

Google™

Making NIH websites
more accessible to
search engine users

Implementing Google Sitemaps
and Web Distribution

Agenda

Government information on the growing web

Sitemaps for search engines

Implementing Google Sitemaps

Success stories

Q&A

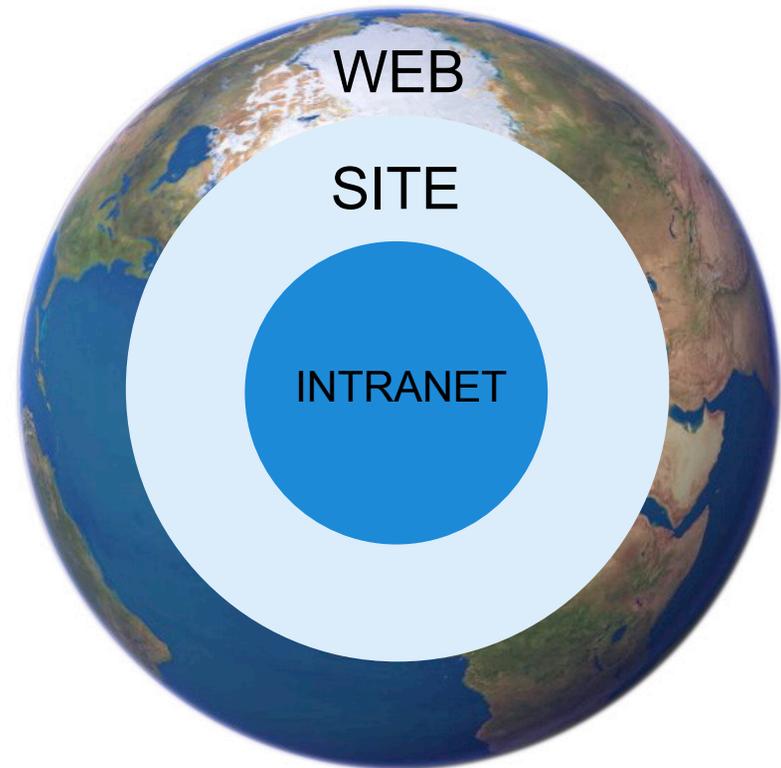
Web distribution, more Q&A

Common Concerns

- No direct cost
- Non-proprietary
- No security risk
- Public content only

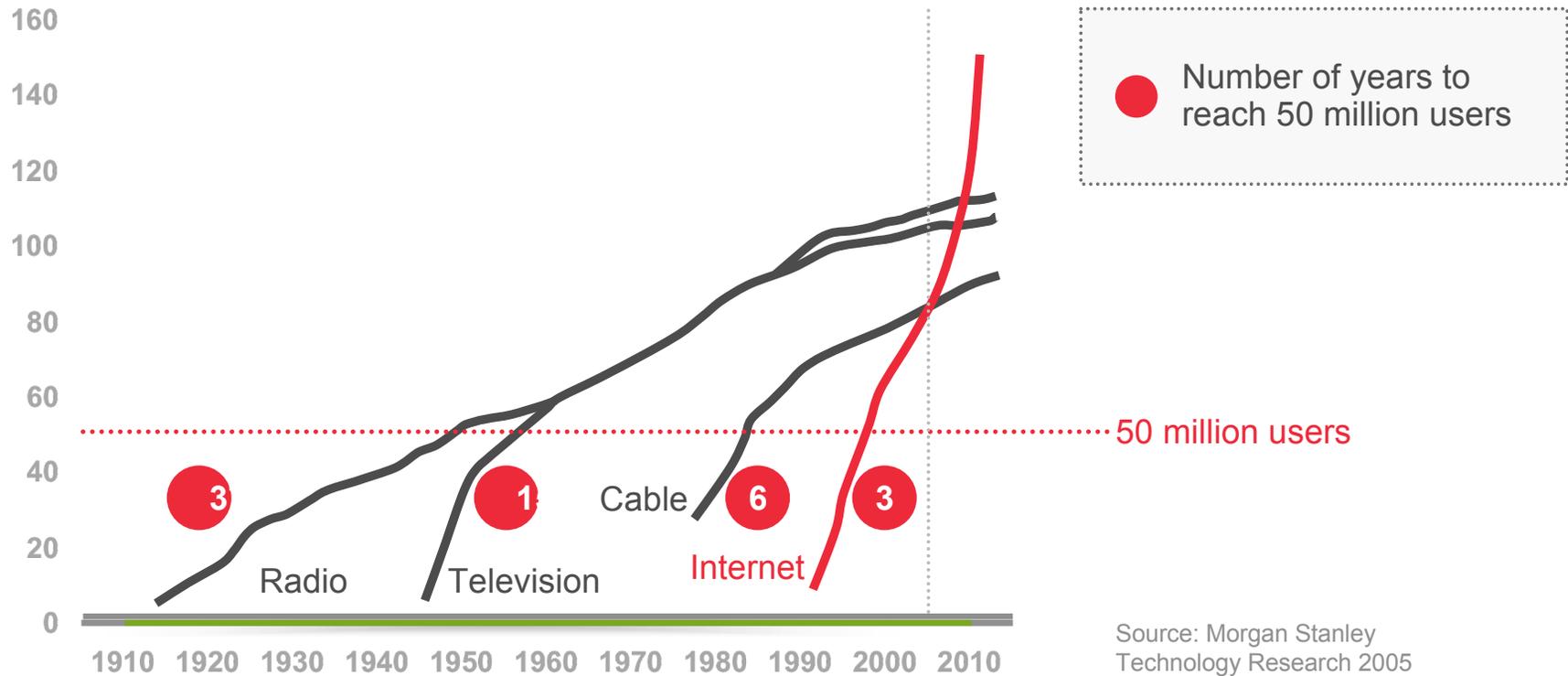
What we're talking about

- **Intranet:** Not your intranet or internal information
- **Site:** Nor search within your public site
- **Web:** We're talking about enabling discovery of your public site on the web



