# Making California government sites more accessible to search engine users

## Implementing the Sitemap protocol

# Why we're talking today

- Sample search 1: ["richard abbott" california accountant license]

  (http://www.google.com/search?hl=en&lr=&safe=off&q=%22richard+abbott%22+california+accountant+license&btnG=Search)

  – What a search engine user does not find:

  http://www2.dca.ca.gov/pls/wllpub/WLLQRYNA$LCEV2.QueryView?P_LICENSE_NUMBER=2543&P_LTE_ID=781

  – Reason: Database uncrawlable

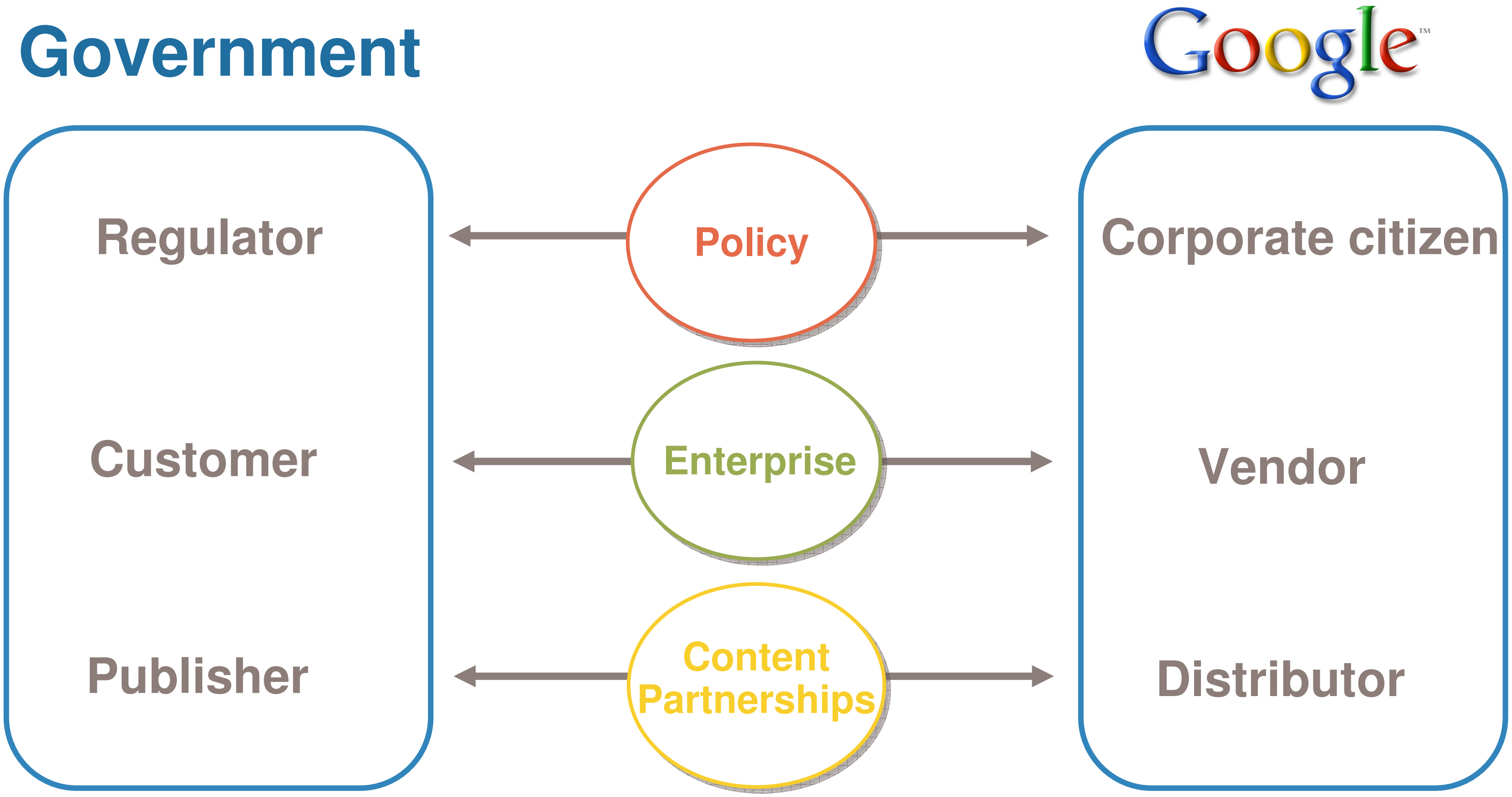
# Why we're talking today

- Sample search 2: [california real estate appraisal licensing courses] (http://www.google.com/search?hl=en&lr=&safe=off&q=california+real+estate+appraisal+licensing+courses+&btnG=Search)
  - What a search engine user does not find: http://secure.dre.ca.gov/publicasp/CEStatutory.asp
  - Reason: Database uncrawlable

