

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

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

Sitemaps.org

An open, industry standard for web search engine crawling



What are Sitemaps?

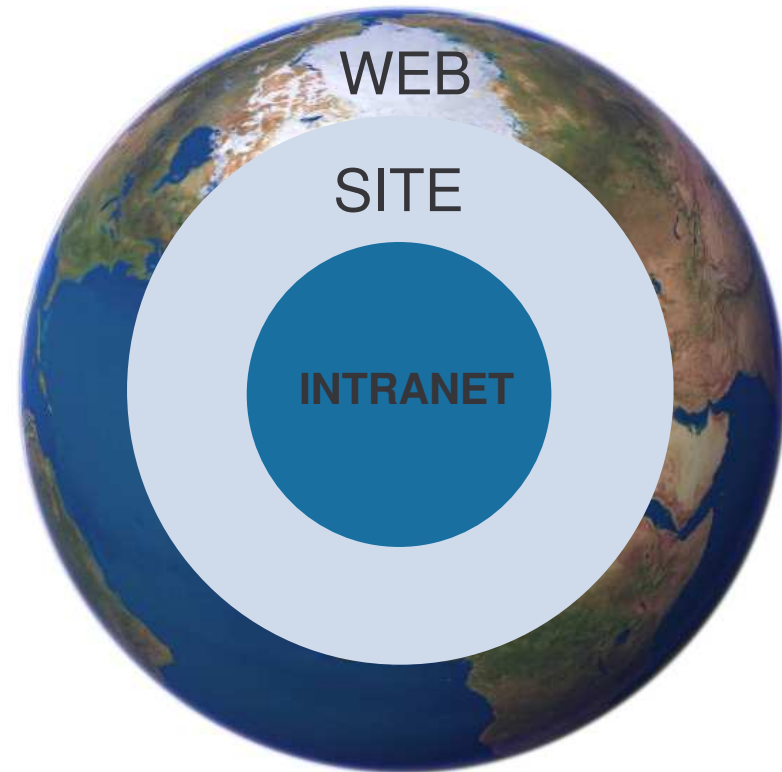
Sitemaps are an easy way for webmasters to inform search engines about pages on their sites that are available for crawling. In its simplest form, a Sitemap is an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how often it usually changes,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

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



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

Citizens increasingly access government through web search engines

National Institutes of Health (nih.gov)

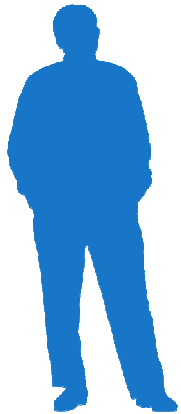
- More than 70% of unique users in July 2006 were referred by web search engines (Google, Yahoo, MSN, AOL, Ask)