Internet has come of age faster than previous media

North American Users/Households (MM)



US Internet user population is diversifying

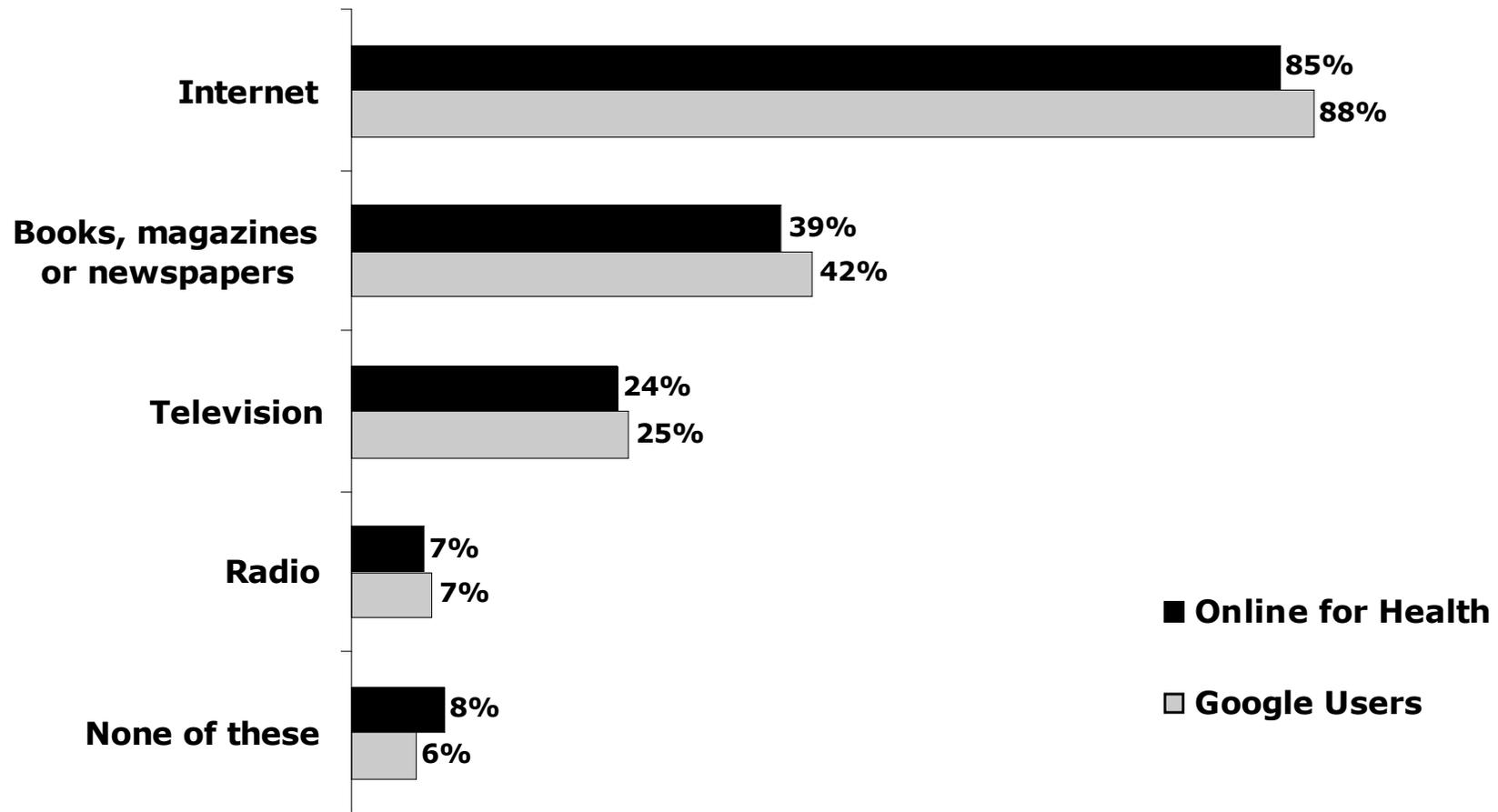
73% of adult population is online

- Not just youngsters: 71% of baby boomers (50–64)
- Not just urban and suburban: 63% of rural residents
- Not just highly educated: 84% with “some college”