# Why we're talking today

- Sample search 3: [cdec station search sensor type] (http://www.google.com/search?hl=en&q=cdec+station+search+sensor+type&btnG=Google+Search)
  - What a search engine user does not find: http://cdec.water.ca.gov/cgi-progs/staSearch; http://cdec.water.ca.gov/robots.txt
  - Reason: Site blocked by robots.txt (database itself also uncrawled)

# Government's relationships with Google



# Agenda

Government information on the growing web

Sitemaps for search engines

Implementing the Sitemap protocol

Google Webmaster Tools

Success stories

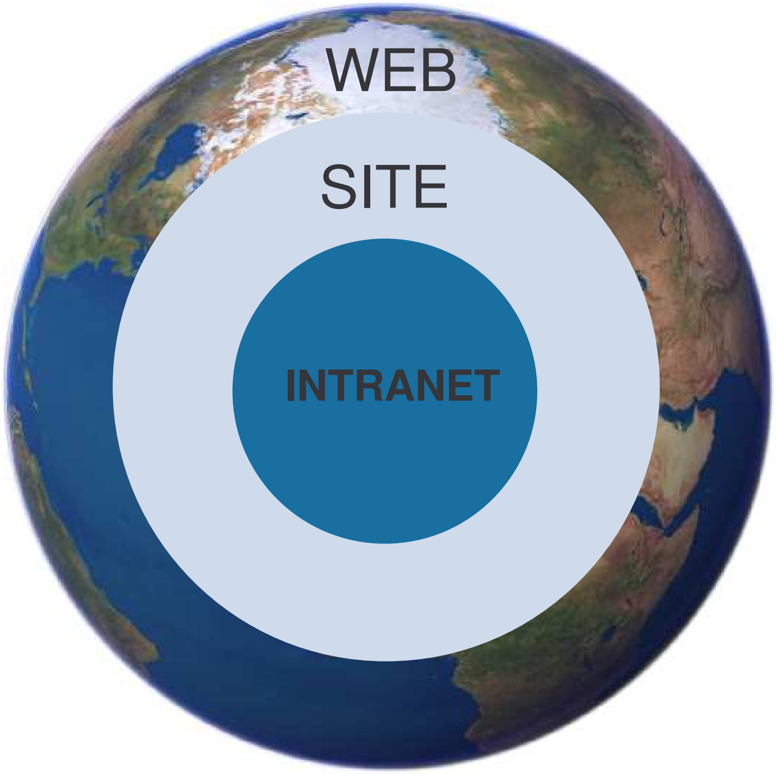
Q&A

# 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: Making content on your site accessible to web search engine users

