Gmail

Michael Chwe <michaelchwe1965@gmail.com>

[HTRC JIRA] Comment posted to request #SUPPORT-27: question about automated searches

Peter Organisciak <htrc-help@hathitrust.org> Wed, Jul 27, 2016 at 12:50 PM
Reply-To: htrc-help@hathitrust.org
To: michael@chwe.net

—Write replies above—

Hi Michael Chwe,

There are automated ways to do it using both our search engine (Solr) and the Bookworm API: however, in both cases our services are currently on public domain works (about 5.5m) and not the entire HathiTrust collection. This will change in coming months, but for now, searching against the full 15m is only possible in the HathiTrust Digital Library, and can't be automated in an official way.

Here's the easiest way that you can do it programmatically against our Solr Proxy, if public domain is fine: set the response to XML or JSON, and you will see a "numFound" field. For example, here's the query for "+ocr:Hi +title:Test" (the text of the book has 'Hi' in it, and the title has 'Test' in it):

http://chinkapin.pti.indiana.edu:9994/solr/ocr/select/?q=%2Bocr:Hi%20%2Btitle:Test&rows=0&wt=json&fl=title

Click that link, and it should be clear where the numFound information is. You can edit the q=.... part of the URL to change the query. A few things to note:

- The numFound is based on **any** of the search parts matching, which is why I used pluses to force it to use **all** search terms. While "+ocr:Hi +title:Test" has 4631 results (obviously 'test' doesn't occur in many titles), "ocr:Hi title:Test" has 2.8 million (!!) because it counts books that have 'Hi' in them but not 'Test' in the title.
- Since we put the query into a URL, the '+' needs to be written as '%2B' because actual pluses mean something different in a web address.
- since you're not interested in the results, I set rows=0, so it doesn't return any of those books.

Hope that helps.

– Peter Organisciak

Michael Chwe created this request. Peter Organisciak is participating.

You can reply directly to this email to add any further comments or attachments.

See request details and updates for #SUPPORT-27 - "question about automated searches"