Michael Kurzmeier

Towards a concept for archiving hacked websites

Introduction

Web defacements as a form of hacktivism are rarely archived and thus mostly lost for systematic study. When they find their way into web archives, it is often more as a by-product of a larger web archiving effort than as the result of a targeted effort. Aside from large collections there also exists a small scene of community-maintained *cybercrime* archives that archive hacked web sites, some of which are hacked in a hacktivist context. It is the purpose of this article to give an overview of the state and possibilities of using both archives for research. Such research may include qualitative and quantitative data analysis from a media and social studies perspective.

The first section of this article will present an introduction and definition of the phenomena of hacktivism on the web, followed by a description of the potential of general, large web archives to be used for the systematic research of hacktivism. This article attempts to arrive at an approximate ratio of *hacked* pages that can be expected to be found, a further subset of this content being pages defaced by hacktivists. In order to make steps towards a systematic model to describe hacked web pages, a Dublin Core Metadata schema will be adapted and described as well as ethical considerations are explored.

Following on from there is a section concerned with the use and usability of community-maintained archives. As this scene has seen many pages disappear since 2014, the focus will be on one of the remaining sites, *Zone-H*. The main point of this section will be the description of Zone-H's collection of hacked websites and metadata, in order to describe the opportunities and challenges in working with this kind of archive. This will be complemented by two brief examples of archive use by the site's personnel as well as academic researchers.

Definitions

Hacktivism

Hacktivism is a phenomena embedded in power structures and technologies. Its understanding is obscured by the narratives that surround it. Hacktivism carries with it the legacy of cyberspace utopia with all its problems.

It is tempting to use the expressions 'hacktivism' and 'hacking' as if they could be described in isolation, without having to refer to any external frameworks. In fact it might be more precise to move away from the conceptual level and not talk about hacking or hacktivism, but about *writing electronic text*. By doing so, it becomes more apparent how this writing is regulated by laws, conventions and design. A subset of this writing, in relation to the principles surrounding it, can be described as antagonistic in the sense that one act of writing overrides the other. Further down in this subset, then, are acts of writing that are not only antagonistic, but violations of one or many of the surrounding laws, conventions and designs. Now we are approaching something that under certain circumstances might be called hacking. Motivations for this type of writing are diverse, but what drives all forms of antagonistic engagement with media is narratives of how media should be. Without this utopian vision, or without an actual cause of concern, none of these practices would be able to mobilize its supporters.

The understanding that hacktivism could be a potential external influence to politics and national infrastructure was defining for writing on the topic during the early 2000s. While US-based agencies in particular attempted to estimate the future influence of hacktivism (National Infrastructure Protection Center 2001), an early strain of activist literature emerged that was equally interested in the possible uses of digital activism. Tim Jordan in his 2002 book *Activism! direct action, hacktivism and the future of society* describes hacktivism as "transgressions of the information infrastructures of 21st-century socio-economies" (2002, 121). On the other side, Mark G. Milone defines hacktivism as "surreptitious computer access or the dissemination of potentially disruptive and/or subversive software" (2002, 77) and advocates for educating hacktivist on how even well-meaning intentions can cause damage to vital Web infrastructure. Leah A. Lievrouw in 2011 presents a more rigid definition of hacktivism as the reconfiguration of media itself and thus qualifies hacktivism as the work of (activist) computer professionals (2011). Finally, Alexandra Samuel in her 2004 dissertation uses the following definition: "hacktivism is the nonviolent use of illegal or legally ambiguous digital tools in pursuit of political ends" (2014, 2). For the context of this article, the act of altering web sites through hacking (defacements) will be the focus of attention while acknowledging other important scholarly works that use sociological

methods to understand the community behind it (Coleman 2013, 2014).

Being so embedded in contemporary technology and the managing of it, hacktivism is closely intertwined with existing structures of power and control. Describing it as *antagonistic writing of electronic text* describes hacktivism in its willing engagement with existing technologies for the creation of electronic text wile driven by a vision influenced by techno-utopian romanticism about what texts should be written.

Web Archiving

As defaced web sites are usually quickly restored, they can be seen as content especially vulnerable to deletion and as such depend on web archiving services to be preserved. There exists a range of different approaches to web archiving, most are specific to their goals and the nature of the material (Brügger 2005; Antracoli et al. 2014; Bragg and Kristine 2013). Specific approaches to ephemeral web material are described by (Healy 2017; McDonald 2015; Ball 2010), with the latter describing a "performative model" for the presentation of complex web content:

> In this model, a researcher does not experience a digital record directly, but instead experiences a performance arising as a result of a process (software, hardware) running on some source data. In the Web context, a researcher looking at a Web page is in fact looking at a rendering of a collection of source files performed by a browser. (Ball 2010, 10)

Ephemeral content is described by Ball as intrinsically competing for limited resources through excess:

> The ephemeral nature of posts [on 4chan] combined with their anonymity confronts the users of boards with a question of meaning – how does a board create a sense of shared experience, something that extends beyond the brief period that individual posts are present? Anonymity, combined with the ephemeral nature of posts, generates a dynamic of competition for the limited resource of attention: a driver of excess and the extreme. (2010, 12)

The described characteristics of ephemeral content on image boards such as 4chan can be used to understand the nature of hacktivist web defacements: Driven by a competition for the limited resources of visibility and attention, archiving defacements also serves as a tool to provide that shared experience.

