What to Do When Google Ignores Your Fedora Objects

Robin Dean, Repository Director, Colorado Alliance of Research Libraries, robin@coalliance.org
Jonathan Green, Chief Technology Officer, Discovery Garden, Inc., jonathan@discoverygarden.ca

Abstract

This session will describe how we used RSS sitemaps and RDFa microdata to improve Google indexing and search result snippets for the Fedora objects in our Islandora repository. Though we are using a custom Islandora solution, the general principles will be applicable to any site using Islandora/Fedora.

Fedora User Group Session Proposal

Background

The Alliance Digital Repository (ADR) Islandora repository solution was developed with a custom JavaScript viewer to display metadata and content files from Fedora objects. Though this interface worked well for end users, it presented difficulties for having our objects indexed in Google with meaningful search results snippets. Most of our objects weren't being indexed, and those that were being indexed were showing search results snippets that weren't compelling.

In order to make repository objects discoverable in Google, we had to find a way to identify each of the Fedora objects as a unique "page" to be indexed by Google. Once Google could find the unique objects, we worked on improving what was indexed from each object page to make the search results snippets more relevant and appealing.

Getting Objects Indexed

In order to improve indexing of our Fedora objects, we generated and submitted sitemaps to Google in RSS feed format. Using RSS had several advantages:

1. The RSS feeds generated by Islandora automatically update when new content is added to Fedora.

2. Some of our sites have objects restricted with XACML (access control metadata) that should not be indexed in Google. The RSS feeds are permissions-aware and expose only the public objects to search engines.

3. By removing any limits on the RSS feed, we can generate lists of all the objects in a repository (in one case, as many as 50,000 objects).

Issues with Search Results

After successfully submitting our RSS feeds as sitemaps, we noticed that the number of objects that Google indexed was much lower than the total public Fedora objects listed in each of the sitemaps. Our theory was that the viewer code was making all the Fedora object pages look too similar and hence "not worth indexing" to Google.

Another problem was the quality of the search results snippets. The unique metadata content of each Fedora object page was buried within the page source. The early snippets tended to show only the page's breadcrumbs, to repeat the object title as the first line of the snippet, or display the labels on the metadata fields in addition to content of the fields.

Improving Search Results

In order to improve what Google displayed for each Fedora object's search results, we took two steps:

1. **Added the MODS <abstract> field to the HTML META description field on each object page.**

   This helped improve what was displayed in the Google search results snippet.

2. **Added RDFa markup to each MODS element on each object page.**

   We used RDFa to mark up the HTML table of MODS metadata that was already being rendered and indexed on each object page. This helped separate the metadata field labels from metadata field content, emphasized the unique metadata on each Fedora object page, and assigned meaning to the MODS elements to improve the relevance of what Google was indexing from our pages.

We will include graphs of objects indexed over time from Google Webmaster Tools, and will show some examples of search results snippets before and after.

Conclusions

After making these changes, we noticed that the number of objects that Google indexed from our repository sites increased considerably over the next two months. The snippet descriptions also improved, with the MODS <abstract> from the META description being privileged over the other, irrelevant page elements.

We will conclude with a discussion of other search engine optimization (SEO) efforts in Islandora, such as the optimization of institutional repository (IR) sites for indexing in Google Scholar.