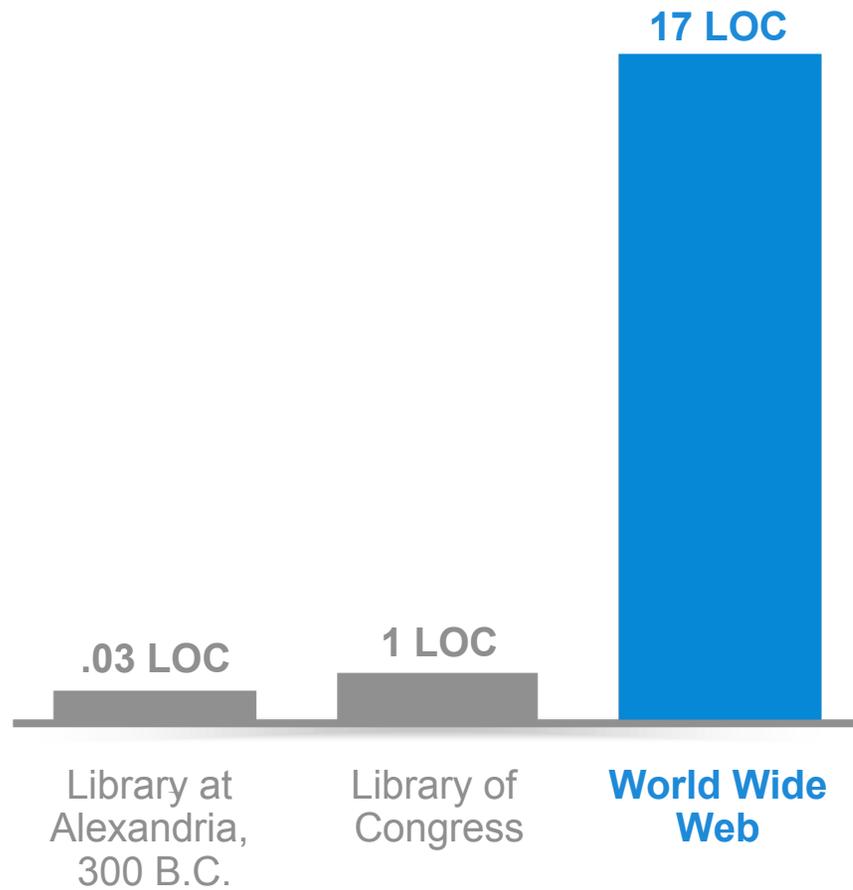
Source: Pew Internet & American Life Project (2006)

The Internet is the Leading Media Source of Health, Medical, and Prescription Drug Information



The growing web

7 million new pages every day



Source: Peter Lyman and Hal Varian (2003)



Library of Congress (LOC)
17M Books

Not all information is created equal

The value of government content – a pillar of the web



Citizens increasingly access government through search engines

National Institutes of Health (nih.gov)

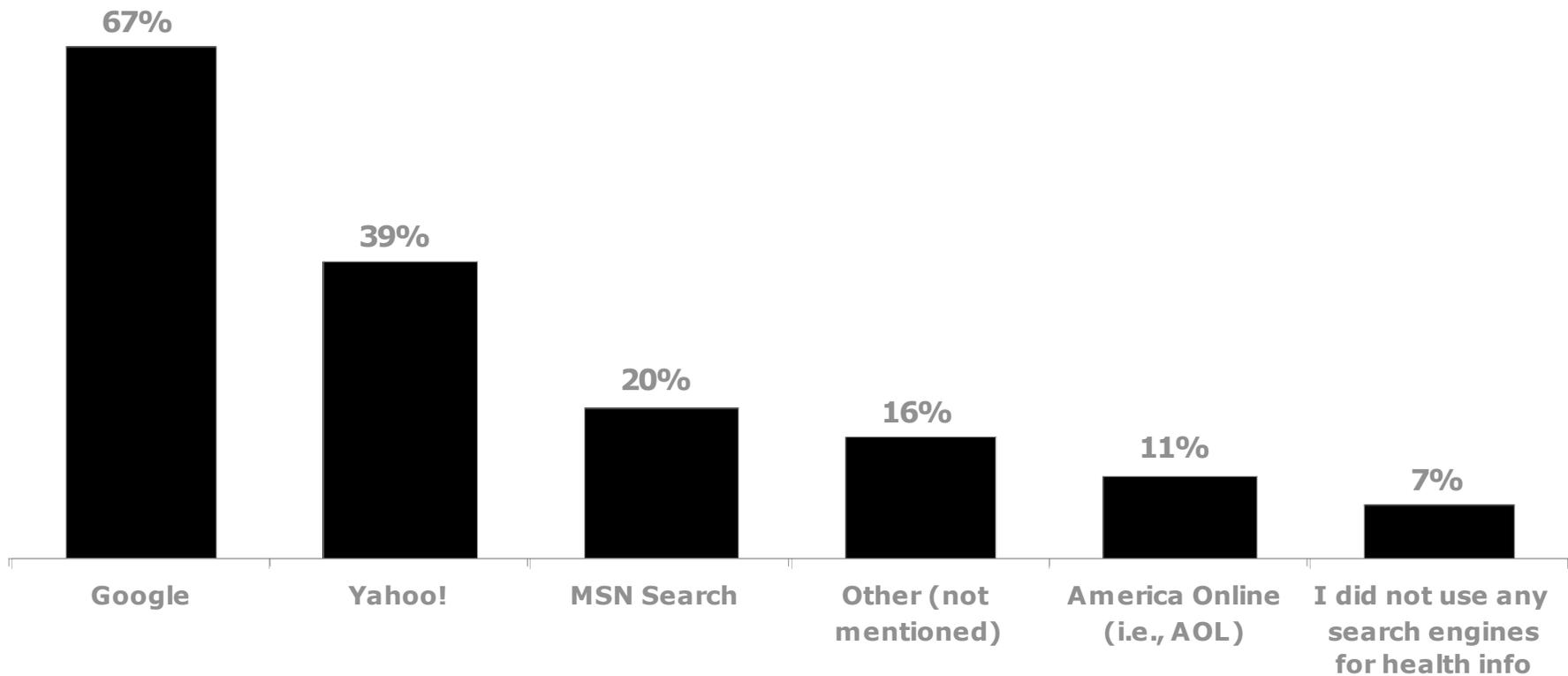
- 70% of unique users in July 2006 referred by search engines (Google, Yahoo, MSN, AOL, Ask)