This explains why web archiving is important in this context, and also hints at why community-maintained archives exist within the scene.

Finding traces of Hacktivism

Traces of hacktivism means that in the context of antagonistic writing, one piece of electronic text was replaced with another, and that this new piece remained there while the site was archived. We can not assume that whoever archived the page had any knowledge of this. In this context, finding traces of hacktivism that coincidentally were archived within larger grabs can be compared to searching for marginalia in books. Yet the analogy with a phenomenon borrowed from textual scholarship (cf. Jackson 2001) is only partially correct. If hacked websites are seen as part of a larger debate within a society, spanning from websites to social media to traditional media forms, it is true that this form of marginal writing can be "a truncated and imperfect witness to the remarkable popularity of [a] [...] topos or commonplace that circulated widely..." (Bawcutt 2015). But looking at the individual site, the idea of marginal writing falls short of describing hacking and its communicative function.

While it is certainly true that hacktivism can be a form of engagement with a site's original content in the form of a remix or parody, its style of engagement is much more aggressive than that of leaving notes at page margins. We must also take into account that the majority of content on hacked sites is exclusive to the original content; one cannot be where the other one is. For these reasons, a better analogy to borrow from textual scholarship is the *Palimpsest*, a pergamene page that has been scraped off and rewritten (cf. Lyons 2011). Research on palimpsets includes forensic examination to restore the overwritten text and tries to establish a connection between the original and the new text. Similarly, forensic examination of digital media treats the existing data as palimpsets and aims to reveal the overwritten (Kirschenbaum 2012).

Even if we are to take a step back and remove any notions of *intent* from web archiving efforts, we still have to account for the fact that hacked websites are, while certainly not uncommon, far outnumbered by uncompromised websites. Amongst this minority, a great part of which will be taken up by spam and malware, hacktivist activity is an even smaller subset.

To exemplify the magnitude of this, a brief look at the Internet Archive's *Wayback Machine* in comparison to a dedicated collection of hacked websites shows to what extent problems of scale affect the usage of large, general web archives for the research of hacktivist activity. Comparing the growth of the Internet Archive and Zone-H, one of the few remaining site defacement archives, over the last 8 years shows that the ratio of hacktivist material one can realistically expect to find in larger, unspecialised web

archiving projects is far below one percent:

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Growth IA (Billion)	0	0	0	223	57	49	31	-224	58
Growth ZH (Billion	0.145	0.151	0.11	0.14	0.11	0.09	0.08	0.08	0.05
Relation	0.000966667	0.001006667	0.000733333	0.000375335	0.000255814	0.000187891	0.000156863	0.00027972	0.000145349

Fig. 1 Reported yearly growth rate of Internet Archive (IA) and Zone-H (ZH)[I]

This is a somewhat crude comparison which does not take into account factors such as availability and changing efforts in web preservation, hacking or archiving of hacked sites.

I am aware of further problems and limitations of this example. For one, the Internet Archive counts *web objects*, i.e. html pages, text and pdf files, Zone-H counts whole *sites*. A comparison of the total sites available on the Internet Archive in 2016 (361 Billion) with the highest pageId in the Zone-H archive (32187275) gives a ratio of approx. 0.09.

Also pageIds on Zone-H are not reused, so that if a site is deemed not eligible and removed from the archive, its pageId will not be reassigned, leaving an "invalid defacement" error behind. Manually reviewing and requesting sites for archival is the third and most important reason to be sceptical of the above data. While the Internet Archive uses automated crawlers to follow links to new pages, re-visit old ones and thus expand the archive, entries on Zone-H are checked by a human operator and, if not valid defacements, removed.

Finally, Zone-H is dedicated to archiving hacked websites. It does not specialize in hacktivism, nor is there any way to sort the archive by messages left behind. Websites get hacked for many reasons, be it an automated attack to distribute spam or malware, be it a kind of tagging to make a name for yourself in the hacking scene, be it (sometimes) a combination of the latter and some kind of political message. From my own research using the Zone-H archive, I assume that less than 10% of content in the Zone-H archive qualifies as hacktivist material by any given academic definition as most of the content lacks any political ambition.

1 Data sourced from from http://www.zone-h.org/stats/ymd (ZH) respectively using the Wayback Machine on itself to get the number of pages held on each 1st January. In 2010, the IA held 150 Billion pages and did not increase that number until 2013.

This example shows how complicated the use of large, unspecialised web archives can be for the research of hacktivism. Even a superficial comparison of two collections shows that a low percentage of hacktivist material is present even in a specialized archive. In a general archive, research must assume an even smaller percentage and can hardly rely on them to access a range of hacked sites. Further, general archives such as the Wayback machine allow users to search by URL and date and thus might be used to confirm claims of sites having been hacked in the past rather than discovering new content.

A more concrete example of approaching a large web archive in the search of traces of hacktivism may be seen in efforts to describe hacktivist activity on the deprecated service *Geocities*.

Yahoo Geocities started as a service called *Beverly Hills Internet* and similarly to projects such as the *Amsterdam Digital City* it owed its name to its organizational structure which was modelled after a city with streets, squares, lots and houses. While the *Amsterdam Digital City* was originally based in and around the city of Amsterdam, Geocities had a global scale and hosted over 100,000 websites making it the fifth most popular site on the web of 1997. When acquired by Yahoo! for $4.6 Billion in 1999, Geocities had upwards of one million users (Miligan 2017, 139). The service was eventually taken offline in 2009 after the static content was considered not profitable enough for the newly dubbed web 2.0.

