Statement of Ari Schwartz
Deputy Director
Center for Democracy & Technology
before the
Committee on Homeland Security and Governmental Affairs
on E-Government

December 11, 2007

Chairman Lieberman, Ranking Member Collins, and members of the Committee, thank you for holding this hearing on E-Government. I am Ari Schwartz, Deputy Director for the Center for Democracy & Technology (CDT).

CDT is a non-profit public interest organization founded in 1994 to promote democratic values and individual liberties for the digital age. CDT works for practical, real-world solutions that enhance free expression, privacy, universal access and democratic participation. We are guided by our vision of the Internet as a uniquely open, global, decentralized and user-controlled medium. We believe the Internet has unprecedented potential to strengthen democracy and encourage citizen participation by placing powerful information and communications technology in the hands of individuals and communities.

The Role of the E-Government Act of 2002

For five years, the E-Government Act has promoted improvements in the federal government's use of information technology, and has resulted in government information resources being more readily available to the public. The law showed great foresight in focusing on issues such as accessibility of information, privacy and security, which remain of central concern to the public.

The principal role of the E-Government Act has been to promote best practices among agencies in important areas and to solidify the federal government's technology management structure. Unquestionably, the E-Government Act has changed the way that the public interacts with the government for the better. For instance, a citizen can look up pending regulations, corporate filings, and search the federal agency websites through the USA.gov portal, regulations.gov and other appropriate sites.

We have also learned a great deal from agency implementation of the law about what areas can be improved. Five years of experience, technological progress, and changes in user expectations should guide revisions to the E-Government Act to facilitate availability of public resources to the public, and privacy protections for new technologies.

Making Government Information Searchable

The Pew Internet Project has found that commercial search engines are the most popular means to find government information.[1] This is true for several reasons. First, citizens don't necessarily know which agency holds the information they seek, but they often know how to search for it. Also, commercial search engines have simply become the most efficient and effective route to find information online. Government agencies must recognize that taxpayers will not find the information that is made available unless this information can be found on commercial search engines. Some agencies have public information resources that are not immediately accessible via search engines due to relatively minor technical problems that the agencies should quickly remedy.

Today, the Center for Democracy & Technology and OMB Watch are releasing a report demonstrating the types of government information that are not available through search engines and why. The full report is attached as an Appendix to this testimony, but I will offer a quick summary of the most important points.

In order to find online information, commercial search engines continually index the Internet via simple programs called crawlers. These crawlers face certain technological limitations that often prevent them from indexing information. Luckily, there are relatively simple ways to make information more accessible to search engines, and help government sites ensure that the most relevant information is provided to the public. Two easy ways to ensure that government information is indexed are to adopt the Sitemaps protocol, which guides search engines to content, and to limit the use of robots.txt files, which ask search engines not to crawl certain content. Unfortunately, CDT and OMB Watch found many important federal government agencies offering information and services that were not being indexed for search because they did not use these protocols well. Select examples of information that cannot be fully found by citizens using commercial search engines include:

- Federal Emergency Management Agency databases: including Flood Map Modernization project at FEMA, which maps out flood hazards.
- Other Department of Homeland Security databases: including topics like environmental radiation monitoring.
- FedBizOpps.gov database: listing approximately 200 government business opportunities within the field of telecommunications.
- Central Contractor Registration database: listing who does business and receives moneys from the federal government
- Federal Procurement Data Services database: includes data on all government contracts, including all telecommunications contracts.
- Smithsonian Institute resources: including many online content collections, including the Smithsonian Institution Research Information System.
- National Oceanic and Atmospheric Administration databases: including databases used to monitor environmental data and research.

---

[1] John B. Horrigan, "How Americans Get in Touch with Government" Pew Internet Life, May 24, 2004 — http://www.pewInternet.org/pdfs/PIP_E-Gov_Report_0504.pdf

- Bureau of Labor Statistics databases: including many of the statistics and collections of information hosted on the BLS site.

It is unclear to CDT and OMB Watch whether these agencies know that their information is not publicly searchable and have not taken the adequate steps to change their practices or whether the agencies simply do not know that this important information is not being crawled. In either case, our findings show that this is a systematic problem that should be addressed as soon as possible.