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

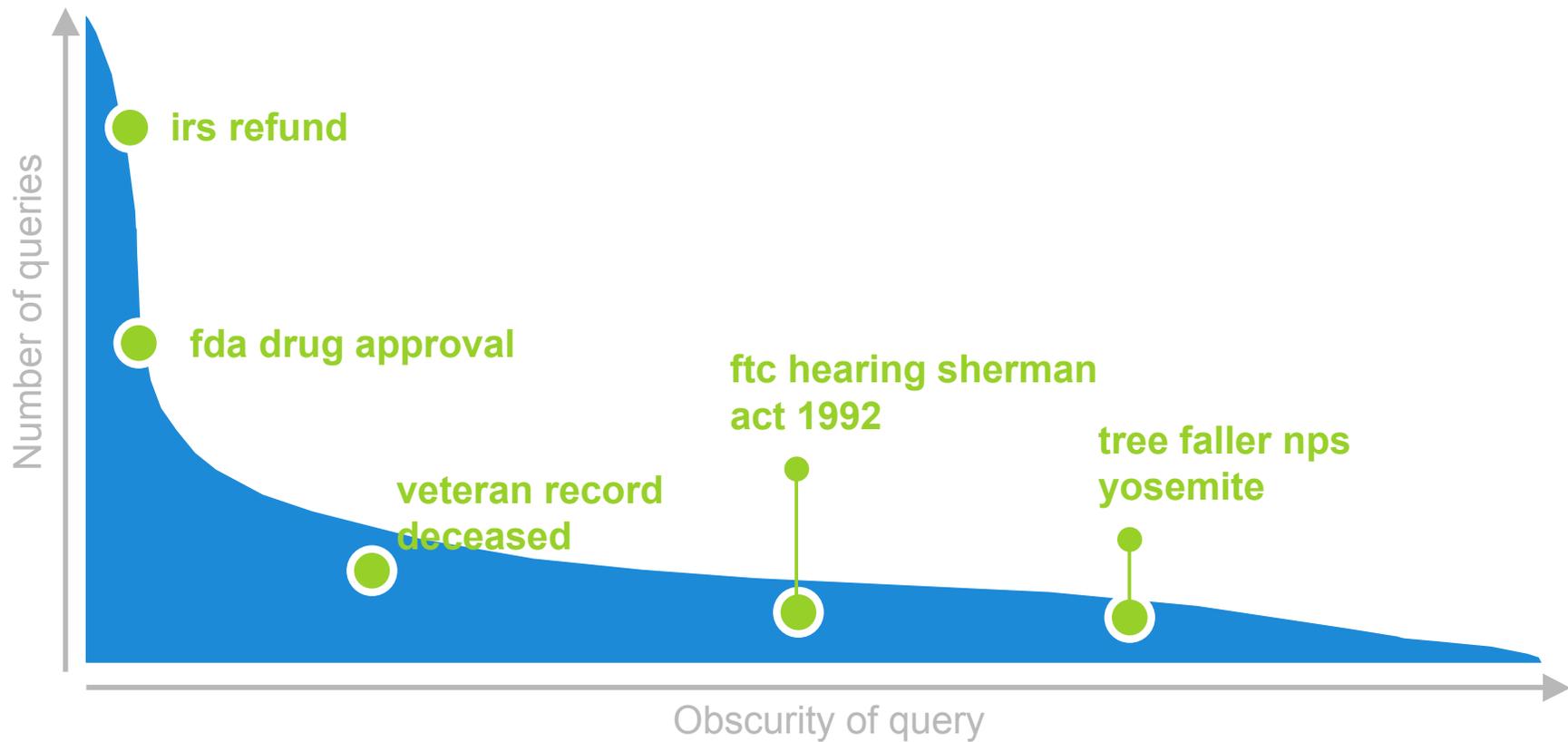
Source: ComScore Research, July 2006

Search Engines Used When Looking for Health Info Online

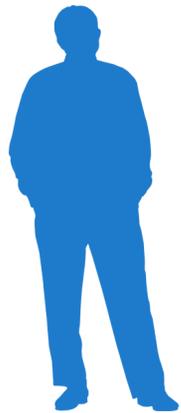


And they expect to find everything

The long tail of federal government information



Search engines are the point of departure, government sites are the destination



Federal

 **Internal Revenue Service**
DEPARTMENT OF THE TREASURY



State

 **virginia.gov**



Localities


King County


City of Dallas

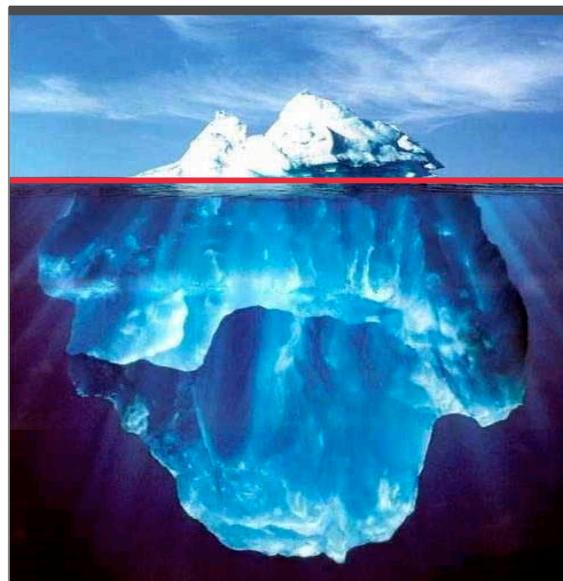
Government information on the growing web – recap