The history of preservation of Geocities shows the immense difficulty in identifying and preserving a project of that scale. Even though the shutdown was announced in advance and efforts to archive Geocities had been existing before that, it proved difficult to index everything, let alone to archive it (Miligan 2018b, 2018a). As a result, most of the available archives estimate their crawlers have the majority of content, but there is no general index, no checksum or catalogue to refer to. This puts the available archives closer to media archaeological investigations than plain storage of a deprecated service. Snapshots were taken throughout Geocities' existence, but owing to the size of the project this produced more fragments than histories. We may know that just before the shutdown, Geocities had 7 Million sites (Fletcher 2009), but we are oblivious as to where traffic went.

Similar to an archaeological investigation is the search for any hacktivist content on what remains of Geocities. The following will show the difficulties and possibilities in surveying the field for further investigation.

The methods used were a general keyword search to survey the archive and an attempt to cross-reference some well-known hacker groups from the Zone-H archive:[2]

Keyword	Results
Hacking	6100
Hacked	10400
Hacktivism	44
Cyberarmy	18

Fig. 2 Keyword search on geocities.ws

A full text search for some general keywords reveals the archive holds sites about hacking, hacked sites and even sites about hacktivism. If we put those results into relation to the size of the archive (40 Million pages), the prevalence of potentially relevant content is 0.0004125. This is within the ranges described earlier. Again, the comparison of pages and sites leads to a very low ratio of relevant content but still fits within the ratio bands the comparison between the *Internet Archive* and *Zone-H* yielded.

General approaches like this are of course only useful to survey the field and to prove the general existence of potentially relevant content. They are problematic for two main reasons: First, there is no standardization within the hacking community because of its disparate and multifacted nature. Not all hacked sites will feature the word "hacked" and very few sites defaced by hacktivists will contain words like "hacktivism" or "defacement". This might explain the low yields for the general keyword search. Second, while it is convenient to have a search engine built into a public-accessible archive such as geocities.ws, the user is unable to confirm the completeness of the index and the search parameters. This creates a black box effect where the user dependent on the tool to function correctly, without having insight into its configuration. Any in-depth study should avoid reliance on black boxed tools and instead use local copies and configurable search engines where possible.

2 The archive used is geocities.ws since it offers a large collection of sites as well as content search.

Cross-referencing

Cross-referencing can be a useful method to overcome limitations of general keyword searches. In this case, cross-referencing uses the lack of standardization that complicates the keyword search to achieve more precise results. Cross-referencing requires a reliable source of information and is also reliant on time-dependent vocabulary. The following table shows the selected results of a cross-reference between the top 50 defacers on *Zone-H* and the respective search results on the Geocities archive.

Name	Defaced Sites on ZH	Search results on geocities.ws
Ashiyane Digital Security Team	1407	1
oroboruo	1799	1
iskorpitx	786	2
uykusuz001	38	2
MCA-CRB	374	4
Panataran	24	4
ISCN	102	4
chinafans	1	4
Mafia Hacking Team	331	24
RxR	706	28
Iran Black Hats Team	418	68
By_aGReSiF	804	104
Turkish Energy Team	320	180
DeltahackingSecurityTEAM	232	183
Digital Boys Underground Team	190	336
1923Turk	475	614
Triad	397	1500
Fatal Error	2751	1790
[#elite top team]	573	1850

Fig. 3 Results of cross referencing the top 50 defacers from Zone-H (ZH) with the geocities.ws archive

Cross-referencing can show continuities across time and virtual space, it can circumvent the often too broad scope of a keyword search and it can make use of specific names and expressions. Its limitations are the often the fast pace at which digital underground communities change, so that it is necessary for the two archives to have some overlap or at least similarity in the covered time period. Some data such as the name "Fatal Error" is simply too general to be used in this context, other depends too much on search engine parameters to be used in a black box scenario. "HighTech" for example will return all results that feature the words high and tech on the same page, while the query ignores spelling such as H1ghT3ch. Again, this is where a local, configurable search engine is especially helpful.

Adopting a Dublin Core Metadata schema

The Dublin Core Metadata Initiative (DCMI) offers metadata sets that can readily be adopted for the description of web pages. While models exist to describe general web pages and scraped web data, the archiving of hacked websites poses a problem that requires a specialized set of metadata tags. In the following I am going to describe an outline that can be adopted into a schema to describe both sites taken from larger archives as well as sites grabbed from the original source.

The goals of this effort are to provide a metadata model for a possible standardization of web defacements in the context of hacktivism and increased interoperability between collections and increased sustainability of research data. The schema is supposed to be quite general in its description, this is done with a view towards a possible automated description by keyword search within the HTML files as well as an computerized image recognition system.

Overview of required tags for a DCMI scheme

• dc:title

Used for the title of the item. It is recommended to refer to the URL here, since hacks often include changing the site's header as well.

• dc:creator

The name of the individual or group claiming authorship. If none is given, "unknown" may be used.

• dc:subject

The general subject of the message. Recommended uses are claims to authorship ("Hacked by"), solidarity with a cause, political messages of general nature, group-based defamations ("Down with X"), or specific memory-related content ("Remember these atrocities"). The use and frequency of these classifications is dependent on the kind of content that is in scope for analysis.

