Comments of Google Inc.

June 19, 2009

To: Office of Science and Technology Policy
Attn: Open Government Recommendations, 725 17th Street,
Washington, DC 20502
opengov@ostp.gov

RE: Open Government Initiative
Federal Registrar Volume 74, Number 97

Pablo Chavez
Managing Policy Counsel
Google Inc.
1101 New York Avenue, NW
Second Floor
Washington DC, 22201
202-346-1100

Introduction and Summary

Google appreciates the opportunity to submit comments on the Administration's Open Government Initiative.

Citizens use the Internet to reach out to the government to solve problems and answer basic questions. Government agency websites contain some of the most authoritative, valuable, and helpful information that citizens access when they are looking for government information online. In fact, a recent Pew Internet and American Life study found that of the adults using the Internet, 79% of them experienced a problem that government could help answer questions and provide assistance and 58% turned to the Internet to seek answers on issues that included health and care, funding for education, taxes, Medicare/Medicaid, food stamps, social security, voting information, and immigration. [1] Of the individuals accessing the Internet in search of government information, 38% turned directly to government office or agency websites. Given the value of government data and the significant amount of public interest in government information, it is important for agencies to make sure that their information is accessible to citizens online. It is equally important to ensure that the information published by agencies is accessible through open formats, and is timely and relevant to citizens.

Making government content accessible to citizens

To make more public government information accessible to citizens, it is important for Google and other search engines (such as Yahoo!, Bing, Ask) to be able to crawl agency websites. Google and other search engines index the web regularly, much of it daily. However, search engines cannot find everything, and if we can't find all public government content, citizens can't either.

The web is big. Last year Google systems that process links on the web to find new content hit a new milestone of 1 trillion unique URLs.[2]

[1] Rainie, Lee, Governing as Social Networking, Pew Internet & American Life Project, April 22, 2009. http://authoring.pewinternet.org/Presentations/2009/12-Governing-as-Social-Networking.aspx

[2] The Official Google Blog: "We knew the web was big" 7/25/2008 http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Given the volume of information online, the ability for users to find and discover relevant content is critical.

Search engines are an important way that Internet users find information that is relevant to their particular needs. According to a 2008 survey almost half of Internet users begin their online experience with a search engine.[3] Typically, citizens perform a search and click on the result that best addresses their immediate query. For example, if a citizen is looking for specific information on the safety of organic food, it is more likely that a user will type in their query into a search box (*e.g.*, 'organic food safety') than navigate directly to a relevant federal government website like that of the U.S. Department of Agriculture.

Citizens assume that search queries that are returned by search engines are complete. Therefore, it is critical for agency websites to be easily indexed and crawlable by search engines. If government websites do not allow search engines to crawl, or certain documents on websites are hidden by robots.txt files or behind databases, the results available to citizens are incomplete and not as helpful as they otherwise could be.

It is important for government websites to make their data accessible to search engines so that they can do their best to ensure that citizens receive relevant and timely results to citizens. We therefore believe that government agencies should move aggressively to ensure that non-sensitive and non-proprietary information is crawled and accessible to search engines.

One of the first steps that agencies can take in making their website more accessible to search engines is to implement a XML Sitemap. There are two different types of sitemaps: HTML and XML. The first type of sitemap is an HTML page listing the pages of your site, often by section, and is meant to help users find the information they need. The other approach, XML Sitemaps (with a capital S), is a way for websites to inform search engines about pages on their sites that are available for crawling.[4] In its simplest form, a Sitemap is an XML file that lists URLs (or web addresses) for a site along with additional metadata about each URL (such as when it was last updated, how

[3] Fallows,Deborah, Search Engine Use, Pew Internet & American Life Project, August 6, 2008. http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx

[4] Sitemap's website: http://www.sitemaps.org/

often it usually changes, and how important it is, relative to other URLs in the site) so that search engines can more intelligently crawl the site. Sitemaps are an effective way for webmasters to identify previously unknown URLs for search engines.

Unfortunately, not all government agencies have implemented the Sitemap protocol.

For those agencies who have adopted Sitemaps, the protocol has enabled them to deliver information more efficiently and provide citizens with access to deep content. States have led the way with adopting the Sitemap protocol. Virginia, California, Utah, Michigan and Arizona have all adopted the Sitemap protocol. For example, in less than 50 technical staff hours, Arizona's Government Information Technology Agency made hundreds of thousands of public records and other web pages crawlable to search engines and visible in search results.[5] In addition to states, many federal government agency websites have implemented Sitemaps including the Office of Scientific and Technical Information under the Department of Energy. OSTI implemented Sitemaps in 2007, which resulted in an increase of 400% in full-text downloads from Information Bridge (one million per month).