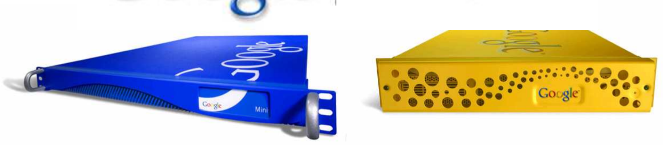


# Web search vs. site search

## Supporting the two levels of search

Google

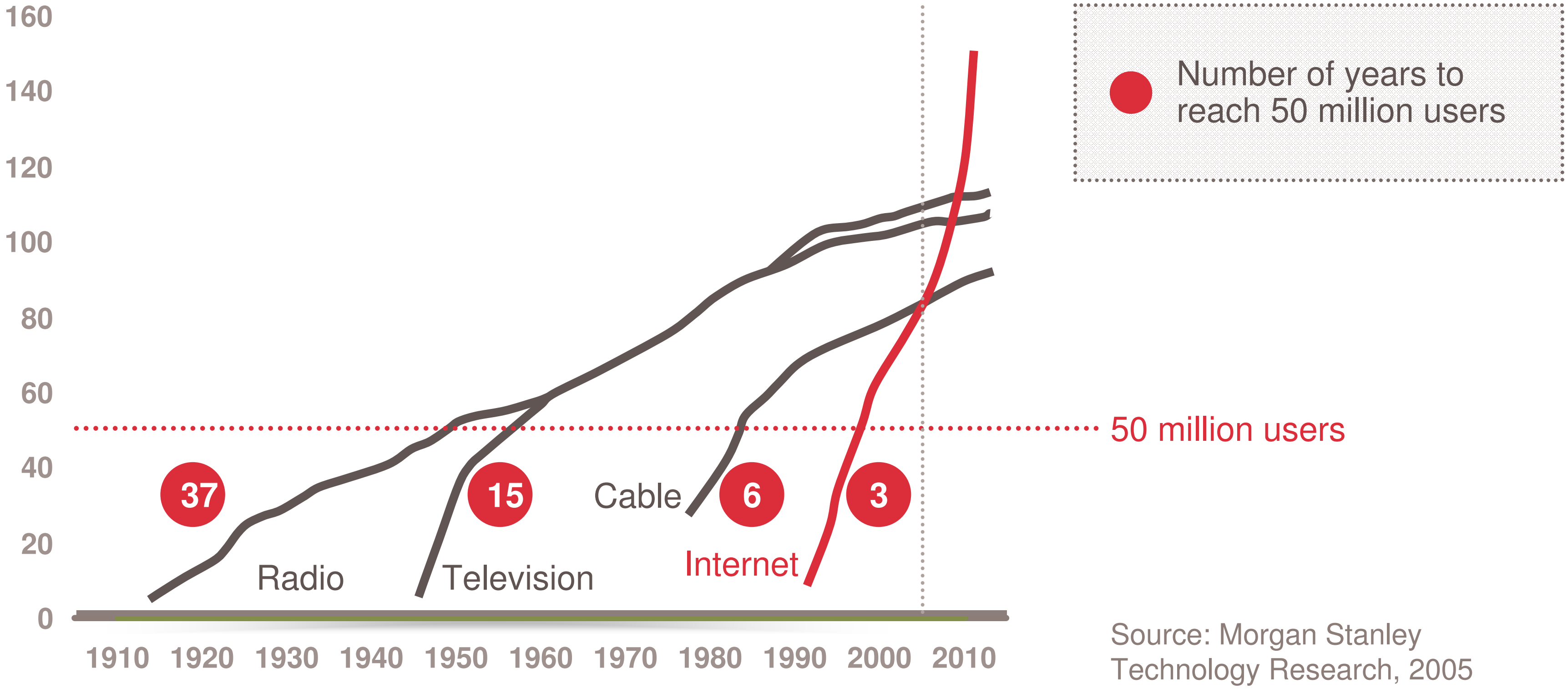
Google Enterprise



| | | |
|---|---|---|
| All of the open and accessible deep web | **Search scope** | A segment of your public site's content |
| Citizens and professionals | **User** | Professionals and citizens |
| Googlebot's 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 |

# The internet has come of age faster than previous media

## Internet adoption by North American users/households



Source: Morgan Stanley Technology Research, 2005

# US Internet user population is diversifying

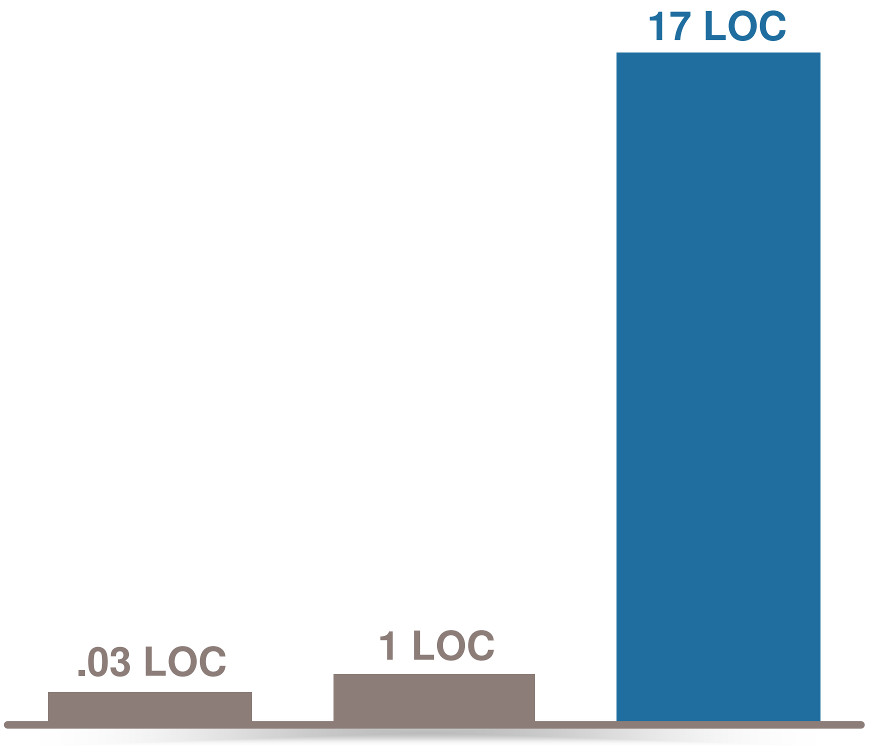
73% of adult population is online

- 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 growing web

## 7 million new pages every day



Library at Alexandria, 300 B.C.