• dc:description

A free text description of the site's contents. This field should include the use of images, sounds, video etc. It can also be used to specify whether the attack was in relation to the site's original content.

• dc:publisher

This field is dependent on the way the site was obtained. If taken from a larger archive, it should be referenced here. Otherwise, "taken from original source" should be used here.

• dc:date

If taken from an archive: State the original date if possible. Otherwise, "unknown" should be used. If captured from a live site, the date of the capture should be used.

• dc:source

If taken from an archive or live capture, this can be the same as dc:publisher. If the data is taken from another already curated set, it should be specified here.

• dc:language

The language used. If mixed, country codes can be combined (en; pl)

• dc:relation

The relation between the new and original content. In most cases, it will be a complete replace of the original site, but also parodies of sites and subversive content insertions can be marked up using this tag.

• dc:rights

When taking material from an archive, relate to the licence agreement the content is under. Note this is a complicated situation as the archival work may be under the archive's licence while the actual site content may be not.

Sample DCMI set

Property	Range	Value String	Comments	Tag example
dc:title	literal	YES	Used for the URL	<dc:title>www.test.ie</dc:title>
dc:creator	literal	YES	Name of individual or group	<dc:creator>hack0r</dc:creator>
dc:subject	non-literal	YES	Message subject, political, defamatory, etc.	<dc:subject>Political</dc:subject>
dc:description	non-literal	YES	Freetext description of the hacked site	<dc:description>Gaza</dc:description>
dc:publisher	literal	YES	Reference to archive, replace if scraped from original source	<dc:publisher>Zone-H</dc:publisher>
dc:date	literal	YES	date of the hack	<dc:date>2002-05-01</dc:date>
dc:type	literal	YES	type of source	<dc:type>Web page</dc:type>
dc:identifier	non-literal	YES	Assigned identifier	<dc:identifier>SRC001_P001</dc:identifier>
dc:source	literal	YES	Used to refer to the data source	<dc:source>Zone-H</dc:source>
dc:language	non-literal	YES	country code	<dc:language>eng</dc:language>
dc:relation	non-literal	YES	relation to original content, defacement, subversion	<dc:relation>defacement</dc:relation>
dc:rights	non-literal	YES	licence situation	<dc:rights>Attribution-Noncommercial-No Derivative Works 3.0 Unported</dc:rights>

Fig. 4 Suggested DCMI set

Using specialized archives

Specialized archives in relation to hacking are independent collections of hacked and defaced websites. They are usually maintained by people with an interest in computer security, be it white hat (lawful) or black hat (unlawful) hackers or internet activists themselves. Some of those archives are part of larger collection of sites based around computer security:

> Attrition.org [...] is a computer security web site dedicated to the collection, dissemination and distribution of information about the security industry for anyone interested in the subject. They maintain one of the only open and honest grim look at the industry, reminding everyone that we must strive to be better than we have been historically. The crusade to expose industry frauds and inform the public about incorrect information in computer security articles is a primary goal of the site.
>
> Previously, Attrition.org maintained the largest catalogs (sic) of security advisories, text files, and humorous image galleries. They are also known for maintaining the largest mirror of Web site defacements ... ('Attrition - about Us' n.d.)

This critical view of "the industry" is often part of these site's self-understanding as the outsiders who "get it" while the mainstream is ignorant or oblivious. This ties in with countercultural ideas of identity rooted in mainstream culture's failure to anticipate developments:

> When a marginal social movement accurately anticipates in the public eye a significant historical failure of judgment on the part of leadership, the effects can be powerful. Being right about something when the powers that be were wrong, for example, was a central collective experience of the 1960s counterculture; by 1969, the world had watched the television networks, the New York Times, and many members of the political establishment change their position on the Vietnam War. (Streeter 2011, 163)

This idea of being right where the established IT security is wrong, of seeing what the masses can not or choose not to see is important to many of these independent archives. It may explain the immense dedication that went into building and maintaining these archives, not to mention checking every submission.

Overview

These specialized independent archives are disappearing from the Internet. Samuel in her 2014 dissertation on hacktivism as a form of political participation draws from archives that have since gone offline, while also dealing with a scarcity of sources herself:

> Thanks to the volume of defacements, the biggest mirrors (Attrition and alldas) have stopped archiving defacements. alldas has gone offline entirely; Attrition stopped maintaining its archive in April 2001 [...] but has preserved its records of defacements from 1995-2001. (Samuel 2014, 8)

The mentioned mirror page, Attrition, has ceased to archive any new content but published a list of recommended mirror sites:

Name	URL	Original Description	Status
Alldas Mirror	https://www.alldas.org/	None	Message "offline"
Zone-H Mirror	http://www.zone-h.org/	None	Online, accepting submissions
Blackhat.info Mirror	http://www.blackhat.info/	None	Offline, empty directory
Turkeynews	http://www.turkeynews.net/Hacked	Turkish Defacements	Offline, empty page
Flashback Mirror	http://www.flashback.se/hack/	None	Online, last update 2012, Swedish site
Hysteria Mirror	http://hysteria.sk/hacked/	Czech defacements	Offline, Error 404
Hackzone	http://www.hackzone.ru/hacked/hacked.html	Russian defacements	Online, last update 2008
Hax0r	http://deface.hax0r.hu/	Hungarian defacements	Offline, server not found
Influence	http://influence.org/~bedlam/misc/defaced-all.html	Irish defacements	Offline, server not found
Philippine Defaced Webpages Archive	http://www.ezroot.com/archive/	None	Offline, no connection
Onething Archive	http://www.onething.com/archive/	No longer updated	Offline, no connection
Beard Mirror	http://www.netrus.net/users/beard/pages/hacks/	No longer updated	Offline, Error 404

Fig. 5 Status of archives mentioned at attrition.org

Availibility

Availability here refers to any of the remaining site's long-term prospects. This includes funding, hosting and content review. Looking at the above list, 9 out of 12 sites have already disappeared from the web. Many of them do not allow automated web crawlers to archive their domains, so that research can not rely on larger, unspecialised collections to contain them. A further two archives are in a state of suspended activity and have not released any new content for seven (Flashback Mirror) and eleven years (Hackzone).

