Making your agency's sites more accessible to web search engine users

Implementing the Sitemap protocol

Agenda

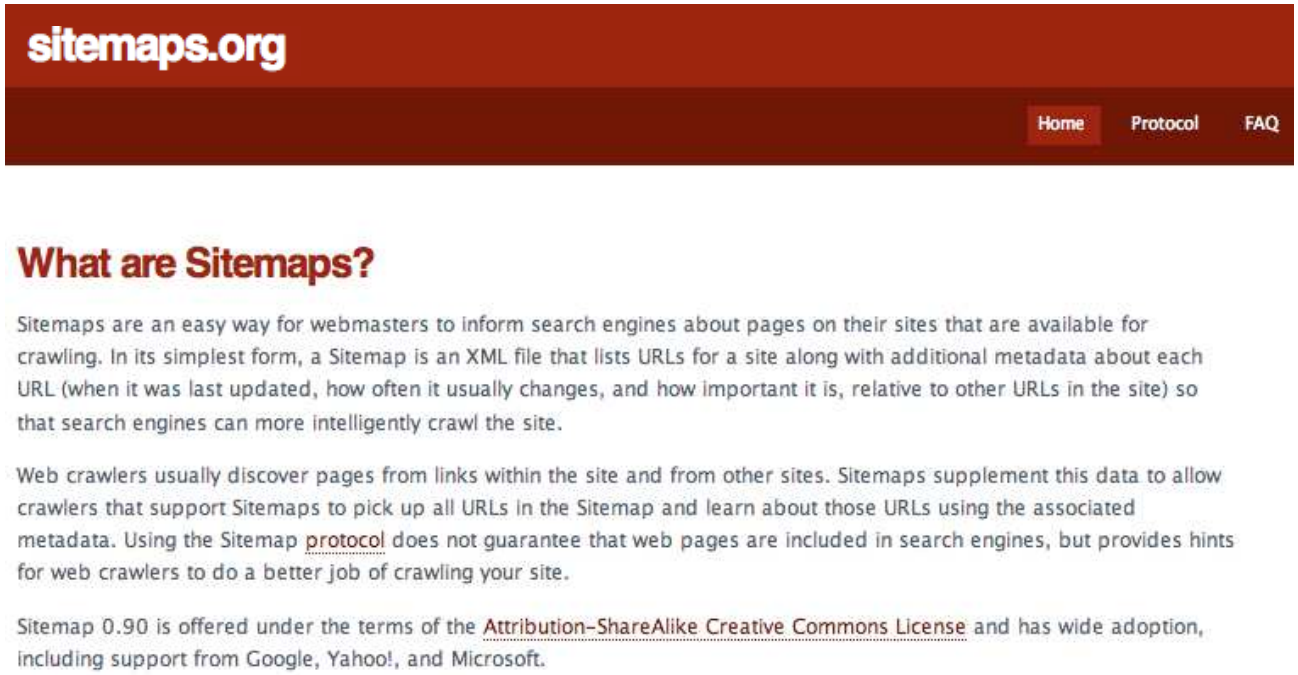
- Common barriers to web search engine crawling
- Supporting the two levels of search
- The Sitemap protocol
- Q&A
- Success stories
- More Q&A

Google

See slides illustrating “Common barriers to web search engine crawling”