Library of Congress

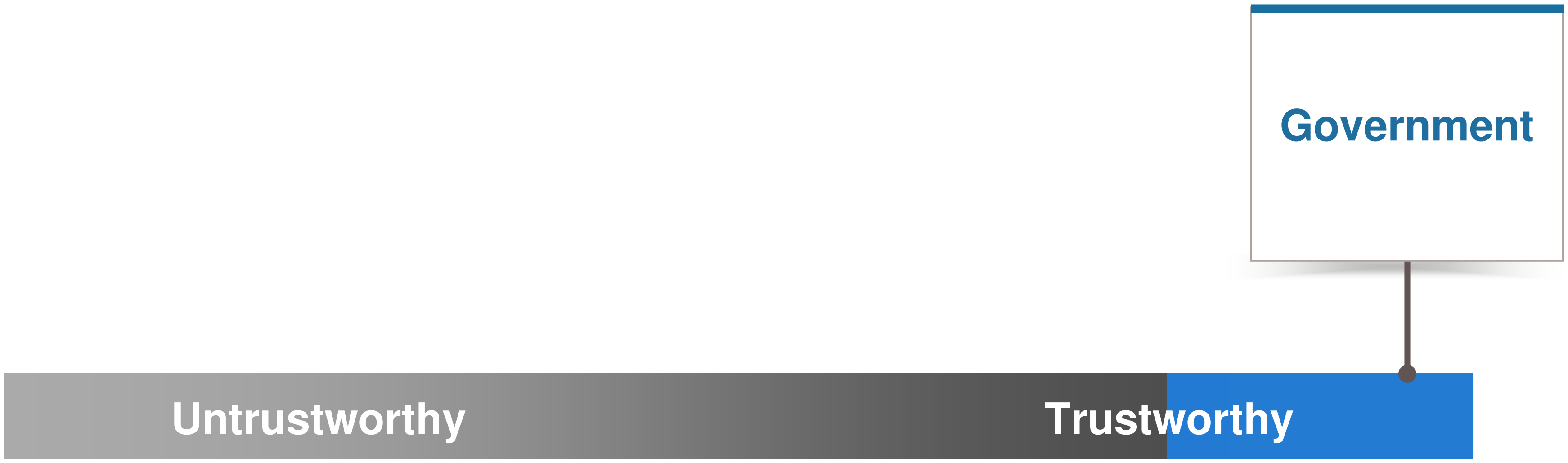
**World Wide Web**

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



Government

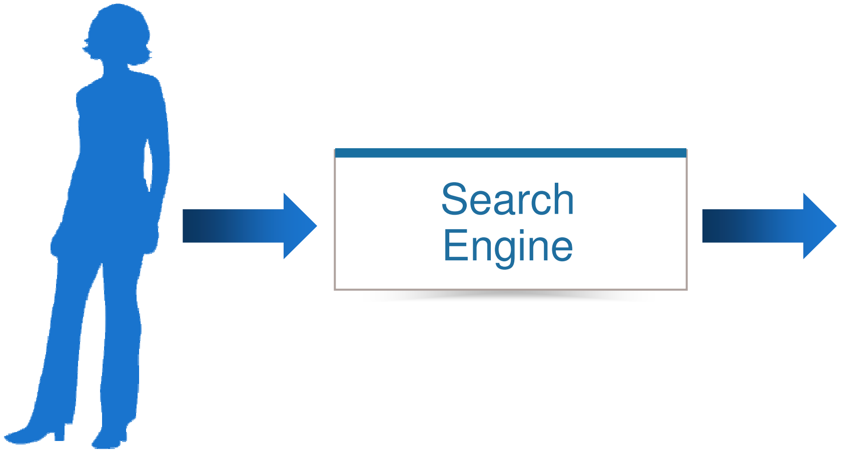
Untrustworthy

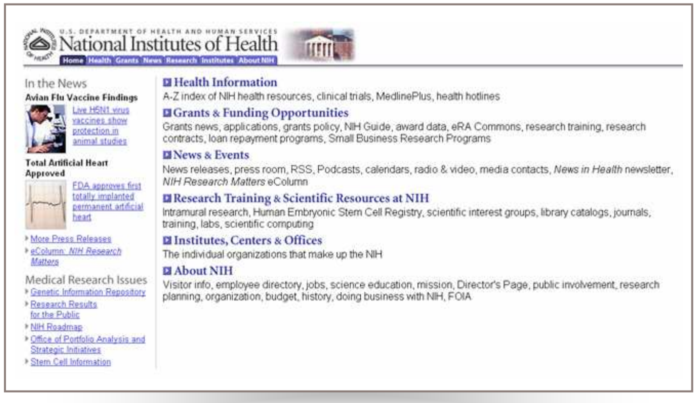
Trustworthy

# Citizens increasingly access government through search engines

## National Institutes of Health (nih.gov)

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





U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

In the News

**Avian Flu Vaccine Findings**

Live H5N1 virus vaccines show protection in animal studies

**Total Artificial Heart Approved**

FDA approves first totally implanted permanent artificial heart

- More Press Releases
- eColumn: *NIH Research Matters*

Medical Research Issues

- Genetic Information Repository
- Research Results for the Public
- NIH Roadmap
- Office of Portfolio Analysis and Strategic Initiatives
- Stem Cell Information

