Journal of Western Archives

Volume 8
Issue 2 *Web Archiving Special Issue*

Article 3

2017

Using RSS to Improve Web Harvest Results for News Web Sites

Gina M. Jones
Library of Congress, gjon@loc.gov

Michael Neubert
Library of Congress, mneu@loc.gov

Using RSS to Improve Web Harvest Results for News Web Sites

Cover Page Footnote
Opinions expressed here are not necessarily those of the Library of Congress.

Using RSS to Improve Web Harvest Results for News Websites

Gina M. Jones
Michael Neubert

ABSTRACT

In the last several years, the Library of Congress web archiving program has grown to include large sites that publish news – over more than a year we learned they present serious challenges. After thinking through the use cases for archived online news sites, we realized that completeness of harvest was paramount. As we developed our understanding of deficiencies in the completeness of these kinds of sites we began to test use of RSS feeds to build customized seed lists for shallow crawls as the primary way these sites are crawled. Over time we discovered that while completeness of harvest was greatly improved, we had a new problem with the ability to browse to all harvested content. This article is a case study describing these iterative experiences that are a work in progress.

In the relatively young field of web archiving, the Library of Congress has one of the largest programs in operation, with the total amount of web content harvested and managed as collections since 2000 now exceeding one petabyte. While many aspects of the work have become routine, one of the continuing challenges is harvesting larger websites that are within our collecting scope. For large news sites in particular, there is an added difficulty because those sites update, revise and add new content frequently. We continue to review our harvest results in the hope that we have worked out a successful approach of combining less frequent crawling of the entire site with the use of aggregator feeds, often referred to as RSS (Really Simple Syndication), to build custom seed lists that assure more complete results as we move forward.

The goal of this paper is to describe the thinking and sequence of efforts in developing a methodology to satisfactorily harvest large websites in a knowledgeable manner which may be useful to other Web archiving programs. This article is not describing an approach that has been fully perfected at the time of publication.

Background

The Library of Congress (LoC) Web archiving program began early in the 21st century as a pilot program that developed into the present Web archiving program.[1] Since LoC did not (and does not) have a legal mandate to harvest websites from U.S. publishers under existing copyright or other law, the program began slowly, with a focus on collections related to particular themes or events. Early on, the most significant collecting effort was the biannual national election cycle, harvesting the sites of congressional, gubernatorial, and presidential campaigns, which continues today. There were also collections related to particular events such as the Internet's reaction to the September 11 attacks and the 2003 War in Iraq. The Congressional Web Archive (House and Senate Web domains) began in 2002 as a one-time harvest, and in 2003 was the first collection that the Library began harvesting on a continuous basis. After 2010, the program began an effort systematically and regularly to crawl all the sites of the legislative branch within the U.S. federal government, adding non-congressional house.gov and senate.gov domains such as clerk.house.gov, as well as smaller agency sites such as that of the Architect of the Capitol, as well as the Library of Congress' own loc.gov site. A few of these sites were quite large—in particular, the LoC's own website which has multiple subdomains and aliases.

About our use of the word "large:" we use "large" because we are not aware of a better, more technical term in use by the Web archiving community and because it accurately describes the problem—a site may have more content than can be fully harvested with a typical level of effort. For "large" sites, we have tried various measures with varying success, most recently the RSS-driven approach described here.

In 2012, the LoC Web archiving program expanded a collection known as the Public Policy Topics Web Archive (https://www.loc.gov/collections/public-policy-topics-web-archive/about-this-collection/) which included websites for the U.S. Department of State (http://state.gov), the White House (http://whitehouse.gov), and other websites too large to harvest fully with a crawl lasting eight to ten days. We realized that monthly method of harvesting for only eight to ten days resulted in a shallow harvest of these very large sites and our archives would thus be incomplete. We tabled the expansion of selecting and collecting other large agency sites while we evaluated the significance of these fragmented and incomplete results for future users.

