Web archiving: The online presence of *The University Observer*

Class: Digital Preservation
Lecturer: Kara Van Malssen
Students: Michael Grant, David Neary, Lorena Ramirez-López

INTRODUCTION

As the focus for our web crawl, we selected *The University Observer*, a student-run publication at University College Dublin, Ireland, and contacted the paper's editor to enquire about its web-based assets. The editor of *The University Observer* changes annually, and the newly appointed editor, Cormac Duffy, finds himself not only at the helm of Ireland's largest student newspaper, but also the custodian of *The University Observer*'s print and digital archive. Recently, on the 20th anniversary of the newspaper's founding, the UCD Students' Union, who finance the paper under a promise of editorial independence, pledged to fund a massive digitization project that would allow for previously print-only editions of the newspaper to be made available in a large online archive.

When we approached Mr. Duffy, he was delighted at the idea that the *Observer*'s website, universityobserver.ie, a cornerstone of the newspaper's recent successes, could similarly be saved as it appears today. The first online edition of *The University Observer* was published in the 1998/99 academic year, but was not maintained and was discontinued. The current website has been in operation since 2009 and currently contains online copies of all major news, sport and comment articles from the newspaper from 2008 to 2014 (Volumes XV through XXI). In addition, PDFs of every edition from Volume XVI onwards have been saved digitally and made available online on the website issuu.com.

As additional issues of the newspaper are added to issuu.com, moving forward as well as looking backwards as earlier volumes are digitized, Issuu will help handle the archiving of *The University Observer* as it appears in print. But the website has its own display method for articles that are produced by students volunteering their writing, which allows for search engine access as well as article linking, and even providing space for comments by readers.

Although the content of the print and online editions is largely identical, the look and layout of the two, and the functional interfaces by which they are experienced, barely resemble one another. Going back and reading an archived news story does not impart the same feeling, or even all of the same *information*, as accessing an archived copy of the website, complete with its carefully laid out main page, and reading the story in that context. And so, with Mr. Duffy's blessing, we began a web crawl using Archive-It to archive universityobserver.ie as it looks today, in addition to the online PDFs at issuu.com, using the seed issuu.com/universityobserver/.

SCOPE

We aimed with our crawl to archive everything related to the newspaper's website, universityobserver.ie, and additional web pages related to the paper.

universityobserver.ie

On the surface, our crawl of *The University Observer*'s main website appears to have gone off almost entirely without a hitch. The site displays in Archive-It exactly as it does on the live web, with the exception that the panels showing how many times articles have been shared on Twitter, Facebook, and Reddit are partially non-functional. We attempted a patch crawl, in case it was a difficulty with the embedding, but this has not yet corrected the problem. We are hopeful that it is a result of robots.txt files on certain of the social networking sites, and that we can thus overcome the brokenness of the panels by overriding those blocks.

Additionally, we found some surprising external webpages archived on the basis of the universityobserver.ie/ seed, including one from the *Hollywood Reporter*, and a blog post about fan fiction. On investigation, however, we found that the *Hollywood Reporter* URL was an image embedded in a movie review, and that the fanfic article passingly referenced an article on the same subject on the *Observer*'s website (although we have not managed to locate where the link is reciprocated, and therefore why *precisely* it was included in the crawl). So in spite of the seeming peculiarity of these URLs, they do fall legitimately within the scope of our project.

The Archive-It help pages contain a page of advice on scoping online newspaper searches, which point to event calendars, printer-friendly story displays, "email to a friend" URLs, and log-in pages, as traps into which a crawler on a news page might fall. None of these was troublesome for the *Observer*, but we similarly found that URLs associated with each story had been archived for RSS feeds and for "trackbacks", which notify the paper when an external site links to the story. As these do not represent any version or aspect of the paper that can be viewed through the archived version, they should be scoped out for future crawls.

Issuu

Issuu.com is a digital publishing platform for magazines, catalogs, and newspapers. It was launched at the end of 2007 in Copenhagen, Denmark, by "a bunch of geeks who love type, code, magazines and table tennis".[1] It provides readers with more than 19 million publications from around the world; topics range from fashion and arts to global affairs. "Issuu has grown to become one of the biggest publishing networks in the industry. It's an archive, library and newsstand all gathered in one reading experience."[2] A user may create an account complete with username and password for one of the four billing plans: basic, plus, premium, and custom. Each plan is able to link via Facebook, Google, LinkedIn, Tumbler, Twitter, Pinterest, or email.

The website's functions and interactive nature make crawling the site very difficult. It was noted that javascript is one of the limitations that affect not

[1] issuu.com/issuustaff

[2] issuu.com/about

only Heritrix, but other web crawlers as well. The entire "interactive newsstand" display, as well as the majority of the actual publications on the site, are not able to be archived. The Wayback Machine returns a "0: undefined error". Yet the information and images from the top section of the layout are archived by the crawler. This metadata can be found within the script text of the <!doctype html> by viewing the source. The actual content seems to be within CSS texts and more complicated javascripts, which explains why the crawler has difficulties in archiving it. This issue is similar to the results when archiving Instagram, as we will see below.