Sitemaps.org

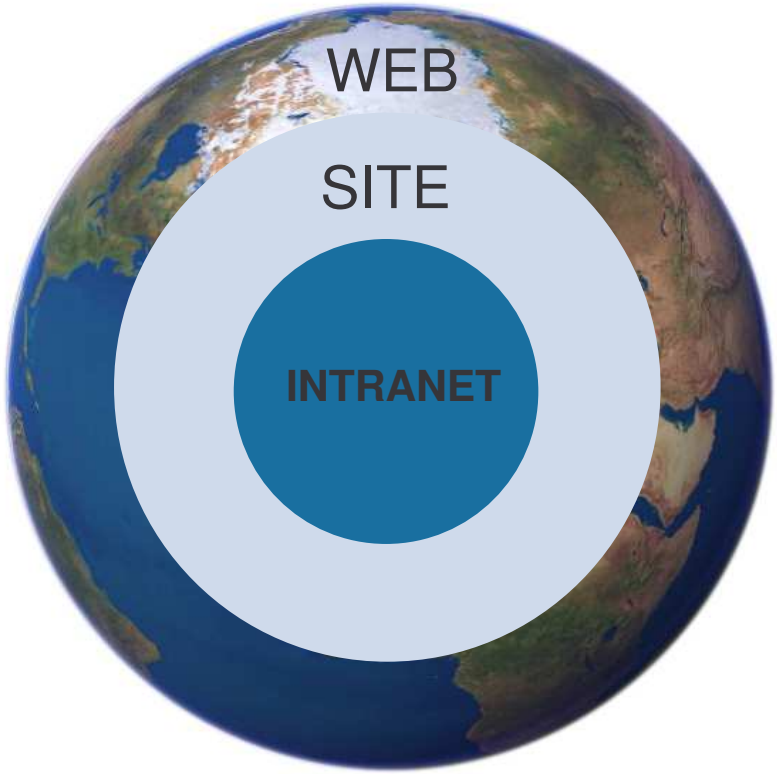
An open, industry standard for web search engine crawling



Google

Clarifications

- Non-proprietary
- No direct cost—nothing for sale
- No security risk
- Web search, not site search (e.g. Google Search Appliance)
- Public content only



Google

Web search vs. site search

Supporting the two levels of search

All of the open and accessible deep web	**Search scope**	A segment of your public sites' content
Citizens and professionals	**User**	Professionals and citizens
Search engine crawling intervals	**Freshness**	Customizable
Limited by robots.txt, dynamic content	**Crawling**	Limited by server capacity and cost
High-level stats	**Reporting tools**	More detailed, all facets
Free	**Cost**	Varies

Google

Citizens increasingly access government through web search engines

National Institutes of Health (nih.gov)

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

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

Source: ComScore, 2006

Google

Web search engines are the point of departure, government sites are the destination