Our first strategy to deal with large sites was to change from monthly crawls of eight to ten days to quarterly crawls of a much longer period of time. At that point,

1. Abbie Grotke, "Web Archiving at the Library of Congress," Information Today, published December, 2011, http://www.infotoday.com/cilmag/dec11/Grotke.shtml (accessed August 15, 2016). Although this article was published in 2011, it provides a thorough and accurate overview of the program's first decade.

contractually, we had two options for crawling, every week for seven days or every month for eight to ten days. We determined from the study of crawl reports and other available data that many URLs of digital objects were identified by the crawler for large sites but not retrieved because the crawler ran out of time. Our first attempt to achieve complete crawls was to institute a quarterly crawl for large sites that would run for eight to ten *weeks*, instead of days. From a work management basis, this meant we had taken on a process of managing the seed nomination and crawl quality review processes of two crawls (weekly, monthly) and to that, added another major effort, this new quarterly crawl. These three crawls total approximately 4,500 selected websites for approximately 250 terabytes annually.

We also realized over the last five years that future researchers would expect the Library of Congress to have in its collections Web-based general news publications that have the extensive readership and original content found on HuffingtonPost.com, DailyKos.com, or Townhall.com (as a few examples). We added several such news sites to our Public Policy Topics Web Archive collection, harvesting some on a monthly basis and others weekly. We were particularly disappointed with the results of the initial weekly harvests and decided that we should identify our desired end results in order to help us develop the best possible solutions. The first step was to review any existing use cases so we consulted the website of the International Internet Preservation Consortium (IIPC, http://netpreserve.org), a non-profit organization that began in 2003 with the goal of helping organizations preserve the Web.

The IIPC provides a list of ten use cases (also referred to as "case studies") on their website.[2] One use case that is of particular interest for our efforts, *News in the 21st Century*, notes "Libraries and archives have always collected newspapers, these are the core collections of many local historical societies. If the news that is distributed online is not preserved there will be a huge hole in our collective memory."[3] This brief statement on the IIPC site is just the beginning of a useful use case for online news.

All the use cases for Web archives on the IIPC site assume that the Web crawler delivers an acceptable harvest without identifying the characteristics of an acceptable harvest for the particular use case is. For online news, the completeness of capture in particular is not mentioned, which was where we identified a major failing in our efforts—many if not most news items were not captured even once over multiple harvests of a news site. We knew that users of pre-Internet print newspaper resources typically fit into one of several "use cases" (although they didn't think in terms of "use cases"): find a particular news item on the basis of a citation or other information; or

2. "Case Studies," International Internet Preservation Consortium (IIPC), http://netpreserve.org/web-archiving/case-studies (accessed August 15, 2016); Emily Reynolds, "Web Archiving Use Cases," published March 7, 2013, http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf (accessed August 15, 2016).

3. "News in the 21st Century," International Internet Preservation Consortium (IIPC), http://netpreserve.org/case-study/news-21st-century (accessed August 15, 2016).

search for news items related to a particular event or topic published in a particular period of time; or perform some other broader subject or name search. Some of these activities would be done using indexing services external to the published newspaper. In the context of print news, libraries understood a complete cache was important for their users and so they needed to provide as complete a collection of the newspaper titles being acquired as possible. A collection of newspapers that was merely a representative sampling of published content would have very little research value for future researchers.

After talking to reference and collection development staff at the Library of Congress, we concluded that the use cases, along with the importance of complete holdings for print newspapers, transfers to use of online Web-based news sites. Users will want to be able to retrieve news stories for which they have a "citation" (URL), search or browse for news on particular topics or about particular people and events during particular time periods, or do larger searches for information on particular topics within the Web archive of a particular news site (or sites). There will be additional use cases for archived news sites, including text mining, which has been demonstrated by the nGram viewer for digitized newspaper content in Chronicling America (http://bookworm.culturomics.org/ChronAm/). All of these use cases suggest the importance of trying to capture each published news item (webpage, typically) *at least once*. This can be contrasted with the "traditional" Web archiving approach that will often capture some webpages on multiple occasions over time in order to document changes that took place in a particular page, but in the case of large and rapidly changing sites, some content will be missed repeatedly, never to be harvested.[4] We will now look at how we assessed the size of news websites and the completeness of capture.

The first technique is a simple one that provides a rule-of-thumb estimate of a website's size using Google search's "site:" command—its usefulness is not limited to news sites, but can be used with any website.[5] Entering any URL, typically dropping "http://" and any version of "www" after the typed-in search limit "site:" in Google without a search term (search argument) will return from Google an approximate count of the number of digital objects (files) indexed by Google under that URL. Thus "site:loc.gov" on July 5, 2016 returned 8,360,000 results. While we do not have a

4. NewsDiffs, http://newsdiffs.org/ (accessed August 15, 2016). Capturing all the changes made to online news texts is an art unto itself, best represented by the NewsDiffs project; it would be impossible for any conventional Web archiving program to track all such textual edits that may happen minute to minute during some periods. The NewsDiffs site "archives [textual] changes in articles after publication" for several online news sites, including NYTimes.com, WashingtonPost.com, and several others. It is intended specifically (and narrowly) to provide access to all changes made to an article's text, allowing full review of successive versions of these texts as they appeared on their respective sites over time.