It is reasonable to assume that these sites too will eventually disappear.

The reason for this decline is unclear. A larger analysis of web defacements suggests web defacements have lost some of their appeal as a political tool (Balduzzi et al. 2018, 59). While this is likely to be one factor, a change towards web content being distributed through centralized platforms rather than individual web pages is certainly another.

Cross-site availability describes the process of copying one archive's contents into another, especially used when sites gracefully stop archiving. For example, Zone-H holds material from Alldas that was copied over before the site went offline. While this can not be expected to be the norm, it may at least prolong the availability of some of the material.

As none of the above sites is or were affiliated with any institutions, they are completely reliant on volunteer work. This of course makes them very vulnerable as any drop in interest might lead to a lack of people to maintain and run the archives. While staff may work for free, servers and bandwidth need to be paid for. Potential sources of income are anonymous donations and ads on the sites. Both of these sources are extremely unreliable and naturally only be a small number of businesses would be interested in being associated with hacked web sites. Financial problems are intensified by exclusion from larger ad networks such as Google AdWords.

The lack of any long-term funding for these archives is most likely the main reason they keep disappearing. Volunteers loose interest, submissions quickly pile up if not tended to daily. Currently 342,807 submissions are awaiting review on Zone-H. Those specialized archives, often build by people who are (maybe a bit too) enthusiastic about IT security, are valuable resources for the study of marginal, subcultural web. Looking at the loss rate throughout the last 5 years, it must be assumed that they will disappear or stop archiving material before long. To exemplify the difficulties in using these independent, specialized archives such as Zone-H or attrition for research purposes, this next section is going to evaluate the FAIR Data Principles against the Zone-H archive. This is not done to belittle the tremendous amount of work and dedication volunteers have put into the making and maintaining the site, but solely to show difficulties in using this and other similar sites for systematic research.

Usability – a short case study

FAIR Data Principles stipulate data must be Findable, Accessible, Interoperable and Re-usable (Force11 n.d.). Further described are 15 sub points:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.
- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta) data.
- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.
- R1.3. (meta)data meet domain-relevant community standards.

Status Zone H

The FAIR criteria allow for research to build tools that scrape archives like Zone-H. Scraping, and thereby creating a local copy, is one of the strategies to get around the limitations of the site and data. One example for such a limitation would be the site's lack of a function to search defacements by date. While time and date is present in the metadata, a user can not search the archive for all defacements submitted during, say, March 2016. Because the metadata is available in html format, however, it is easily possible to use a software tool to create a local copy of all metadata, sort it by date and return a list with the requested items.

In a similar fashion, web scraping can be used to download the defaced web sites together with their metadata and create a local archive to perform a keyword search. The two examples show how, despite falling short of fulfilling the FAIR data requirements, availability and open file formats can at least to some extent circumvent the limitations of the archives.

Criteria	Status	Comment
F1	Not met	Local Id assigned, not referenced in metadata
F2	Partially met	Some metadata about attacker, time and targe
F3	Partially met	Some metadata is searchable
F4	Not met	Identifier not referenced
A1	Partially met	Data accessible via html request
A1.1	Met	Data stored in html format
A1.2	Partially met	Archive is public, site allows login
A2	Met	Metadata is independent from mirrored site
I1	Met	Metadata is clear and concise
I2	Partially met	Where applicable or possible
I3	Not met	Sites are unconnected
R1	Partially met	Content description missing entirely
R1.1	Met	Licence defined
R1.2	Not met	Metadata in isolation from objects
R1.3	Partially met	In accordance with other, similar sites

Fig. 6 Status of FAIR criteria for Zone-H

Accessibility

As the previous sections already mentioned web scraping, this section is going to give further details on the accessibility of the mentioned specialized archives. Zone-H allows access exclusively through http, there is no API or any (legit) way to directly access the full collection. Using web scraping tools is complicated by the site's usage of captchas which occur around every 500 page requests. It has also occurred that the server closed the connection after too many requests in a 24h period. Even browsing the archive using the unique local IDs assigned to every defacement is not something that is intended to be part of the site's functionality.[3]

3 URLs are build according to the schema: https://www.zone-h.org/mirror/id/{ID}. By inputting different values for {ID}, a user can access different defacements. Zone-H acknowledges this can constitute a security risk when sensitive data can be accessed through this method. (Maslowski 2007)

Most likely it is a known and tolerated feature since no sensitive data can be access using this method and the design allows for a maximum of 30 pages á 25 entries to be skipped.

Because of these limitations, it becomes quite difficult to use web-based archiving services to extract data from Zone-H's collection. The site's robot.txt excludes automated services such as the Internet Archive's Wayback machine. While there is a search function included, users can only search for *domains* and not sort the results in any way.