Google

Federal

Internal Revenue Service
DEPARTMENT OF THE TREASURY

NASA

State

virginia.gov

utah dot gov

Localities

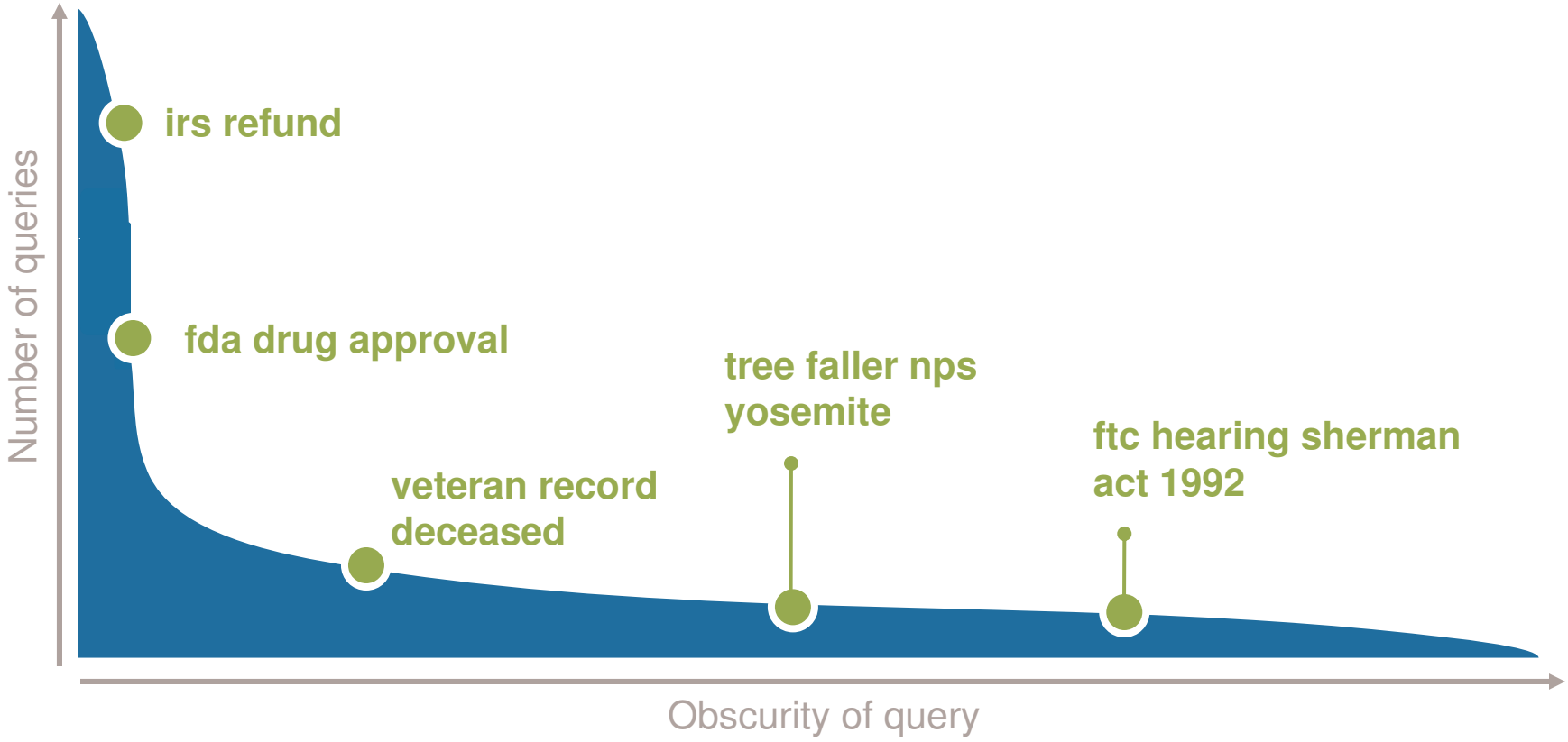
King County

City of Dallas

Google

And they expect to find everything

The long tail of federal government information

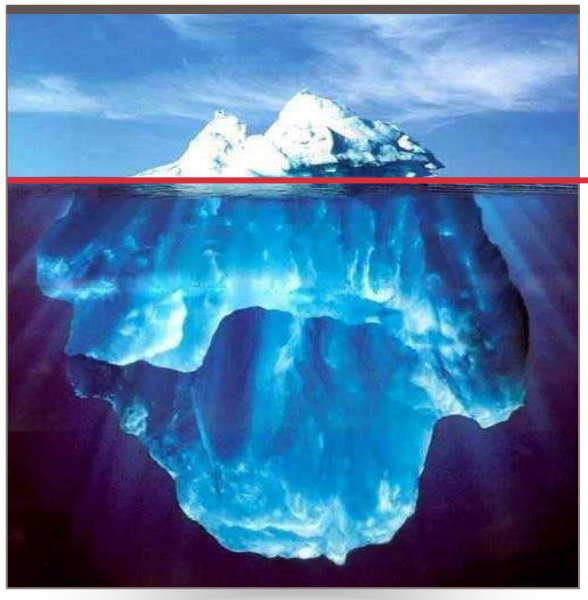


Google

Barriers to web search engine crawling

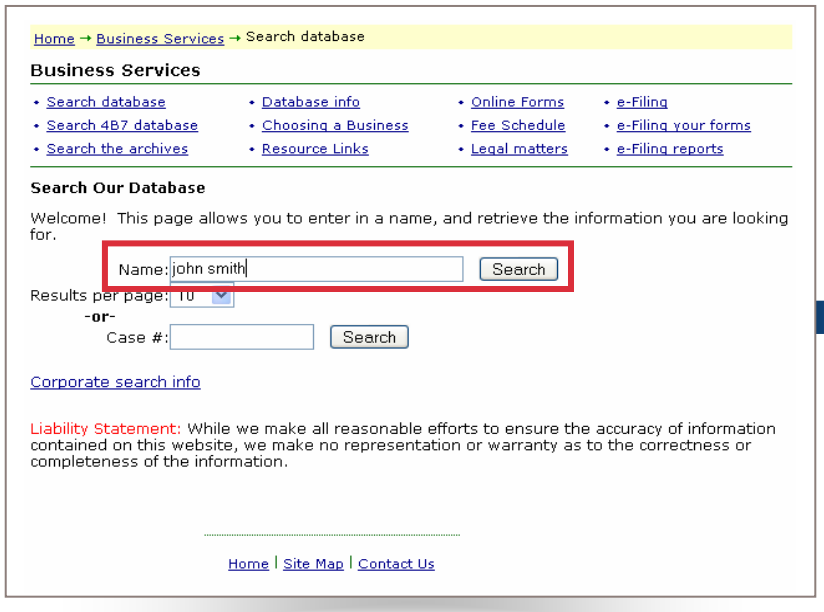
What can make a site effectively invisible to search engine users

- Content “hidden” behind search forms
- Non-HTML links
- Outdated robots.txt crawling restrictions
- Server errors (crawler times out when fetching content)
- Orphaned URLs
- Rich media: audio, video
- Premium content

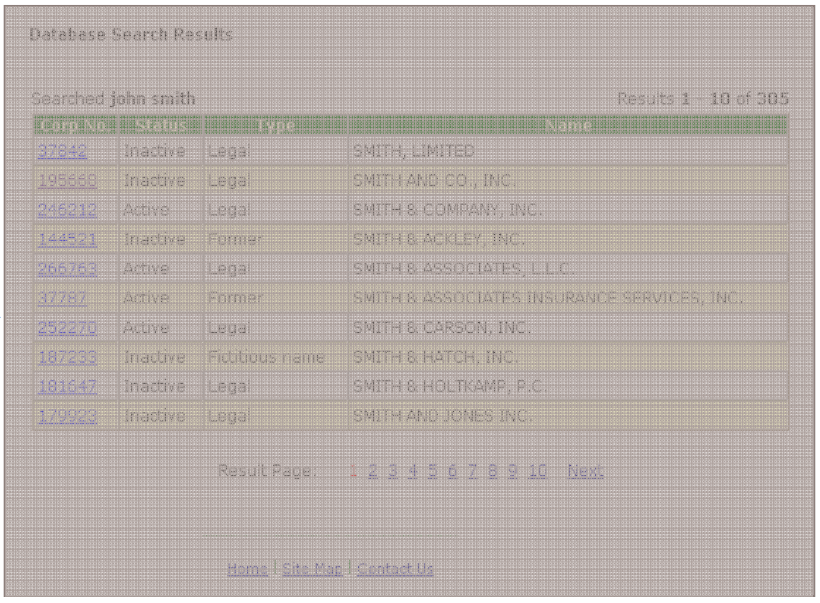


Google

Crawlers cannot navigate search forms



When crawled



Search results are invisible

Google

The solution: Sitemaps

The Sitemap protocol enables a web publisher to proactively manage web search engine crawling

"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

- Sitemap protocol developed by Google in June 2005 and released under Creative Commons License
- Adopted as an industry standard in November 2006: www.sitemaps.org

Google

Sitemaps for users

A browse index or site map that enables a user to navigate throughout a site

SITE INDEX

To view or print the PDF content on this page, download the free Adobe® Acrobat® Reader®.

NEWS

Treasury Deputy Secretary Kimmitt Travels to Asia this week to Discuss Compact with Iraq

KEY TOPICS

General Interest
Law Enforcement
International
Taxes
Financial Markets
Currency & Coins
Small Business
Accounting & Budget
Technology

PRESS ROOM

Public Schedule

OFFICES

Office of Domestic Finance
- Debt Management
- Advanced Counterfeit Deterrence
- Office of Financial Institutions
- Federal Financing Bank
- Financial Institutions
- Financial Markets
- Fiscal Service

