Ensuring government is only one search away

Implementing the Sitemap protocol
Google’s relationships with government

Government

Regulator

Policy

Enterprise

Content Partnerships

Publisher

Corporate citizen

Vendor

Distributor
Sitemaps.org

An open, industry standard for search engine crawling

What are Sitemaps?

Sitemaps are an easy way for webmasters to inform search engines about pages on their sites that are available for crawling. In its simplest form, a Sitemap is an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how 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.

Web crawlers usually discover pages from links within the site and from other sites. Sitemaps supplement this data to allow crawlers that support Sitemaps to pick up all URLs in the Sitemap and learn about those URLs using the associated metadata. Using the Sitemap protocol does not guarantee that web pages are included in search engines, but provides hints for web crawlers to do a better job of crawling your site.

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.
<table>
<thead>
<tr>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barriers to search engine crawling</td>
</tr>
<tr>
<td>Scope of the issue</td>
</tr>
<tr>
<td>How citizens access government online</td>
</tr>
<tr>
<td>The Sitemap protocol</td>
</tr>
<tr>
<td>Success stories</td>
</tr>
<tr>
<td>Meeting citizens’ expectations of government</td>
</tr>
</tbody>
</table>
The three levels of search
Many government web services are effectively invisible to search engine users

Barriers to search engine crawling:

- Content “hidden” behind search forms
- Non-HTML links
- Outdated robots.txt crawling restrictions
- Server errors
- Orphaned URLs
- Rich media: audio, video
- Paid/subscription-only access
Representative barriers

• 1: US Census Bureau Question & Answer Center (https://ask.census.gov/):
  – [census america's largest minority group]
  – Invisible page: http://tinyurl.com/yqbja7
  – Barrier: Dynamic database
Representative barriers

• 2: ATF website (http://www.atf.gov):
  – Sample search: [atf small business program]
  – Invisible page: http://tinyurl.com/2c2qam
  – Barrier: Robots.txt
Representative barriers

• 3: USAJOBS.gov job search (http://jobsearch.usajobs.opm.gov/):
  – [statistician jobs federal government]
  – Invisible page: http://tinyurl.com/2yja46
  – Barriers: Agency information access policy and dynamic database
Scope of the issue

• US:
  – Federal government: an informal “Scope and Progress” report
  – EPA:  
    http://spreadsheets.google.com/ccc?key=pUb62ZKHnzggEoGF4L Ff3Gw
  – State of Texas:  
    http://spreadsheets0.google.com/ccc?key=pUb62ZKHnzgpiqOMS qZj-AA

• UK central government: initial analysis

• Organization of American States:  
  http://spreadsheets.google.com/ccc?key=pUb62ZKHnzgo2eHJ US9bBRg&pli=1
Public sector information is a pillar of the web

Generally authoritative, unbiased, free and trustworthy

Untrustworthy Trustworthy

Government
The majority of citizens access government through search engines

National Institutes of Health (nih.gov):

- More than 70% of unique users in July 2006 were referred by search engines (Google, Yahoo, MSN, AOL, Ask)

- Only 4% of unique users came directly to nih.gov sites

Source: ComScore, 2006
And citizens expect to find everything through search

The long tail of US federal government information
The good news

• The Sitemap protocol provides a solution:
  – Can make all pages, documents, and records in a web service accessible to search engine crawlers (and thus visible to search engine users)
  – Does not require website redevelopment or redesign, just a comparably modest time investment
  – Keeps the site’s owner in control
XML Sitemap

- A comprehensive list of a site’s URLs tagged with:
  - Location
  - Last modification date
  - Change frequency
  - Priority

XML sitemap

```xml
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.84">
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=12&amp;desc=vacation_haui</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=73&amp;desc=vacation_new_zealand</loc>
    <lastmod>2004-12-23</lastmod>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=74&amp;desc=vacation_newfoundland</loc>
    <lastmod>2004-12-23T16:00:15+00:00</lastmod>
    <priority>0.3</priority>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=50&amp;desc=vacation_usa</loc>
    <lastmod>2004-11-23</lastmod>
  </url>
</urlset>
```
One solution for all search engines