5. The Google "site:" command will only retrieve results available to it based on any robots.txt restrictions for a particular URL. Most news sites that want users to be able to find their content via search engines will have relatively non-restrictive robots.txt.

benchmark or standard for our use of this technique, we consider that a site with over a million addressable digital objects as counted by Google is "large" and over five million, "very large." The larger its size, the more a site presents serious challenges to harvest it completely (see Figures 1).

When we have identified a particular site as being large, we can then look at our success at harvesting the site if we have already done so. The Library of Congress Web archiving curator tool includes information on the URLs identified by the crawler as it traversed a target website in its initial pass and then the number of these URLs "left in queue" before the crawler gave up and moved on to the next assigned target (seed) URL. In Figure 1, the Total number gives some sense of the size of the site as assessed by the crawler within the time allotted. Unfortunately, there is no formula for correlating that figure with what the Google "site:" command reports for a particular site. For large sites, a significant gap between these two numbers would be a sign of failure to gather all the available URLs for "crawlable" content. We often observe in large sites a significant gap between what was identified to be crawled and what was actually crawled, indicating that the crawler ran out of time to adequately crawl a site before needing to move on (see Figure 1, "URLs not Crawled").[6] Of note, numbers are based on just what the crawler had time to discover, not the number of digital objects (files) on the site.

Capture Date	Response Code	URLs Crawled			URLs Not Crawled
		Total	New (%)	Duplicate (%)	
Mar 2, 2016	200	69,394	92%	8%	257,795
Feb 3, 2016	200	38,818	95%	5%	254,253
Jan 6, 2016	200	66,179	93%	7%	247,515

Figure 1. Weekly capture data for three crawls for Huffingtonpost.com

With large news sites that rapidly and continuously add new content (though often it seems rarely if ever remove any), an additional and arguably more significant test is how the archived site performs with simple clicking on the homepage and beyond to access different articles that were within scope of the crawl. Our first attempts to harvest HuffingtonPost.com used a weekly harvest schedule (or

6. The Library of Congress uses a curator/URL scoping software tool, known as the DigiBoard, that was built by Library of Congress software developers to Library of Congress requirements for managing its Web archiving that is performed by a contract.

frequency); however this approach only allowed a crawl of a few days for each crawl before the next weekly crawl cycle would commence, starting over. Of note, most crawlers harvest using a "politeness" (time interval) factor so as not to overwhelm Web servers. The maximum crawl result, and in turn its completeness, is a function of the time available for the crawl and a given number of documents that can be captured per second. More content harvest could be achieved simply by turning up the speed of the crawl in real time but our primary goal in harvesting is to not cause issues on websites so politeness is critical for continued relationships with content owners.

In analyzing what was crawled and comparing to what was not crawled (the URLs remaining), the reports suggested that materials were not harvested, resulting in what we would have in our collections as incomplete. Anecdotally, in what could be called a hands-on "click-click" quality review of the result, too many documents turned up "Not in Archive" (see Figure 2) as we randomly clicked through pages in the archived version of HuffingtonPost.com.

Figure 2. An example of a "Resource Not in Archive" in the Library of Congress Web Archives

One aspect of news sites that is different from many other sites is that many news sites use a publishing mechanism to inform subscribed users about new content called Rich Site Summary, also known as Really Simple Syndication or RSS. RSS uses a family of standard Web formats to publish frequently updated information on websites. An RSS document (which may be called a "feed," "web feed," or "channel") includes full or summarized text, metadata, and a URL of the published news item on the publishing organization's site.[7] http://www.huffingtonpost.com/feeds/index.xml is an example of an RSS feed that the Huffington Post publishes for all news published across its site. RSS feeds from news sites are not created with Web

7. "RSS," Wikipedia, last modified January 12, 2017, https://en.wikipedia.org/wiki/RSS (accessed August 15, 2016).

archiving support in mind, of course, but we considered adapting them for this purpose, specifically to build shallow seed lists of published pages to be harvested one -by-one.

