Electronic Records Modules

Electronic Records Committee

Congressional Papers Roundtable

Society of American Archivists

Setting up and Scoping Web Crawls with Archive-It 5.0

Gregory Wiedeman

University of Albany, SUNY

gwiedeman@albany.edu

Date Published: **July 2016**

Module#: ERCM005

Setting up and Scoping Web Crawls with Archive-It 5.0

June 2016

Overview and Rationale

The web is a huge, complicated place and crawling a website is not as simple as it first appears. As we'll see, running a basic crawl can yield a large amount of data outside of your intended scope. Thus, we use the feedback from test crawls to see where the crawler is going and then give it some instructions that will contain the crawler while still assuring that it reaches all the pages we wish to save.

Procedural Assumptions

This workflow is intended to guide archives, with limited staff time devoted to web archiving, who wish to make topical crawls with one or a handful of seeds using Archive-It 5.0.

This workflow is designed to get a basic topical web archiving program up and running with a very limited time commitment. Thus, this workflow focuses on being efficient with staff time over being efficient with Archive-It's data budget or developing a perfect web archive. A larger-scale program may choose to commit more resources to developing more through scoping procedures.

Hardware and Software Requirements

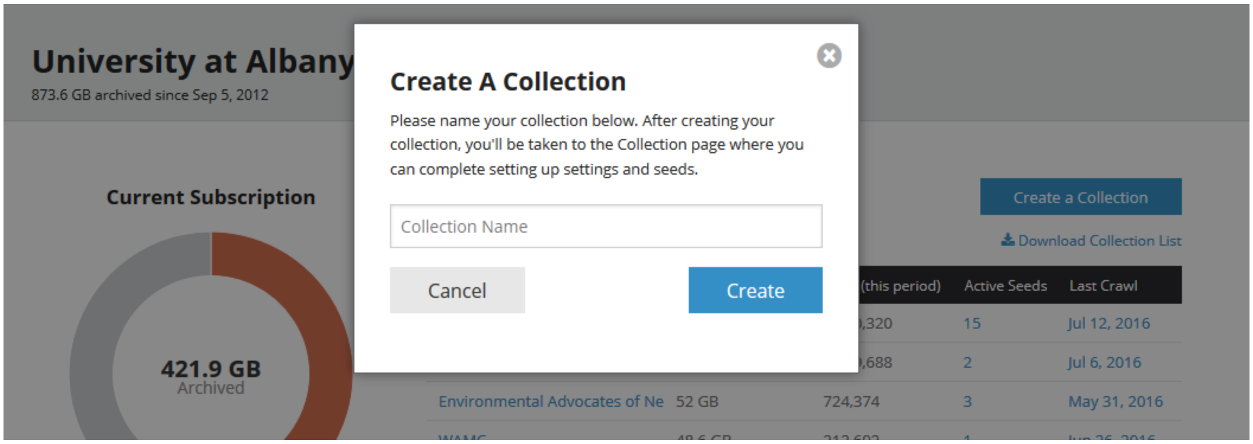
This workflow requires an Archive-It subscription and is focused on version 5.0. Previous versions use a different interface and lack some features like saving test crawls.

Workflow and Examples

Table of Contents:

Defining Collections

After Partners, Collections are the highest level of description within Archive-It. From the Home page after you log-in, you just select "Create a Collection" and enter a name. Each collection gets an Archive-It ID number that you can see later in Wayback URLs. Keep in mind that here, Collections are not synonymous with the idea of collections in archival description. It's generally always a good idea to *respect de fonds*, but web archives can come from a number of different domains. Perhaps it could be a good idea to divide collections according to organizations who administer the site you wish to collect, yet many web archives are also be topically-focused.



Create A Collection

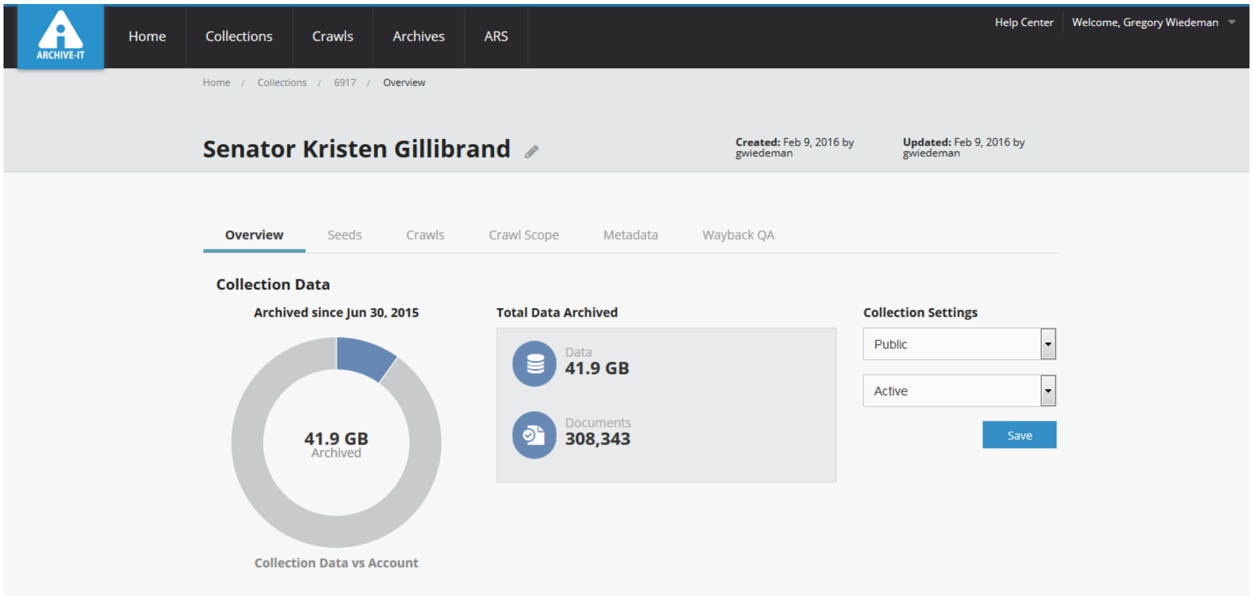
Please name your collection below. After creating your collection, you'll be taken to the Collection page where you can complete setting up settings and seeds.

		(this period)	Active Seeds	Last Crawl
		,320	15	Jul 12, 2016
		,688	2	Jul 6, 2016
Environmental Advocates of Ne	52 GB	724,374	3	May 31, 2016

Figure 1: Creating a Collection

Selecting Seeds

Seeds are URLs that represent the starting point for Archive-It web crawling: they point the crawler in the right direction. Typically this is the root domain of the website or websites you wish to preserve. For instance, if your seed is www.yahoo.com, the crawler will save that page and then move to every link that is listed on that page, like mail.yahoo.com, www.flickr.com, and then store every link that is listed on those pages, and then every link on those pages, etc., until it hits some type of boundary. This boundary could be a crawl time limit, or even your entire data budget. It is easy to see how web crawling can get out of control. Thus, we use sets of rules to limit what pages the crawler saves and how far it limits itself. The process of developing these rules is called "scoping."