It should also be noted that even the government's own search engine is directly impacted by this problem. The USA.gov site utilizes Microsoft Live Search to run its search capability. Therefore, it is subject to exactly the same inability to search these important sites as other commercial search engines.

Fortunately, the E-Government Act recognized the importance of the availability and accessibility of information. Section 207 of the Act was meant to improve the organization and categorization of government information. OMB was directed to require that agencies proactively improve access to government information and services. As President Bush said in his signing statement for the E-Government Act, "[t]he Act will also assist in expanding the use of the Internet and computer resources in order to deliver Government services, [...] for a citizen-centered, results-oriented, and market-based Government."[2] Recently, by passing the searchability provision of the Reauthorization of the E-Government Act, this Committee helped to ensure that this provision was modernized to include promote best practices that could be used to tackle this problem.

We urge the Committee to work with us to encourage agencies that have not made public information available to search engines to do so immediately and to oversee proper implementation of the search provisions of the Reauthorization Act to ensure prompt compliance.

Privacy Impact Assessments

The increased ability to find information brings with it the challenge to better manage, protect and secure the personal information of individuals held by government that could inadvertently be made public if proper steps are not taken. Congress clearly understood this concern when it passed the E-Government Act. Section 208 of the Act was specifically designed to "ensure sufficient protections for the privacy of personal information as agencies implement citizen-centered electronic Government."[3] The method to achieve this goal was to increase transparency about how the government collects, manages and uses personal information about individuals through Web privacy notices and privacy impact assessments (PIAs).

---

[2] http://www.whitehouse.gov/omb/egov/g-3-statement.html
[3] PL 107-347, Section 208

The E-Government Act required that agencies perform PIAs before adopting new technology or using collections of personally identifiable information. These PIAs are public documents, containing a description of the project, a risk assessment and a discussion of potential threats to privacy, and ways to mitigate those risks. PIAs ensure that universal privacy concerns are considered as part of these decisions, and that the public has access to this element of the decision making process.

Over the past five years, PIAs have become an essential tool to help protect privacy. They are sometimes called "one of the three pillars" of the US government privacy policy.[4] Unfortunately, as with the other privacy laws, the federal government has unevenly implemented even the basic transparency requirement of PIAs across agencies.

The guidance issued by OMB pursuant to the Act with respect to PIAs was vague and has simply not provided agencies with the tools they need to successfully implement the PIA requirement unless they already had privacy experts on staff. While some agencies, like the Department of Homeland Security (DHS),[5] have set a high standard for PIAs and have continued to improve them over time, the lack of clear guidance has led some agencies to create cursory PIAs or none at all. For example, despite the major privacy implications of the use of RFID in passports, the US Department of State gave the issue only cursory consideration in its PIA, a document of only ten sentences.[6] Even more troubling is the finding that some agencies simply do not perform PIAs on as many as half their qualifying technologies.[7] Other agencies, even those that prepare in depth PIAs, too often complete them after a project has been developed and approved. PIAs are supposed to inform the decisionmaking process, not ratify it.

The inconsistent implementation of PIAs should be of great concern to this committee. The work of the agencies that have taken the mandate to develop PIAs seriously and used them as a tool for analysis and change should be used as a starting point for developing best practices for all federal agencies. CDT hopes that the provision included in the E-Government Act Reauthorization bill that passed out of this committee last month that

---

[4] DHS Chief Privacy Officer Hugo Teuffel, *Presentation before the European Commission's Conference on Public Security, Privacy and Technology*, November 20, 2007 Brussels, Belgium. Mr. Teuffel suggested that the three current pillars are the Privacy Act of 1974, Section 208 of the E-Government Act and the Freedom of Information Act.

[5] The DHS Website on Privacy Impact Assessment offers a range of resources to DHS components and to other agencies —
 http://www.dhs.gov/xinfoshare/publications/editorial_0511.shtm

[6] http://foia.state.gov/SPIAS/20061.DOS.PIA.Summary.Passport-cleared.pdf Also see CDT's letter May 2, 2007 letter to Secretary of State Rice on the agencies failure to provide adequate PIAs for this and a related project —
http://www.cdt.org/security/identity/20070502rice.pdf

[7] OMB FY2006 Report to Congress on Implementation of the Federal Information Security Management Act of 2002, at
www.whitehouse.gov/omb/inforegreports/2006_fisma_report.pdf

would specifically requires OMB to create best practices for PIAs across the government will help to address these problems.

