Lightning rounds: Archiving news at the Internet Archive

Mark Graham, director of the Wayback Machine, Internet Archive | Oct. 14, 2016 | Charles E. Young Research Library, UCLA | Dodging the Memory Hole 2016: Saving online news

---

MARK GRAHAM: ***[00:10]*** Hi, I'm Mark Graham. I have the honor of working at the Internet Archive managing the Wayback Machine. Most recently, I worked at NBC News Digital. I also remember that 25 years ago I was publishing FAIR, Fairness and Accuracy In Reporting online. So, this is the Internet Archive; we're a nonprofit library based in San Francisco. We're about to have our 20th anniversary on October 26. Please come if you can. There are five different ways we're archiving the news. One is a focus crawl that we do ourselves. It generates about 30 million URLs a week. Another is, we work with a series of volunteers whose mission is to archive all online news.

[00:52] We're currently capturing 50 million URLs a week with this project called NewsGrabber. There's a public webpage on it, and you can go in and see where all the sources for our seeds are, and you can add to them, as well. We also work with Kalev [Leetaru], who will be speaking here on his GDELT project and currently capturing 25 million URLs a week with that project. The fourth way we're capturing and saving news is using our Archive-It subscription service.

[01:25] There are many of our partners that are using us, specifically, to archive the news. This is one particular crawl that I set up, archiving North Korean news. I am capturing something like forty-some North Korean sources every day. And then, finally, Save Page Now; this is where anyone can come to our service, click one button and save a page. This button is pressed 50 times a second, 24/7. An example of something that got saved via this method and, to the best of our knowledge, the only way this was saved, was someone boasting on a Russian social media site that they had downed an aircraft at the time and location of MH17.

[02:08] OK, now a little bit about what we're doing to make all this more accessible. And when I say all this, we right now have more than 500 billion URLs captured on the Wayback Machine. Next week, if all goes according to plan, we'll be launching our next version of the Wayback Machine with site search. Here's an example where I typed in, "New York Times," and see, "NY Times" came up, so that worked. We also are surfacing the provenance for each of our captures. So, in this case I've done a rollover on a particular capture, and you can see the why, why we have that capture.

[03:45] Among the more than 500 crawls a day that we conduct at the Archive, each of them has a number of captures, and so you can find out why we have a particular capture. And then, finally, a new feature is a summary of the mime-type breakdown for a host, a domain, and/or a top-level domain. We've got a little project we call "No More 404s." There's a few things we're doing around that; most recently, we started running the PURL.org permanent URL service of OCLC. We are working in collaboration with Harvard on their AMBER project, and obviously working with WordPress and the Jetpack plug-in.

[03:25] We right now use this to monitor, actively, more than 20 million WordPress sites and are actively archiving them. And then, finally, we partnered with Mozilla Foundation on a plugin for Firefox that tens of thousands of people have installed over the last few weeks. In the time that I've been speaking, we have archived more than 30,000 news URLs and more than 300,000 URLs overall. But, you know, we could really use your help.

[03:55] So, I just want to put it out there, if anyone has any ideas about how you could work with us to help us do a better job of preserving and presenting news, and frankly everything that is published via the public web, please contact me — Mark@archive.org — and please check web-beta.archive.org and give us any feedback. Thank you very much.