Senator Kristen Gillibrand

Created: Feb 9, 2016 by gwiedeman

Updated: Feb 9, 2016 by gwiedeman

Data 41.9 GB

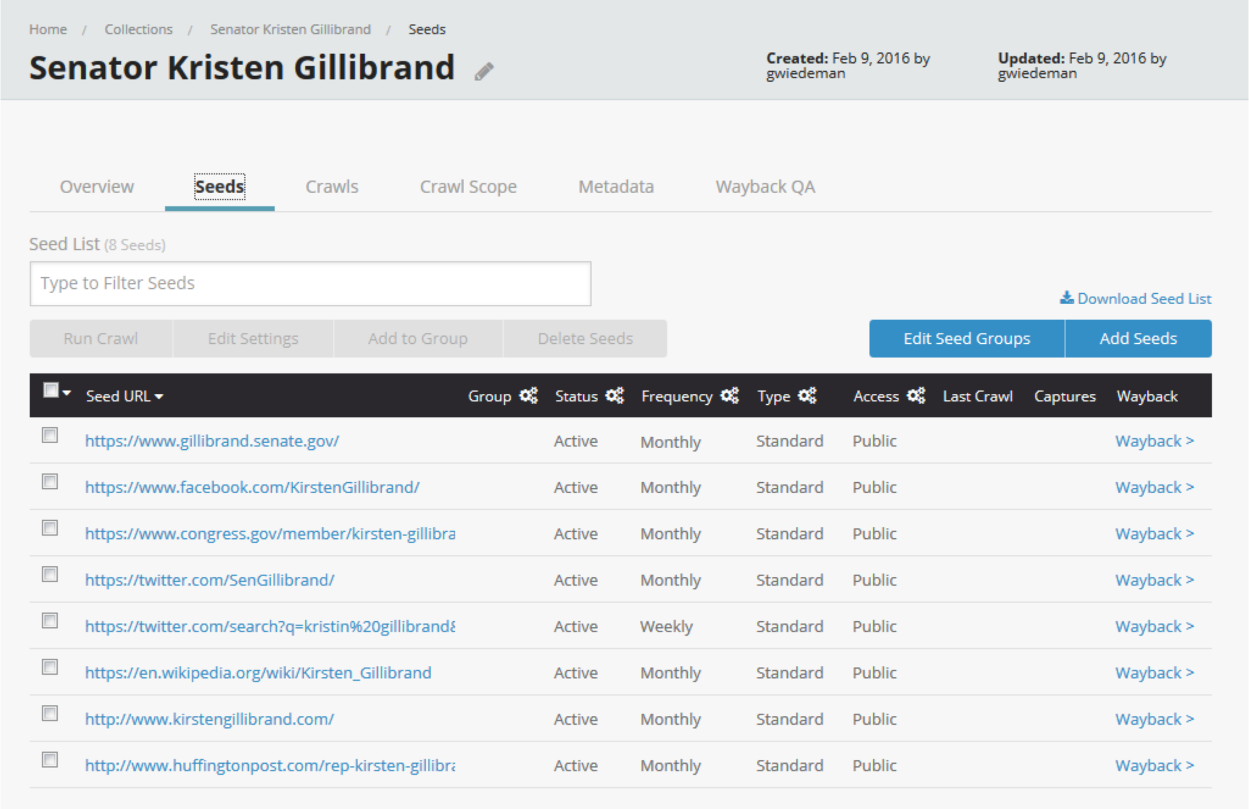
Documents 308,343

Public

Active

Figure 2: Example of the Overview Tab for a Collection. From here you can navigate to the Seeds Tab, Crawls Tab, or Crawl Scope Tab.

Probably the simplest way to find seeds is just to run a Google search of what you would like to crawl. Here some examples of seeds for topical collections:



Senator Kristen Gillibrand

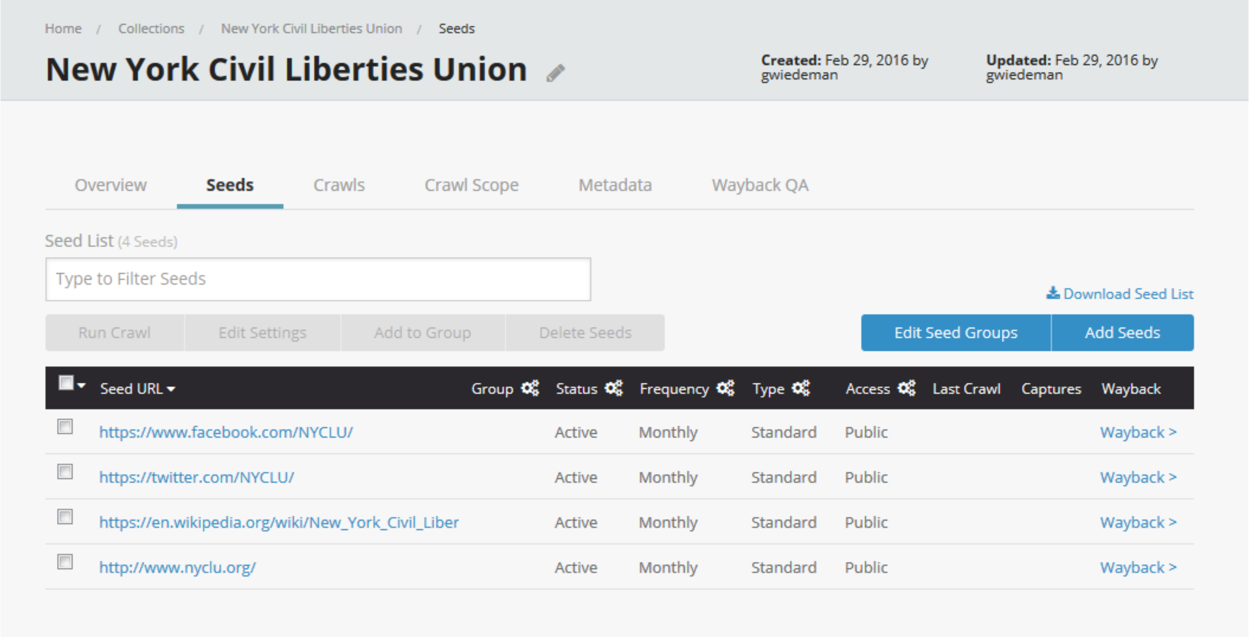
Created: Feb 9, 2016 by gwiedeman

Updated: Feb 9, 2016 by gwiedeman

Seed List (8 Seeds)

	Seed URL	Group	Status	Frequency	Type	Access	Last Crawl	Captures	Wayback
□	https://www.gillibrand.senate.gov/		Active	Monthly	Standard	Public			Wayback >
□	https://www.facebook.com/KirstenGillibrand/		Active	Monthly	Standard	Public			Wayback >
□	https://www.congress.gov/member/kirsten-gillibra		Active	Monthly	Standard	Public			Wayback >
□	https://twitter.com/SenGillibrand/		Active	Monthly	Standard	Public			Wayback >
□	https://twitter.com/search?q=kristin%20gillibrand&		Active	Weekly	Standard	Public			Wayback >
□	https://en.wikipedia.org/wiki/Kirsten_Gillibrand		Active	Monthly	Standard	Public			Wayback >
□	http://www.kirstengillibrand.com/		Active	Monthly	Standard	Public			Wayback >
□	http://www.huffingtonpost.com/rep-kirsten-gillibr		Active	Monthly	Standard	Public			Wayback >