- ✓ Internet becoming dominant medium for accessing government
- ✓ Users value government information
- ✓ Users prefer to access government through search engines

The web is growing deeper and more dynamic

Challenges to web crawling or indexing are growing and multiplying

- Outdated robots.txt crawling instructions
- Non-html links
- Content “hidden” behind search forms
- Server errors (crawler times out when fetching content)
- Orphaned URLs
- Rich media: audio, video
- Paid/premium content databases



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 305

Corp No.	Status	Type	Name
37842	Inactive	Legal	SMITH, LIMITED
195668	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.
187233	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

Examples of uncrawlable NIH sites/databases

- NCI, RAID Project Information:

<http://dtp.nci.nih.gov/icssweb/raidservlet?searchtype=namesort&wheretype=all&outputformat=html>

- NHLBI, GENELINK:

<https://genelink.nhlbi.nih.gov/Linkage/studygroupinfo.jsp?studygroup=HyperGEN>

- NHGRI, Sarcoma Database:

http://watson.nhgri.nih.gov/sarcomadb/?rm=gene_details;exprs=log2ctr;geneid=1799

- NIA, Mouse Gene Index:

<http://lgsun.grc.nia.nih.gov/geneindex/mm3/bin/giU.cgi?genename=U062158>

More examples of uncrawlable NIH sites/databases

- NIDDK, Reference Collection:
<http://catalog.niddk.nih.gov/resources/detail.cfm?pubid=2171&disp=full&result=50&record=495&searchterms=insulin&databases=1&searchtype=basic>
- NIEHS, Collaborative Centers for Parkinson's Disease Environmental Research, segment blocked by robots.txt:
https://www-apps.niehs.nih.gov/spires/ccpubs/ccpder_pubs.cfm?grantno=U54ES012068
- NLM/NCBI, several databases blocked by robots.txt:
<http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=9031&CHR=1>
- NLM/NCBI, GENSAT Database, “Gensat (db of genes in mouse CNS) results use information from Gene database (which is indexed) and info from nucleotide database (which is not indexed). In addition the database contains GENSAT images which are not indexed” :
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gensat&doptcmdl=Detail&term=58992>