As the Government Accountability Office and others have pointed out, OMB has not provided real leadership on privacy issues.[8] Best practices on PIAs can be a starting point for OMB to begin providing such leadership.

Even then, the transparency provided by PIAs must not be viewed as a full solution. Congress needs to begin to address more fundamental privacy issues within government agencies to ensure the trust of the American people. This should begin with a review of the Privacy Act of 1974 and a look into whether the law is adequate to address how the federal government today is using personal information. We look forward to working with this committee to help address these critical privacy issues in more detail in the near future.

Conclusion

The five years of experience in implementing the E-Government Act has provided valuable lessons in how to move government information services forward. In the short term, changes in the way people use the Internet mean that public government information online must be made accessible to search engines. Privacy impact assessments can be improved across the federal government based on the good work that has been done. In the long term, we will need leadership from OMB to protect privacy and security of Americans. We urge this committee to continue its leadership in adapting policy to fit the changing landscape and in oversight of that policy.

---

[8] Government Accountability Office, *Privacy Act: OMB Leadership Needed to Improve Agency Compliance*, June 2003 — http://www.gao.gov/new.items/d03304.pdf . Also see Kenneth A. Bamberger and Deirdre K. Mulligan, Privacy Decisionmaking in Administrative Agencies, University of Chicago Law Journal (forthcoming). A draft is available at— http://www.law.uchicago.edu/Lawecon/events/bamberger.pdf

Appendix: Center for Democracy and Technology and OMB Watch report
Hiding in Plain Sight: Why important Government Information Cannot Be Found
Through Commercial Search Engines

Hiding in Plain Sight:
Why Important Government Information Cannot Be Found
Through Commercial Search Engines

1. Summary

In 2002, the E-Government Act was signed into law with noble goals, including "to promote access to high quality Government information and services across multiple channels" and "to make the Federal Government more transparent and accountable."[1] In many respects, the law has been successful, including encouraging agencies to work together to build Web sites that allow users to find information by its content and not only where it is housed in the bureaucracy.   However, as more individuals use commercial search engines to find government information, making information accessible to search by various sources has become an important goal. Unfortunately, many important information sources within the federal government are essentially hidden from the very search engines that the public is most likely to use.

In this report, we examine search queries that we believe Americans would expect to result in authoritative and trustworthy government information showing up prominently in their search results. In an examination of Google, Yahoo, Microsoft Live and Ask and the search function provided by USA.gov, we confirmed that many of these searches miss critical information simply because of the manner in which the government agency has published the information. For example:

- A search for "New York radiation" does not find basic FEMA and DHS information about current conditions and monitoring.
- A search to help grandparents with a question about visitation of their grandchildren in any search engine does not turn up an article of the same title located on the Web site of the Administration for Children & Families.
- A search for "small farm loans" turns up the commercial offers for loans, and statistics about government loans, but not most of the major federal government programs designed to help fund small farms.

We have several recommendations for the federal government. Each of these would encourage greater accessibility of government information by making it more searchable.

- Congress should pass the E-Government reauthorization act, which would require the Office of Management and Budget (OMB) to create best practices to encourage searchability of federal Web sites.

---

[1] PL 107-347

- OMB should officially recognize the importance of commercial search engines to Internet users and work with the CIO Council to adopt policies to help users find information.
- Agencies should adopt an information policy that makes public accessibility of online content and resources a priority.
- Agencies should create Sitemaps of content on their sites, with special attention given to materials stored in databases and accessible only through drop-down menus. For example, many agencies have FAQ databases that are not accessible to search crawlers but contain very succinct and useful answers to common questions.
- Agencies should review their use of robots.txt files in order to ensure they are used in the least restrictive way possible. Every effort should be made to include, rather than exclude, materials from the website, whether materials were excluded purposefully or accidentally in the past.

This report serves only to spotlight a critical gap in the accessibility of government information; we don't seek to punish or embarrass government agencies here. We also do not know whether some agencies purposefully choose to exclude their information from search engines, or whether the agencies don't know how to make this information more available. We hope that this report will call attention to this issue and encourage federal agencies to review their information policies.