Figure 3: Seeds for the Senator Gillibrand Collection



New York Civil Liberties Union

Created: Feb 29, 2016 by gwiedeman

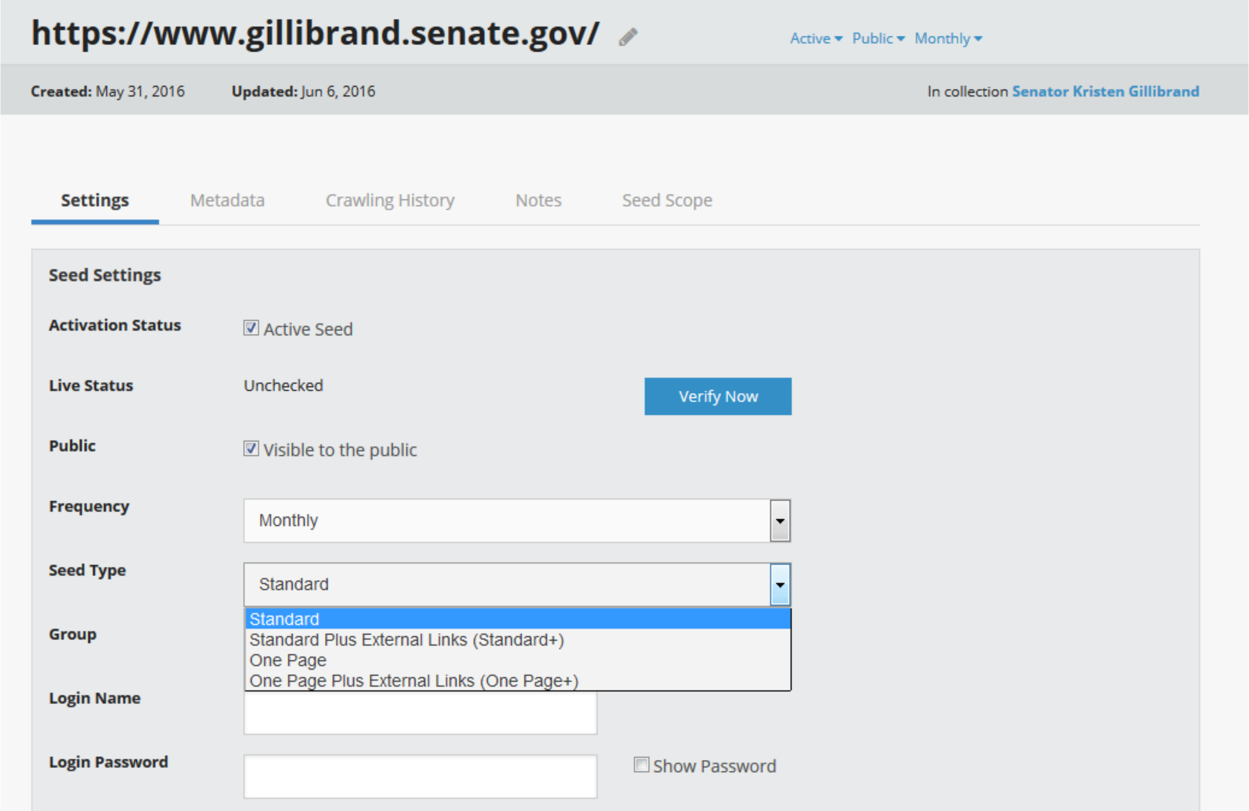
Updated: Feb 29, 2016 by gwiedeman

Seed List (4 Seeds)

	Seed URL	Group	Status	Frequency	Type	Access	Last Crawl	Captures	Wayback
□	https://www.facebook.com/NYCLU/		Active	Monthly	Standard	Public			Wayback >
□	https://twitter.com/NYCLU/		Active	Monthly	Standard	Public			Wayback >
□	https://en.wikipedia.org/wiki/New_York_Civil_Liber		Active	Monthly	Standard	Public			Wayback >
□	http://www.nyclu.org/		Active	Monthly	Standard	Public			Wayback >

Figure 4: Seeds for the NYCLU Collection

When you select seeds, you select a basic set a rules that will govern each crawl. You can select if a seed it private or available to the general public, the frequency you wish to crawl at, and the type of crawl. You can crawl seeds only one-time, or tell Archive-It to crawl automatically anywhere from daily or every 12 hours, up to annually. Standard crawls are the default crawl type and store all pages that are linked from the source seed until it hits a crawl limit. Standard+ crawls hit crawl limits and go one page beyond. One Page crawls only save the seed page, and One Page+ crawls store only the seed and pages that are directly linked to it.



https://www.gillibrand.senate.gov/

Active ▾ Public ▾ Monthly ▾

Created: May 31, 2016 **Updated:** Jun 6, 2016

In collection **Senator Kristen Gillibrand**

Settings
Metadata
Crawling History
Notes
Seed Scope

Seed Settings

Activation Status ☑ Active Seed

Live Status Unchecked

Public ☑ Visible to the public

Frequency Monthly

Seed Type Standard

Standard
Standard Plus External Links (Standard+)
One Page
One Page Plus External Links (One Page+)

Group

Login Name

Login Password □ Show Password

Figure 5: Editing Basic Seed Settings

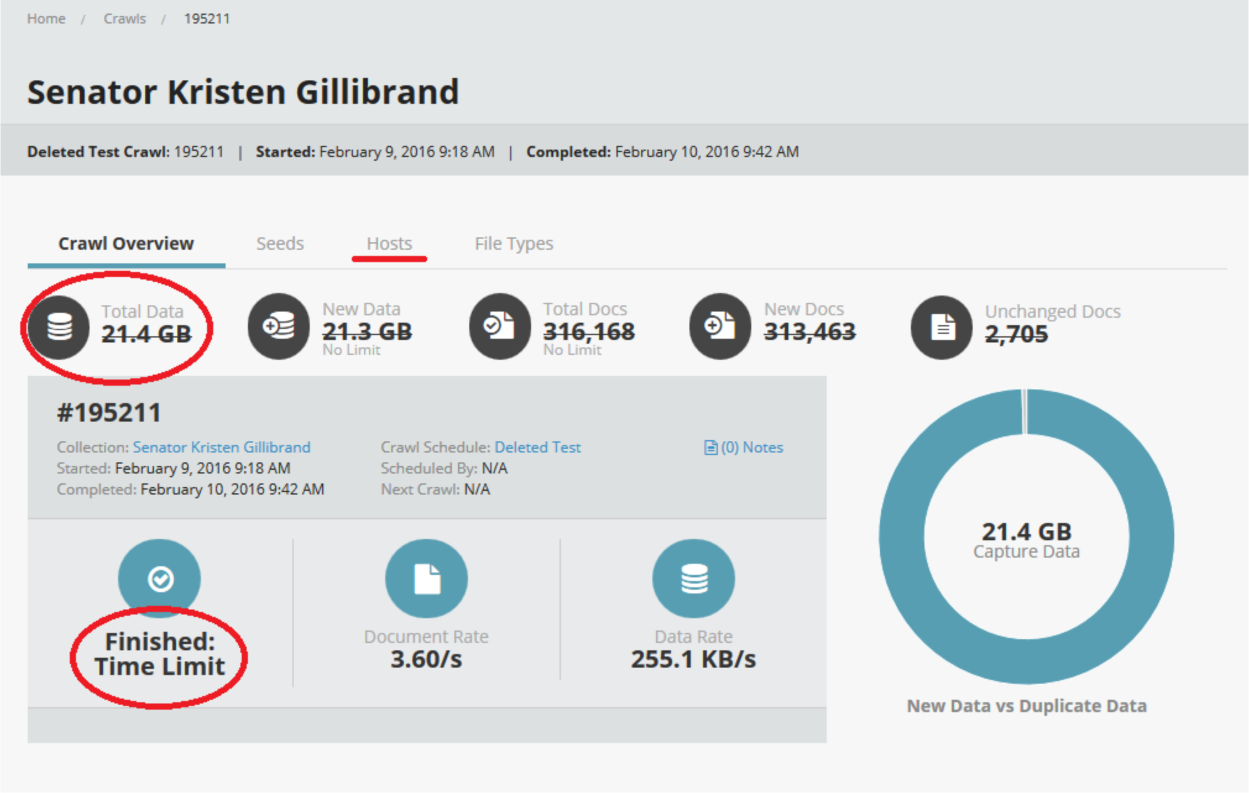
Crawling

Once you select seeds for a collection, you can select one or more seeds with the checkbox on the left hand side and select “Run Crawl” to actually start creating your web archive. It’s always a good idea to start with a Test Crawl. This lets you review the effect of your crawl and see what it gathers before preserving it. While you may want to apply some basic crawl rules first under the Crawl Scope tab, it is tough to scope a collection without seeing what you get.