In "regular" Web archiving, the crawler starts harvesting on a specified page (seed URL), typically a site's homepage, then downloads all relevant content to create an archived version of each page, including images, JavaScript, HTML and other elements. URLs waiting to be crawled are managed by what is called the "frontier", an activity manager for the crawling. All new URLs are entered into the frontier where they are scheduled for crawling. When a URL comes up for crawling, it is emitted by the frontier. At any given time, there may be multiple URLs being crawled. Each emitted URL then passes through a chain of processors. This chain begins with preparatory work (e.g. check robots.txt) and proceeds to actually fetching the URL, link extraction, and WARC writing.[8] Eventually links discovered in the downloaded document are processed.[9] The content then may be post-processed to remove duplicate digital components already harvested in previous crawls.

For news sites, our revised strategy was to use the RSS to build what amounts to a shallow seed list of newly published pages to be crawled. The crawler first visits the RSS feed and discovers the new content through that publishing mechanism, the RSS feed. The crawler crawls the RSS not to build an archived version of the RSS but to create a focused seed list from the article URLs given in the RSS. Harvesting RSS in this way is not a new idea. The National and University Library of Iceland (Landsbókasafn Íslands Háskólabókasafn) has published an add-on to the crawler Heritrix (https://github.com/internetarchive/heritrix3), the result of an IIPC project initiated in 2003 and led by the Internet Archive, that developed support to handle RSS feeds in this way. The add-on, Crawl RSS (https://github.com/kris-sigur/crawlrss), can be scheduled to visit RSS feed pages multiple times per day and then in turn, harvests new content from the target site based on the quasi-seed list built on the results of crawling the RSS.

The Library of Congress began an initiative in early 2015 to begin capturing RSS feeds for selected websites twice-daily (every twelve hours) and in turn harvest only the pages listed in the feed items. The harvester stays narrowly focused and does not crawl more than that one link off from the feed page. It captures all files necessary to rebuild and represent the pages identified in the RSS.

An RSS feed-driven crawl will only support the crawling of pages as they are added to the defined site and announced via RSS; it will not support the pages that already make up a website and may have been announced via RSS feed prior to this

8. "Web ARChive," Wikipedia, last modified May 12, 2016, https://en.wikipedia.org/wiki/Web_ARChive (accessed November 9, 2016).

9. Kristin Sigurdsson, "Implementing CrawlRSS," Kris's Blog, posted March 10, 2015, https://kris-sigur.blogspot.com/2015/03/implementing-crawlrss.html (accessed August 15, 2016).

initiative. For this reason, the Library of Congress continues to crawl these sites either monthly or quarterly in addition to using the RSS-driven approach.

Our analysis indicates that augmenting regular harvests of websites with the additional harvest of selected RSS feeds for larger sites provides far more complete results. Figure 3 shows the difference in documents captured before and after the Library started using the RSS harvest strategy. At the aggregate level and over a period of time, RSS should provide for more completeness; complementary "regular" website harvesting would provide the framework of links to click into content.

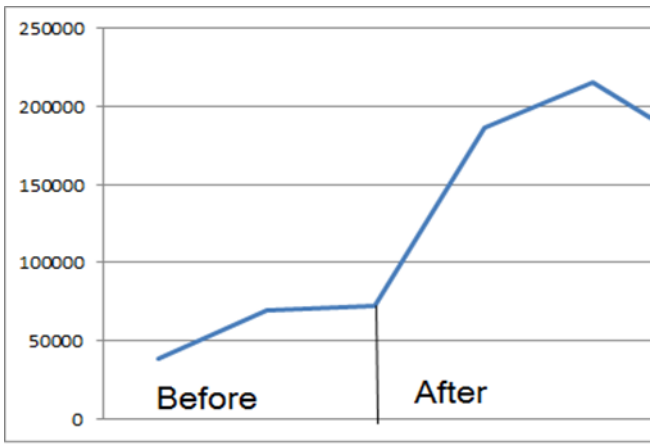

Figure 3. Document count for Huffingtonpost.com before and after RSS

The overall aggregate as shown in Figure 3 is somewhat misleading; Huffington Post has multiple RSS feeds that we used as the basis for our harvesting which overlap with slightly different URLs in different feeds for the same article. Huffington Post apparently does this in part to track user clicks by subject (RSS feed). For the harvested version of the site, it would mean we harvest as many versions of a particular news page as the news publisher chooses to provide different URLs based on their business needs. For example, in our archive we have one same story with five different URLs, reflecting the appearance of this one news item in multiple Huffington Post RSS feeds that we then used to drive our harvest of that very large site.

