
* Then found a “directory” of the files
* But that directory didn’t have all the filenames (e.g. epub (and probably others) seemed to be missing)
* So had to scape yet another page
* Then use threading (or perhaps await) to download them in parallel.

TO BE CONTINUED…

Overview

My friend Gary recently emailed me that many issues of Galaxy Magazine were now available on archive.org. This document describes a program I wrote to automate downloading “all” (see later) of them.

This document describes some relevant notes about the program. See below for a sort of 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.

You can download the issues in a number of 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.

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.

Oh, one thing that’s not currently being worked on, but which should be easy to add, would be to help organize the files a bit better.

For example, right now all the Galaxy issues would go into C:\Magazines[[2]](#footnote-2)\Galaxy, C:\Magazines\World of IF, etc. But with Galaxy magazine having 355 archived issues, I thought it would be nice if they could go into C:\Magazines\Galaxy\1950\\*.pdf, C:\Magazines\Galaxy\1951\\*.pdf etc.

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 seem out of order. I have no plans to fix this.

\* It assumes that all the issues have URLs of the form

https://archive.org/download/galaxymagazine-{year}-{month:00}/Galaxy\_{year}\_{month:00}

But (a) Galaxy had a peripatetic publishing schedule; sometimes is was monthly, sometimes bi-monthly, and sometimes something else. So there are a lot of "Can't find the file" error messages. These are mostly benign. But (b), some of the files gave a "Forbidden" error. I assume this was a mistake by whoever uploaded these things. And (c) some of the issues are either missing, or don't have the same format URL as I'm looking for. If you run across any issues with a different format URL, let me know.

\* Sometimes I've seen a timeout error on a download. The good news is that the program checks to see if the file is already there, and if so, doesn't try to download it again. So you can rerun the program with impunity. It'll still give "Can't find issue" errors, but if it missed something before because of a timeout, it should get it again this time.

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 5 magazine collections.
* A <Formats> section that defines the filetypes (e.g. .pdf, .epub, .mobi, etc) that the program supports.

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 edit these. Just edit the .xml file directly.

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. See <https://archive.org/details/pulpmagazinearchive>

Things to Come

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.

But the professional in me wanted to run as much as possible in parallel, so I’m currently working on a way to run many web requests in parallel. But doing so without tripping over one’s toes isn’t trivial (especially with the messy code that has ensued from massive code changes along the way). But if you’re reading this, maybe I’ve already done it.

User Guide

Note: The program requires the .NET runtime version 4.7, which must be installed on the user’s computer, for example from <https://www.microsoft.com/en-us/download/details.aspx?id=55170>.

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. [↑](#footnote-ref-1)
2. Or whatever target directory is specified in the config file. [↑](#footnote-ref-2)