If you are trying to topically crawl a large website like Twitter or Wikipedia, it’s a good idea to add some sort of document limit even before your test crawl. These sites are hundreds of thousands (or more?) of interlinked pages, and will skew your test crawl report.

Scoping

The first thing and most basic thing to look at to evaluate a crawl is how it finished. Blind three day crawls can time out even with just one or two seeds. Hitting that time limit is problematic because that means the crawl expired before reaching all of the pages you targeted. You could just extend the crawl limit, but that would include a ton of data that is likely outside of your aims. So I look at the Host tab to see what the crawl found.



Home / Crawls / 195211

Senator Kristen Gillibrand

Deleted Test Crawl: 195211 | **Started:** February 9, 2016 9:18 AM | **Completed:** February 10, 2016 9:42 AM

Crawl Overview | Seeds | Hosts | File Types

Total Data ~~21.4 GB~~

New Data ~~21.3 GB~~ No Limit

Total Docs ~~316,168~~ No Limit

New Docs ~~313,463~~

Unchanged Docs ~~2,705~~

#195211

Collection: Senator Kristen Gillibrand
Started: February 9, 2016 9:18 AM
Completed: February 10, 2016 9:42 AM

Crawl Schedule: Deleted Test
Scheduled By: N/A
Next Crawl: N/A

(0) Notes

Finished: Time Limit

Document Rate 3.60/s

Data Rate 255.1 KB/s

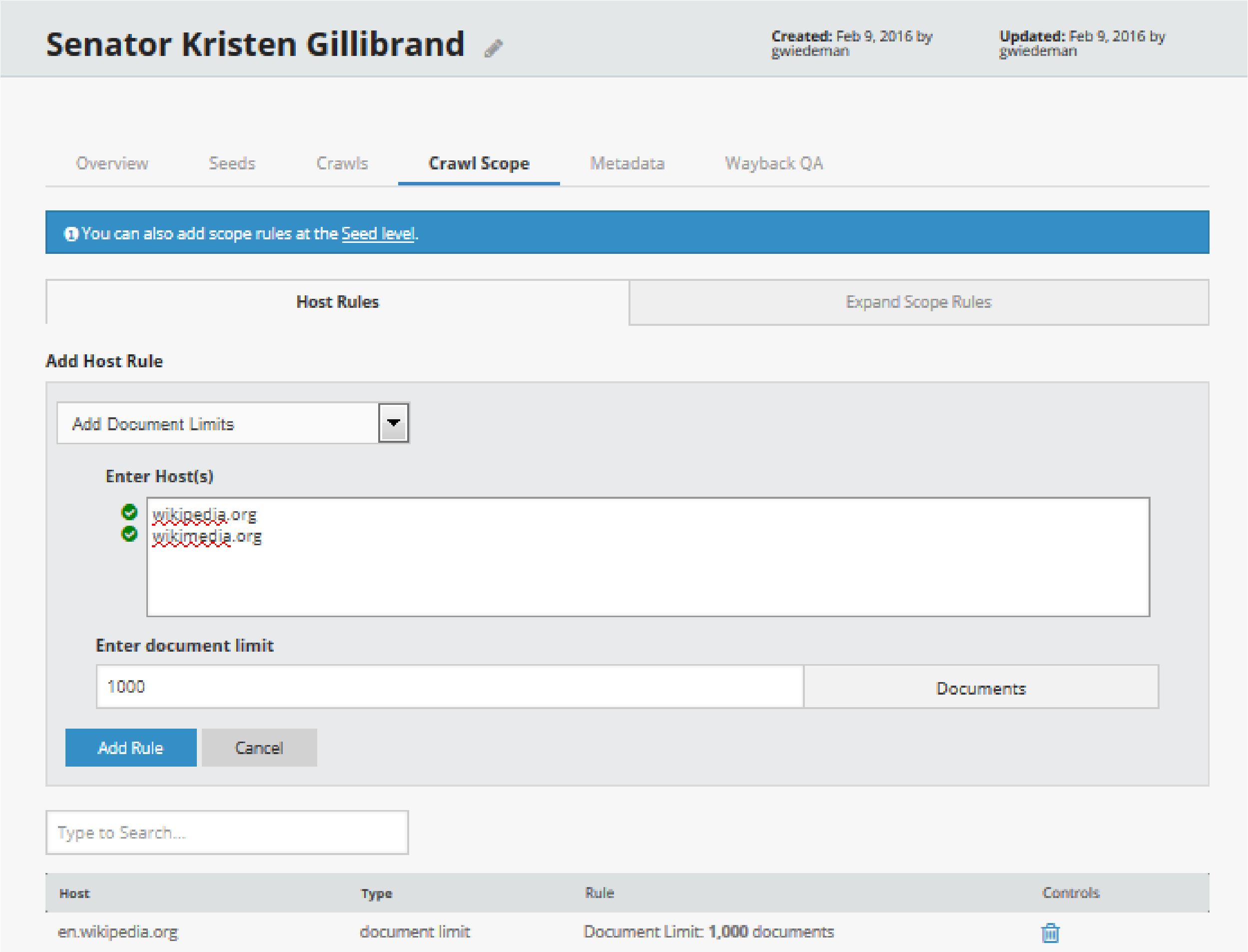
21.4 GB Capture Data

New Data vs Duplicate Data

Figure 6: An Overview of a completed test crawl that has been deleted

Here is an example of a blind test crawl which collected over 21 GB and discovered over 190,000 unique hosts! This is a good example of what happens when you use a Wikipedia or Twitter page as a seed without any scoping limits. Since Wikipedia is a huge corpus of interconnected links, it effectively sends our crawler down a never ending rabbit hole. Here, unless the aim is to download all of Wikipedia, we need to set some crawl limits. Since this was a test crawl the data is deleted automatically, unless you choose to save it.

In Archive-It 5.0 there are two primary levels where you can apply crawl rules: to certain seeds or whole crawls. This provides you with a lot of flexibility, but one thing to keep in mind is that you can only add document limits at the Crawl Scope-level, not the Seed Scope-level. For example, you can add a 1,000 document limit to anything from the Wikipedia.org and Wikimedia.org domains. You can also block a URL if it contains a certain phrase or expand to include it if it would otherwise be out of scope.



Senator Kristen Gillibrand

Created: Feb 9, 2016 by gwiedeman

Updated: Feb 9, 2016 by gwiedeman

You can also add scope rules at the Seed level.

