Archive.org About Search

Introduction

The Internet Archive is pleased to provide API endpoints for search access to items in the archive; in particular, by searching against the metadata stored for items.

advancedsearch.php

The traditional method for API access for search is the Advanced Search API. The advanced search page describes the formats provided, and the query language for searching.

We limit the number of sorted paged results returnable to 10,000. Paged sorted results are supported only until the 10,000th result. For example, the search:

https://archive.org/advancedsearch.php?q=subject:palm+pilot+software&output=json&rows=100&**page=5**

should be fine, but requesting `page=10000` is rejected.

Scraping API

To provide the ability to page deeply into the items at the Archive, we are now providing a *scraping API*. The scraping API uses a *cursor* approach. One makes a scraping API call, which will return a list of results and a *cursor*. The cursor can then be used to search again with the search continuing where the cursor left off.

The following pseudo-code (it's really Python) shows a yield loop that loops through all of the results of a search by using a cursor:

```
def yield_results(basic_params):
        result = requests.get(endpoint, params=basic_params)
        while True:
            if (result.status_code != 200):
                yield (None, result.json())
                break
            else:
                result_obj = result.json()
                yield (result_obj, None)
                cursor = result_obj.get('cursor', None)
                if cursor is None:
                    break
                else:
                    params = basic_params.copy()
                    params['cursor'] = cursor
                    result = requests.get(endpoint, params=params)
```

The scraping API can be found at: https://archive.org/services/search/v1/scrape.

Its parameters are:

- `q` : the query (using the same query Lucene-like queries supported by Internet Archive Advanced Search.
- `fields` : Metadata fields to return, comma delimited
- `sorts` : Fields to sort on, comma delimited (if `identifier` is specified, it must be last)
- `count` : Number of results to return (minimum of 100)
- `cursor` : A cursor, if any (otherwise, search starts at the beginning)
- `total_only` : if this is set to `true` , then only the number of results is returned.

A description of these fields (in Swagger/OpenAPI format format) can be found at htttps://archive.org/services/search/v1.

Example

Here is an example CURL session scraping for items in the NASA collection: https://archive.org/services/search/v1/scrape?fields=title&q=collection%3Anasa

```
    > curl "https://archive.org/services/search/v1/scrape?fields=title&q=collection%3Anasa"
    {"items":[{"title":"International Space Station exhibit","identifier":"00-042-154"} ... ],
    "total":198879,"count":100,"cursor":"W3siaWRlbnRpZmllciI6IjE5NjEtTC0wNTkxNCJ9XQ=="}
    > curl "https://archive.org/services/search/v1/scrape?fields=title&q=collection%3Anasa&cursor=W3siaWRlbnRpZmllciI6IjE5NjEtTC0wNTkxNCJ9XQ=="
```

Note that there is no absolute guarantee that every item will be returned, or that every item returned will remain in the Archive. Additions and deletions happen all the time, and it's possible that an item could be added or deleted during the scraping loop. The `total` value returned is the number of items that match the search criteria (including the `count`). If a cursor is provided, the total refers, not to the total of all items, but the total of all items from the point of the cursor.

Internet command-line tool

The Internet Archive command line tool, named `ia`, is available for using Archive.org from the command-line. It also installs the internetarchive Python module for programatic access to archive.org. Version 1.0.0 and above use the scraping API transparently. Older versions must be upgraded for search operations to work properly.