#Creating a Community-based Web Archive

##The Project

Over a year ago I began a web archiving project intended to preserve the creative work of the blogs/websites of marginalized people. At this moment in time, the archive is dark/inaccessible, mainly because of technological and ethical considerations that I'll be discussing later.

I do quarterly web crawls, mostly because I don't really have the resources to do them more frequently. At the moment, I'm not using any software fancier than wget and a shell script -- since versions of wget greater than 1.14 include the ability to create warc files -- and a simple bash script I wrote to read the URLs of the blogs/webpages I crawl and execute the wget command. Some of the websites I crawl less frequently, since some have been abandoned by their owners (in the sense that they are no longer regularly updated).

At present, I have more than 5GB of data already archived from a list of 15-20 blogs or websites. By and large, these are currently just sitting on a hard-drive on my desk. Not the best place for them, I know, but the best I can do, for now.

For the sake of clarity, when I say 'marginalized people,' I'm primarily talking about people of colour. This is the only shared axis of oppression amongst the people I've been seeking to archive. Within this fairly broad category of people, I've been focusing on trans and cis women of colour, non-binary people of colour, queer people of colour, disabled people of colour, and several other overlapping identities.

Thus, later, when I talk about people 'outside' of the community, I generally mean white people, but not always. Since 'community' within this paper also doesn't necessarily refer to a group of people who explicitly identify as belonging to the community, more so a collection of individuals who interact with each other (but are often called a community by outsiders -- 'the social justice community', 'Black tumblr', 'Black twitter', 'tumblr feminists', and so on).

##Motivation

This project was created at a moment, within a community of bloggers, when several cases of plagiarism was prompting many important thinkers to consider deleting their blogs. Most often, it was journalists plagiarising the ideas and discussions occurring within social media communities, but the occassional academic would was also responsible for exploiting the people in some fashion or other.

This was the first time that I was present (as in a participating member of an online community) when this issue of plagiarism and exploitation came to the forefront. Unfortunately, this is not the first time that this has been an issue for marginalized people on the internet. Even more unfortunately, this will continue to be a problem so long as there are marginalized people and oppressors happy to exploit.

I haven't been participating within these types of communities that long (perhaps only three or so years), but even within my short time, I've seen wonderful writers bullied, harassed, and threatened into silence. And since it is so easy to delete your tumblr, twitter, blog, etc. their work simply disappears. But the echoes of their work is often heard for years afterwards, in the impact they made on the people around them.

And in case anyone is wondering why it is necessary that I undertake this archiving when the Internet Archive exists, the Internet Archive cannot archive everything. Tumblr, for example, has an option on their settings page that you can toggle to 'prevent search engines from indexing this site'. The Internet Archive relies on the same web crawling software that search engines do, thus, any blog that has this option turned on is not being preserved by them (I'll be going into technical details later).

Many mariginalized people on the internet rely on pseudonyms to protect their safety and privacy. Tumblr, twitter, and wordpress.com are all popular places for marginalized people because none of them have 'real name' policies. Wordpress likewise has an option to prevent search engines from indexing your blog. Even if (for example) tumblr users do not have this option turned on, many users frequently use the feature that allows them to change their url/username whenever the mood strikes them (for example, during the month of October, a lot of tumblr users will use temporary Halloween themed urls).

These blogs and this work is the very definition of 'ephemeral': but these blogs are also critically important, which is why journalists and academics remain so interested in exploiting the content and their creators.

Thus, this talk is called 'community-based web archiving' not only because I am, myself, an active member of the communities I'm discussing, but also because having this be a community initiated project allows me to practice and embody a lot of the different ethical and moral concerns raised by these issues.

##Ethical concerns

I'm not going to speak too much about the ethics of journalism, since that is beyond my own experience and knowledge. Instead, I'll be focusing more on the academic side of things, since I have more experience with this ethical domain. Moreover,

academics are one of the main groups who frequently exploit marginalized communities online.

One of the biggest difficulties when discussing the ethics of researching online communities, is the fact that what I'm about to discuss isn't actually supported by the ethics committees of most institutions. Indeed, I've had friends who wanted to conduct their research within the parameters I'll be discussing, but were actually told by their ethics committees and/or supervisors that it was unnecessary and too much work.

The primary issue is informed consent (or even just regular ol' consent). In the example of my friend, they wished to obtain consent from each of the bloggers they'd be discussing. But it was considered 'unneccessary' because the posts and such were 'public' and, thus, no consent was required. This notion of 'public' tends to be the grounds on which most academics (and journalists) use to justify their exploitation of marginalized voices.

They reason that since many of the people do use pseudonyms, it can be difficult for their readers to tie these pseudonyms to offline, 'real' identities, thus providing one of the critical provisions for research involving human subjects. Informed consent isn't needed because these individuals are willingly and of their own volition sharing their ideas/writing/lives on publicly accessible blogs or websites. And so all this information and data is a free-for-all academic buffet, allowing many of them to sidestep the more stringent ethics requirements for research involving human subjects when done offline.

This all seems well and good, until you actually consult the people often exploited by this loophole in academic research ethics. Nowadays, it has become more and more common to see notes or disclaimers on many different people's (often marginalized) blogs explicitely stating that they do not consent to being a research subject for any kind of academic research. But this is only the most obvious and explicit tactic used by marginalized people to protect themselves online from academic exploitation.

Other things they'll do is, as mentioned earlier, use pseudonyms. Ironic, given that the use of pseudonyms (and attendant anonymity) is often a justification for exploitation. However, it is clear, in many cases and for many marginalized people, that pseudonyms are used as a means to protect their privacy. So too, is toggling the 'do not allow search engines to index my blog' option. Is it ethical to disregard these clear indications that the person has made to protect their privacy?