Once implemented, Sitemaps can be made available to all major search engines

“The launch of Sitemaps is significant because it allows for a single, easy way for websites to provide content and metadata to search engines”
— Tim Mayer, Senior Director of Product Management, Yahoo Search

“We are 100% behind this protocol -- this kind of collaboration will help improve the search experience for all of our customers”
— Ken Moss, General Manager, Live Search
Success stories

- US federal government:
  - PlainLanguage.gov
  - OSTI, Department of Energy
- State of California
- National Archives of the United Kingdom
PlainLanguage.gov success story

- Plain Language Information and Action Network (PLAIN), a federal inter-agency volunteer working group that encourages clarity in government communication to the public through PlainLanguage.gov

- Before-and-after examples of government documents served dynamically, thus uncrawlable
PlainLanguage.gov success story

• Web manager successfully implemented sitemap in ~8 hours, using available resources and through trial and error

• As new examples are added to the database, the sitemap is regenerated and submitted
OSTI success story

• Department of Energy agency that “makes R&D findings available and useful, so that science and technological creativity can advance”

• Web manager submitted sitemaps for Energy Citations and Information Bridge services, opening 2.3M bibliographic records and full-text documents to crawling

• Sitemap standard assures search engines have “a complete picture” of information in OSTI services
• Benefits include better representation in search results and reduced load on servers (by limiting duplicate crawling)

• First implementation completed in 16 staff hours -- can now be easily replicated across search engines
Google partners with states of Arizona, California, Utah and Virginia

“Transparent digital government is a top priority in Virginia. Our goal is to provide easy, quick and intuitive citizen access to every government resource. Google is helping us to achieve this.”

-- Tim Kaine, Governor of Virginia
California Department of Health Services implemented Sitemaps
Now citizens can easily find more of the information on the agency’s public websites.
National Archives of the UK

• ~20 million records across multiple services inaccessible to search engines
• Sitemaps successfully implemented to open several million records
• [needham national archives uk]
Meeting citizens’ expectations of government

<table>
<thead>
<tr>
<th>Google Benefits</th>
<th>Agency Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Expands the reach of Google’s search services</td>
<td>• Ensures all public information is visible to all potential users</td>
</tr>
<tr>
<td>• Incorporates more authoritative, trustworthy information</td>
<td>• Maximizes the investment e-gov services and the search engine channel</td>
</tr>
<tr>
<td>• Increases crawling efficiency</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>• More and better information from a trusted source</td>
</tr>
<tr>
<td>• Quicker access to and navigation of government information and services</td>
</tr>
<tr>
<td>• Serendipitous discovery</td>
</tr>
</tbody>
</table>
Toward more transparent government

• What are the long-range information policy goals that remain unmet by current eGovernment strategies?
  – Effective, rather than just functional, access to all public information, through citizens’ chosen channels (search engines)

• What new dynamics of information services and web publishing marketplace should governments consider in setting eGovernment strategy?
  – Search engines have become citizens’ preferred gateway to all levels of public information
Toward more transparent government

• How can governments leverage open standards (Web protocols, data formats, etc.) to achieve public sector information goals and better interoperability?
  – By fully adopting the Sitemap protocol, a government can ensure that all public documents are only one search away

• Are there any significant differences between Web services created in commercial or other non-governmental environments, and those designed specifically for government application? If so, what can we learn from these differences about the unique requirements of eGovernment services?
  – To enable navigation over its corpus of information, and unleash its value, a government must fully embrace web search and structured data distribution
Join in

• Get informed: www.google.com/publicsector

• Track progress by US federal government agencies at Federal Sitemaps wiki: http://tinyurl.com/3byhy7

• Advocate for the Sitemap protocol

• Query: John Lewis (JL) Needham, Manager, Public Sector Content Partnerships, jlneedham@google.com