The solution? A sitemap

Navigational sitemap

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

Sitemaps for search engines

- HTML
- Simple text
- XML

Simple text sitemap

Simple text: a comprehensive list of URLs

<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

XML: a comprehensive list of URLs in XML

- Tagged with each URL's location, last modification, change frequency and priority

```
<?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 Sitemaps & Google Webmaster Tools

Free resources and tools for making your agency site more accessible to search engine users

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: <https://www.google.com/sitemapsgov>

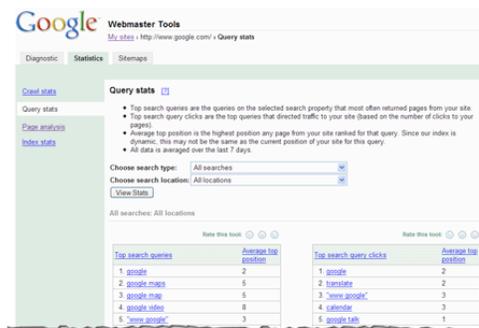
Step 1: Log into Webmaster Tools with your Google™ account

Google Webmaster Tools

Google's [webmaster tools](#) provide you with a free and easy way to make your site more Google-friendly. Using our tools, you can:

Get Google's view of your website, and diagnose potential problems.
See how Google crawls and indexes your site and learn about specific problems we're having accessing it.

See how your site is performing.
Learn which queries drive traffic to your site, and see exactly how users arrive there.



Share info with us to help us crawl your site better.
Tell us about your pages: which ones are most important to you and how often they change. You can also let us know how you would like the URLs we index to appear.

Get started today -- it's free!
Simply log in with your Google Account and [add your site URL](#) to get started. It's an easy and free way to have a more interactive experience with Google.

Sign in to Google Webmaster Tools with your Google Account

Email:

Password:

Remember me on this computer.

[I cannot access my account](#)

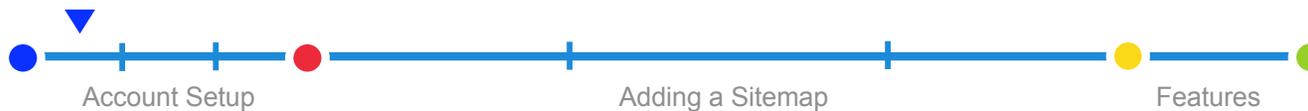
← Login

Not using Gmail or other Google Account services?

[Create a Google Account](#)

Learn more about Google webmaster tools:

- [About Google webmaster tools](#)
- [Google webmaster central](#)
- [Webmaster help center](#)
- [Google webmaster discussion group](#)



Step 2: Add a site to verify ownership

Google webmaster tools are an easy way for you to submit all your URLs to the Google index and get detailed reports about the visibility of your pages on Google. To get started, simply add the URL of your site. You'll start to see information about your site right away.

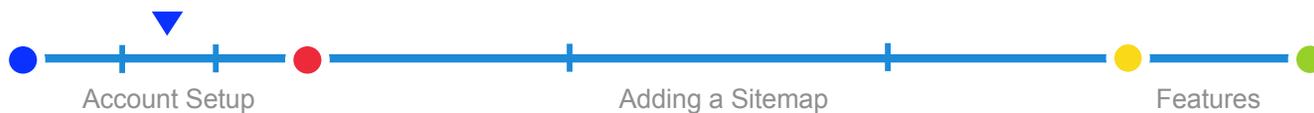
Add Site:

Example: <http://www.google.com/> [\[?\]](#)

Read more about Google webmaster tools:

[Learn more about the Google webmaster tools program](#)

[Learn more about Google Mobile Sitemaps](#)



Step 3: Verify your site

Google Webmaster Tools
mygovsite@gmail.com | My Account | Help | Sign out
My sites > http://mygovsite.googlepages.com/ > Summary

Diagnostic | Statistics | Sitemaps

Summary

Crawl errors

- Web crawl
- Mobile Web

Tools

- [robots.txt analysis](#)
- Manage site verification
- Crawl rate
- Preferred domain
- Enhanced Image Search

Summary

Rate this tool: ⓪ ⓪ ⓪

✓ Your site has been added to your account. Verify your ownership to view detailed statistics and errors for it.

⚠ No pages from your site are currently included in Google's index. Indexing can take time. You may find it helpful to review our [information for webmasters](#) and webmaster guidelines. [?]

⚠ We may not know about all the pages of your site. [Submit a Sitemap](#) to tell us more about your site.

» **Next Step**

[Verify your site](#). By verifying your site you can access comprehensive statistics and crawl errors about the pages in your site

Verification status



Verification status: NOT VERIFIED

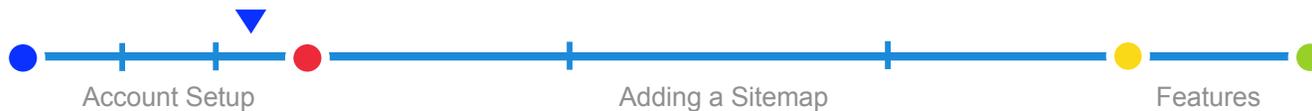
Once you verify that you're the site owner, we can provide you with comprehensive statistics and error information about the pages in your site. If you're unable to verify, you can still use the webmaster tools, submit Sitemaps, and see detailed information about those Sitemaps as well as basic information about your site. [?]