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

Source: ComScore, 2006

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



Federal



State



Localities



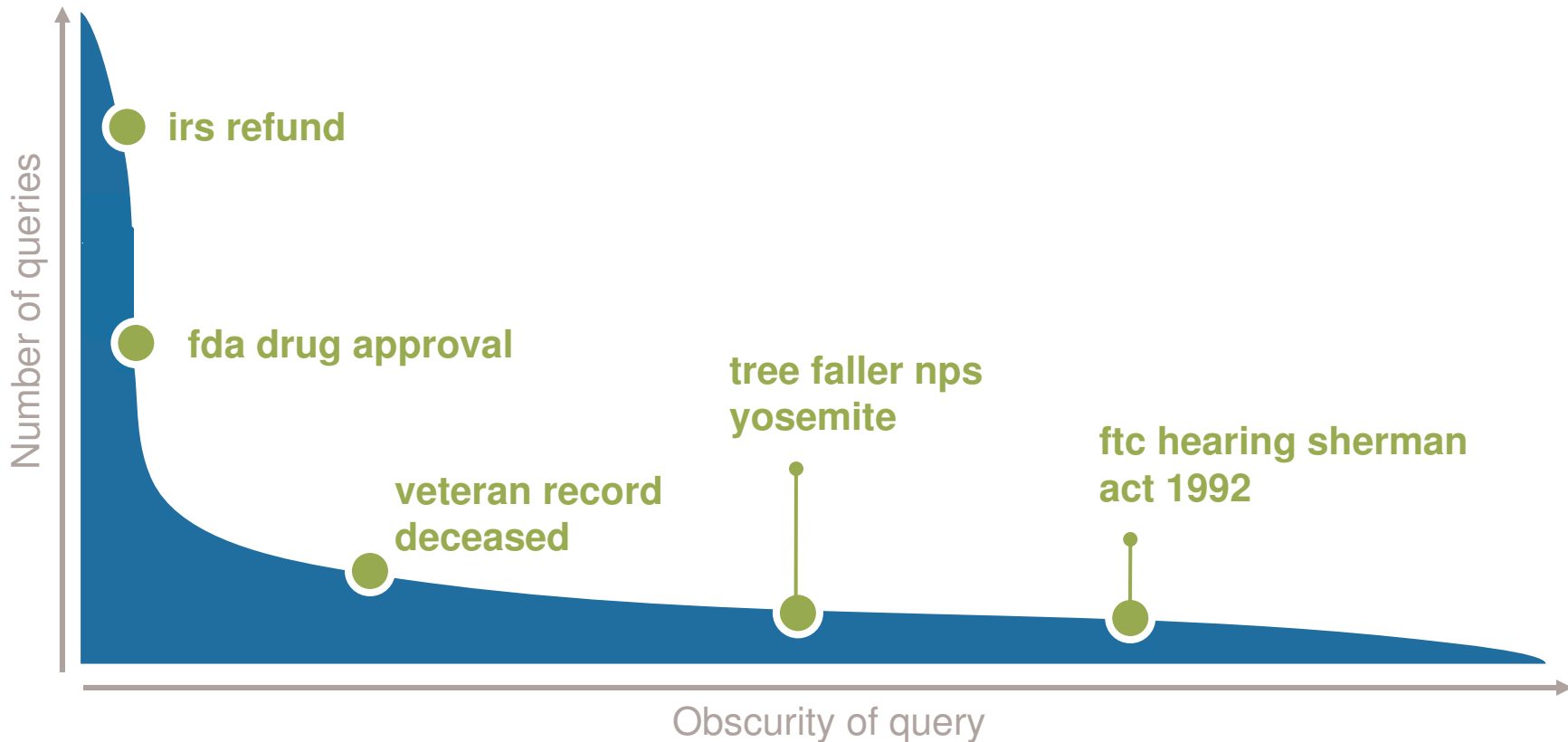
King County



City of Dallas

And they expect to find everything

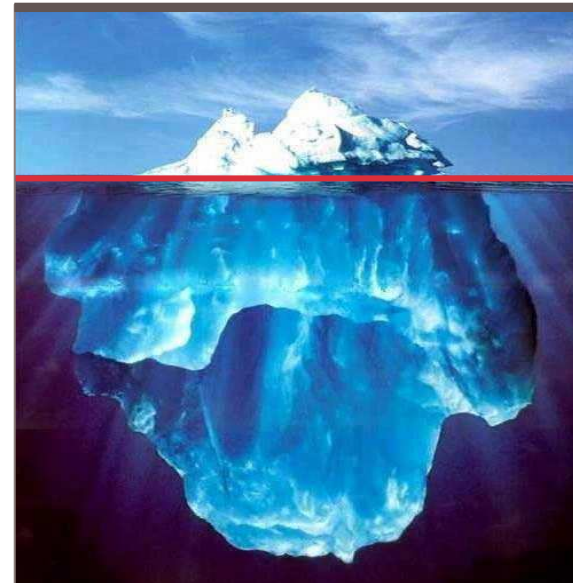
The long tail of federal government information



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



WEB
Searchable

DEEP WEB
Not searchable

Crawlers cannot navigate search forms

When crawled

[Home](#) → [Business Services](#) → Search database

Business Services

- [Search database](#)
- [Search 4B7 database](#)
- [Search the archives](#)
- [Database info](#)
- [Choosing a Business](#)
- [Resource Links](#)
- [Online Forms](#)
- [Fee Schedule](#)
- [Legal matters](#)
- [e-Filing](#)
- [e-Filing your forms](#)
- [e-Filing reports](#)

Search Our Database

Welcome! This page allows you to enter in a name, and retrieve the information you are looking for.

