Bookmarks, favs, likes - backfilling years of gaps

What do you do when you have years of hoarded internet on your computer and you want to put them into one place, to make it searchable? You put them on your website, while retroactively like, favorite, and bookmark them at their source.

Update: This idea lead to an insanely large website, with content rather alien to myself. As a result, I ended up removing most of what was imported, strictly limiting my site to content somehow made or created by me. The process of this is documented under the article Content, bloat, privacy, archives

The vanishing internet

I've saved things from the internet. Art, pictures, videos, texts. Some were allowed, encouraged even, others not directly discouraged, others forbidden to be republished.

Fast forward a few years: the canonical source is gone. Images, videos, texts, thoughts, life fragments deleted. Domains vanished, re-purposed, sold. There are no redirects, just a lot of `404 Not Found`. And, in countless cases, this wasn't even on purpose: the provider deleted them, declared bankrupt, went on 'an incredible journey' of acquisition[^1], and now that piece of art, those last words from a DnD set is nowhere to be found.

On top if this, we now have social networks: walled gardens of marvels. They have an amount of content never before was gathered to a single provider - but it's forbidden goods.[^2]. *You're not from here, you can't play with us!*

So you register. You bookmark. You 👍, you ❤, you ★. But that's all inside, all locked away, and you won't even be able to find them again[^3].

When I decided to clean up my Facebook account, I went through everything - including comments on a friends post[^4] - it's a rare, world-readable Facebook note, mostly about how social media is reducing useful interactions on content. So I sent him the link, asking if anything changed in ~2 years and I got this reply:

> bazzeg, mennyit kerestem ezt a posztomat!! :D (és ez így mennyire kemény mondat a kontenthez mérten :D) // fuckit, I've been looking all over for this post!! :D (and this, knowing the content, is a devastating statement :D)

Which highlights the insanity of Facebook: even though the text exists, it's impossible to find it, even for it's owner.

Would it be easier to search for something in text files? Hell yes. It may require a friend with arcane knowledge[^5], but it's certainly something doable.

The place of social networks on the web

I've been "stalking" content via feeds - RSS and Atom - for many years, and, in the old blogging world, this was completely fine: comments and responses were the main indicator, not follower or like counts. Nobody was able to tell how many followers you had, and that was good.

What I recently discovered is that **followers and reactions matter to many** nowadays: these can decide popularity, even profitability, which makes reacting an important way of appreciation.

Besides this, the open web, the decentralized web, the real web comes with a big, so far unsolved problem: **discovering interesting content on the web is really hard**.

Even those who mostly create like to sometimes sit down at enjoy what others made; articles, paintings, music. Technorati used to index blogs[^6], but without it, the blogsphere fell apart, because people never figured out how to discover without central hubs.

I believe that some social networks, especially content- and art-sharing ones, like Tumblr, DeviantArt, 500px, SoundCloud, are irreplaceable for finding and sharing whatever interests you.

There are ongoing attempts to address discovery: some want to bring blogrolls back[^7], which is the old world equivalent of today's listing of who and what you follow; others are experimenting with recommendation systems for their sites, but these efforts are not there yet.

**Still, they come at a price: whatever you discover in these walled gardens, these silos, will stay in there. If they are gone, all of that magnificent content will be gone as well, and you'll never find it again

- unless they are present somewhere else as well.**

Dust off your website and make it your personal archives

The solution to prevent loosing all those likes, bookmarks, entries, as idea, is simple: **instead of storing all that data only within services, have the content on your own website as well**.

The reasoning is simple: first, when it comes to your site, you are in charge of the data, you own it. Second: you don't exclude people who can't or don't want to access a certain social network. Remember, there are 1 billion+ people in China, who cannot use Facebook or Google.

When it comes to cross-posting content between systems there are two main approaches to it: POSSE[^8] *(Publish on your Own Site, Syndicate Elsewhere)* and PESOS[^9] *(Publish Elsewhere, Syndicate to your Own Site)*. The main difference is the direction of how data flows: with POSSE you first post on your site and push to the other services; with PESOS you first publish in a network and pull it with your site.

There a remarkable solutions already: WordPress plugins[^10]; Known[^11], a website engine, that has this built-in - which means there is really nothing preventing this to be a relatively easy reality to have a backup of your activity.

Unfortunately, by definition[^12] both methods only address present and future post; the process of getting the old data out from the services is either importing or backfilling. Thankfully many of the PESOS plugins are capable of importing past content besides the new ones.

Sounds interesting, but I'm new to this, where should I start?

It depends what do you have right now: if you already have a website, and you're familiar with basic webprogramming, go to the Getting Started[^13] wiki page.

If you don't, then the easiest way to try the idea out is probably with WordPress.com: they offer an addon, called `Publicize`[^14], and all you have to do is sign up with them to see it working.

- go to https://wordpress.com
- create a new blog (sign up), or, if you already have one: sign in
- take a note of your WordPress.com address - it usually looks something like `something.wordpress.com`

- go to `https://wordpress.com/sharing/something.wordpress.com` (replace) `something.wordpress.com` with your WordPress.com address
- enable the services you want to cross-post content to: click on 'Connect' and follow the instructions
- post content!

That's all it takes to try it out. If you want a deeper dive: the movement behind these ideas is called indieweb[^15] and there are lots of things to read about on the indieweb wiki.