We offer two methods of verification. You can either upload an HTML file with a name we specify, or you can add a META tag to your site's index file. Choose your preferred method below. [?]

Two verification options



Choose verification method... ▾
Choose verification method...
Add a META tag
Upload an HTML file

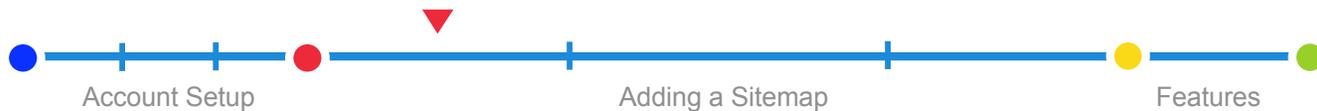


Step 4: Add a sitemap

- A. Create a sitemap with Google's Sitemap Generator or any third-party tool
- Use any available third-party tools (http://code.google.com/sm_thirdparty.html)
 - For custom, dynamic environments, you may need to develop internal scripts to generate a list of URLs

The screenshot shows the Google Webmaster Tools interface. At the top, it says "Google Webmaster Tools" and "My sites". On the right, there is a user profile for "mygovsite@gmail.com" with links for "My Account", "Help", and "Sign out". Below this, there is an "Add Site:" input field with an "OK" button and a "Tools" button. A table lists the site "http://mygovsite.googlepages.com/". The table has columns for "Site", "Sitemap", and "Site Verified?". The "Sitemap" column contains a link "Add a Sitemap", which is highlighted with a red box. A blue arrow points from the text "Add a Sitemap" below to this link. Other links in the table include "Site", "Manage", "Delete Selected", "Download this table", "Site Verified?", and "Verify".

Add a Sitemap

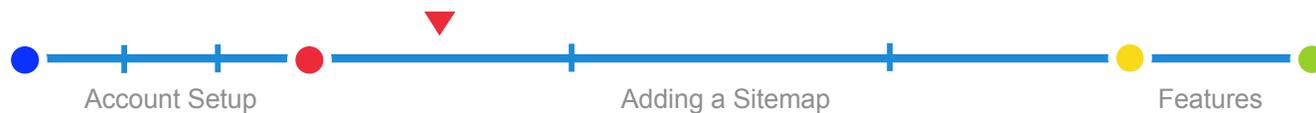


Step 4: Add a sitemap

- A. Create a sitemap with Google's Sitemap Generator or any third-party tool

Which sitemap is best for your site?

	Simple Text	XML
pros	+ Easy to create + Acceptable format	+ Provides detailed information for smarter, efficient crawling + Tags are optional
cons	- No gains in efficiency	- Entails additional steps

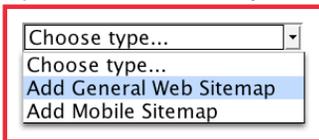


Step 4: Add a sitemap

- B. Upload the sitemap file to your site
- C. Add the sitemap URL to your account
 - Add at the highest level in your website directory structure that you want crawled
 - Review the status of the sites and sitemaps in your account

Add Sitemap

You can add a Sitemap to your account to provide us with additional information about your site. We will process your Sitemap and provide information on any errors in the Sitemaps tab. [?](#)



A screenshot of a dropdown menu with a red border. The menu is open, showing three options: "Choose type...", "Add General Web Sitemap", and "Add Mobile Sitemap". The "Add General Web Sitemap" option is highlighted in blue.

Lists pages that are meant to be accessed by desktop browsers.

1. I've created a Sitemap in a supported format. [?](#)
2. I've uploaded my Sitemap to the highest-level directory to which I have access.

Enter sitemap URL →



A screenshot of a text input field with a red border. The text "3. My Sitemap URL is:" is above the field. Below the field is an example URL: "Example: http://mygovsite.googlepages.com/sitemap.xml".

Add Web Sitemap



Features: Query Stats identifies popular queries

- See your **top 20** search queries and search query clicks
- **Top position** shows you where your pages were listed per search query
- Easily export a report with **CSV** download feature

Search queries = impressions in search results

Query stats

- Top search queries are the queries on the selected search property that most often returned pages from your site.
- Top search query clicks are the top queries that directed traffic to your site (based on the number of clicks to your pages).
- Average top position is the highest position any page from your site ranked for that query. Since our index is dynamic, this may not be the same as the current position of your site for this query.

All data is averaged over the last three weeks.

Choose search type: Web Search

Choose search location: (France) google.fr

View Stats

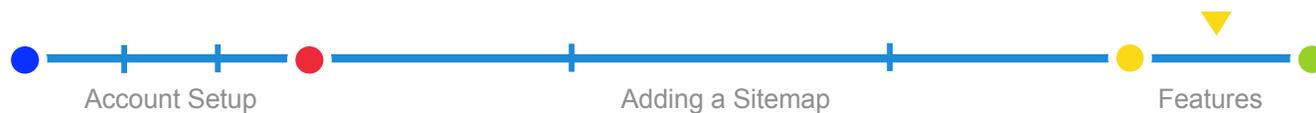
Web Search: (France) google.fr

Top search queries	Average top position
1. google	3
2. sexyloo	10
3. jacquieetmichel	8
4. chat nrj	9
5. google map	7
6. jacquie et michel	8
7. google maps	5
8. france examen	10

Top search query clicks	Average top position
1. google	3
2. translate	4
3. google arabe	2
4. google calendar	2