Social Media on Archive-It

In order to fully capture *The University Observer*'s online presence, it felt necessary for us to include its social media presence. Social media has become indisputably integral to the operations of media outlets, and student newspapers are by no means an exception to this rule. Since the advent of *The University Observer*'s website and the repercussions of the financial crisis, recent years have seen the newspaper's print run drop from 10,000 to 6,000. Thus drawing attention to news stories on the website is now more important than ever. Mr. Duffy agreed it would be a great benefit to the project to have *The University Observer*'s social media presence recorded to the greatest extent possible. Our results were mixed.

Twitter:

Our first crawl of the *Observer*'s Twitter feed was quite successful, but shockingly large at more than 58,000 URLs. Examining the list of Twitter URLs archived, we found that the crawl had not gone too far afield – all of the tweets that we sampled were written or retweeted by the *Observer* account – but a great deal of the list was taken up with non-public facing pages to govern the structure and display of Twitter, and with thirty-seven different language pages for each tweet.

The former are obviously necessary to make the archived page operate, but the language pages (which do not translate the tweet into different languages, but only display the Twitter-wide language on each page ["Have an account? **Sign in**"; "Meron ng account? **Mag-sign in**"; "Heb je al een account? **Inloggen**"; "Đã có tài khoản? **Đăng nhập**"]) are needlessly redundant.

To cut out these extra pages, we followed the instructions on the social media help page of Archive-It and added a host constraint to block all pages with a language modifier ("block URLs if URL contains ?lang="), but allow the tweets' main URLs. This was not successful, and the next crawl returned even *more* extra Twitter results, including the additional language views. For our last crawl, which will finish between the time that this paper goes to press and when we present our results to the class, we tried a regular expression to constrain the crawl, as explained on the help page on modifying crawl scope.

This entire issue presents an interesting question, however: among the language tags are "en" and "en-gb", English and English-Great Britain. Because we are archiving-it from the U.S.A., the version of Twitter automatically accessed is formatted in American English ("en"), while a user or crawler accessing Twitter in Ireland will presumably receive it in "en-gb". The difference between "en" and "en-gb" on a public-facing, non-signed-in Twitter feed is extremely small, but the difference between "favoriting" a Tweet and "favouriting" one is

not negligible. For purposes of this project we are only archiving the standard feed, but if the *Observer* wishes to pick up the project where we leave off they will need to make a decision whether to capture the raw Twitter feed (which, despite their Irish IP address, will likely still be American, as Archive-It is based here in the U.S.A.), or to capture specifically the en-gb edition.

Additionally, we found when reviewing the results of our crawls that a great many files were saved from the mobile edition of Twitter, in all thirty-seven languages. Since this content was duplicative of the main Twitter feed, and since the *Observer* does not optimize their feed especially for mobile devices, we attempted to use Scope-It to remove the mobile page from further crawls. This was similarly unsuccessful, but evidence of our attempt can be seen in the below image.

	Host	Block All	Block URLs if		Doc Limit	Ignore Robots.txt
x	akamaihd.net	☐		Add Rule	click me	☑
x	fbcdn.net	☐		Add Rule	click me	☑
x	mobile.twitter.com	☑				☐
x	twitter.com	☐	URL Contains ?lang=	Edit x	click me	☐
			URL Contains ^.*?lang=.*$	Edit x Add Rule		
x	www.facebook.com	☐		Add Rule	click me	☑

Figure 1 Attempts to add limitations to the Archive-It crawl

Facebook:

In our initial crawls, we found that *The University Observer*'s Facebook page was blocked by robots.txt. Investigating the situation on the Help Wiki, we learned that Facebook blocks everything on its website, and that we could contact Archive-It and have the restriction ignored by the crawler for the *UO*'s specific Facebook URL.

Instagram:

We encountered some troubles, insurmountable for now, with our crawl of the *Observer*'s Instagram page. The front page displays beautifully, but when you click to enlarge a picture, the screen dims in anticipation of the coming image, but the image never comes; the loading wheel simply spins until you close the tab or navigate back to the previous page.

Seeking a workaround, we found on Archive-It's help pages that although their engineers have "improved [Archive-It's] ability to capture and replay Instagram pages", "you may not be able to enlarge images or see the comments and likes for an image". The engineers are "working towards further improving our capabilities for crawling Instagram", though, so we are optimistic that future crawls will be able to display that content. And for now, the crawler is faithfully capturing the basic presentation of the feed, which – assuming the *Observer* staff don't go on an image-purging rampage – is the information most likely to be lost before Archive-It's Instagram capabilities improve.