These difficulties in accessing material show mainly two key aspects. One, how the site, much like similar sites, is very difficult to access for a structured analysis. In fact, any structured analysis can only be done with the material held in the archive, not through the site as a gateway to the material. Second, how the site's nature as a scene-maintained hacking database shapes its structure. It seems only reasonable, looking at it from that perspective, to offer as few services as possible in order to not become an entry in their own list of hacked web sites. Adding to this is also the before mentioned perspective of the knowledgable outsider. It would be contradictory to maintain this identity while at the same time opening up the site for large cooperation with research or, worse, industry.

Use cases

While there might be a reluctance to opening up archives, there is evidence of past projects utilizing specialized archives, as will be described in the following. This section is going to be divided into external and internal projects, depending on their authors. While projects and cooperation with external authors shows the use of an archive for research and art projects, internal engagement with the archived material can be seen as defining the self-understanding of a site and shows the level of awareness of the material held in storage.

On February 3rd 2014, the following message was posted on Zone-H:

> I'm a specialist in graffiti, one of the oldest and most interesting forms of human expression.
>
> As a researcher at a major North American University, I've been carefully following the most recent development in the long history of graffiti: the move to the digital environment and the rise of what I call Virtual Graffiti – my name for the work you do and for what is preserved here at Zone-H.
>
> I'm interviewing virtual graffiti artists as part of The Virtual Graffiti Project, an exhibition that I'm putting together with some

> friends and colleagues, to introduce virtual graffiti to the broader world. I work on Skype or Gmail chat, and record interviews on a digital recorder, carefully protecting the anonymity of my sources (see below).
>
> In the exhibition and accompanying book, I want to highlight the interesting works of virtual graffiti I've discovered, as well as comments from the artists, to introduce both to the world, and establish links between them and the 'traditional' graffiti we find in the 'real' world.
>
> The project has been approved by my university – which will guarantee the confidentiality of our interview. I provide all participants with a formal letter outlining the measures taken to guarantee the privacy and confidentiality of everything discussed over the course of the interview. [...] (Hopkins 2014b)

This request for people to come forward and contact Tedd Hopkins shows that some level of cooperation must have existed between the Zone-H admins and a north American researcher. This cooperation ultimately contributed to a 2014 PhD thesis titled *Virtual Graffiti: Dyscribing Humans*. Its acknowledgements start with:

> I would like to thank the people who participated, some directly and some indirectly, in my ethnographic research: the LeetBoys, 1337Mir, Gantengers Crew, PhantomGhost, CoupdeGrace, Mr.WWW, 4prili666h05T, NoFace, and many other virtual dyscriptors and makers of marks on the digital cave walls; 'Vympel' and 'Siegfried' from the zone-h archive—both of whom have seen more virtual graffiti than anyone on earth; 'Pavel' and 'Bruce' and the other webmasters willing to talk to me about the breaches at their sites ... (Hopkins 2014a, sec. III)

The thesis itself features a qualitative approach to digital graffiti, thus "tracking the emergence of a new, 'digital-borne' analogue to traditional graffiti praxis". (Hopkins 2014a, ii) However, no details on the nature of cooperation with Zone-H or attrition.org are given. Despite this, the work described in the thesis shows that research has acknowledged these specialized archives as valuable sources of information and that successful cooperation is possible. Whether similar efforts are still possible today (due to the shrinking number of archives and defacements) is unclear.

In addition to scholarly engagement with Zone-H and similar sites, there exists reflective engagement with the archive's contents. This is yet another form of engagement with those community archives that deserves to be mentioned as this type of introspection allows insight into the site's self-understanding and showcases some types of usage that those archives were designed for.

In the following excerpt, posted on the News section of Zone-H, the author reflects on the political issues that drive hacktivists:

> But what happened in the cyber-world? [referring to Mahmoud Ahmadinejad's election victory in 2009] Did we witness any sort of digital protest as we used to see in the past like those related to the Kashmir dispute? Or something like the Estonian bronze statue protest? Or even the Prophet Mohammed cartoon protest? [...]
>
> Surprisingly (or not?) the voice of cyber-protesters is still focused on:
>
> wishes of death to Israel, Usa and Denmark (still, after 3 years!) [...] Armenian and Azerbaijan rants [...] Turkish hackers against Israel [...] Turkish hackers against Armenians [...] wishes against Israel [...] wishes of a free Iraq, Afghanistan and Palestine while showing the finger to USA and Israel [...] whishes [sic] of a Kosovo independency [...] and absolutely nothing, zero related to the recent Iranian happenings. Yet another statistical anomaly? (Preatoni 2009)

This excerpt shows how Roberto Preatoni is well aware of hacktivism as a phenomena within the spectrum that Zone-H attempts to capture and preserve. He continues to identify political agendas that hacktivists are focussed on and provides links for each, while being surprised about the lack of content regarding the Iran elections. The post ends with an update:

> UPDATE: I received a message from an Iranian defacer. He says that he indeed protested during the election days. [...] I didn't notice it as... I don't speak Farsi language. (Preatoni 2009)

The situation is very much owed to the structure of the archive itself, such as no description of the content and nature of the hack, so that material gets lost in the volume of submissions and may be very difficult to access due to language barriers. These issues are worth mentioning as they affect researchers in the same way. Yet this list of topics also shows a great level of awareness and communication between the author and defectors.