5. google translate	2
6. google talk	3
7. google analytics	2
8. google usa	2

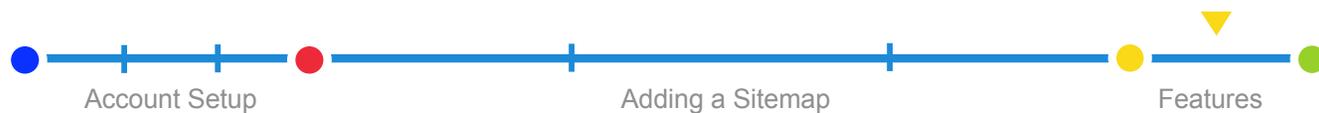
Position per query in results

Query clicks = traffic



Overview of features: More resources and tools

- Crawl Errors → shows you which pages were problematic
- Query Stats → shows queries that drive traffic to your site
- Diagnostic → tab reports help you troubleshoot crawl errors
- Robots.txt → helps to improve your coverage
- Page Analysis → shows how Google sees your pages
- Index Stats → shows how your pages are indexed



Success stories

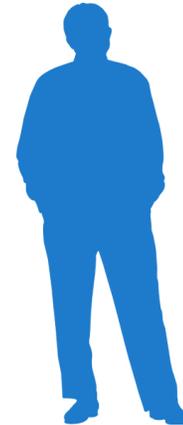
- A small federal agency that provides directory and statistical content resources
- A federal institution that maintains one of the world's largest networks of sites, consisting of content in all formats and many, many databases
- A federal department that centrally manages hundreds of agency sites marked by routine content updates

Some questions to consider

- Publishing system:
 - What database applications (Oracle, SQL Server, flat files, etc) do you operate?
- System management:
 - Can you download and install third-party tools on your web server?
- URL structure:
 - Can you list and explain a few combinations of how your site URL is constructed?

Common concerns about Google Sitemaps, revisited

- **No direct cost:** Generally only a time investment
- **Non-proprietary:** Google Sitemaps protocol based on open standard: <https://www.google.com/webmasters/sitemaps/docs/en/protocol.html>
- **No security risk:** Implementation occurs on public side of site
- **Public content only:** Limited to content on your public site



Making your site more accessible to search engine users

- Implementing Google Sitemaps can **enhance**, but **does not replace**, the Google crawl
- It does not guarantee inclusion, but **helps** to deliver more information and **fresher results** to users

- **It's free** and can be easy to implement
- Makes Google crawling **more efficient**, reducing demands on servers
- Reporting tools **uncover and pinpoint** technical errors

- Ensures **all** your public information is discoverable by **all** potential users
- **Accelerates** the inclusion of new information in search results
- Sitemap protocol released as an **open standard**

The virtuous circle

Google Benefits

- **Expands** the reach of Google's search services
- Incorporates more **authoritative, trustworthy, fresh** content
- Increases crawling **efficiency**

User Benefits

- **More and better** information from a trusted source
- **Quicker** access to and navigation of government information and services
- Enables **serendipitous discovery**

Next steps, Q&A

- To get started, and for access to the list of NIH sites/databases to sitemap, sign up at www.google.com/sitemapsgov
- Verify your site
- Generate and upload your sitemap
- Begin tracking your progress

Sitemaps sign up:

<http://www.google.com/sitemapsgov>

Webmaster Central:

<http://www.google.com/webmasters>

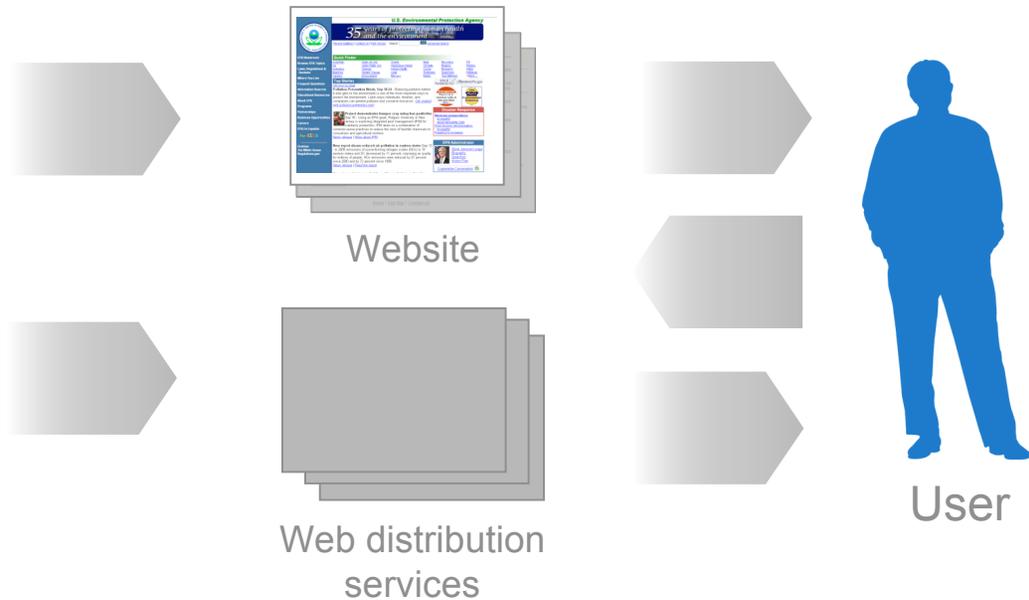
Contact us at:

sitemap-partners@google.com

Think web distribution, not just website

Content formats

- Audio
- Data (geospatial, statistical)
- Images
- Subject knowledge
- Text
- Video



Among the web distribution vehicles

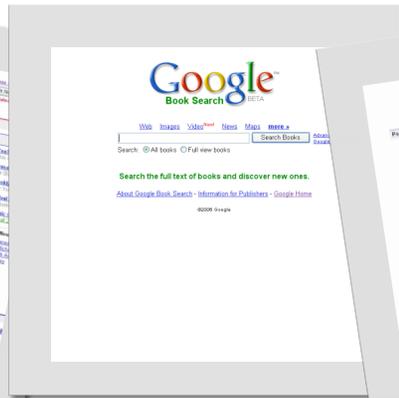
Mobile



News



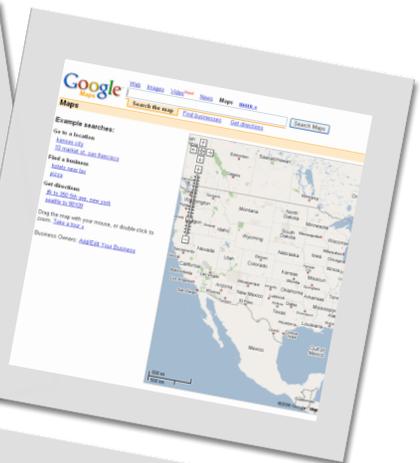
Books



Video



Maps



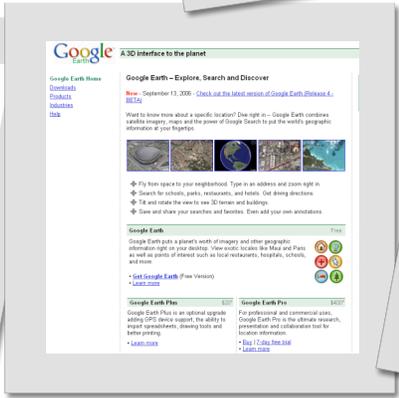
Base



Scholar



Earth



Co-op



Why web distribution?

- Can be much more efficient than seeking to serve as sole source for agency content
- The content is more likely to reach users in the context where they seek it or can use it
- It helps ensure the message (alert or report) is received without mediation
- In the case of Google, a link to the source publisher is always prominent – can significantly increase traffic to the agency site

Some examples

- Google Co-op
- Google Base
- Google Images
- Google Mobile
- Google Scholar

Next steps, Q&A

- Google Sitemaps: Sign up at www.google.com/sitemapsgov
- Google web distribution services:
John Lewis (JL) Needham, jlneedham@google.com