**Health Information**
A-Z index of NIH health resources, clinical trials, MedlinePlus, health hotlines

**Grants & Funding Opportunities**
Grants news, applications, grants policy, NIH Guide, award data, eRA Commons, research training, research contracts, loan repayment programs, Small Business Research Programs

**News & Events**
News releases, press room, RSS, Podcasts, calendars, radio & video, media contacts, *News in Health* newsletter, *NIH Research Matters* eColumn

**Research Training & Scientific Resources at NIH**
Intramural research, Human Embryonic Stem Cell Registry, scientific interest groups, library catalogs, journals, training, labs, scientific computing

**Institutes, Centers & Offices**
The individual organizations that make up the NIH

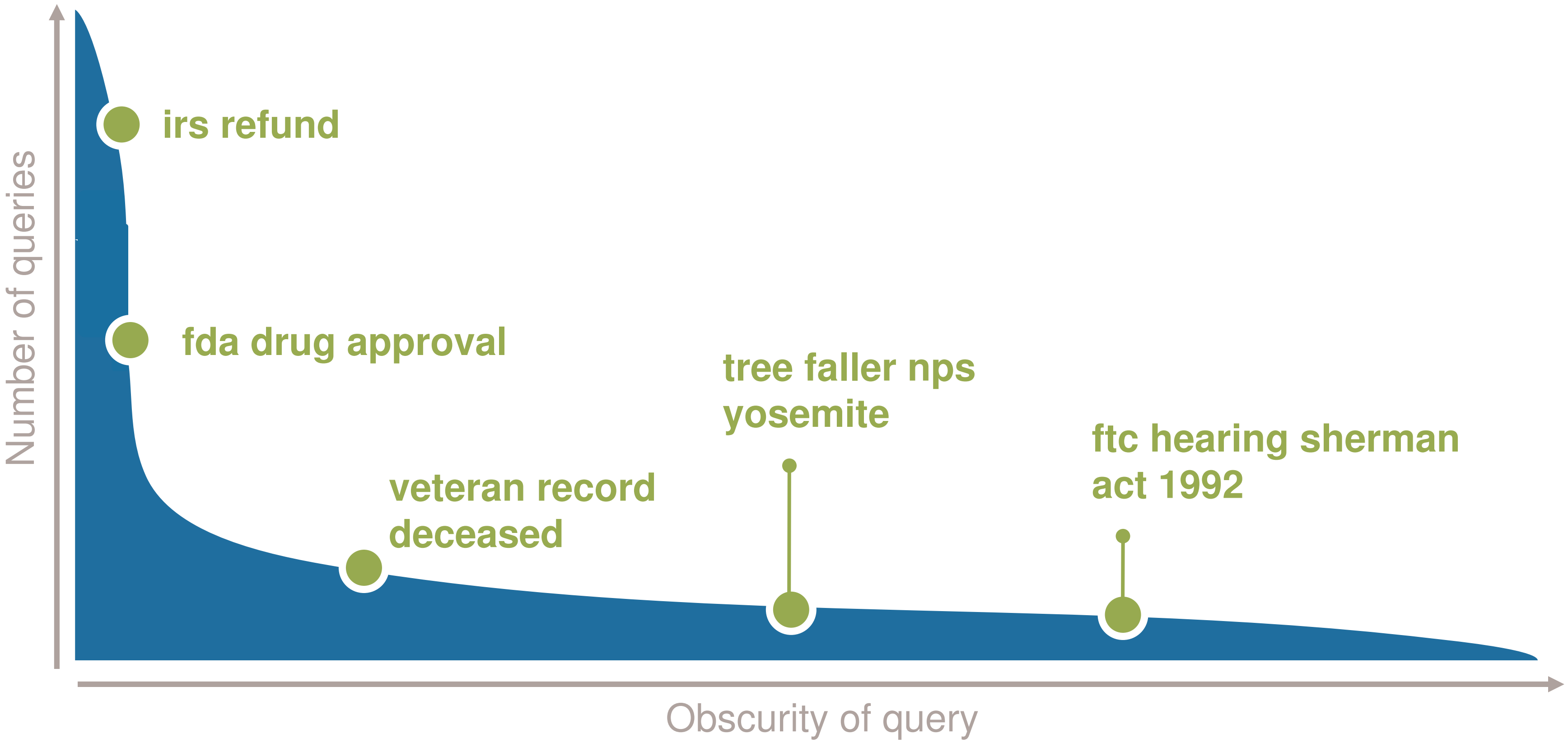
**About NIH**
Visitor info, employee directory, jobs, science education, mission, Director's Page, public involvement, research planning, organization, budget, history, doing business with NIH, FOIA