1. http://www.huffingtonpost.com/2015/11/02/daphne-rubin-vega-calls-out-racism-ageism-and-sexism-in-show-business_n_8473648.html

2. http://www.huffingtonpost.com/2015/11/02/daphne-rubin-vega-calls-out-racism-ageism-and-sexism-in-show-business_n_8473648.html?utm_hp_ref=black-voices&ir=Black+Voices

3. http://www.huffingtonpost.com/2015/11/02/daphne-rubin-vega-calls-out-racism-ageism-and-sexism-in-show-business_n_8473648.html?utm_hp_ref=fifty&ir=Fifty

4. http://www.huffingtonpost.com/2015/11/02/daphne-rubin-vega-calls-out-racism-ageism-and-sexism-in-show-business_n_8473648.html?utm_hp_ref=latino-voices&ir=Latino+Voices

5. http://www.huffingtonpost.com/2015/11/02/daphne-rubin-vega-calls-out-racism-ageism-and-sexism-in-show-business_n_8473648.html?utm_hp_ref=women&ir=Women

Although this issue exists to some extent in the regular harvests, RSS subject-oriented harvests by publisher-defined topics significantly increase the possible number of duplicate harvests of the same stories. Fortunately, most news producers have fewer separate feeds with less overlap. As with much of Web archiving, the burden is on the organization doing the harvesting to work through these issues since the publisher is not tailoring its RSS feeds to support use of Web archiving but rather to serve their own organization's publishing goals.

RSS-Driven Crawling and Challenges for "Traditional" Web Archive Navigation

There are a number of challenges with replay and RSS harvesting. The goal of RSS feed pages is to supply users with links that take them directly to content on webpages without browsing through a website to find that material. Most of our Web archive users today however have to browse in order to access the archived pages. We presently are harvesting news site webpages that will result in not being linked back to them from other pages in the archive. For example, if the following article is selected from the RSS feed for the Indian Country Today Media Network, http://indiancountrytodaymedianetwork.com/department/american-indian-history/feed, it is not likely that the archived pages that LoC has for this site will have any browsable page or pages that have a link to this news item.

William Howard Taft: Let Derogatory 'Wild West' Movies Slide
Tuesday, July 05, 2016
Editor's note: Voters this year will elect the 45th president of the United States….

Lacking browsable linkages from other pages in the archive, this and other harvested articles are unfindable with our present Wayback Machine approach (see Figure 4).

Figure 4. Typical Indian Country story page (http://indiancountrytodaymedianetwork.com/)

The Library of Congress hopes to offer full-text search of its Web archives that would provide the missing ability to leapfrog to webpages that have been archived based on RSS feeds, however this type of search is at least several years off.[10] Our present goal is to achieve greater completeness in what we acquire from news sites despite the possible access limitations to some of these materials. One short-term solution would be to provide links to the harvested RSS feeds as well as the homepage of the archived website, in order to navigate to these materials. In the long-term, we think full-text search will be the most efficient way to find such content in archives with billions of documents and millions of pages. It also seems likely that the absence of browsable links for such sites will be unimportant for advanced use cases such as textual analysis using nGram viewers.

An even more dramatic example of the absence of browsability is provided by our RSS-driven harvest of PR Newswire (http://www.prnewswire.com/). While a limited amount of the site's press release-type content is available via browsing, most of it is announced via third party sites or discoverable via Google. For example, our regular harvesting effort discovers and harvests on average, 50-100 prnewsire.com articles per

10. Vinay Goel, "Beta Wayback Machine – Now with Site Search!," Internet Archive Blogs, posted October 24, 2016, https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/ (accessed November 7, 2016). This is a good example of a recent and promising development in full-text site search.

crawl. In a one month RSS crawl, we harvested over 20,000 prnewswire.com articles. The only reliable way to acquire a significant portion of the content is via harvest of the RSS feed that the site supplies. For example, http://www.prnewswire.com/rss/policy-public-interest/public-safety-news.rss (retrieved August 16, 2016) had an article "TruGreen Partners with American Red Cross - Supporting Effort to Reduce the Nationwide Drowning Rate among Children" published on Tuesday, August 16, 2016 at 10:00 AM EST. One can find this article on the sites in Figure 5, but the Library is not harvesting any of those sites, so the only way to archive this is through the RSS crawls.