2. Introduction

When Americans look for information online, they generally start by using a commercial search engine. According to industry figures, Americans used commercial search engines over 9 billion times in September alone.[2] Search is also the starting point for locating government information online, whether people are looking for information about the safety of drinking water, legislation on domestic spying, or the availability of government jobs. But very often, searches come up short.

Spurred in part by the E-FOIA Amendments of 1996 and the E-Government Act of 2002, the federal government is putting more information and services online, but a considerable amount of government information is, for all practical purposes, invisible to many users. Many federal agencies operate Web sites that are simply not configured to enable access through popular search engines. These Web sites don't allow search engines to "crawl" them, an industry term for indexing online content, and sometimes even block sites from being found by search engines.

With as many as 80 percent of Internet users accessing government information through third party search engines, these uncrawlable sites pose a significant problem. Many Americans are failing to find authoritative government sources, or worse, concluding that the information or service does not exist.

Often, Web sites will be specifically created to allow the public to access a wide swath of valuable government information, such as forms.gov or regulations.gov. However, when an Internet user searches a commercial search engine for the forms and regulations in these databases, they often do not show up.

Ironically, because a commercial engine powers the search at the heart of the federal government's "Official Web Portal" at USA.gov, the same sites that are not found using commercial search engines are likewise not found using the official government search. By opening government databases to commercial search, agencies can also ensure that they will be indexed for USA.gov.

The reasons that government sites are often inaccessible through search vary. In some cases, government agencies may be unaware that their technical decisions have limited the accessibility of the information they control. Agencies may not realize the simple steps they can take to make sure their information is accessible. In other cases, agencies have a policy of making their information unavailable to search engines.

---

[2] "U.S. Search Engine Rankings, September 2007," Enid Burns, *Search Engine Watch*, Nov 20, 2007. — http://searchenginewatch.com/showPage.html?page=3627654

3. Frustrating Searches

Before examining the technical barriers to searching government information or proposing solutions, it is instructive to illustrate what type of information is missing today from the viewpoint of Internet users.

Below, we set out a number of typical scenarios that would lead a individual to search for government information and show the data sources that remain hidden after thorough searches, because agencies have not taken the requisite steps to open them to indexing.

Often the agencies mentioned operate tens or hundreds of dynamic databases that cannot be indexed and searched. This is not a comprehensive list of agencies with non-searchable, but useful, content. We have chosen these examples to illustrate the usefulness of some of the resources that are currently inaccessible.

Environmental Protection

A resident of New York City is investigating the environmental hazards in her neighborhood. Concerned about her children's health and safety, she wants to investigate radiation levels in New York. She types "New York radiation levels" into a search engine but is unable to access the Environmental Protection Agency's database on radiation levels or the Department of Homeland Security's monitoring information.

Search terms: "New York radiation levels"
What she doesn't find:

- Environmental Measurements Laboratory: This division of DHS has installed monitoring equipment in New York, and recent data is accessible via its Radiological Emergency Management System. (See Figure 1)
  (http://www.eml.st.dhs.gov/Homeland/)
  DHS has also maintained a historical database of fallout measurements.
  (http://www.eml.st.dhs.gov/databases/fallout/Fallout_Data_Searchform.cfm)

- Environmental Protection Agency: has a searchable database of environmental radiation monitoring.
  (http://www.epa.gov/narel/radnet/eramsdbase.html)

Other agencies hosting inaccessible environmental content include the National Oceanographic and Atmospheric Administration, the U.S. Fish and Wildlife Service, and others.