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

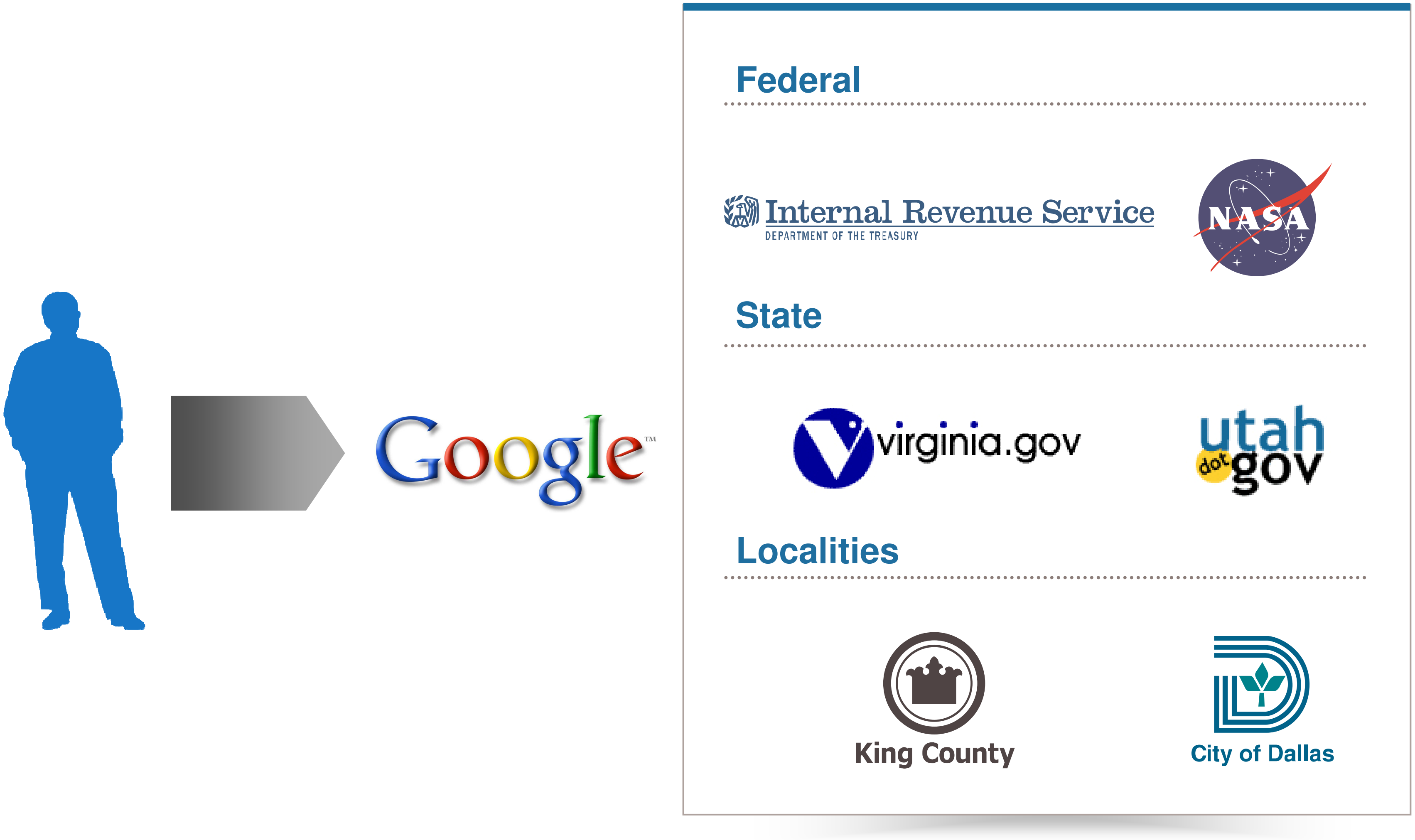
Source: ComScore, 2006

# 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

State

Localities

# Government information on the growing web – recap

✓ Internet becoming dominant medium for accessing government

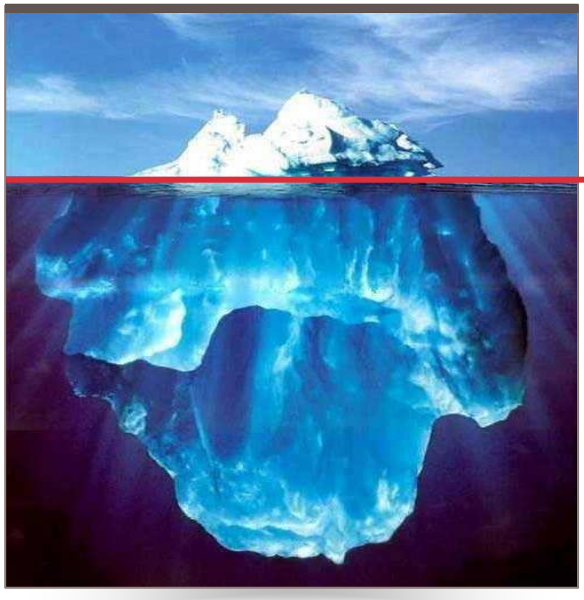
✓ Users value government information

