Save Page Now 2 Change Log

vangelis@archive.org

Please see also SPN2 public API docs:

https://docs.google.com/document/d/1Nsv52MvSjbLb2PCpHlat0gkzw0EvtSgpKHu4mk0MnrA/edit

2021-11-01

- Outlinks are filtered using https://github.com/Intsights/braveblock.

2021-10-22

- https://archive.org/services/wayback-gsheets has been updated. Improvements:
    - Performance is faster when capturing outlinks. Previous code would check the status of multiple outlink captures one by one using multiple status requests. The new code checks multiple captures with a single HTTP request.
    - The email report includes the capture options used. (@mark thanks for the suggestion)
    - The formatting of big numbers is improved (embeds and outlinks numbers are written like 1,000,000 instead of 1000000).
    - Embeds counter bug when using maxhops=1 / 2 reported by @mark is fixed.
    - Internal code cleanup and refactoring.

2021-08-06

- Extract outlinks from epub files using https://pypi.org/project/EbookLib/.

2021-07-03

- Anonymous clients who are listed in https://www.spamhaus.org/sbl/ or https://www.spamhaus.org/xbl/ lists are blocked. Tor exit nodes are excluded.

2021-06-28

- MRSS support for outlink extraction.

2021-05-23

- Anonymous users have lower concurrent captures limit (limit=3) compared to authenticated users (limit=5).
- The limit of daily captures for anonymous users is reduced from 8k to 5k.
- The size of screenshots is limited to 4MB. Bigger screenshots are not allowed due to system overload.
- If a target site returns HTTP status=529 (bandwidth exceeded), we pause crawling that for an hour. If a target site returns HTTP status=429 (too many requests), we pause crawling that for a minute. All requests for the same host in that period get a relevant error message. Previously, we started these captures later, adding a delay of 20-30sec.
- There are 10+ new error codes to give users a better idea on why captures fail. Ref: https://docs.google.com/document/d/1Nsv52MvSjbLb2PCpHlat0gkzw0EvtSgpKHu4mk0MnrA/edit#heading=h.2bu3y5mtzecu

2021-05-04

- New SPN2 API option: `delay_wb_availability=1`. With this option, the capture becomes available in the Wayback Machine after ~12 hours instead of immediately. This option helps reduce the load on the Wayback Machine.

2021-04-10

- Improve PDF outlink extraction performance and resilience using PyMuPDF.
- Apply limits to spn@archive.org service to handle the increased load. The system can process up to 10 emails per day per user. Any more emails will be simply discarded.

2021-03-05

- Use the latest Google Chrome stable (v89) to make SPN captures.

2021-02-23

- https://archive.org/services/wayback-gsheets now displays the user's capture statistics in real time: (Currently running, limit, daily captures, daily captures limit).

2021-01-10

- Set daily capture limits for all users to avoid system abuse. The default limit for authenticated users is 100k captures per day and for anonymous users its 8k captures per day.

2021-01-07

- As a general rule, we archive URLs every 30 min to reduce redundant captures. To improve this, we reduce the archiving frequency for secondary web page resources. Selected mime types like favicons, fonts, text/css & text/javascript are currently archived every 4 hours.

2020-12-28

- Fix problem with capturing http / https redirects of the same URL. They were not captured correctly due to invalid internal cache keys.

2020-11-31

- Add new error code: `error:too-many-requests` which means that the target URL has returned HTTP status=429. subsequent requests to the same host are delayed to avoid having more of this error.
- Keep track of the number of current captures to the same host. If they exceed a limit (currently 50), delay following captures to avoid overloading the target host.

2020-10-23

- Add many new error codes to provide more detailed information on capture errors. All error codes are documented in the SPN2 doc.

2020-10-13

- Add a new endpoint http://web.archive.org/save/status/user that displays the current active and available sessions for your user account. Clients which run multiple parallel sessions could use this to control their capture pipeline in a better way.
- Increase limits to 100 outlinks per capture.

2020-10-04

- Increase limits to 80 outlinks per capture and 5 concurrent captures per user.
- Better exception handling which results in less failed captures.
- Bug fix: SPN2 capture would timeout and fail because we try to fetch embedded HTML5 <video> or <audio> URLs if available and many times, these are infinite streams. Now, when that happens, we handle the exception and continue the capture.

2020-09-02

- Improved https://archive.org/services/wayback-gsheets/:
    - A capture task now has a Tracking URL. You can bookmark/share this. Now you aren't supposed to keep the app open all the time to see capture progress.