Sitemap 0.90 is offered under the terms of the Attribution-ShareAlike Creative Commons License and has wide adoption, including support from Google, Yahoo!, and Microsoft.[6] The protocol's wide adoption saves webmasters the time and money that it would take to create specialized maps for each individual crawler. Most search engines offer free Sitemap generator tools.[7] Using the Sitemap protocol does not guarantee that web pages are included in search engines, but the protocol does provide hints for web crawlers to do a better job of crawling websites.[8]

Agencies should also be encouraged to do a review of the use of robots.txt files on their website. Like Sitemaps, this often requires little technical time. The robot exclusion standard, also known as the Robots Exclusion Protocol or robots.txt protocol, allows webmasters to prevent

[5] Arizona State Sitemap case study: http://www.google.com/publicsector/arizona.html

[6] Attribution-ShareAlike Creative Commons License: http://creativecommons.org/licenses/by-sa/2.5/

[7] Google's sitemap generator: http://code.google.com/p/googlesitemapgenerator/

[8] Sitemap protocol: http://www.sitemaps.org/protocol.php

search engine’s web spiders from accessing all or part of a website which is otherwise publically available. Many agency websites contain both information that can be made publicly available and information that is sensitive and proprietary. Many webmasters of these sites take the extreme approach by blocking these sites entirely, which prevents search engines from crawling and indexing these websites. Webmasters have the ability to be selective about the pages on their website that they want to restrict search engines from crawling. Pages that contain sensitive information or are under construction can be individually blocked by using a robots.txt file or a no index meta tag. There is no need to block the website entirely when you can build in granular public access through code. Many search engines offer free robots.txt analysis tools that can assist Webmasters identifying pages that are being blocked from search engines.[9] Agency webmasters should be encouraged to do a review of their use of robots.txt files to identify additional web pages and information that can be made available to the public.

We have worked with many state agencies to help them identify and remove unnecessary robots.txt files, no index and no follow tags. These efforts have included the Florida Department of Education, Michigan Department of Human Services and the State of Michigan Workers Compensation Appellate Commission. This has resulted in tens of thousands of new pages containing deep content that are now discoverable to citizens through search.

Creating a sitemap and removing unwarranted robots.txt files is only part of the solution in creating an accessible website for search engines and citizens. The following are some other low cost easy technical suggestions that agencies should consider to ensure all of the content that they want to be discoverable is crawled and indexed:

- Update your title tags and meta descriptions to ensure that users are receiving a concise description of your website's content (This text appears on search results pages).

- Annotate all of the images on your website with ‘alt-text’ description and use descriptive name files. Many search engines’ crawlers do not recognize text contained in images.

[9] Google’s robot.txt tool: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35237

- Ensure critical information found in online databases is accessible by ensuring that each data point has its own unique URL. Alternatively, these databases should be made available via the website Data.gov, rather than maintained behind online forms where they cannot be crawled and indexed by search engines.

- Create a user-friendly sitemap (with a lower case 's') that appears under your website in the search results and that features links that help users navigate to the most important parts of your website.[10]

- Use a text browser such as Lynx to examine your site, because most search engine spiders see your site much as Lynx would. If fancy features such as JavaScript, cookies, session IDs, frames, DHTML, or Flash keep you from seeing all of your site in a text browser, then search engine spiders may have trouble crawling your site.

- Ensure that the structure of your website has a clear hierarchy and text links. Every page should be reachable from at least one static text link.

- Determine which keywords citizens would use in a search query to find the content on your website, and include these keywords in your site. Many search engines offer free webmaster tools to see what queries are driving traffic to your website and what position your website is showing in the search results against those keywords.

Making government data useful and relevant

Citizen participation and collaboration help make the government more transparent, accountable and relevant. Technology provides numerous low-cost, simple solutions for citizen participation and feedback, and the government should adopt them. One of the easiest ways to engage with citizens is by publishing data in open formats. When websites publish their data in formats that are open, citizens can aggregate, analyze and engage with the data and build the online tools necessary to make the information useful to them and others.

[10] Environmental Protection Agency's sitemap is a good example: http://www.epa.gov/enviro/html/sitemap.html

However, not all government data are the same. Citizens want access to timely and relevant information. Static, obscure, and dated information is not useful to most citizens. Third parties have little incentive to build useful tools around this data because citizens are looking for timely and high-quality information. Citizens are looking for access to timely information and data that is relevant to the political decisions happening in Washington, DC, as well as at the state and local levels.

Because of the resulting benefits to citizens, federal agencies should publish their current data in open formats. Non-proprietary, freely redistributable formats are available publicly and free of royalties. This means that the data can be used by any service provider without requiring purchase of additional licenses, thus saving money and reducing barriers to changing service providers. This cost savings and administrative flexibility is in the public interest. It is understood that some data cannot and should not be made available to the public, due to privacy, national security, and other important reasons. However, for data that can be made publicly available the following are some suggestions for principles of non-proprietary and freely redistributable formats:

- The format is interoperable among diverse internal and external platforms and applications.
- The format is fully published and available royalty-free.
- The format is implemented by multiple vendors.
- The format is controlled by an open industry organization with a well-defined inclusive process for evolution of the standard.

Conclusion:

As we embark upon modernizing the U.S. government’s technology we must make a realistic assessment of where we are today. We should go back to the basics and make the content that already exists accessible to citizens online. Making government content accessible and useful through search is an important step in opening up the government, making it more transparent, and making it more accountable to citizens. Federal agencies’ efforts to publish and make more accessible timely, relevant data in open formats so citizens can easily find and benefit from public government information is a significant step towards better serving American citizens and improving the transparency and accountability of government.

Many in the Administration and Congress are demonstrating every day that they know the power of new media and Web.2.0 tools to engage in dialogue with their constituents. We look forward to the day when all government agencies will embrace new ways to covey information, obtain feedback and showcase to the world the democratizing power of information.