✓ Users prefer to access government through search engines

# Growing deeper and more dynamic

## Challenges to web crawling are growing and multiplying

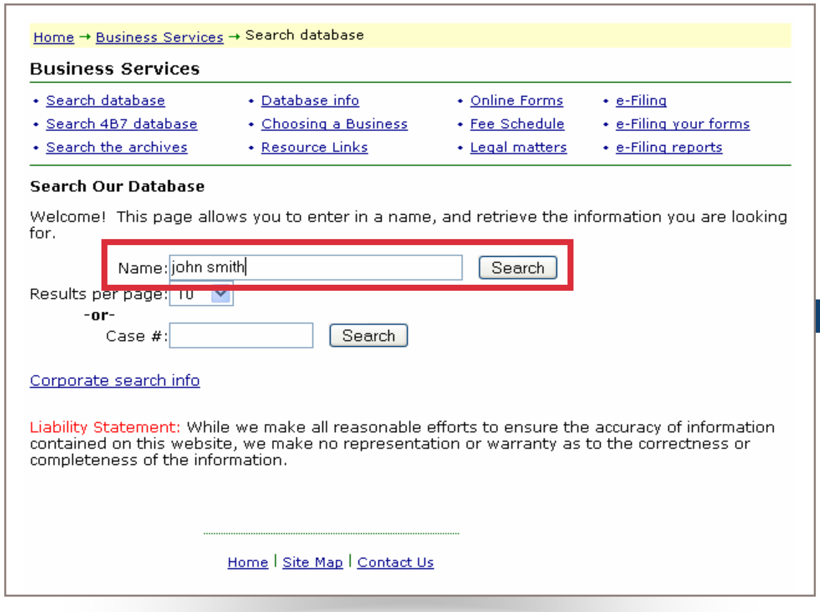
- Outdated robots.txt crawling instructions
- Non-html links
- Content “hidden” behind search forms
- Server errors (crawler times out when fetching content)
- Orphaned URLs
- Rich media: audio, video
- Paid/premium content databases



