

Sunday, Nov 23, 2025

Nov 23, 2025, 2:03 PM

Hey guys, I just saw Professor Scott's email. I was going through the sites more closely to see what was getting missed and I think I want to add some rules/seeds when doing the Brozzler crawls to make sure we get everything. But if we can just save the test crawls, I'm thinking I can just run one more Brozzler test crawl with the New Bev, the Frida, the Egyptian, and the Bay Theatre and see if that gets everything. Does that seem like a good idea? I'm hoping that should work better than running multiple individual Brozzler crawls, because the server requests should be spread out more evenly among the domains

👍 1

Nov 23, 2025, 2:06 PM

I trust your judgment 100%, Michelle.

👍 1

Nov 23, 2025, 2:14 PM

that makes sense! and i was happy to see that we can use the information we have from non production crawls! so we should be able to meet tuesday then

👍 1

I just started the new test crawl. I did set it to 3 days instead of 1 day to try to make sure it captures things correctly. But hopefully it will finish before then so we can meet Tuesday. I just wanted to give us a buffer time-wise since I added some seeds to try to capture things that the standard crawler wasn't picking up

Nov 23, 2025, 10:39 PM

Hmm, I just checked on the crawl, and it's already at 3 GB. A lot of it seems to be MP3's from the Pure Cinema Podcast which is the official podcast of the New Bev. The MP3's are embedded in a bunch of their blog posts. But I'm not sure if we actually need to archive these, since they are also available on a bunch of other platforms, so I added a rule blocking the podcast host site (Libsyn) just so we don't run out of space. Are you guys okay with this?

👍 1

Monday, Nov 24, 2025

Nov 24, 2025, 7:09 AM

YES

thank you

Nov 24, 2025, 9:01 AM

Yep.

Nov 25, 2025, 11:46 AM

Well the good news is, the document numbers haven't changed in many hours, so it's probably done and we could just shut it down now. The bad news is, the Brozzler-crawled Egyptian site still doesn't look too good. It's kind of complicated, but the stuff I tried to make the sites that are only loaded by the hover menus didn't work. I also don't think we'll be able to get the showtimes calendar to work. We can try to add every URL as a seed, but it won't be possible to access them from the main page because those stupid hover menus are the only way to get to them. So I don't know what you guys wanna do. Here's the link so you can see what I'm talking about: <https://wayback.archive-it.org/30617-test/20251124040603/https://www.egyptiantheatre.com/>

capture the sites* I mean

Nov 25, 2025, 11:54 AM

Ok...let me finish making lunch for my daughter and I'll look thru it!

Thank you for identifying those specific issues!

Nov 25, 2025, 12:01 PM

Interesting. When I checked this morning we were at 18 seeds. I am running errands and will be back at 2

2 replies Nov 25, 2025, 11:15 AM

Nov 25, 2025, 12:01 PM

I just asked [Claude.ai](#) if it had any suggestions and it wrote a whole script to use with brozzler to help capture hover menus. Do you want to try that?

Nov 25, 2025, 12:03 PM

<https://claude.ai/share/2a67997b-327d-44cf-bb01-b4afeca08725> Here's some suggestions.

Nov 25, 2025, 12:04 PM

Not 100% sure but I don't think we can run JavaScript behaviors in Brozzler through the Archive-It interface

Nov 25, 2025, 12:04 PM

ohhhh

Nov 25, 2025, 12:14 PM

I guess we write about the limitations of Archive-it and how websites that are built artistically with hover features or other CSS/Javascript designs often have limitations in archiving abilities.

👍 1

Nov 25, 2025, 12:27 PM

Yeah, I think it's technically possible but not with the tools we've been provided with

Another thing I realized is that the embedded YouTube videos on this page didn't seem to get picked up: <https://thefridacinema.org/discover-dtsa/>



Archive-It has some guidelines on how you can tell it to ignore robots.txt on [googlevideo.com](https://www.archive-it.org/en/faq/google-video) which is what serves most videos but I didn't do that on the original crawl

Nov 25, 2025, 12:33 PM

is there a way to like patch that particular crawl? Like go back and say to archive those videos?

Nov 25, 2025, 12:35 PM

It's unclear to me whether adding a collection scope rule will actually affect ongoing crawls, or if you have to start a new one after adding the rule

Nov 25, 2025, 12:39 PM

it is also unclear to me. Based on things I am finding, it seems implied that rules can't be changed mid-crawl. But it does not ever explicitly say that anywhere.

Nov 25, 2025, 12:40 PM

Yeah. I did add some rules mid-crawl but it didn't seem to make the crawler start archiving things it hadn't been

One thing we could try is to finish this crawl and then start a new one just for the Frida because that seems to be the one with the most embedded YouTube videos from their own channel

Nov 25, 2025, 12:41 PM

yeah, I think that's what we'll have to do.

They call it "patching" in the help