Add Host Rule

Add Document Limits

Enter Host(s)

✓ wikipedia.org
✓ wikimedia.org

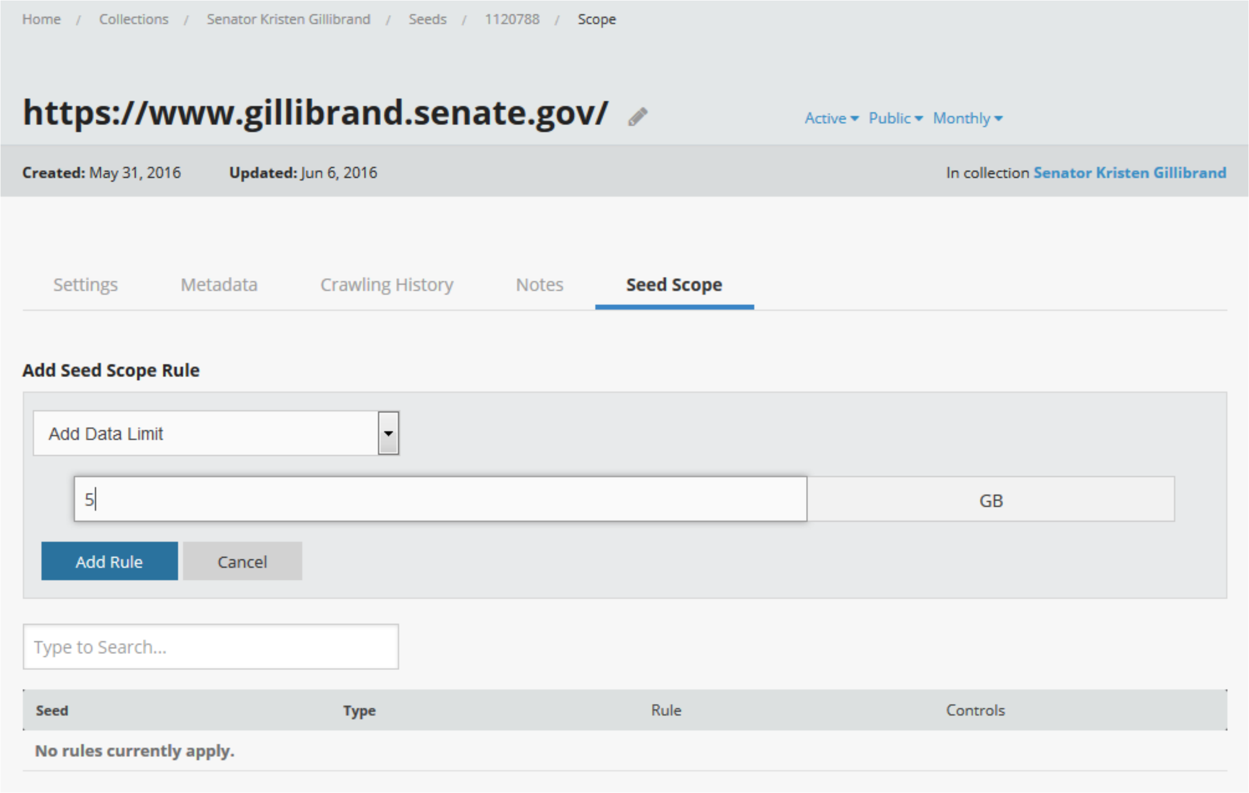
Enter document limit

1000 Documents

Host	Type	Rule	Controls
en.wikipedia.org	document limit	Document Limit: **1,000** documents	🗑

Figure 7: Adding Crawl limits

To add limits to individual seeds, you need to return to the Seed tab and select a seed. Then a new set of tabs appears that includes a Seed Scope tab. Here you can limit your crawl to stop adding content from a Twitter page seed after it reaches a certain amount of data.



Home / Collections / Senator Kristen Gillibrand / Seeds / 1120788 / Scope

https://www.gillibrand.senate.gov/

Created: May 31, 2016 Updated: Jun 6, 2016 In collection Senator Kristen Gillibrand

Add Seed Scope Rule

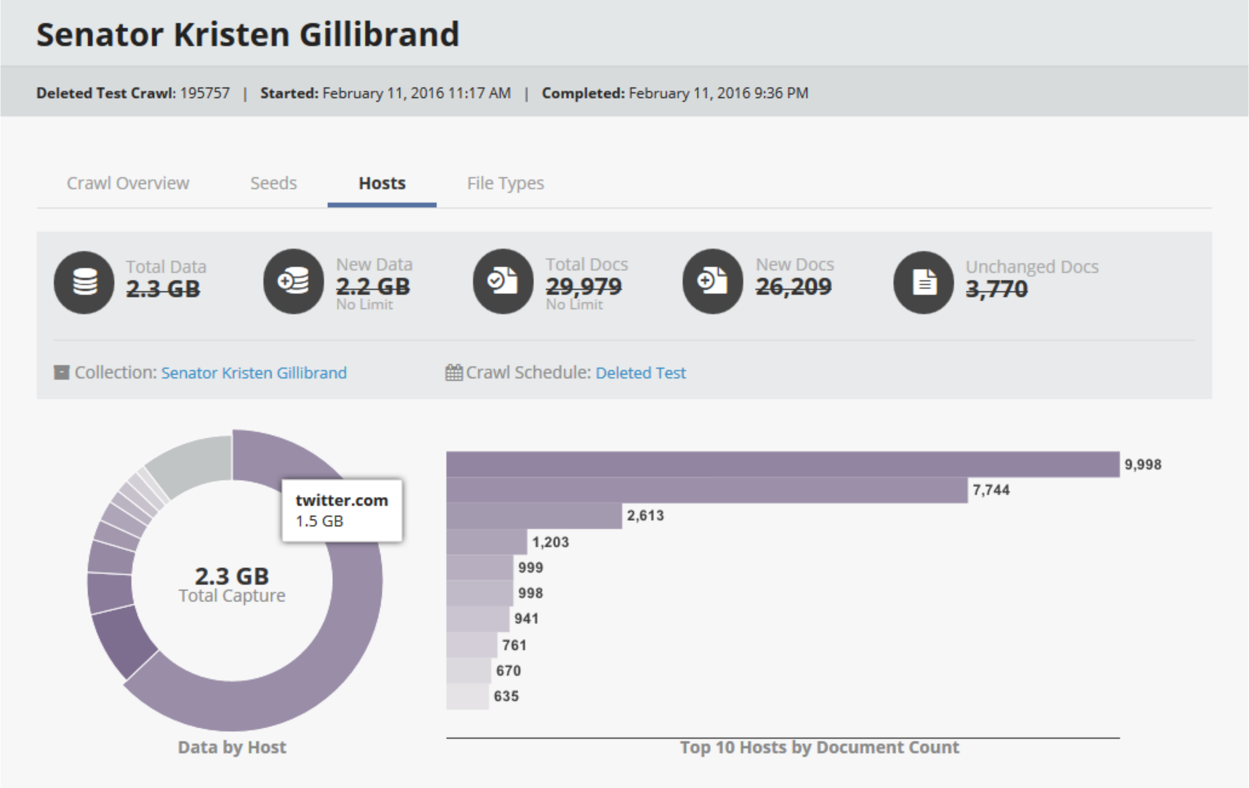
Add Data Limit

5 GB

Seed	Type	Rule	Controls
No rules currently apply.			

Figure 8: Adding limits to seeds

Scoping normally takes more than a few test crawls to get right. Here is the result of another test crawl with the above rules set in place.



Senator Kristen Gillibrand

Deleted Test Crawl: 195757 | **Started:** February 11, 2016 11:17 AM | **Completed:** February 11, 2016 9:36 PM