Salvaging old data

I happened to have that stupid amount of saved data I mentioned at the beginning. In order to bring that content to 2017, and to centralise them somewhere, instead of a myriad folders on my filesystem, I started tracing their origins; **consider these ideas how to find the source of similar, old, dusty content.**

Saved artworks

Deviantart

Back in the days - we're talking ~10 years - I saved a good amount of incredible artwork from deviantArt. Yes, I know it's not nice, though then there were not many way to support artists, unlike today, and many actually had a 'Download' button. The internet was also younger, but since I learnt too well that things do disappear from the web, despite common beliefs.

A long while ago DA used to apply the following naming to uploaded works: `(slug-of-title)-by-(da_username).(extension)` . This made it possible to trace those artworks: go to their page and favorite them within DeviantArt. Once this was done I imported them to my website.

Unfortunately a surprising amount (~30%) belonged to deactivated account or removed artworks, which I can completely understand, given the DA Terms & Conditions[^16].

Fotozz.hu

There used to be a photo sharing website which was active 12+ years ago in Hungary. I had only a handful of saved images from here, but the all followed the `fotozz_(longnumber).jpg` format.

However, when you visit the site - which is still alive! -, you'll notice that an entry looks like this: `http://fotozz.hu/fotot_megmutat?Foto_ID=(longnumber)`

That long number turned out to be the same number, so I could reconstruct the original URLs and make favorite entries on my site with their address.

About 20% was dead from here as well; probably removed work, or, given that the site wasn't upgraded in the past decade at all, it might as well be simply gone.

Tumblr

`tumblr_(hash).jpg`. Searching for `hash` doesn't return anything, EXIF is wiped clean. I had to turn to Google Vision API[^17] and load that - thankfully not too much - image one by one there, trying to find a source.

It's tedious and cweiske's idea of saving the source URL with the image[^18] would have come really handy.

The ghost of dead sites in my RSS reader

> *I can see dead blogs*

I've been reading websites via feeds for a long while now. It started with Thunderbird's built-in feed reader, then I used various things to convert entries to email.

These go back 15+ years, and, as it could be expected, a lot of the sites I received those from are gone, dead, half-dead (more on this later), or, rarely, still happy and alive.

The ones that are alive are easy: just add them as a bookmark, with the date and time in the mail.

The trickier ones were the dead and half-dead blogs: there were sites where the content is gone, or giving you and error, but the images were still loading from the domain. To preserve these I decided to pull what could be pulled from the server and use the content from the mail as main HTML.

The workflow of these entries looked like this:

- is it on a social network and is it alive?
    - go to the silo
    - sign up, if needed
    - favorite, like, etc the entry
    - import them back to my website

- it's not a social network or doesn't have favorites, bookmarks, etc:
    - make a bookmark entry on my site
    - try fetching the original URL
    - if succeeds:
        - save a copy
        - ping archive.org to save it[^19]
    - if it's dead:
        - use the mail content as source HTML
        - save images from that content

The reason why I want a cache of the content is to include it in the search corpus, so when I'm looking for a specific term, say, ZFS, I can find it; however, I'm not allowed to repost the content, so I won't do that.

Articles printed to PDF

Many of the prints contained either a full title or the URL in the top right corner.

Googling the full title between ""s usually resulted in finding the sources, and all I had to do is make a bookmark entry on my site.

Those which I only have in PDF form and are dead online: there isn't much I can do apart from holding on to that PDF.

Addendum: donate to artists, if they made it possible

There is one important thing I wanted to point out: don't forget to donate. **Likes and comments don't pay for bills and coffee, and artists should be paid for their work.**

Artists: provide a way for people to send you money, without silos. There is art I'm not going to "like" in a silo way: I might disagree too much with the silo to sign up, or I don't want that system to know my preferences, or to create a false virtual image of me based on partial data. Set up a Patreon page[^20], a Flattr account[^21], a paypal.me[^22] link, a monzo.me[^23] URL - something, what people can use to actually display their appreciation beyond likes.

In the end, you may end up making money without the trap of likes and followers.

Links

1. https://ourincrediblejourney.tumblr.com/
2. https://archive.org/post/442767/facebook-blocking-archiveorg
3. https://support.500px.com/hc/en-us/articles/236067827-Where-can-I-view-my-Likes-
4. https://mbasic.facebook.com/ninth.tibor/posts/10152041851714707
5. http://www.regular-expressions.info/
6. http://www.bayareatechpros.com/technorati-dead/
7. http://altplatform.org/2017/06/20/building-a-blogroll-in-2017/
8. https://indieweb.org/POSSE
9. https://indieweb.org/PESOS
10. https://indieweb.org/WordPress/Plugins
11. https://withknown.com
12. https://chat.indieweb.org/dev/2017-07-05#t1499290446868000
13. http://indieweb.org/Getting_Started
14. https://en.support.wordpress.com/publicize/
15. http://indieweb.org/
16. http://indieweb.org/deviantart#Selling_Your_Work
17. https://cloud.google.com/vision
18. https://cweiske.de/tagebuch/exif-url.htm
19. https://archive.org/web/#web_save_date_div
20. https://www.patreon.com/
21. https://flattr.com/
22. https://paypal.me
23. https://monzo.me/

Created by Peter Molnar <mail@petermolnar.net>, published at 2017-07-08 16:00 UTC, last modified at 2021-05-11 11:49 UTC , to canonical URL https://petermolnar.net/journal/personal-website-as-archiving-vault/ , licensed under CC-BY-4.0 .