Figure 1: A highly relevant DHS resource – Radiation levels in New York

Figure 2: A search on USA.gov for "New York radiation levels"
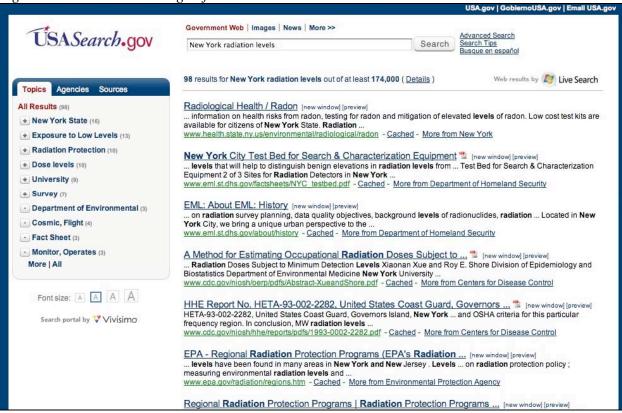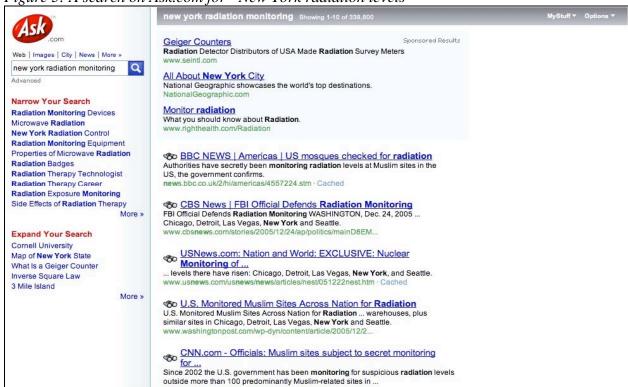

Figure 3: A search on Ask.com for "New York radiation levels"

Federal Business Opportunities

An employee at a telecommunications company is tasked with researching business opportunities with the federal government. He wants to research telecommunications contracts, both historical information about prior contracts and those currently open for bids. He starts out by typing "government telecommunications contracts" into a search engine.

Search terms: "government telecommunications contracts"
What he can't find:

- FedBizOpps.gov: listing approximately 200 government business opportunities within the field of telecommunications.
  (http://vsearch1.fbo.gov/servlet/SearchServlet)

- Export.gov: listing opportunities for telecommunications work overseas.
  (http://www.export.gov/industry/infocomm/)

- GovSales.gov: listing the sale of government property, including the sale of telecommunications equipment.
  (http://www.govsales.gov/fassys/fassys/?function=003000000000)

- Central Contractor Registration: listing who does business and receives moneys from the federal government.
  (http://www.ccr.gov/)

- General Services Administration: lists information about current federal contracts and awarded contracts.
  (http://www.gsaelibrary.gsa.gov/ElibMain/ElibHome)

- Federal Procurement Data Services: includes data on all government contracts, including all telecommunications contracts.
  (https://www.fpds.gov/)

Other General Services Administration sites and individual agency sites list contract information as well, and often can't be indexed.

Museum Collections
A high school student is doing research on African masks and remembers the collection that he saw at the Smithsonian. The student types in "Smithsonian African mask collection" into a search engine and isn't able to access the Smithsonian Institute's online collection of mask images.

Sample search: "Smithsonian African mask collection"

What he can't find:

- Smithsonian Institute resources: Many online content collections, including the Smithsonian Institution Research Information System. In particular, the collection of images of African masks is not indexed.
  (http://www.nmafa.si.edu/collections/divqry1.asp?ClassificationID=13&ObjectTypeID=-1)

- Library of Congress resources: the online catalog of material, as well as many collections of American historical resources online.
  (http://catalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First)

The fantastic array of resources available from the various cultural institutions administered by the U.S. government represent a tremendous absence from the search engines used by the public.

Health and Human Services
A grandmother is upset that she is not being allowed to visit with her grandchildren. She begins her search by typing her question into a search engine: "I'm not allowed to visit with my grandchildren. What can I do?" Even though the Administration for Children & Families has a page in their frequently asked questions with exactly this question, the grandmother does not find it.

Search terms: "I'm not allowed to visit with my grandchildren. What can I do?"
What she can't find:

- Department of Health and Human Services: the frequently asked questions at the Administration for Children & Families.
  (http://faq.acf.hhs.gov/cgi-bin/acfrightnow.cfg/php/enduser/std_alp.php)

Many federal websites have a large collection of frequently asked questions that is entirely inaccessible to search.

Figure 1: A valuable government resource with ACF's answer to the frequently asked question "I'm not allowed to visit with my grandchildren. What can I do?"