- You receive an email on capture start that includes the Tracking URL and the gsheet URL and another email on capture finish that includes detailed statistics.
- The UI has a detailed activity log where you can see the whole progress of your task from start to finish.
- You cannot run more than 1 capture process at a time. The system checks for queued and active tasks from the same user before starting a new one. This way we prevent system abuse.
- Various small bug fixes.

2020-08-17

- The list of outlinks returned by SPN2 when “capture outlinks” is not selected is limited to 1k URLs.

2020-08-13

- It is possible to capture the same URL only 10 times per day.

2020-08-08

- Identify and route HTML page captures and misc file captures to different queues. Misc file captures are much faster than HTML page captures because they don’t require a full browser to run. The result is much faster archiving for misc files because their dedicated queue is processed a lot faster.
- SPN2 already blocked some web page embeds using a list of web trackers coming from Mozilla https://github.com/mozilla-services/shavar-prod-lists. Now, we start using this list to block main capture URLs.
- There is a limit of 100k captures per day for anonymous users (tracking is done via their IP address).

2020-07-21

- Add the new option “skip_first_archive” to skip checking if a capture is a first if you don’t need this information. This will make captures run faster.

2020-07-06

- Capture videos & thumbnails from Tweets in all available formats. Captured URLs appear at the end of the list of capture resources.

2020-06-30

- Track the number of concurrent captures to the same host. If they exceed a threshold, we delay any new captures for that host for 15s to avoid having issues with failed captures (crawl politeness).

2020-06-24

- Extend the list of outlinks we filter out. They can be grouped in the following categories:
  - Login forms, account pages user password reset pages, etc.
  - Create new FB page / pinterest page / youtube channel, etc.
  - Submit content to bookmark services like stumbleupon, digg, etc.

2020-06-21

- Use “User-agent: Googlebot” to capture twitter.com using the “old” UI because there are issues with playing back the “new” UI in the Wayback Machine.

2020-06-15

- Limit the download rate for each headless browser instance to 8MBps. The aim of this change is to increase system stability and avoid situations where a single bad actor can take down a whole server.

2020-06-12

- Add email reporting to https://archive.org/services/wayback-gsheets. When processing of a sheet finishes, the user receives an email report containing useful statistics including processing time, captures seeds, embeds, outlinks, screenshots and errors.

2020-06-10

- Add option “Save screenshot” to https://archive.org/services/wayback-gsheets

2020-06-08

- Major SPN2 optimization when capturing outlinks. When trying to capture a URL using “capture outlinks” and some of the outlinks have been already captured by another capture, we reuse them and we do not start new capture processes. The result is up to 25% faster outlink processing performance.

2020-05-22

- Fix bugs when capturing many outlinks from the same host.The host could be overwhelmed or it could use some kind of defensive mechanism to block too many requests from the same network. To avoid this, we add a minor delay for outlink captures from the same host. The 1st outlink will start after 1 sec, the 2nd outlink will start after 2 sec, etc.

2020-05-03

- Add the time it took for spn@archive.org to process an email. The response now includes text like: *“It took us ~50 minutes to process it”.*
- Improve link extraction from emails. Ignore “unsubscribe” links and resolve bugs with email formatting.

2020-04-28

- Add a thumbnail of the captured webpage (if available) to the email response when using the option “Please email me the result”.
- Block spam emails from spn@archive.org using a blacklist.
- Extend SPN2 auto login capabilities to more sites.

2020-04-14

- Add a new option “Please email the result” that sends an email report for the captured URL and captured outlinks to the user. The option is available to logged in users only.

2020-04-12

- Update capture statistics to report more accurate results. (Grafana problem)

2020-04-03

- Use the new Brozzler option `simpler404` to speed up the capturing of 4xx/5xx pages. We no longer run JS behaviors in these cases.
- Extend URL canonicalization to include common FB URL param `fbclid`.
- Fix PDF outlink extraction crash that was triggered by broken PDF.

2020-03-30

- Implement admin UI to view capture queues and active users in wayback-admin.

2020-03-10

- New SPN2 API option “force_get=1”. Force the use of a simple HTTP GET request to capture the target URL. By default SPN2 does a HTTP HEAD on the target URL to decide whether to use a headless browser or a simple HTTP GET request. force_get overrides this behaviour.
- If emails to spn@archive.org include “capture outlinks” in their subject, SPN2 captures link outlinks and includes them in the email response.

2020-02-20

- There are limits to the number of outlinks & concurrent capture sessions users can run. We can now define “superusers” with custom limits.