Wikipedia

A late suggestion from Mr. Duffy was to add the Wikipedia page about *The University Observer* to our crawl, to preserve evidence of changes made to it over the years as well as to group it with the collection to which it refers.

Archive-It faithfully captured the content of the *Observer*'s Wikipedia page, but displays it in a layout unlike that of the page on the live web. We consulted the help pages on archiving Wikipedia, but they were unhelpful in this matter.

FREQUENCY

During our work on this project we opted to set Archive-It to crawl daily, in order to gain the maximum number of results in the time allotted. For continued crawling, it may make sense to archive less frequently. A weekly crawl for the *Observer* page itself – perhaps the night before a new issue is to go up, in order to capture as many comments as possible before the layout changes – would probably suffice. Issuu crawls, if they can be accomplished, could be conducted even less frequently. Given the dynamic nature of social media, however, the *Observer* may wish to continue with daily crawls of Twitter and Facebook; if the paper were involved in a contentious public debate, for instance, frequent crawls increase the chances of capturing interesting and inflammatory comments and tweets before their authors cool down and delete them!

METADATA

Basic technical metadata was created for each seed web page using Archive-It's Dublin Core pages. In each case we created a suitable list of metadata entries for each website, and added searchable topics and additional metadata such as publisher and coverage, where applicable.

IP/PRIVACY CONCERNS

Since this crawl is not gathering any information that has not been put out for public consumption, we do not regard any of the collected data as presenting a privacy concern. Some tweets and comments (perhaps even some news articles!) may be embarrassing to their authors later on, but not even in Europe is there a right to make public statements and have them ignored.

As for intellectual property, the issues are more interesting: the *Observer* enters into no formal legal arrangements with its contributors, and therefore does not hold the copyright to their work. Indeed, in reviewing the reports of one of our crawls we discovered the blog of one Evan O'Quigley (quiggersblog.wordpress.com), described as "a place to keep my articles, mostly written for the University Observer". But it is tacitly understood between the paper and its writers that their work will be published in print and online; unless the *Observer* begins to find ways to monetize its website, it would be unprecedented for its contributors to make a copyright claim against the paper.

Archiving the website should not present any additional legal challenges, especially given the approval of the editor in doing so. The legality of social media crawls is not well ruled-on, but this job is within the bounds of an ordinary crawl of its kind, and is not capturing anything not freely put out to the public. No legal issues are foreseen.

FINDINGS/PROBLEMS/CONCLUSIONS

This initial crawl has not been without its difficulties: after the initial groping around to find our way in Archive-It, we found that the online help documentation has not always been easy to navigate, and has (as in the case of language exclusions for Twitter) seemingly not always given accurate advice. And it may be partly our lack of expertise, but the issues we uncovered with Issuu are troubling. Archive-It doesn't archive Instagram properly, but its engineers are actively working to correct the problem; there is surely less impetus to work specifically on a site like Issuu, yet there was nothing that we could work out on our end that we could do to manipulate the software to capture the uncaptured content.

The process also led to some interesting discoveries, and insights into how web crawlers – and, by extension, web logic – function. If the organization of the help pages is sometimes not all that helpful in the way they are intended to be, they did lead us to make discoveries on our own that might otherwise have been simply explained in the instructions. Finding the Twitter language problem through reviewing reports, rather than by simply first following an instruction on how to avoid the problem, was a learning experience that made a valuable impression on our brains.

In investigating some of the difficulties that we encountered, we checked for our seeds on the archive.org and made another remarkable discovery: that the Internet Archive had *taken our crawl results*. In the case of the *Observer* Issuu page – which was not even archived with any reasonable degree of success! – it had not been touched by the Wayback Machine in all of 2014, but began to be crawled by it when we started our test crawls. This might seem underhanded if we were paying a less sympathetic organization – Google, say – for archiving services, but given the public spirit of the Internet Archive, it seems far more like a laudably innovative way of simultaneously providing a commercial service to Archive-It's customers, and a public service to users of Archive.org.

Cormac Duffy told us he was delighted with the success of the crawl in archiving universityobserver.ie, and felt that everything else was of secondary importance – Issuu, the Facebook and Instagram pages, none of which were properly archived by our crawl. It would be easy for staff at *The University Observer* to take over responsibility for this crawl if they so wished, capturing snapshots of their evolving website in time. If they were, we would recommend their attempting to correct the issues we encountered, by hassling Archive-It about Issue.com, requesting a Facebook override, and investigating how the Twitter feed might be correctly archived without crawling through every language version as we had to.

The most important thing is that the site is saved as it is now, so that no matter what happens to *The University Observer* in future (whether it is dropped to make up Union expenses or abandoned by the jaded or lazy students of tomorrow), this integral aspect is not lost like the original *University Observer Online* from 1999.