Given that, in these environments, consent isn't considered a necessary condition for studying the community members, we can also understand that the bloggers who've taken the extra step to explicitely state that they do not wish to be used for academic research is more of a plea, than anything else. As far as most academics are concerned, even when consent is explicitly withheld, they have no obligation to actually care about this.

All of this matters for the marginal archive. Particularly since I personally support the 'right to be forgotten' which is only in the nascent stages of being considered as critical to online privacy (some EU privacy laws have begun to legislate this right, requiring that providers of online services actually delete the data of their users when requested). This means that even if I personally feel that a certain blog is important and ought to be preserved, I don't actually archive it unless I can get consent from the creator.

And consent is and always will be of primary importance for the archive. Much in the same way that analog archives will mainly archive items donated, as opposed to attempting to archive everything any given person has produced within their lifetime, this archive is functioning on the same principle. The difference though, is that since it is beyond the technical skills for most of the creators 'donating' their material to archive it themselves, I obtain their consent to archive their blog (or website) before archiving it.

There are, though, some exceptions to this. Some of the blogs I've archived are ones that haven't been updated in about five or six years. I've done my best to contact the creators for the blogs in question, but with little success. In one case, I've archived a blog created by a deceased person and consent is impossible (and it is really unclear as to who, other than the creator in this case, would be in a position to grant consent). The key with these exceptions, though, is that I've not made a decision about whether or not they'll ever be publicly accessible. I believe they are important and ought to be preserved, but this doesn't necessarily translate to unrestricted, public access.

The deceased person I mention here is Mark Aguhar. She was a Filipin@ artist who died a few years ago. After writing this part of the talk, it came to my attention that there are two white people presenting a talk about her at this symposium. Their session is happening concurrent to this one. I actually know some of Mark's friends and asked a fellow Filipin@ about whether or not they thought the white people presenting had obtained permission from Mark's friends or family to present on her life. Given that the title of Mark's blog is "blogging for brown gurls" and that so much of her art and writing was explicit about her general dislike for white people, this presentation is pretty much the exact kind of unethical behaviour I'm discussing.

As far as access is concerned, it will also be up to the individual how open they want their contribution to the archive to be. Or what the parameters are for granting access to groups or individuals.

In other words, there are two key ethical considerations: consent and agency. Consent determines, in most cases, whether or not a blog/website is 'donated' to the archive. Agency is about determining who will be able to access the archive and under what conditions.

##Technological Considerations

Interestingly, the issue of consent actually has important technological implications. Mainly since, as mentioned earlier, users of various free blogging platforms are able to select an option that prevents their blog from being indexed by search engines. What this means, on the surface, is that the blog's content will not be visible in a Google search, for example.

On the technical side of things, however, this actually means that the blog will have a robots.txt file telling web crawlers to ignore the site. A robots.txt file is an optional file that sets the rules for how programs like web crawlers are allowed to interact with the website. It is an old convention created back when servers were significantly less powerful and bandwidth was more costly than it is today, such that a web crawler indexing a site could crash the website (or server).

Note, this is a convention. Web crawlers do not necessarily need to listen to the robots.txt file, but most will. Google will ignore your website if your robots.txt file tells them too. You can also, for whatever reason, have a more granular file. For example, instead of asking all web crawlers to leave your site alone, you could specify that you don't want Google to index your site, but that Bing is allowed. You can also use your robots.txt file to indicate that some parts of your site are off-limits while the rest is open.

Many marginalized people will toggle the option that will create a robots.txt file preventing web crawling. It will ask that no web crawler index the site. Tools like wget that can be used to do web crawls will respect robots.txt files by default, however, you can turn off this option. Doing so means specifically choosing to circumvent a privacy measure engaged by the creator of the blog.

This is why, for this particular project, I'm only archiving blogs 'donated' to the archive. By ignoring the robots.txt file, I'm not only breaking general web conventions but disregarding a marginalized person's attempt to protect themselves.

For long term preservation, I'll be using the repository software, Islandora. Islandora within the last year integrated the ability to ingest warc files. Warc files are the internet standard for archiving websites (warc = web archive). It is actually one of my colleagues, Nick Ruest, who wrote the web archiving solution pack for Islandora. And I know that he is working to integrate the Internet Archive's wayback machine to allow interactive viewing of the warc files within Islandora (at present, you can download the warc file, which isn't useful for most users, since viewing a warc file isn't necessarily the easiest or most obvious task).

Importantly (and despite how much of a pain in the ass Drupal can be), the Drupal frontend for Islandora is a critical aspect for why I decided to go in this direction. One of the strongest features in Drupal is user management, in the sense of having

very fine grained control over user roles and permissions. Thus, it should be possible within the system to give a fair amount of control to the archive contributors over who is able to access their archived material. Both requiring a lot less work on my part but also ensuring that they are able to exersize their agency as much as possible.

##Conclusions

I understand that I sort of breezed over a lot of details in this presentation, particularly the technical ones (but I figured that people would generally be more interested in the ethical considerations vs. the technical ones and also because I've largely been unable to get a working Islandora site on a publicly accessible server).

But if I leave you with nothing else, it is to begin to really consider how digital archiving on the web (and other types of 'scholarly' activity) can and does raise a different set of ethical questions and concerns than are raised by physical materials. Moreover, other ethical questions arise from attempting to do this project within a networked environment where individuals are able to implement privacy measures.