2020-02-16

- Capture FTP URLs. Also capture FTP outlinks (e.g. if you capture an FTP dir using “capture outlinks”, all the files listed there will be captured as well) https://webarchive.jira.com/browse/WWM-1188

2020-02-09

- Upgrade from Python 3.5.2 to 3.7.6. Everything became a bit faster and more efficient.

2020-02-05

- Archive crawl logs in petabox https://webarchive.jira.com/browse/WWM-1189
- Implement FoundationDB plugin for SPN2 warcprox. We can now write Live CDX records in FoundationDB.
- When email to spn@archive.org includes "capture outlinks" in the subject, capture outlinks for all links and add them in the email response of the service.

2020-02-03

- Allow capture outlinks & screenshot for registered users only.

2020-01-25

- Add specific recent check limit for outlinks only. https://webarchive.jira.com/browse/WWM-1191 We can now use different recent check limits for main capture and their outlinks.
- Improve /status endpoint to return HTTP status=429 when reaching user session limit https://webarchive.jira.com/browse/WWM-1190

2020-01-14

- Return an error message when trying to archive files over 2GB. https://git.archive.org/wb/pyspn/commit/829727282d01b9b37d0a84a5fc23ab06f193a129
- Reduce exception errors like “Capture not found”:
- Avoid URL canonicalization errors: https://git.archive.org/wb/pyspn/commit/626274fe9e125c447837e6e85dc72ad1d785faf9

2020-01-10

- Change Kafka client settings to avoid exception “Local queue: Full” http://wwwb-sentry.us.archive.org/ia/spn/issues/114247/events/7853937/ https://git.archive.org/wb/pyspn/commit/b17677089952ddb197316bfa36e65d48240195d2

2020-01-09

- Check for empty PDF data before trying to parse it. https://git.archive.org/wb/pyspn/commit/d2f3b8d095bc8932a65b18b6833b5b6bd3800b36

- Handle UnicodeError in PDF URL extraction. https://git.archive.org/wb/pyspn/commit/79b061e24f9a6c2ef0600e9f5932df6107d7c00a
- Avoid UnicodeEncodeError when capturing extra embeds. Indicative exception: http://wwwb-sentry.us.archive.org/ia/spn/issues/114247/events/7840174/ Also trying to handle exceptions more gracefully https://git.archive.org/wb/pyspn/commit/b3a79d160c9e29d248c53eb88fa8527827ac7691

2020-01-07

- Fix corrupt SPN2 warcs. https://webarchive.jira.com/browse/WWM-1163

2019-12-26

- Improve spn.worker code structure. https://git.archive.org/wb/pyspn/commit/2360ca5b9891ff3a6a1b46a9edfaf5083dab96ff

2019-12-23

- More consistent user session management to enforce max sessions limits better. https://git.archive.org/wb/pyspn/commit/99934a6969bca8287bd5d59a837e5def15dd6cc5

2019-12-22

- Report download size progress for binary file captures. https://git.archive.org/wb/pyspn/commit/621d7704a557da8cd18b93f8918a86e43a88a5dc

2019-12-18

- Experiment with auto scaling the number of workers per machine based on system load. This experiment failed because we had a lot of errors when starting/stopping workers. https://webarchive.jira.com/browse/WWM-1182

2019-12-17

- Move outlinks outside Celery along with other capture info in a different Redis server to reduce `wwwb-celery0` memory overloading. https://git.archive.org/wb/pyspn/commit/019c4abf3ee42d82da18d92050bd27fdb128e669
- Resolve spn@archive.org exception: http://wwwb-sentry.us.archive.org/ia/spn/issues/190491/
- Upgrade to Celery 4.4.0 and Brozzler 1.5.18.

2019-12-16

- Move URL related code to new mode `spn.url` and add more unit tests.

2019-12-15

- Run “First archive” CDX query check asynchronously to speed up capture time. Capture and First archive check now run in parallel.

2019-12-13

- Optimise spn@archive.org to handle the increased load. Run processing in parallel, set limits on the max URLs we process per email msg. https://git.archive.org/wb/pyspn/commit/06e432ca70aa2219c1b6855a02f061ad385f80fc

2019-12-06

- Reorganize kafka related code and use Confluent kafka client instead of kafka-python because its faster and better. We resolve a lot of kafka related exceptions with this improvement.

2019-12-05

- Implement CDXAerospikeWriter warcprox plugin to write CDX data to CDXRedis and Aerospike in parallel.

Activity prior to December 2019 is available at:

- https://webarchive.jira.com/browse/WWM-832
- https://git.archive.org/wb/pyspn/activity