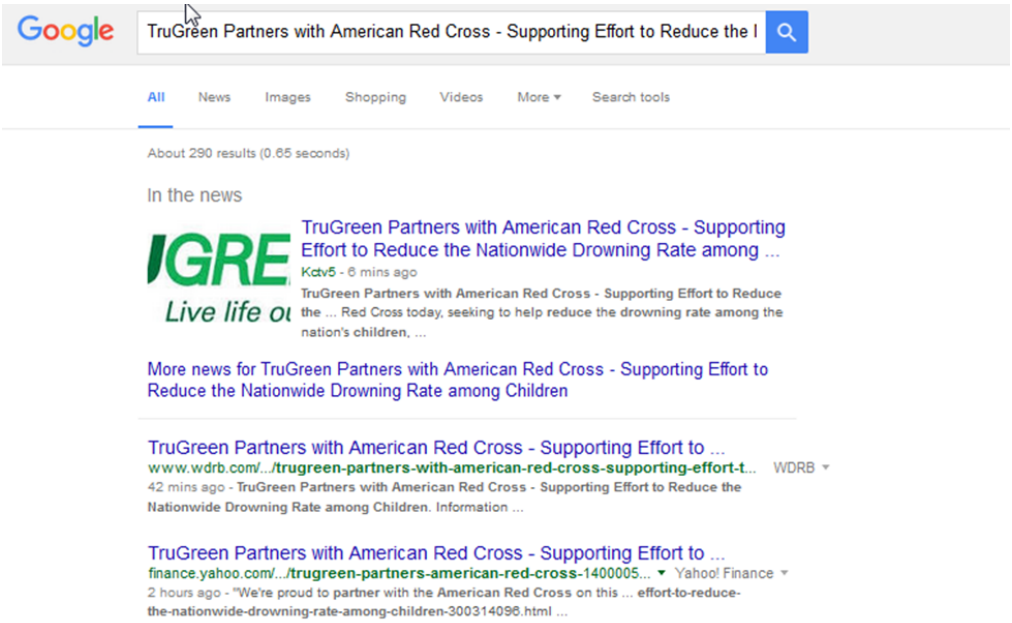

Figure 5. Snapshot of Google search for a PRNewswire article retrieved August 16, 2016

Again, while for now we are not sure how access would be provided, we see clear value in a site where many different corporate and other entities pool announcements that we expect to be useful research collection materials—it is compelling for what it will build up to over time, documenting many different aspects of how the many different organizations and companies that use the platform talk about themselves.

Sites Restricting Crawler Access

Many websites monitor IP addresses for traffic hitting their sites and when a bot is noticed that is perceived as behaving badly, webmasters will stop that traffic. As a result, Web crawlers are configured to be "polite," to throttle back the rate of

interactions with the target site. If despite best efforts at politeness in performing a crawl and a site owner excludes a crawler, then it's necessary to communicate with the organization directly and attempt to resolve the situation. For example, the U.S. Department of State has RSS feeds for a significant number of topical areas (listed on http://www.state.gov/misc/echannels/66791.htm). Adding RSS-driven crawls of these parts of state.gov seemed like a good measure to improve capture of the five million plus documents discovered by the crawler, again with a focus on newly added materials. Unfortunately, about a month after beginning the crawls using the U.S. Department of State's RSS feeds, we began to see these notices (see Figure 6).

Figure 6. State.gov crawler error message due to exclusion of our crawler

We were eventually able to communicate with a high level manager at the Department of State who allowed a restart of our RSS crawling, but such efforts can be difficult—simply identifying the right office and person to communicate with is challenging.

In general, our experience suggests that adding RSS-driven crawling to ongoing regular crawling of a site runs the risk of attracting negative attention from the site owner; that is, someone may notice the additional crawler traffic and decide to block it. Since the alternative is to have a less than useful harvest, we have typically taken the risk.