Crawl Overview | Seeds | **Hosts** | File Types

Total Data **~~2.3 GB~~**

New Data **~~2.2 GB~~** No Limit

Total Docs **~~29,979~~** No Limit

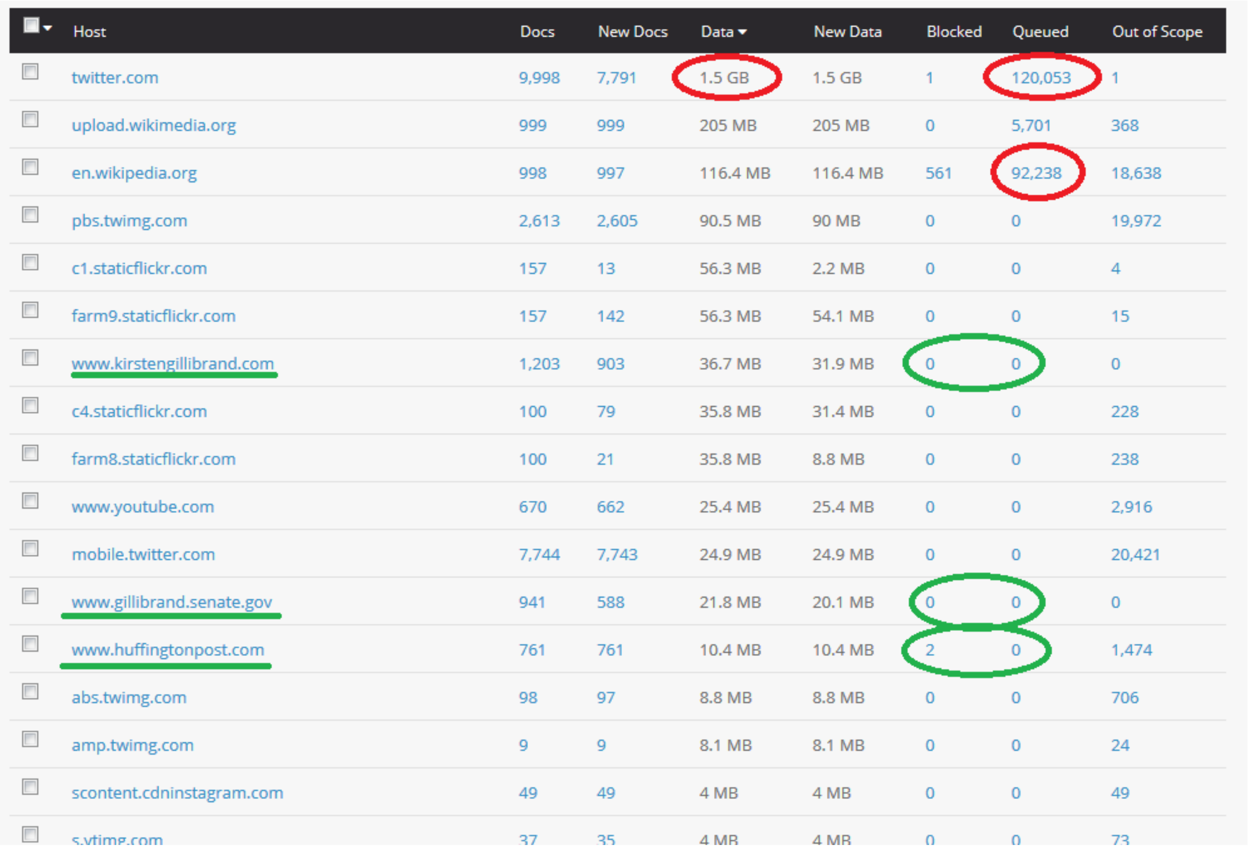
New Docs **~~26,209~~**

Unchanged Docs **~~3,770~~**

Collection: Senator Kristen Gillibrand

Crawl Schedule: Deleted Test

Figure 9: Host breakdown of the second test crawl



☐ ▾	Host	Docs	New Docs	Data ▾	New Data	Blocked	Queued	Out of Scope
☐	twitter.com	9,998	7,791	1.5 GB	1.5 GB	1	120,053	1
☐	upload.wikimedia.org	999	999	205 MB	205 MB	0	5,701	368
☐	en.wikipedia.org	998	997	116.4 MB	116.4 MB	561	92,238	18,638
☐	pbs.twimg.com	2,613	2,605	90.5 MB	90 MB	0	0	19,972
☐	c1.staticflickr.com	157	13	56.3 MB	2.2 MB	0	0	4
☐	farm9.staticflickr.com	157	142	56.3 MB	54.1 MB	0	0	15
☐	www.kirstengillibrand.com	1,203	903	36.7 MB	31.9 MB	0	0	0
☐	c4.staticflickr.com	100	79	35.8 MB	31.4 MB	0	0	228
☐	farm8.staticflickr.com	100	21	35.8 MB	8.8 MB	0	0	238
☐	www.youtube.com	670	662	25.4 MB	25.4 MB	0	0	2,916
☐	mobile.twitter.com	7,744	7,743	24.9 MB	24.9 MB	0	0	20,421
☐	www.gillibrand.senate.gov	941	588	21.8 MB	20.1 MB	0	0	0
☐	www.huffingtonpost.com	761	761	10.4 MB	10.4 MB	2	0	1,474
☐	abs.twimg.com	98	97	8.8 MB	8.8 MB	0	0	706
☐	amp.twimg.com	9	9	8.1 MB	8.1 MB	0	0	24
☐	scontent.cdninstagram.com	49	49	4 MB	4 MB	0	0	49
☐	s.vtimg.com	37	35	4 MB	4 MB	0	0	73

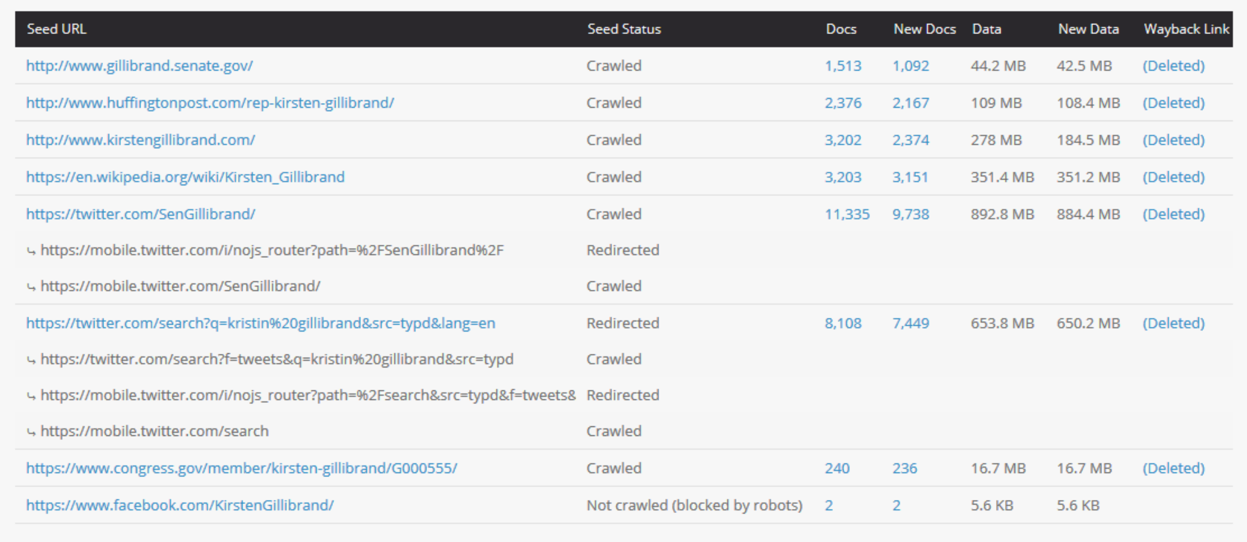
Figure 10: Host report for a second crawl

The Host Report might seem overwhelming, and it really contains the important feedback you need to adjust your crawl. Each host represents a different domain or subdomain, and the Docs column shows the number of pages the crawler collected. The numbers in blue are also links to a text list of the URLs that were collected, blocked, queued, or out of scope. If this list is too long, Archive-It will make you download a .zip file of the listing. You can just paste a sampling of these URLs right in your browser to see what pages you have collected and what kind of pages the crawler missed. You can spend a lot of time evaluating a crawl, but sometimes it can be effective to just select a small sample to see what types of pages the crawl is collecting

This crawl collected 2.3 GB of content, of which 1.5 GB alone is just Twitter data. Yet, there is also good news here too. One problem with the first example was that the crawler was so busy collecting all of Wikipedia and Twitter, it reached its three day limit before getting to all of www.kristengillbrand.com and www.gillibrand.senate.gov, and hundreds of pages remained queued. Here, all but one of the major domains in the targeted seeds were collected completely.