Office of Economic Policy
- Working Papers
- Total Taxable Resources

Terrorism and Financial Intelligence
- Office of Foreign Assets Control
- Executive Order 13324
- National Money Laundering Strategy
- Executive Office for Asset Forfeiture

Google

Sitemaps for web search engines

A comprehensive, machine-readable listing of the site's URLs in:

- HTML
- Simple text
- XML

Simple text sitemap

http://www.firstgov.gov/index.shtml
http://www.firstgov.gov/About.shtml
http://www.firstgov.gov/Citizen/Services/Address_Changes.shtml
http://www.firstgov.gov/Topics/Parents_Adoptive.shtml
http://www.firstgov.gov/Government/State_Local/Ag_Environment.shtml
http://www.firstgov.gov/Citizen/Topics/Environment_Agriculture/Agriculture.shtml
http://www.firstgov.gov/Citizen/Facts/Facts_Agriculture.shtml
http://www.firstgov.gov/Agencies/Federal/Executive/Agriculture.shtml

XML sitemap

- A comprehensive list of URLs in XML
- Tagged with each URL's location, last modification, change frequency and priority

XML sitemap

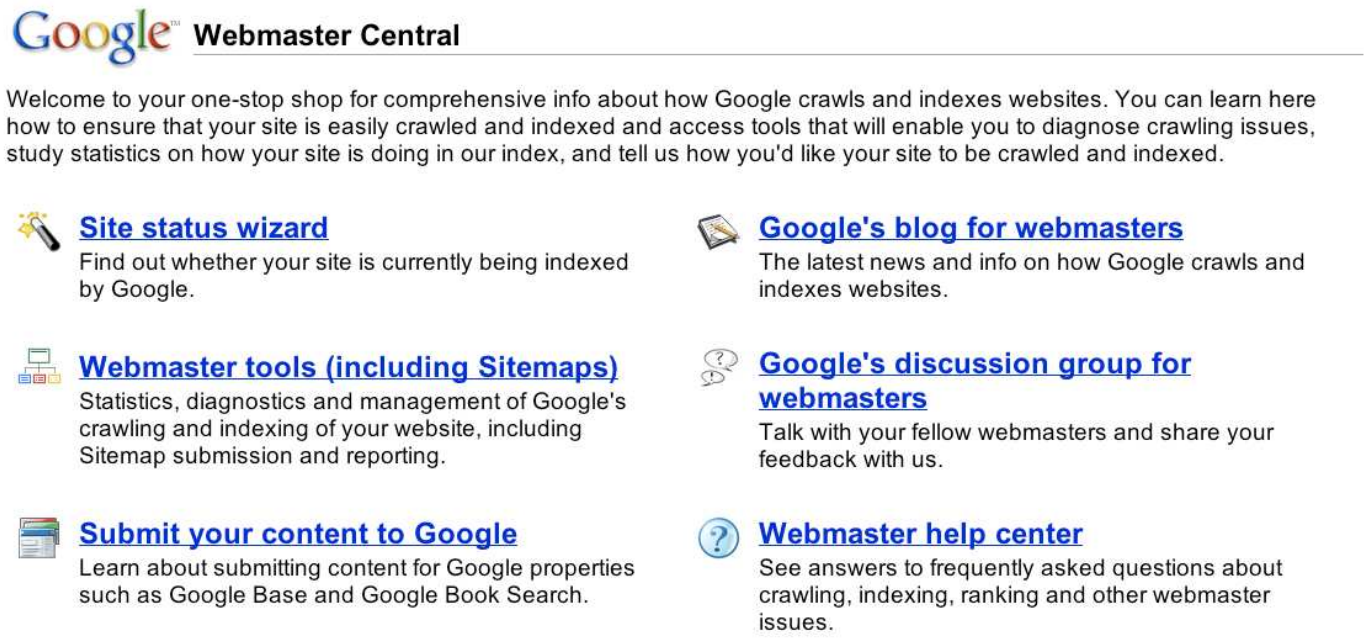
```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.84">

   <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_hawaii</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>2004-12-23T18:00:15+00:00</lastmod>
      <priority>0.3</priority>
   </url>
   <url>
      <loc>http://www.example.com/catalog?item=83&amp;desc=vacation_usa</loc>
      <lastmod>2004-11-23</lastmod>
   </url>
</urlset>
```


