Lightning rounds: Supporting data-driven research using news-related web archives

Jefferson Bailey, director, web archiving programs, Internet Archive | Oct. 14, 2016 | Charles E. Young Research Library | Dodging the Memory Hole 2016: Saving online news

JEFFERSON BAILEY: ***[00:11]*** Before I start, one of our sort of mantras at IA, is, "bits in, bits out." So, I like that Mark did "bits in" and then I'm going to do "bits out." Let's see how it goes.

[00:38] So, I wanted to talk about trying to get web archives into the hands of researchers, especially for computational and data-driven research. What happens with web archives when we crawl the web is they go into these things called WARC files, that's Web Archive Resource Container, I think. It is an ISO standard and a piece of just maddening junk.

[01:18] It's concatenated gzip file of concatenated gzip files, so it's totally arbitrary in what goes in there. The crawler opens the file and it starts pumping in data. You can set a size and it will close off the file, at that point, and then package it up and it's a WARC file. They can have tens of thousands of HTML pages in there, or it can have one big YouTube video file in there. So very arbitrary. That makes it very hard for researchers to work with because it's a somewhat obtuse format. It's also, of course, most of our crawls; we have about 13, 14 petabytes of web data.

[01:55] So that's a lot to work with. And even things like Archive-It accounts are often multi-terabyte collections, and people are very challenged as far as doing data mining on them. So, what we're trying to do is take the stuff — and, of course, the last one is that there's often IP or content restrictions to giving people stuff that is in copyright. So, a lot of partners that we work with in libraries, archives, museums, don't want to hand out full, raw content, but they do want to support data-driven research. What we're trying to do is we make these, what we call "derivative data sets." We're doing things like pulling out specific metadata elements that are not full page text, putting them in raw JSON files instead of in WARC files, and then often building API's and other access methods on top of those.

[02:41] There we have WAT files, which is sort of key resources from each record in the WARC. Things like page title, anchor text, meta tags; obviously all the capture provenance, HTTP information and stuff like that. We have graph files that are "this links to this" with a timestamp for every single capture across a big collection. And then we are doing things like named entity extraction, topic modeling, and other things like that. So, we take all this stuff that is valuable within the WARC, but is not the entire WARC, put it in JSON, and then make API's on top of it and give that to people.

[03:18] We do, sometimes, big transfers of data to people on disk, but we're really excited about things that are facilitating programmatic access to archive content, while getting around some of the size and complexity issues. So, how am I doing on time? OK. Well, I can post a bunch of scripts. I was going to demo like, "you can just do little shell scripts and get crazy awesome information." But, instead, I thought what I would show, because it's relevant, and it's coming out for our 20th anniversary, is that Venagle, who is one of our senior data engineers, has been doing comparative capture snapshots of front pages, and one that we thought would be awesome would be news.

[04:10] And it goes on for a while... It gets better when it gets low. So, it's USA Today, New York Times, and WaPo (Washington Post). So, we'll put this media file up in the next week or two. You can, sort of, play it and pause it. They're all on the same day, so it's really interesting to both track the, sort of, visual change. You know, the Web Archive's 20 years old, so you can look at all the graphic design changes, but also see how different stories appear in, quote unquote, the three most popular U.S. newspapers.

[04:40] Yay, Research stuff!