LAAWS Crawler Service API

An API for managing crawler tasks
More information: https://www.lockss.org/
Contact Info: lockss-support@lockss.org
Version: 1.0.0
BasePath:/
BSD-3-Clause
https://www.lockss.org/support/open-source-license/

Access

1. HTTP Basic Authentication

Methods

[Jump to Models]

Table of Contents

Crawlers

- GET /crawlers/{crawler}
- GET /crawlers/

Crawls

- POST /crawls/
- DELETE /crawls/{jobId}
- DELETE /crawls/
- GET /crawls/{jobId}
- GET /crawls/{jobId}/errored
- GET /crawls/{jobId}/excluded
- GET /crawls/{jobId}/fetched
- GET /crawls/{jobId}/notModified
- GET /crawls/{jobId}/parsed
- GET /crawls/{jobId}/pending
- GET /crawls/

Status

- GET /status

Crawlers

GET /crawlers/{crawler}

Up

Return information about a crawler. (**getCrawlerByName**)

Get information related to a installed crawler.

Path parameters

crawler (required)
Path Parameter — Identifier for the crawler

Return type
crawlerInfo

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
Crawler Info Found crawlerInfo
401
Access Denied.
404
Not Found

GET /crawlers/ Up

Get the list of supported crawlers. (**getCrawlers**)

Return the list of supported crawlers.

Return type
inline_response_200

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The crawler list. inline_response_200
404
No Such Crawler

Crawls

POST /crawls/ Up

Request a crawl using a descriptor (**addCrawl**)

Use the information found in the descriptor object to initiate a crawl.

Consumes
This API call consumes the following media types via the Content-Type request header:

- `application/json`

Request body

body crawlRequest (required)
Body Parameter —

Return type

crawlRequest

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
202
The crawl request has been queued for operation. crawlRequest
400
Bad Request
401
Unauthorized
403
Forbidden
404
Not Found
500
Internal Server Error

DELETE /crawls/{jobId}

Up

Remove or stop a crawl (**deleteCrawlById**)

Delete a crawl given the crawl identifier, stopping any current processing, if necessary

Path parameters

jobId (required)
Path Parameter — identifier used to identify a specific crawl. format: int32

Return type
crawlRequest

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The deleted crawl crawlRequest
401
Unauthorized
403
Forbidden
404
Not Found
500
Internal Server Error

DELETE /crawls/

Up

Delete all of the currently queued and active crawl requests (**deleteCrawls**)

Halt and delete all of the currently queued and active crawls

Query parameters

id (required)
Query Parameter — The crawl id

Responses
501
Not Implemented.

GET /crawls/{jobId}

Up

Get the crawl info for this job (**getCrawlById**)

Get the job represented by this crawl id

Path parameters

jobId (required)
Path Parameter — identifier used to identify a specific crawl. format: int32

Return type
crawlStatus

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The crawl status of the requested crawl crawlStatus
401
Unauthorized
404
Not Found
500
Internal Server Error

GET /crawls/{jobId}/errored

Up

A pagable list of errored urls. (**getCrawlErrored**)

Get a list of errored urls.

Path parameters

jobId (required)
Path Parameter — format: int32

Query parameters

continuationToken (optional)
Query Parameter — "The continuation token of the next page of jobs to be returned."

limit (optional)
Query Parameter — The number of jobs per page. format: int32

Return type
errorPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The requested errored urls. errorPager
400
Bad Request
401
Unauthorized
409
Not Found
500
Internal Server Error

GET /crawls/{jobId}/excluded

Up

A pagable list of excluded urls. (**getCrawlExcluded**)

Get a list of excluded urls.

Path parameters

jobId (required)
Path Parameter — identifier used to identify a specific crawl. format: int32

Query parameters

continuationToken (optional)
Query Parameter — "The continuation token of the next page of jobs to be returned."

limit (optional)
Query Parameter — The number of jobs per page. format: int32

Return type
urlPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The requested excluded urls. urlPager
400
Bad Request
401

Unauthorized
409
Not Found
500
Internal Server Error

GET /crawls/{jobId}/fetched

Up

A pagable list of fetched urls. (**getCrawlFetched**)

Get a list of fetched urls.

Path parameters

jobId (required)
Path Parameter — format: int32

Query parameters

continuationToken (optional)
Query Parameter — "The continuation token of the next page of jobs to be returned."

limit (optional)
Query Parameter — The number of jobs per page. format: int32

Return type
urlPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The requested fetched urls. urlPager
400
Bad Request
401
Unauthorized
409
Not Found
500
Internal Server Error

GET /crawls/{jobId}/notMotified

Up

A pagable list of notMotified urls. (**getCrawlNotModified**)

Get a list of notMotified urls.

Path parameters

jobId (required)
Path Parameter — format: int32

Query parameters

continuationToken (optional)
Query Parameter — "The continuation token of the next page of jobs to be returned."

limit (optional)
Query Parameter — The number of jobs per page. format: int32

Return type
urlPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The requested notMotified urls. urlPager
400
Bad Request
401
Unauthorized
409
Not Found
500
Internal Server Error

`GET /crawls/{jobId}/parsed` Up

A pagable list of parsed urls. (**getCrawlParsed**)

Get a list of parsed urls.