Zone-H offers other posts that explore the archive often from an insider perspective, some featuring surveys as to why people engage in defacing web sites. As often with surveys on the Internet, they should be taken with some grains of salt but serve as clear indicators that Zone-H is interested in its community and does understand itself to be more than just a defacer leaderboard.

Ethics

Research on and with hacktivist material is going to be research on the fringes of the Internet, as the content is only accidentally captured by academic web archiving. Being on the fringes has the tendency to enlarge and distort tendencies from the centre, such as discrimination and sexism. This is evident as hacking conventions frequently report issues of gender discrimination and harassment (Richterich 2018). Therefore, even though an approach may be metadata-based and without personal identifiable data, these dynamics must be understood and must be considered.

It can generally be assumed, when working with material either obtained through a general or specialized archive, that hacking of any web page is done without the owner's consent. This is hardly an epiphany but should be kept in mind when deciding how to engage with hacktivist content in a scholarly context.

The first question derived from this is the medium itself. Maynooth University Ethics Policy states that:

> The consequences of research may reverberate at many levels, including the local community of participants, the professional community and the wider society. Researchers should be cognizant of this and sensitive to issues arising from inequalities of power. (Maynooth University Research Ethics Policy and Committee 2016)

This responsibility explicitly extends to the professional and private community of owners of hacked web sites. It should be considered, especially in the case of smaller websites, whether the URL needs to be provided to understand the hack or whether the owner could be exposed through it. In many cases, it will be sufficient to briefly describe the site to provide context. It its a key question in web research whether information is to be treated as text or as personal expression. Stine Lomborg acknowledges that:

> [for] the study of the development of specific sub-cultures on the web (e.g. sexual minority groups, political extremist debate fora or camgirls' websites), using archival data on the participants and their online communications may be more problematic, if only because of the risk of exposing and causing harm to specific private individuals, based on their prior affiliation to a sub-culture. The case-by-case ethical judgement starts with reflections about the textuality or humanness of the data in question. (Lomborg 2018, 102)

According to Lomborgs's argumentation here, the use of hacked sites could potentially be problematic insofar as individuals might be – against their will – affiliated with cybercrime. This can be mitigated by shifting the focus more towards the hacktivist material and less to the original page.

The second aspect is the posted material. It is entirely possible to come across personal information that was posted for no other reason than to expose and humiliate a person. Although such sites will hardly be hacktivist content by definition, it is also a requirement of the research design to exclude any identifiable information as consent can not be sought or assumed. In the same way, other identifiable information such as photographs must be dealt with great care and consideration. It is in the absence of ethical guidelines for engagement with hacktivism that those principles had to be derived from general ethics guidelines.

Conclusions

Traces of hacktivism are hard to find on the web. The vast majority of it is lost for research purposes. What remains are fragments that are occasionally found in general web archives such as the *Wayback Machine* or in very large grabs such as the *Geocities* archive. It is hard to estimate how much or how little material is present in those general archives, but actual hacktivist content probably occurs at a prevalence of less than 0.01. General archives such as the Internet Archive's Wayback Machine are also dependent on their search engines to deliver results, which in some cases adds more obscurity through a black box effect where search parameters are not always made transparent.

This scarcity of material is further amplified by the lack of a metadata scheme to effectively describe hacked websites. Such a scheme, as proposed in this article, would allow researchers to engage with the content and message beyond individual collections.

Specialized archives are better suited for the systematic study of hacktivism, yet suffer from:

a) lack of support and thus a dwindling number of archives that remain active

b) lack of standardization and lack of easy access

Also most of the archives mentioned in the article are not *hacktivist* archives per definition, they are self-described *cybercrime* archives, thus covering a wide range of phenomena from web defacements to hacking for the *lulz*. They have, however, shown some cooperation with researchers in the field even though this does not seem to have included any kind of privileged access to the data.

Complimenting these external analysis, Zone-H also features a range of news entries that show a kind of internal engagement with the material held in the archive. This introspection shows that hacktivism is acknowledged as part of the collection and that efforts were made to find and describe sites hacked by hacktivists.

There are ethical considerations when dealing with content that was placed on a 3rd party web site without their permission, even more so when that content is offensive or abusive. Ethics policies require research to be aware of possible implications in the private, public and professional sector, all of which might be in scope when dealing with hacktivist material.

What ties all sectors together is their lack of a common base to stand on. Research on hacktivism has been and probably will be research on fringe material either captured by chance during a large archiving effort or reported and archived by a volunteer site with uncertain lifespan. In the light of this, it becomes even more important to acknowledge existing material and work towards a robust standardization in the description of hacked web sites.

Work Cited

Antracoli, Alexis, Steven Duckworth, Judith Silva, and Kristen Yarmey. 2014. 'Capture All the URLs: First Steps in Web Archiving'. Pennsylvania Libraries: Research & Practice 2 (2): 155–70. https://doi.org/10.5195/PALRAP.2014.67.

'Attrition - about Us'. n.d. Attrition.Org. Accessed 15 February 2019. http://attrition.org/attrition/.

Balduzzi, Marco, Ryan Flores, Lion Gu, and Federico Maggi. 2018. 'A Deep Dive into Defacement: How Geopolitical Events Trigger Web Attacks'. Trend Micro Forward-Looking Threat Research (FTR) Team.

Ball, Alex. 2010. 'DCC State of the Art Report: Web Archiving'. University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council, 2010. Edinburgh Research Archive (era). http://hdl.handle.net/1842/3327.