Figure 2:A search for "I'm not allowed to visit with my grandchildren. What can I do?" on Yahoo.com
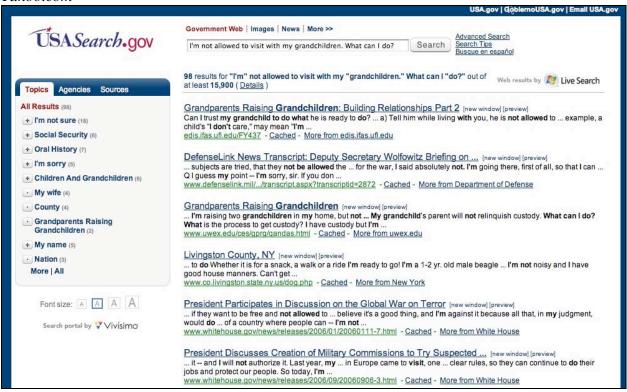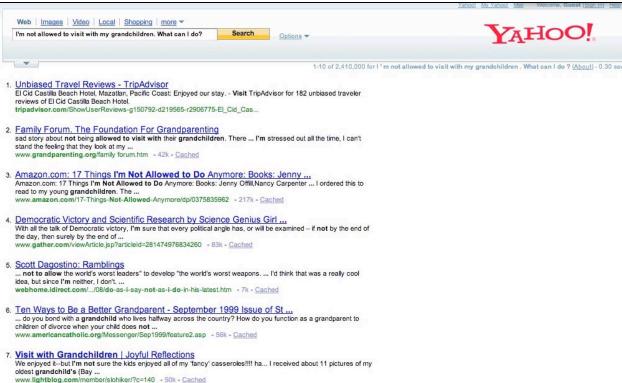

Figure 3:A search for "I'm not allowed to visit with my grandchildren. What can I do?" on Yahoo.com

Emergency Response Resources

A resident of Seattle, WA, is considering a move and is investigating what residential neighborhoods in Seattle are in a flood zone. He types "Seattle flood zone" into a search engine and doesn't find the Federal Emergency Management Agency's flood-mapping tool, Department of Homeland Security's data, or U.S. Geological Survey flooding resources.

Search terms: "Seattle flood zone"
What he can't find:

- Federal Emergency Management Agency: unable to access the Flood Map Modernization project at FEMA, which maps out flood hazards.
(https://hazards.fema.gov/femaportal/wps/portal)

SEC Filings

A retiree is doing background research on some of the companies in which he has invested. He is trying to find the Securities and Exchange Commission filings of General Motors and types "SEC filing General Motors" into a search engine. He is unable to access the SEC's database of company filings.

Search terms: "SEC filing General Motors"
What he can't find:

- U.S. Securities and Exchange Commission: database of all company SEC filings, formal documents submitted to the SEC used by professionals, investors, and the public to gather information about companies.
(http://www.sec.gov/edgar/searchedgar/companysearch.html)

Small Business Programs
An independent farmer in Nebraska is interested in applying for a small farm loan, and he
types "small farm loan" into a search engine. He is unable to access data from the Small
Business Administration or the Department of Agriculture's Farm Loan Programs.

Search terms: "small farm loan"
What he can't find:

- Small Business Administration: This uncrawlable page lists the contracts that are open in
  many agencies, but this information is not accessible on search engines.
  (http://fbo.gov/spg/)

- GovLoans.gov: This website has a collection of loans available from the government, but
  because it is not accessible to search engines, it does not appear in a search. (See Figure 1)
  (http://www.govloans.gov/)

Many agencies run loan programs, or provide other benefits, but these programs can be hard to
find and therefore are hard to utilize.