Name:

Results per page: 10

-or-

Case #:

[Corporate search info](#)

Liability Statement: While we make all reasonable efforts to ensure the accuracy of information contained on this website, we make no representation or warranty as to the correctness or completeness of the information.

[Home](#) | [Site Map](#) | [Contact Us](#)



Database Search Results

Searched john smith Results 1 - 10 of 385

Case No.	Status	Type	Name
37842	Inactive	Legal	SMITH, LIMITED
195660	Inactive	Legal	SMITH AND CO., INC.
246212	Active	Legal	SMITH & COMPANY, INC.
144521	Inactive	Former	SMITH & ACKLEY, INC.
266763	Active	Legal	SMITH & ASSOCIATES, L.L.C.
37787	Active	Former	SMITH & ASSOCIATES INSURANCE SERVICES, INC.
252270	Active	Legal	SMITH & CARSON, INC.
187293	Inactive	Fictitious name	SMITH & HATCH, INC.
181647	Inactive	Legal	SMITH & HOLTkamp, P.C.
179923	Inactive	Legal	SMITH AND JONES INC.

Result Page: 1 2 3 4 5 6 7 8 9 10 Next

[Home](#) | [Site Map](#) | [Contact Us](#)

Search results are invisible

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

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	OFFICES
Treasury Deputy Secretary Kimmitt Travels to Asia this week to Discuss Compact with Iraq	Office of Domestic Finance
KEY TOPICS	Debt Management
General Interest	Advanced Counterfeit Deterrence
Law Enforcement	Office of Financial Institutions
International	Federal Financing Bank
Taxes	Financial Institutions
Financial Markets	Financial Markets
Currency & Coins	Fiscal Service
Small Business	Office of Economic Policy
Accounting & Budget	Working Papers
Technology	Total Taxable Resources
PRESS ROOM	Terrorism and Financial Intelligence
Public Schedule	Office of Foreign Assets Control
	Executive Order 13324
	National Money Laundering Strategy
	Executive Office for Asset Forfeiture