It's also important to look for any blocked pages. Sometimes these pages are just 404 pages that are not worth collecting anyway. Other times large swaths of important content may be blocked by robots.txt files. When that happens, you must either contact the administrators of the sites you are trying to crawl and ask them to permit the Archive-It crawler or contact Archive-It to request the ability to ignore this basic type of crawler block.

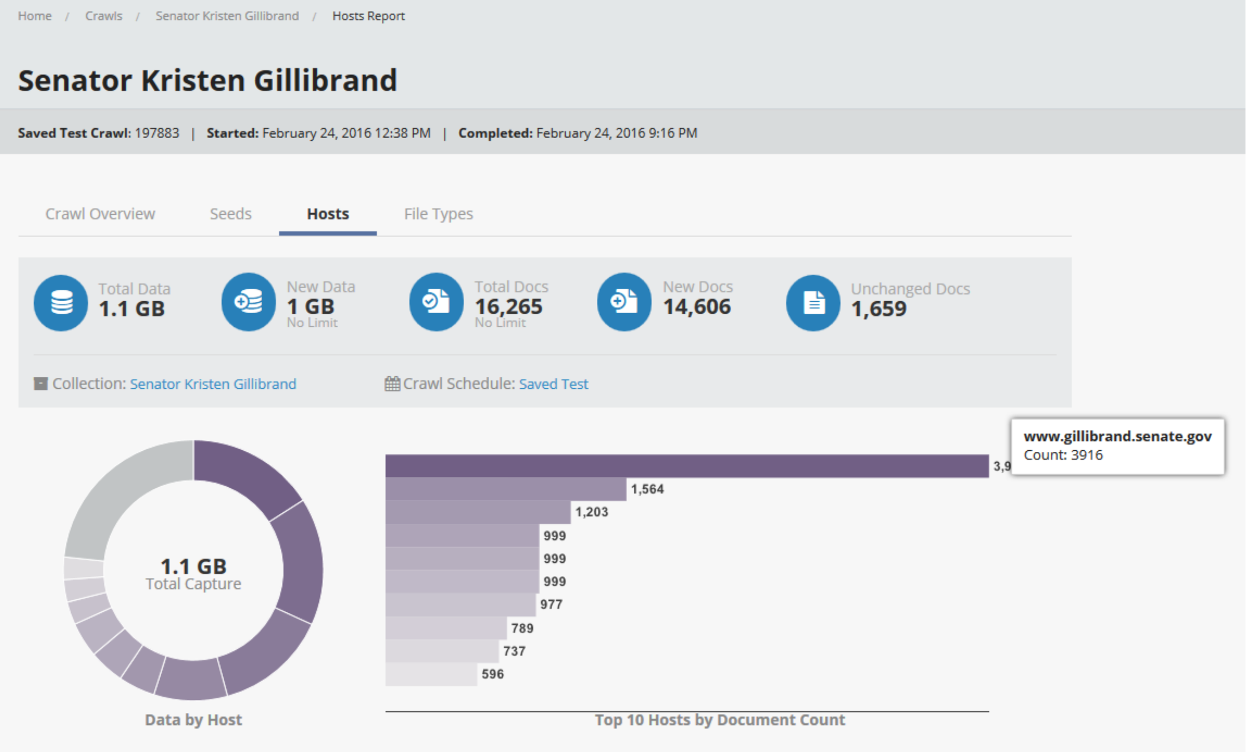
Another thing to look out for is crawler traps. If you have more Docs—either downloaded or queued—then you expected, you can look through the URL list to see if there are any pages like calendars of links where a page was downloaded for every day of the last century. Typically the Archive-it crawler is good at avoiding things like calendars if they employ more modern techniques.



Seed URL	Seed Status	Docs	New Docs	Data	New Data	Wayback Link
http://www.gillibrand.senate.gov/	Crawled	1,513	1,092	44.2 MB	42.5 MB	(Deleted)
http://www.huffingtonpost.com/rep-kirsten-gillibrand/	Crawled	2,376	2,167	109 MB	108.4 MB	(Deleted)
http://www.kirstengillibrand.com/	Crawled	3,202	2,374	278 MB	184.5 MB	(Deleted)
https://en.wikipedia.org/wiki/Kirsten_Gillibrand	Crawled	3,203	3,151	351.4 MB	351.2 MB	(Deleted)
https://twitter.com/SenGillibrand/	Crawled	11,335	9,738	892.8 MB	884.4 MB	(Deleted)
↳ https://mobile.twitter.com/i/nojs_router?path=%2FSenGillibrand%2F	Redirected					
↳ https://mobile.twitter.com/SenGillibrand/	Crawled					
https://twitter.com/search?q=kristin%20gillibrand&src=typd&lang=en	Redirected	8,108	7,449	653.8 MB	650.2 MB	(Deleted)
↳ https://twitter.com/search?f=tweets&q=kristin%20gillibrand&src=typd	Crawled					
↳ https://mobile.twitter.com/i/nojs_router?path=%2Fsearch&src=typd&f=tweets&	Redirected					
↳ https://mobile.twitter.com/search	Crawled					
https://www.congress.gov/member/kirsten-gillibrand/G000555/	Crawled	240	236	16.7 MB	16.7 MB	(Deleted)
https://www.facebook.com/KirstenGillibrand/	Not crawled (blocked by robots)	2	2	5.6 KB	5.6 KB	

Figure 11: Crawl results by seed

You can also move to the Seed tab within the Crawl page (which is different from the Seed tab within a collection page) that breaks down the results of the crawl by seed.



Home / Crawls / Senator Kristen Gillibrand / Hosts Report

Senator Kristen Gillibrand

Saved Test Crawl: 197883 | **Started:** February 24, 2016 12:38 PM | **Completed:** February 24, 2016 9:16 PM