Google

Introducing Google's Webmaster Tools

Free resources and tools to help you implement sitemaps and improve your sites' visibility in Google search results



Learn more at: http://www.google.com/sitemapsgov

Google

Q&A

Google

Success stories

Questions to consider

- For what web services have you implemented sitemaps?
- Why did you implement the Sitemap protocol?
- What benefits have you observed or do you anticipate?
- What steps did the implementation involve?
- What was the biggest challenge?

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

Google

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

Google

Library of Congress success story

- THOMAS, a Library of Congress service that provides access to the proceedings of Congress -- one of many sites LOC has opened to search engine users

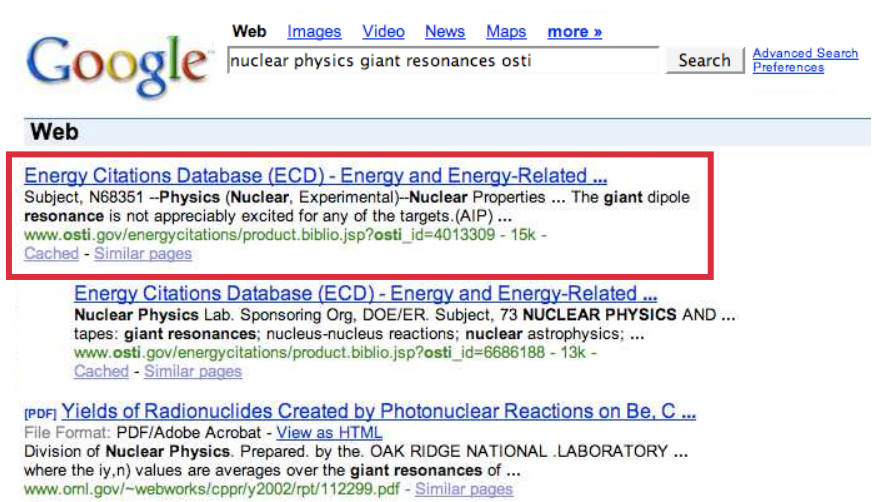
Google

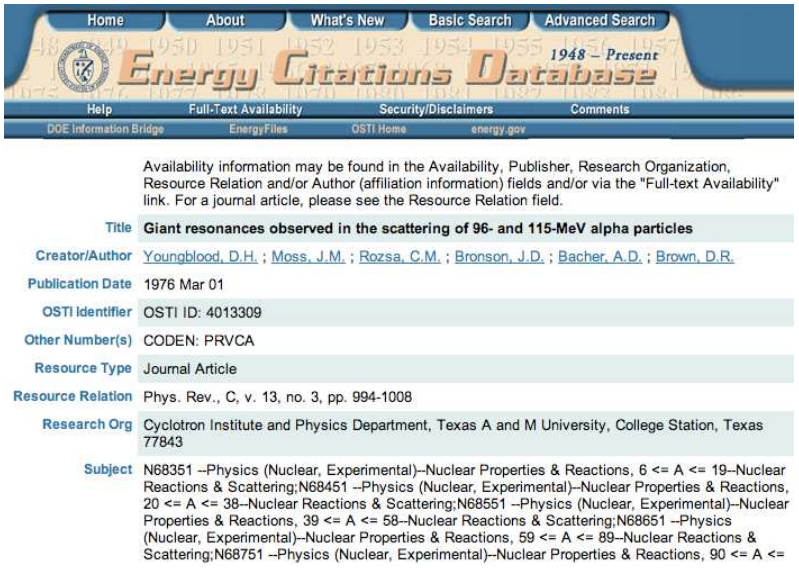
OSTI success story

- 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 web search engines have “a complete picture” of information in OSTI services