Bawcutt, P. 2015. 'The Source and Significance of a Marginal Inscription in the Buik of Alexander (c.1580)'. Notes and Queries 62 (1): 56–58. https://doi.org/10.1093/notesj/gju226.

Bragg, Molly, and Hanna Kristine. 2013. 'The Web Archiving Life Cycle Model'. The Archive-It Team, Internet Archive. https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

Brügger, Niels. 2005. Archiving Websites: General Considerations and Strategies. Århus, Denmark: Centre for Internet Research.

Coleman, E. Gabriella. 2013. 'Anonymous in Context: The Politics and Power behind the Mask'. Internet Governance Papers, September 2013. https://www.cigionline.org/sites/default/files/no3_8.pdf.

———. 2014. Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous. London ; New York: Verso.

Fletcher, Dan. 2009. 'Internet Atrocity! GeoCities' Demise Erases Web History'. Time Magazine, 9 November 2009. http://content.time.com/time/business/article/0,8599,1936645,00.html.

Force11. n.d. 'Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0'. https://web.archive.org/save/https://www.force11.org/fairprinciples. Accessed 21 February 2019. https://www.force11.org/fairprinciples.

Healy, Sharon. 2017. 'The Web Archiving of Irish Election Campaigns: A Case Study into the Usefulness of the Irish Web Archive for Researchers and Historians'. In . University ofLondon. http://netpreserve.org/wac2017/abstracts/#_abstract74.

Hopkins, Todd. 2014a. 'Virtual Graffiti: Dyscribing Humans'. Ottawa, Canada: Carleton University. https://curve.carleton.ca/system/files/etd/9ef84295-5316-4532-a776-8ceeecb3fcfa/etd_pdf/db2d82e78e60f-268c42e2964109eeed/hopkins-virtualgraffitidyscribinghumans.pdf.

———. 2014b. 'The Virtual Graffiti Project'. Zone-H. 3 February 2014. https://www.zone-h.org/news/id/4745.

Jackson, H. J. 2001. Marginalia: Readers Writing in Books. New Haven: Yale University Press.

Jordan, Tim. 2002. Activism! Direct Action, Hacktivism and the Future of Society. Focus on Contemporary Issues. London: Reaktion Books.

Kirschenbaum, Matthew G. 2012. Mechanisms: New Media and the Forensic Imagination. 1. paperback ed. Cambridge, Mass: MIT Press.

Lievrouw, Leah A. 2011. Alternative and Activist New Media. Digital Media and Society Series. Cambridge, UK ; Malden, MA: Polity.

Lomborg, Stine. 2018. 'Ethical Considerations for Web Archives and Web History Research'. In The Sage Handbook of Web History, edited by Niels Brügger and Ian Miligan, 99–111. Thousand Oaks, CA: SAGE Inc.

Lyons, Martyn. 2011. Books: A Living History. Los Angeles: J. Paul Getty Museum.

Maynooth University Research Ethics Policy and Committee. 2016. 'Maynooth University Research Ethics Policy'. https://www.maynoothuniversity.ie/sites/default/files/assets/document/Maynooth%20University%20Research%20Ethics%20Policy%20June%202016_0.pdf.

McDonald, Kevin. 2015. 'From Indymedia to Anonymous: Rethinking Action and Identity in Digital Cultures'. Information, Communication & Society 18 (8): 968–82. https://doi.org/10.1080/1369118X.2015.1039561.

Miligan, Ian. 2017. 'Welcome to the Web: The Online Community of GeoCities during the Early Years of the World Wide Web'. In The Web as History, edited by Niels Brügger, 137–57. UCL Press. https://doi.org/10.14324/111.9781911307563.

———. 2018a. 'Exploring Web Archives in the Age of Abundance: A Social History Case Study of GeoCities'. In The Sage Handbook of Web History, edited by Niels Brügger and Ian Miligan, 344–58. Thousand Oaks, CA: SAGE Inc. 978-1-4739-8005-1.

———. 2018b. 'Ethics and the Archived Web Presentation: "The Ethics of Studying GeoCities"'. Digital History, Web Archiving, and Contemporary History (blog). 27 March 2018. https://ianmilligan.ca/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/.

Milone, Mark G. 2002. 'Hacktivism: Securing the National Infrastructure'. The Business Lawyer 58 (1): 383–413.

National Infrastructure Protection Center. 2001. 'Cyber Protests: The Threat to the U.S. Information Infrastructure'. National Infrastructure Protection Center. http://www.au.af.mil/au/awc/awcgate/nipc/cyberprotests.pdf.

Preatoni, Roberto. 2009. 'One Sided Hacktivism (Updated)'. Zone-H. 22 June 2009. https://www.zone-h.org/news/id/4713.

Richterich, Annika. 2018. 'Tracing Controversies in Hacker Communities: Ethical Considerations for Internet Research'. Information, Communication & Society, July, 1–18. https://doi.org/10.1080/1369118X.2018.1486867.

Samuel, Alexandra. 2014. 'Hacktivism and the Future of Political Participation'. Cambridge, Massachusetts: Harvard University. http://www.alexandrasamuel.com/dissertation/pdfs/Samuel-Hacktivism-entire.pdf.

Streeter, Thomas. 2011. The Net Effect: Romanticism, Capitalism, and the Internet. Critical Cultural Communication. New York: New York University Press.