Search Cornell University Library Websites with Google

Google Search Appliance at Cornell

According to Uncle Ezra it was in March 2006 that Cornell replaced the Inktomi search engine it had been using for University website searching with a Google Search Appliance. The Google Search Appliance (GSA) is administered by the Office of Web Communications. The GSA now supports the search you see on the Cornell Identity Banner on most Cornell University web pages.

The Office of Web Communications has provided instructions for modifying the Identity Banner search dialog to limit results to a single domain. This is the Unit Search capability. This could work for searching domains ending in 'library.cornell.edu', but many of the Cornell Universiy Library digital collections have different domain names.

The Office of Web Communications also allows orgainzations to create a 'Google Search Appliance Collection'. This Collection is a list of the domains and paths you want to search. You maintain the list using an administrator interface that lets you add and remove items, provides statistics, and allows you to tell the indexing robots to check a certain domains in the collection. You can then point your search dialog at the indexes in your collection so only the paths in your collection will be searched.

The effect of these collections is to filter the results that would have come from the overall Cornell search, allowing only the subset of results that correspond to your list of domains to be reported. The advantage is that with the normal seach dialog you can only specifiy a single domain to filter on - with a collection it can be many domains. *Adding a domain to your collection that is outside of the Cornell master list will not cause it to be indexed!* You may be wondering what domains are in Cornell's master list.

The Cornell University Web Knowledgebase has some good information about the Google Search Appliance here...

Cornell University Library Websites Google Search Appliance Collection

In September 2006 I asked Lisa Cameron-Norfleet at The Office of Web Communications to set up a Google Search Appliance Collection that I could use for searching Cornell University Library websites. She kindly (and quickly) created the collection and told me how to get in to the administrator interface. I added all the Cornell Universiy Library digital collections , the list of individual libraries, the Registry of Digital Collections, and a few other library websites. Here are the domains and paths currently in the collection.

Once a collection is established it's easy to use. Here is a simple search form using the collection:

```
<form action="http://web.search.cornell.edu/search" method="GET" name="gs">
<label for="search-form-query">SEARCH Cornell University Library Websites:<br /></label>
<input type="text" name="q" value="" size="50" maxlength="256" id="search-form-query" />
<input type="submit" name="btnG" value="go" id="search-form-submit" />
<input type="hidden" name="sort" value="" />
<input type="hidden" name="ie" value="UTF-8" />
<input type="hidden" name="gsa_client" value="default_frontend" />
<input type="hidden" name="oe" value="UTF-8" />
<input type="hidden" name="site" value="libraries" />  <!-- note: 'libraries' instead of 'default_collection' --
>
</form>
```

I pointed the search dialogs on several websites at this collection, and used some special code to display the search results. Here are the search pages that use the Google Search Appliance to search Cornell University Library websites:

- commonspot.library.cornell.edu
- astech.library.cornell.edu
- www.glopad.org

What you can find with the Cornell Library GSA Collection

- English words in web pages, like canoe
- Words or phrases in UTF-8 characters, like ?????? (Unfortunately, Confluence does not play well with Japanese!)
- Phrases in pdf documents linked to web pages, like Engr Math PSL Vet* ACCEL
- Anything* in dspace, dlxs, or vivo - like dog for example (* not really - just things that show up on web pages that are linked to pages in the collection. The OCR text of articles inside dlxs, for example, is not available for searching this way, but dlxs index pages are.)

Statistics from Library Collection

The Google Seach Appliance Collection administration interface has a report and statistics section that can tell you things like 'How many pages are being crawled on each site?' or 'What were the top 100 search phrases in the month of March?'

Google Search Appliance Links

http://www.google.com/enterprise/gsa/index.html

Google's page describing the device.

Cornell Google Search Appliance web page.

Example Searches

Recently http://www.digitalhimalaya.com/ was added to the collection.

The Oxford Bön Project library gateway search

> When I first wrote this page the following search returned no results. I had http://www.digitalhimalaya.com/ in the libraries collection, but it was not in the overall Cornell collection. Now since The Office of Web Communications has added it to the Cornell collection I do get results:

The Oxford Bön Project Google Search Appliance of CUL sites search

Full web search for The Oxford Bön Project

Search for 'library hours' to find which libraries are searched:
GSA Collection Search for library hours

Library Gateway Search for library hours

Extra

The 'Search Library Pages' link on the Library Gateway page is using a search on indexes provided by Nutch - it finds things in 'library.cornell.edu' and 'mannlib.cornell.edu' and 'www.ilr.cornell.edu/library/catherwood/', but not things in some of the digital collections.

Here is an expiremental link to Luna Metadata for the Political Americana collection to check an issue with the GSA search.