Google

OSTI success story





- 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 web search engines

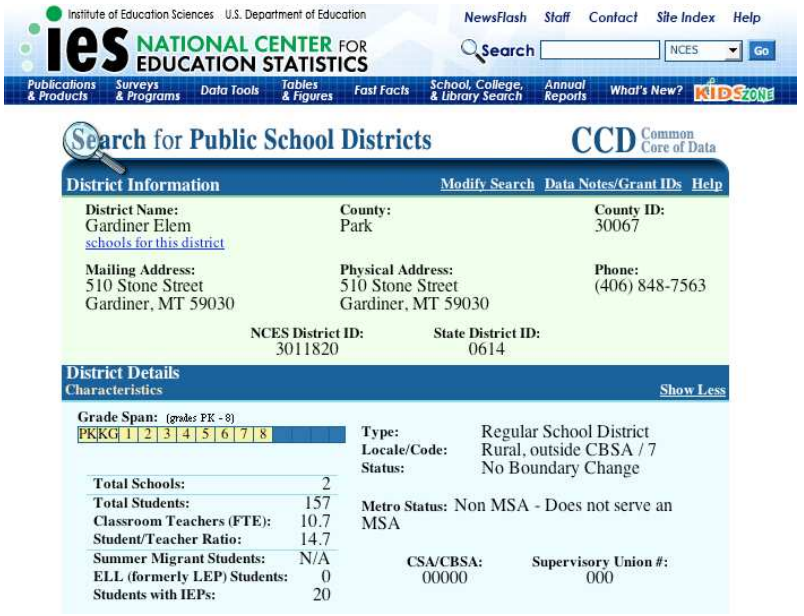
Google

NCES success story

- Department of Education agency that provides statistical information about districts, schools, and other educational facilities
- Using freely available tools, web manager submitted sitemaps to open five dynamic databases to crawling, adding 180K URLs