Path parameters

jobId (required)
Path Parameter — format: int32

Query parameters

continuationToken (optional)
Query Parameter — "The continuation token of the next page of jobs to be returned."

limit (optional)
Query Parameter — The number of jobs per page. format: int32

Return type
urlPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200

The requested modified urls. urlPager
400
Bad Request
401
Unauthorized
409
Not Found
500
Internal Server Error

GET /crawls/{jobId}/pending

Up

A pagable list of pending urls. (**getCrawlPending**)

Get a list of pending urls.

Path parameters

jobId (required)
Path Parameter — format: int32

Query parameters

continuationToken (optional)
Query Parameter — "The continuation token of the next page of jobs to be returned."

limit (optional)
Query Parameter — The number of jobs per page. format: int32

Return type
urlPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The requested modified urls. urlPager
400
Bad Request
401
Unauthorized
409
Not Found
500
Internal Server Error

GET /crawls/

Up

Get a list of active crawls. (**getCrawls**)

Get a list of all currently active crawls or a pageful of the list defined by the continuation token and size

Query parameters

limit (optional)
Query Parameter — The number of jobs per page format: int32

continuationToken (optional)
Query Parameter — The continuation token of the next page of jobs to be returned.

Return type
jobPager

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The requested crawls jobPager
400
Bad Request
401
Unauthorized
500
Internal Server Error

Status

`GET /status` Up

Get the status of the service (**getStatus**)

Get the status of the service

Return type
apiStatus

Produces
This API call produces the following media types according to the Accept request header; the media type will be conveyed by the Content-Type response header.

- `application/json`

Responses
200
The status of the service apiStatus
401
Unauthorized
500
Internal Server Error

Models

[Jump to Methods]

Table of Contents

1. apiStatus

2. counter
3. crawlRequest
4. crawlStatus
5. crawlerInfo
6. errorPager
7. inline_response_200
8. jobPager
9. mimeCounter
10. pageInfo
11. status
12. urlError
13. urlPager

apiStatus Up

The status information of the service

version
String The version of the service

ready
Boolean The indication of whether the service is available

counter Up

A counter for urls.

count
Integer The number of elements format: int32

counterLink
String A link to the list of count elements or to a pager with count elements.

crawlRequest Up

A descriptor for a LOCKSS crawl.

auId
String The unique au id for this crawled unit.

crawlKind
String The kind of crawl being performed. For now this is either new content or repair.
Enum:
newContent
repair

crawler (optional)
String The crawler for this crawl

repairList (optional)
array[String] The repair urls in a repair crawl

crawlStatus Up

The status of a single crawl.

key
String The id for the crawl.

auId
String The id for the au.

auName
String The name for the au.

type
String The type of crawl.

startUrls
array[String] The array of start urls.

startTime
date The time the crawl began in ISO-8601 format: date

endTime
date The time the crawl ended in ISO-8601 format: date

status
status

isWaiting
Boolean True if the crawl wating to start.

isActive
Boolean True if the crawl is active.

isError
Boolean True if the crawl has errored.

priority
Integer The priority for this crawl. format: int32

fetched (optional)
counter

excluded (optional)
counter

notModified (optional)
counter

parsed (optional)
counter

sources (optional)
array[String] The sources to use for the crawl.

pending (optional)
counter

errors (optional)
counter

mimeTypes (optional)
mimeCounter

bytesFetched (optional)
Integer The number of bytes fetched. format: int62

depth (optional)
Integer The depth of the crawl. format: int32

refetchDepth (optional)
Integer The refetch depth of the crawl. format: int32

proxy (optional)
String The proxy used for crawling.

crawlerInfo Up

Information about a specific crawler.

isEnabled
Boolean Is the crawler enabled

isRunning
Boolean Is the crawl starter running

infoMap (optional)
array[String] key value pairs specific providing configuration information.

errorPager Up

A Pager for urls.

pageInfo
pageInfo

errors
array[urlError] An list of urls

inline_response_200 Up

A list of crawlers.

crawls (optional)
array[String] An array of crawler names

jobPager Up

A display page of jobs

jobs
array[crawlStatus] The jobs displayed in the page

pageInfo
pageInfo

mimeCounter Up

A counter for mimeTypes seen during a crawl.

mimeType
String The mime type to count.

counts
Integer number of urls of mimeType. format: int32

pageInfo Up

The information related to pagination of content

totalCount
Integer The total number of elements to be paginated format: int32

resultsPerPage
Integer The number of results per page format: int32

continuationToken
String The continuation token

curLink
String The link to the current page

nextLink (optional)
String The link to the next page

status

Up

The existing state of a job

code
Integer The numeric value for the current state format: int32

msg
String A text message defining the current state

urlError

Up

The detail for an error

url
String The url which caused the error.

errCode
Integer The code for the error. format: int32

detail
String A detailed message related to the error.

urlPager

Up

A Pager for urls.

pageInfo
pageInfo

urls
array[String] An list of urls