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&desc=vacation_hawaii</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=73&desc=vacation_new_zealand</loc>
    <lastmod>2004-12-23</lastmod>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=74&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&desc=vacation_usa</loc>
    <lastmod>2004-11-23</lastmod>
  </url>
</urlset>
```

Introducing Google's Webmaster Tools

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

Google™ Webmaster Central

Welcome to your one-stop shop for comprehensive info about how Google crawls and indexes websites. You can learn here how to ensure that your site is easily crawled and indexed and access tools that will enable you to diagnose crawling issues, study statistics on how your site is doing in our index, and tell us how you'd like your site to be crawled and indexed.



[Site status wizard](#)

Find out whether your site is currently being indexed by Google.



[Google's blog for webmasters](#)

The latest news and info on how Google crawls and indexes websites.



[Webmaster tools \(including Sitemaps\)](#)

Statistics, diagnostics and management of Google's crawling and indexing of your website, including Sitemap submission and reporting.



[Google's discussion group for webmasters](#)

Talk with your fellow webmasters and share your feedback with us.



[Submit your content to Google](#)

Learn about submitting content for Google properties such as Google Base and Google Book Search.



[Webmaster help center](#)

See answers to frequently asked questions about crawling, indexing, ranking and other webmaster issues.

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

Q&A

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



PlainLanguage.gov success story

Google Web Images Video News Maps more »
over-the-counter drug label Search Advanced Search Preferences

Web

[The New Over-the-Counter Medicine Label: Take a Look](#)
All nonprescription, **over-the-counter** (OTC) medicine **labels** have detailed usage and ... The new **Drug Facts labeling** requirements do not apply to dietary ...
www.fda.gov/cder/consumerinfo/OTClabel.htm - 34k - [Cached](#) - [Similar pages](#)

[Over-the-Counter Medicines: What's Right for You?](#)
Read the **label!** **Drug labels** change as new information becomes available. That's why it's important to read ... Consumer Education: **Over-the-Counter** Medicine ...
www.fda.gov/cder/consumerinfo/WhatsRightForYou.htm - 40k - [Cached](#) - [Similar pages](#)
[[More results from www.fda.gov](#)]

[Understand Over-the-Counter Drug Labels](#)
Reading the product **label** is the most important part of taking care of yourself or your family when using **over-the-counter** (OTC) medicines (those that are ...
www.webmd.com/content/article/61/67580.htm - 40k - [Cached](#) - [Similar pages](#)

Plain Language: Over-the-counter drug label DocID: 6
Examples>[Over-the-counter drug label](#). [Over-the-counter drug label](#) Food and Drug Administration, No-Gobbledygook Award Winner. No None Version available ...
www.plainlanguage.gov/test/Examples/indexExample.cfm?record=6&search=BA - 13k - [Cached](#) - [Similar pages](#)



Plain Language.gov
Improving Communication from the Federal Government to the Public

Home What is PL? Why PL? Using PL

[Examples](#)>Over-the-counter drug label

Over-the-counter drug label
Food and Drug Administration

[Over-the-counter drug label](#) (pdf)

- 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

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 Web Images Video News Maps more »
federal funding accountability transparency act Search Advanced Search Preferences

Web

[Federal Funding Accountability and Transparency Act of 2006 ...](#)
The **Federal Funding Accountability and Transparency Act** of 2006 (S. 2590) [1] is an act that requires the full disclosure of all entities or organizations ...
en.wikipedia.org/wiki/Federal_Funding_Accountability_and_Transparency_Act_of_2006 - 43k - [Cached](#) - [Similar pages](#)

Search Results - THOMAS (Library of Congress)
Federal Funding Accountability and Transparency Act of 2006 (Introduced in Senate)[S.2590.IS] 2 . **Federal Funding Accountability and Transparency Act of ...**
thomas.loc.gov/cgi-bin/query/z?c109:S.2590: - 4k - [Cached](#) - [Similar pages](#)

The LIBRARY of CONGRESS THOMAS

The Library of Congress > THOMAS Home > Bills, Resolutions > Search Results

THIS SEARCH	THIS DOCUMENT	GO TO
Next Hit	Forward	New Bills Search
Prev Hit	Back	HomePage
Hit List	Best Sections	Help
	Contents Display	

There are 4 versions of Bill Number S.2590 for the 109th Congress

- 1 . Federal Funding Accountability and Transparency Act of 2006 (Introduced in Senate)[[S.2590.IS](#)]
- 2 . Federal Funding Accountability and Transparency Act of 2006 (Reported in Senate)[[S.2590.RS](#)]
- 3 . Federal Funding Accountability and Transparency Act of 2006 (Engrossed as Agreed to or Passed by Senate)[[S.2590.ES](#)]
- 4 . Federal Funding Accountability and Transparency Act of 2006 (Enrolled as Agreed to or Passed by Both House and Senate)[[S.2590.ENR](#)]

THIS SEARCH	THIS DOCUMENT	GO TO
Next Hit	Forward	New Bills Search
Prev Hit	Back	HomePage
Hit List	Best Sections	Help