Frequency of Capture

It may be useful for those wishing to understand the RSS-driven harvesting technique to understand how we set the frequency of crawls for different sites where we are using this approach. We decided to crawl RSS feeds twice a day after some study of the different RSS feeds provided by different news sites. An RSS feed is a series of time-stamped entries going backwards in time, with the RSS creator setting some limit after which old entries are no longer part of the feed. It is fairly straightforward to look at an RSS and see when the oldest entries in the feed are and verify that a crawl every twelve hours will acquire all the items in the feed. For our approach to work, the twice-daily capture of the RSS must be more frequent than RSS items enter and leave the RSS page. Here are some examples:

1. The Library of Congress RSS feeds, for example http://www.loc.gov/rss/law/reports.xml and http://www.loc.gov/rss/pao/news.xml, retain content for an average of two months. So twice daily harvests are clearly adequate.

2. Some Congress.gov feeds update daily (https://www.congress.gov/rss/house-floor-today.xml) and weekly (https://www.congress.gov/rss/most-viewed-bills.xml). Again, RSS feed updates are less than twice daily.

3. Know Your Meme's newsfeed has consistently kept ten items on http://knowyourmeme.com/newsfeed.rss and RSS items are kept longer than 12 hours, but http://knowyourmeme.com/photos.rss seems to drop stories in less than 12 hour increments. So twice-daily for this second feed is not sufficient. Hopefully the quarterly crawl discovers what we may be missing from only twice-daily crawls. Also, for collection development reasons, we are less concerned about completeness for this part of the site.

4. The main feed for Buzz Feed (https://www.buzzfeed.com/index.xml) also updates and drops stories in less than 12 hours, but provides for 43 section feeds such as http://www.buzzfeed.com/geeky.xml and https://www.buzzfeed.com/omg.xml that update and keep stories for greater 12 hours. We therefore choose to also harvest the separate section feeds as well as the main feed page to ensure a more complete coverage.

We looked closely at the coverage of typical news site feed updates in order to set a minimum frequency for our harvesting of such RSS feeds—we decided on twice daily for RSS harvesting. After the initial analysis, we added Huffington Post (huffingtonpost.com) to the RSS crawl and eventually determined that some of its section feeds update more frequently than ever twelve hours. With the number of RSS feeds in the crawl, the crawler was not able to scrape each RSS feed in the four hours available and complete the harvest so the Library has opted to continue twice-daily RSS feed crawls for now even as it appears to miss some materials for this one site. As we add more feeds to the RSS crawl, we will have to monitor completeness for the twice-daily crawl.

Conclusion

Using RSS to acquire seeds at the page or article level for very large news sites clearly improves completeness of harvested results over a "traditional" Web archiving approach, even when the "regular" harvest is allowed to go on for an extended period of time. In some cases, we have archived content for which users will not be able to browse in a conventional "click on links" way, which is not a situation that we are used to with Web archiving. This suggests the need for different and better access tools for archived Web content rather than a deficiency in the RSS approach itself that more completely captures content.

Although perhaps it goes without saying, we have made a significant level of effort beyond what is normally required as we implement RSS crawling of selected

sites. That is, for any site that we choose to harvest using RSS feed(s), it takes more manual human effort than a standard site nomination. This effort includes analysis of the RSS feed(s) provided by the target site, creation of separate curator tool records for each RSS feed to be captured, a determination on what frequency of "regular" crawling to continue with in addition to the RSS crawl, and (much) more nuanced examination of the results.

In terms of crawling capacity, we believe this approach is more efficient since we are trying to completely harvest very large sites as a way of acquiring the content less often—crawls that as noted often don't provide good results. That is, there will be less crawling activity overall in order to have an extended crawl of a site four times a year and otherwise only crawl what is identified by the RSS on a shallow, one-hop basis. However we admit we are not sure that this efficiency would substantively benefit a Web archiving program overall when balanced against the greater level of human work to implement.

As we have developed familiarity and comfort with use of selected RSS feeds to create shallow crawl seed lists for news, we then circled back to use it selectively for certain large sites that are large but that have news or other sections that publish frequently, particularly U.S. government agency sites with significant news sections that might be missed by our new deep (extended) quarterly crawl approach. We have decided that this technique implemented narrowly for news sites has utility in assuring more complete harvesting of fast changing parts within other large sites, in particular large federal government agency sites.

At the twenty-year mark for Web archiving, which began with the first crawls performed by the Internet Archive in 1996, we are nevertheless still apparently in an early period for developing new techniques to perfect results which we look forward towards.