**WEB**
Searchable

**DEEP WEB**
Not searchable

# Crawlers cannot navigate search forms



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: john smith Search

Results per page: 10

-or-

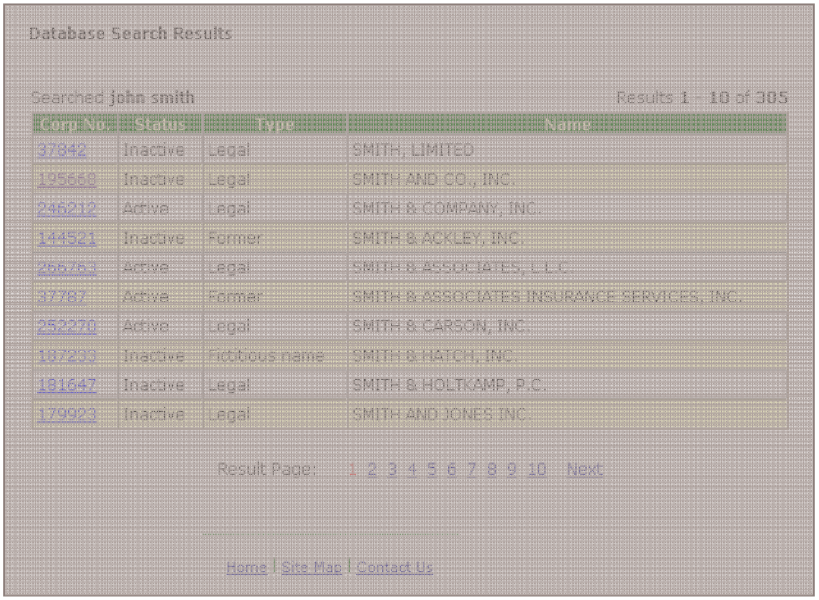
Case #: Search

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

## When crawled



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**

# The solution: Sitemaps

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

Google™ YAHOO!® Microsoft®

“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

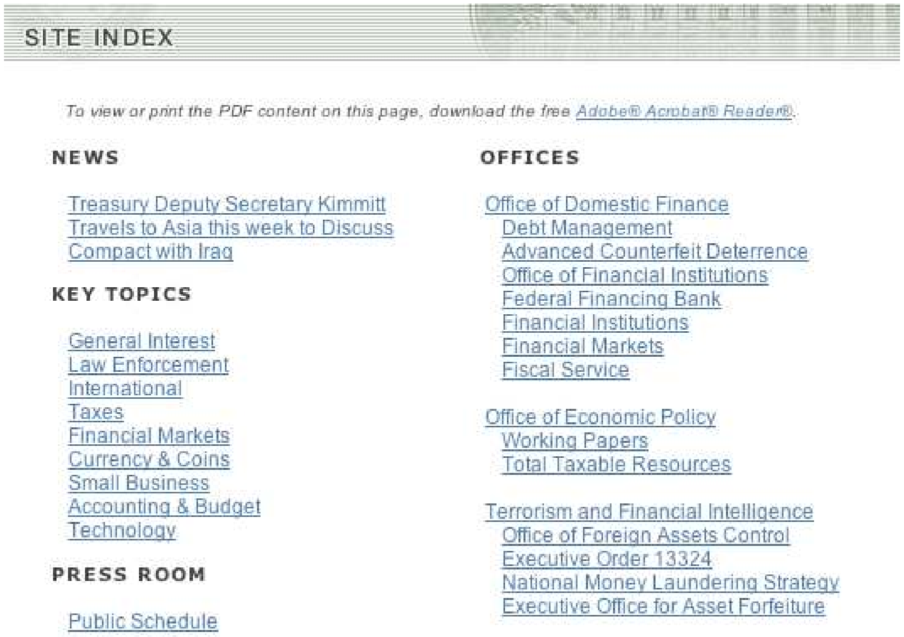
“We are 100% behind this protocol - this kind of collaboration will help improve the search experience for all of our customers”

—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

# 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**

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

# Sitemaps for search engines

- HTML
- Simple text
- XML

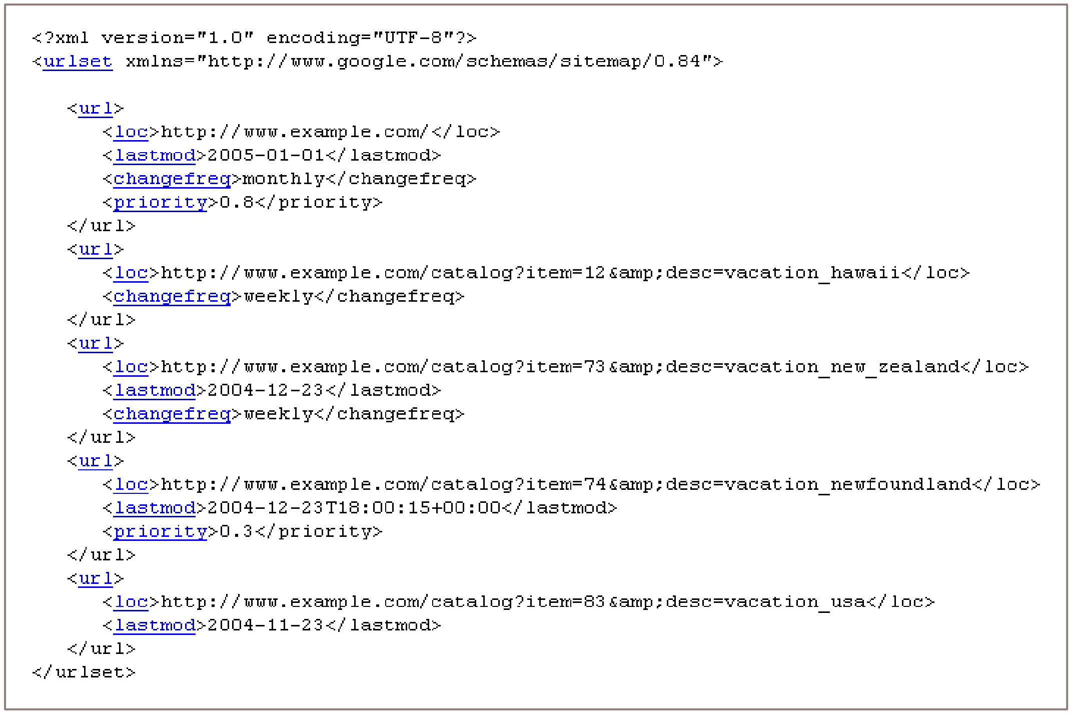
# Simple text sitemap

## 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

- 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&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>
```