	Contents Display	

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



OSTI success story

Google Web Images Video News Maps more »

nuclear physics giant resonances osti Search Advanced Search Preferences

Web

Energy Citations Database (ECD) - Energy and Energy-Related ...
 Subject, N68351 --Physics (Nuclear, Experimental)--Nuclear Properties ... The **giant dipole resonance** is not appreciably excited for any of the targets.(AIP) ...
www.osti.gov/energycitations/product.biblio.jsp?osti_id=4013309 - 15k -
[Cached](#) - [Similar pages](#)

Energy Citations Database (ECD) - Energy and Energy-Related ...
 Nuclear Physics Lab. Sponsoring Org, DOE/ER. Subject, 73 **NUCLEAR PHYSICS** AND ...
 tapes: **giant resonances**; nucleus-nucleus reactions; **nuclear** astrophysics; ...
www.osti.gov/energycitations/product.biblio.jsp?osti_id=6686188 - 13k -
[Cached](#) - [Similar pages](#)

[PDF] Yields of Radionuclides Created by Photonuclear Reactions on Be, C ...
 File Format: PDF/Adobe Acrobat - [View as HTML](#)
 Division of **Nuclear Physics**. Prepared by the OAK RIDGE NATIONAL LABORATORY ...
 where the (y,n) values are averages over the **giant resonances** of ...
www.ornl.gov/~webworks/cppri/2002/rpt/112299.pdf - [Similar pages](#)



Home About What's New Basic Search Advanced Search

1948 - Present

Energy Citations Database

Help Full-Text Availability Security/Disclaimers Comments

DOE Information Bridge Energy Files OSTI Home energy.gov

Availability information may be found in the Availability, Publisher, Research Organization, Resource Relation and/or Author (affiliation information) fields and/or via the "Full-text Availability" link. For a journal article, please see the Resource Relation field.

Title Giant resonances observed in the scattering of 96- and 115-MeV alpha particles

Creator/Author Youngblood, D.H. ; Moss, J.M. ; Rozsa, C.M. ; Bronson, J.D. ; Bacher, A.D. ; Brown, D.R.

Publication Date 1976 Mar 01

OSTI Identifier OSTI ID: 4013309

Other Number(s) CODEN: PRVCA

Resource Type Journal Article

Resource Relation Phys. Rev., C, v. 13, no. 3, pp. 994-1008

Research Org Cyclotron Institute and Physics Department, Texas A and M University, College Station, Texas 77843

Subject N68351 --Physics (Nuclear, Experimental)--Nuclear Properties & Reactions, 6 <= A <= 19--Nuclear Reactions & Scattering;N68451 --Physics (Nuclear, Experimental)--Nuclear Properties & Reactions, 20 <= A <= 38--Nuclear Reactions & Scattering;N68551 --Physics (Nuclear, Experimental)--Nuclear Properties & Reactions, 39 <= A <= 58--Nuclear Reactions & Scattering;N68651 --Physics (Nuclear, Experimental)--Nuclear Properties & Reactions, 59 <= A <= 89--Nuclear Reactions & Scattering;N68751 --Physics (Nuclear, Experimental)--Nuclear Properties & Reactions, 90 <= A <=

- 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

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



NCES success story

Google [Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

gardiner elem school district [Advanced Search](#) [Preferences](#)

Web

Search for Public School Districts - District Detail for Gardiner Elem
 Use the Search For Public **S**chool Districts locator to retrieve information on all US public **s**chool districts. This data is collected annually directly from ...
nces.ed.gov/ccd/districtsearch/district_detail.asp?ID2=3011820 - 41k -
[Cached](#) - [Similar pages](#)

Montana School, School System and School District Information - MT
GARDINER ELEM GARDINER, MT CITY School District and Schools Information.
 Elementary Schools, Middle Schools and High Schools in **GARDINER, MT** ...
www.schoolmatch.com/ppsi/schools/txtschmt.cfm - 94k - Nov 29, 2006 -
[Cached](#) - [Similar pages](#)



Institute of Education Sciences U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION STATISTICS

NewsFlash Staff Contact Site Index Help

Search NCES

[Publications & Products](#) [Surveys & Programs](#) [Data Tools](#) [Tables & Figures](#) [Fast Facts](#) [School, College, & Library Search](#) [Annual Reports](#) [What's New?](#) [KIDSZONE](#)

Search for Public School Districts **CCD** Common Core of Data

[Modify Search](#) [Data Notes/Grant IDs](#) [Help](#)

District Information

District Name: Gardiner Elem schools for this district	County: Park	County ID: 30067
Mailing Address: 510 Stone Street Gardiner, MT 59030	Physical Address: 510 Stone Street Gardiner, MT 59030	Phone: (406) 848-7563
NCES District ID: 3011820	State District ID: 0614	

District Details [Show Less](#)

Characteristics

Grade Span: (grades PK - 8)
 PK|KG|1|2|3|4|5|6|7|8

Type:	Regular School District
Locale/Code:	Rural, outside CBSA / 7
Status:	No Boundary Change
Metro Status:	Non MSA - Does not serve an MSA
CSA/CBSA:	00000
Supervisory Union #:	000

Total Schools:	2
Total Students:	157
Classroom Teachers (FTE):	10.7
Student/Teacher Ratio:	14.7
Summer Migrant Students:	N/A
ELL (formerly LEP) Students:	0
Students with IEPs:	20

- 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

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>

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.”

Q&A

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

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