Google

NCES success story

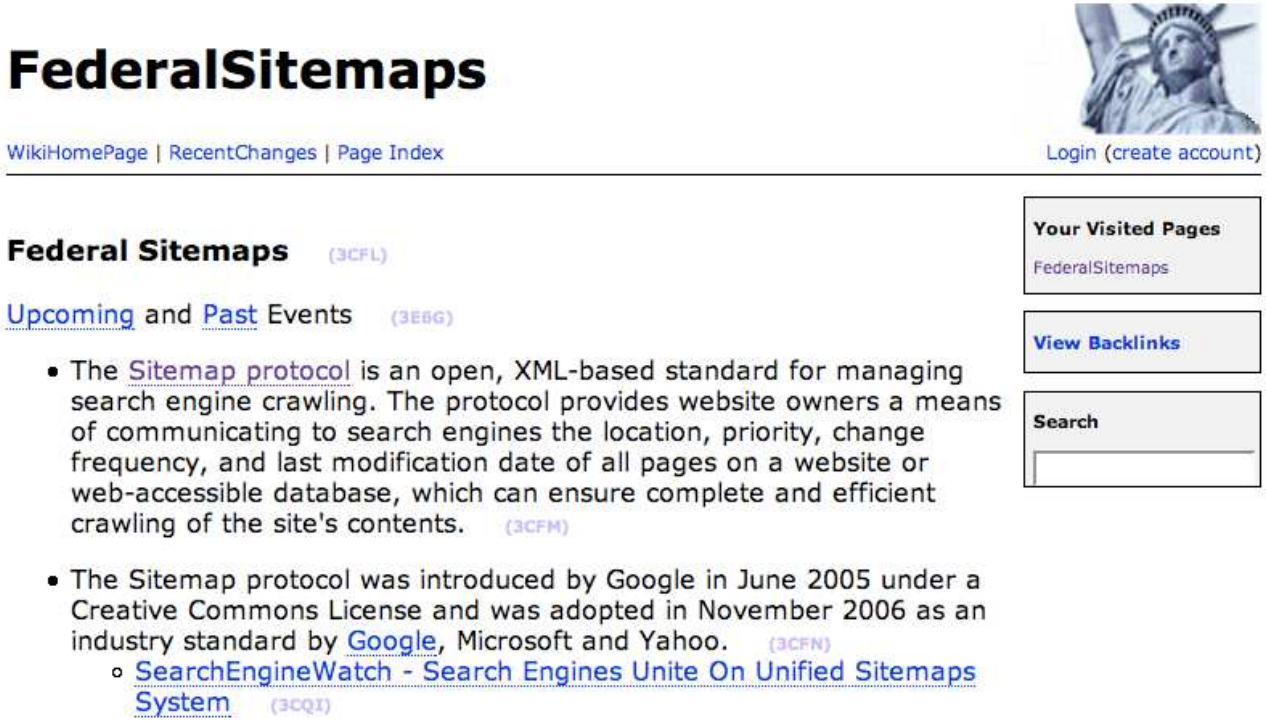


- Now surfacing tens of thousands of potential web search hits with links to NCES services
- Helping to ensure citizen users gain access to the latest data from the original source

Google

Federal Sitemaps wiki

An initiative to help federal agencies make their websites more accessible to search engine users



FederalSitemaps

WikiHomePage | RecentChanges | Page Index

Login (create account)

Federal Sitemaps (3CFL)

Upcoming and Past Events (3E6G)

- The Sitemap protocol is an open, XML-based standard for managing search engine crawling. The protocol provides website owners a means of communicating to search engines the location, priority, change frequency, and last modification date of all pages on a website or web-accessible database, which can ensure complete and efficient crawling of the site's contents. (3CFM)
- The Sitemap protocol was introduced by Google in June 2005 under a Creative Commons License and was adopted in November 2006 as an industry standard by Google, Microsoft and Yahoo. (3CFN)
  - SearchEngineWatch - Search Engines Unite On Unified Sitemaps System (3CQI)

Your Visited Pages

FederalSitemaps

View Backlinks

Search

http://colab.cim3.net/cgi-bin/wiki.pl?FederalSitemaps
Or http://tinyurl.com/3byhy7

Google

Relevant legislation and OMB policy

- The Sitemap protocol supports the **E-Government Act of 2002** requirements to:
  - "Organize and categorize information intended for public access and ensure it is searchable across agencies…[using] sophisticated Internet search functions (including their crawl and index mechanisms)..."
  - "…publish your information directly to the Internet…expos[ing] information to freely available and other search functions [that] adequately [organize] and [categorize] your information."
  - "…[When] disseminating significant information dissemination products, advance preparation, such as using formal information models, may be necessary to ensure effective interchange or dissemination. This procedure is needed when freely available and other search functions do not adequately organize and categorize your information."
- The Sitemap protocol also supports the **Federal Enterprise Architecture's Data Reference Model 2.0** requirements to:
  - "Identify how information and data are created, maintained, accessed, and used…[and] Define data and describe relationships among data elements used in the agency's information systems."

Google

Q&A

Google

Next steps for web managers: Prepare

- Audit your agency's sites to identify uncrawlable elements
  - Google can provide support with analysis: **sitemap-partners@google.com**
  - Sample sitemapping target list: http://spreadsheets.google.com/pub?key=pUb62ZKHnzgqEoGF4LFf3Gw
- Get trained:
  - Attend Google's webinar on technical steps to implementing sitemaps: Thursdays, 3:00 EST
  - Or arrange dedicated webinar for your agency

Next steps for web managers: Implement

- Sign up at www.google.com/sitemapsgov:
  - Verify your sites' ownership
  - Produce and upload sitemaps
- Get answers:
  - At Webmaster Central: **www.google.com/webmasters**
  - Or directly: **sitemap-partners@google.com**
- Track your progress

Google

Making your agency's sites more accessible

- Implementing sitemaps can enhance, but does not replace, a web search engine's crawling
- It does not guarantee inclusion, but helps to provide users more information and fresher results

- The Sitemap protocol is an open, industry standard
- Ensures all your agency's public information and services are discoverable by all potential users
- Also accelerates the inclusion of new information in search results

- Makes web search engine crawling more efficient, reducing demands on servers
- Most sitemaping tools are free and can be easy to implement
- Can be readily replicated across web search engines

Google