Total Data 1.1 GB

New Data 1 GB No Limit

Total Docs 16,265 No Limit

New Docs 14,606

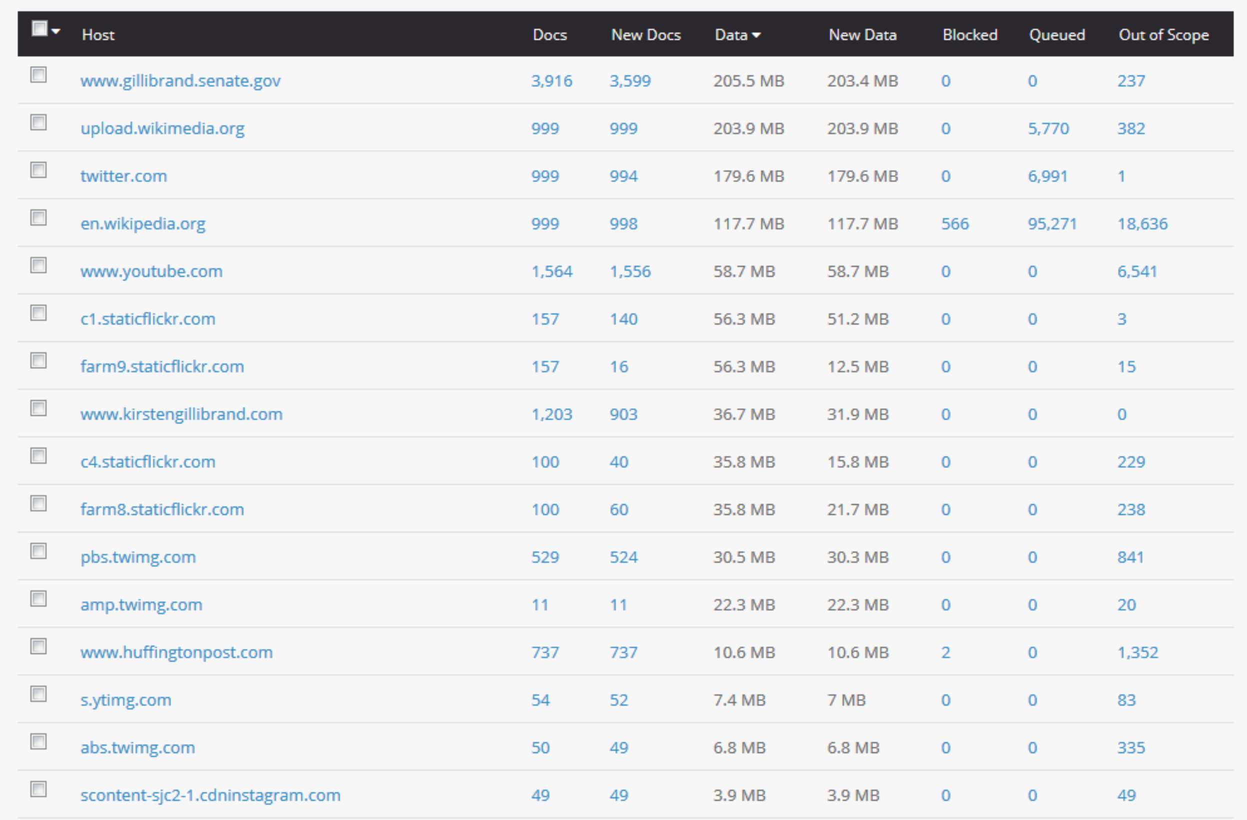
Unchanged Docs 1,659

Collection: Senator Kristen Gillibrand

Crawl Schedule: Saved Test

Figure 12: Results of a third test crawl

Once you have eliminated larger pages, your test crawls are likely to yield much more promising results. For this example the crawl collected about 16,000 pages for just over 1 GB of data. If we look at the Host report, the domains we entered as seeds are featured prominently and were captured completely. This is a probably more tolerable amount of noise and a sustainable amount of data to collect monthly. Additionally, future crawls are likely to produce a significant amount of documents that were not changed or altered since the last crawl, so there is a potential to reduce the data footprint in the future with deduplication.

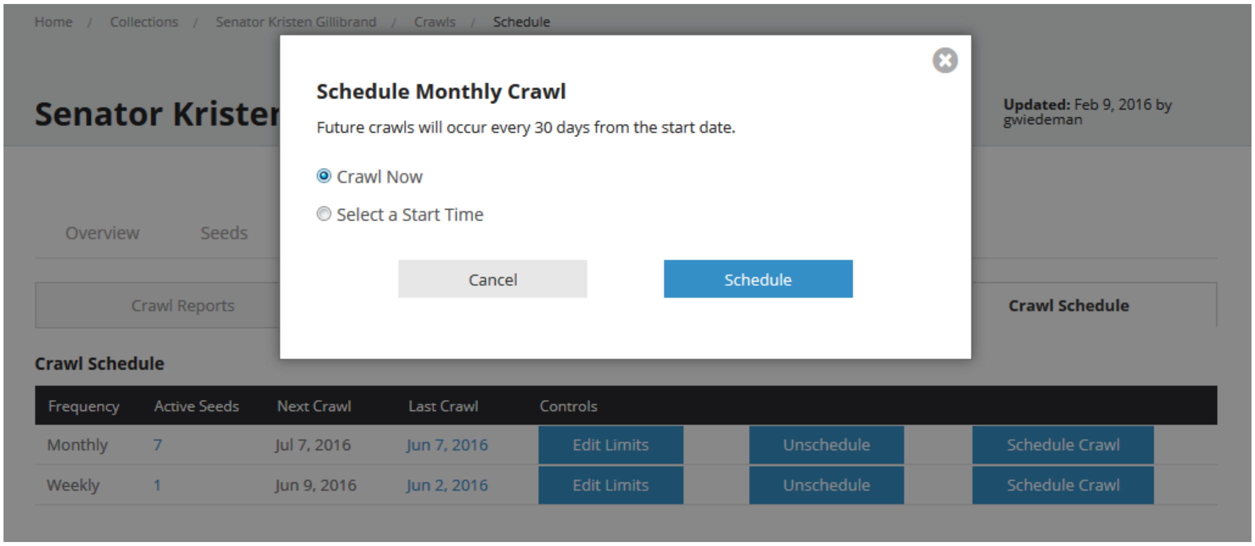


☐ ▾	Host	Docs	New Docs	Data ▾	New Data	Blocked	Queued	Out of Scope
☐	www.gillibrand.senate.gov	3,916	3,599	205.5 MB	203.4 MB	0	0	237
☐	upload.wikimedia.org	999	999	203.9 MB	203.9 MB	0	5,770	382
☐	twitter.com	999	994	179.6 MB	179.6 MB	0	6,991	1
☐	en.wikipedia.org	999	998	117.7 MB	117.7 MB	566	95,271	18,636
☐	www.youtube.com	1,564	1,556	58.7 MB	58.7 MB	0	0	6,541
☐	c1.staticflickr.com	157	140	56.3 MB	51.2 MB	0	0	3
☐	farm9.staticflickr.com	157	16	56.3 MB	12.5 MB	0	0	15
☐	www.kirstengillibrand.com	1,203	903	36.7 MB	31.9 MB	0	0	0
☐	c4.staticflickr.com	100	40	35.8 MB	15.8 MB	0	0	229
☐	farm8.staticflickr.com	100	60	35.8 MB	21.7 MB	0	0	238
☐	pbs.twimg.com	529	524	30.5 MB	30.3 MB	0	0	841
☐	amp.twimg.com	11	11	22.3 MB	22.3 MB	0	0	20
☐	www.huffingtonpost.com	737	737	10.6 MB	10.6 MB	2	0	1,352
☐	s.ytimg.com	54	52	7.4 MB	7 MB	0	0	83
☐	abs.twimg.com	50	49	6.8 MB	6.8 MB	0	0	335
☐	scontent-sjc2-1.cdninstagram.com	49	49	3.9 MB	3.9 MB	0	0	49

Figure 13: Host report for the third test crawl

Archive-It certainly allows for more detailed scoping. The File Types tab produces a breakdown of the files you crawled, so you can see the proportion of HTML text files to images, audio, or video files you collected. From the host list you can spend hours examining precisely what you are collecting by pasting URLs into a browser. Archive-It certainly has no limits to how granular you can get with crawl limits.

Keep in mind that all of the websites you crawl will eventually change, along with the structure and the technology of the web itself. Scoping is certainly not a final process and routine monitoring of collections is often necessary.


Home / Collections / Senator Kristen Gillibrand / Crawls / Schedule
Senator Kristen
Updated: Feb 9, 2016 by gwiedeman
Overview
Seeds
Crawl Reports
Crawl Schedule
Cancel
Schedule

Schedule Monthly Crawl

Future crawls will occur every 30 days from the start date.

● Crawl Now

○ Select a Start Time

Crawl Schedule

Frequency	Active Seeds	Next Crawl	Last Crawl	Controls		
Monthly	7	Jul 7, 2016	Jun 7, 2016	Edit Limits	Unschedule	Schedule Crawl
Weekly	1	Jun 9, 2016	Jun 2, 2016	Edit Limits	Unschedule	Schedule Crawl

Figure 14: Finally, setting the crawl schedule

Once you are satisfied with your test crawl you can now select the option to save that data. To set up a crawl schedule, you go back up to the Collection Page and—under Crawl tab—you should made sure to start the schedule for how often you want to crawl certain seeds. Archive-It will take it from here and you will begin to create a web archive.