Figure 1: A valuable resource listing agricultural loans on GovLoans.gov

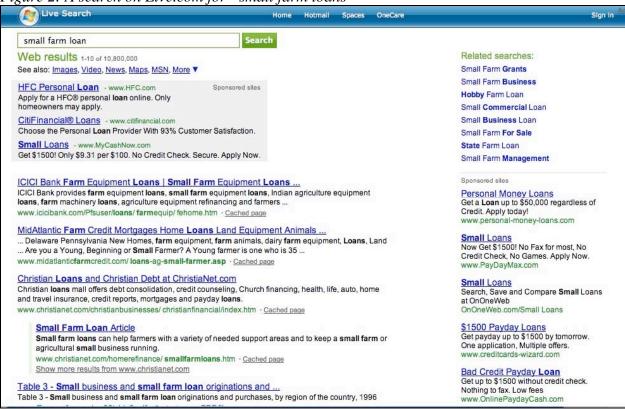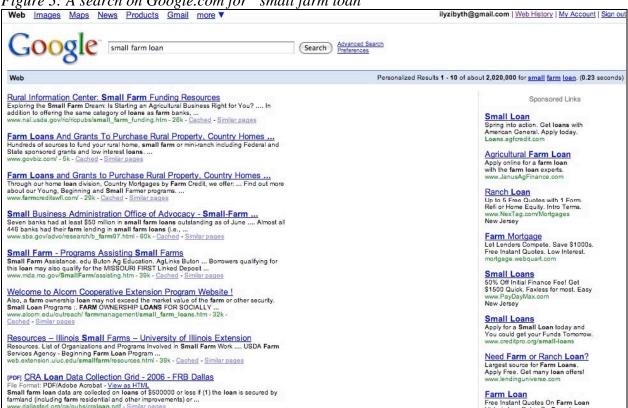*Figure 2: A search on Live.com for "small farm loans"*

Figure 3: A search on Google.com for "small farm loan"

4. Web Crawlers

Search engines need to index the massive amount of content that exists online, so they use automated programs to crawl the World Wide Web. Web crawlers identify, analyze, and add information to search engine indexes.

Web crawlers are invaluable tools for indexing content on the Internet. However, they are not equipped to handle technical hurdles posed by dynamic databases and specialized interfaces. For example, databases provide answers based on the queries submitted by users. However, content that can't be accessed and indexed by search crawlers can't be found by search engines, and therefore, appears invisible to the typical search engine user.

With so many Americans using the major search engines as their main entry to Internet content, this is a critical obstacle to better, more complete information. The irony here is that these databases can be easily configured to make their content available to users through search engines.

5. Robots.txt Files

Site operators may mark some links so that they are ignored by Web crawlers – typically, for the purpose of keeping some section of a Web site invisible to casual searchers. The most common means of doing this is to create a file in a standard location, technically called a "robots.txt" file, that lists a set of locations or directories that the crawler is asked not to index. It is completely voluntary for companies to follow this protocol, but all of the major search engines do.

There are legitimate reasons to use a robots.txt file to stop information that, while available on the Web, may not be appropriate for wide distribution, or to prevent copyrighted material from being cached in search engines. A robots.txt file also can be used to prevent duplicate content from being crawled, or to protect non-robust applications used on the Web site. However, robots.txt can be misused, too, over-blocking content and preventing search engines from crawling the site. For instance, much has been said about the whitehouse.gov robots file[3] and other agencies such as ATF have added wide swaths of their websites to the list of hard to find information with just a few lines of code.

Federal government Web sites contain public information and resources that should be readily available. The widespread use of robots.txt on federal government Web sites is a questionable practice that serves to limit the availability of information, as shown in our previous examples.

6. Sitemapping

The Sitemap protocol is an open and freely available standard that can be used to create a document that allows search engines to effectively crawl and index Web sites. Sitemaps are, in some ways, the opposite of robot.txt files. Like robots.txt, the protocol uses a file in a well-known location. However, rather than listing locations that the crawlers should not index as

---

[3] "White House says blocking Iraq Web documents was 'mistake'," Declan McCullagh — http://www.news.com/8301-13578_3-9773721-38.html

found in robots.txt, sitemap.xml is a list of locations that the crawler should index, but might not find.

The leading search engines – Ask, Google, Microsoft Live, and Yahoo – have adopted the Sitemap protocol. Government agency implementation of the Sitemap protocol allows them to make exhaustive lists of content so that all participating search engines can easily find it.

The E-Government regulations have established the Web site at http://www.USA.gov as the portal for government information. The search engine used by USA.gov is provided by a major commercial search engine and, thus, is subject to the limitations of all search crawlers — it cannot access most government databases, because of the way that they have been implemented. While this is simply a complication for the commercial search engines, it is a major problem for the USA.gov search. USA.gov's tagline is "Government Made Easy," but in this case, it is just as hard to find this information on its search as others. With the implementation of the Sitemap protocol, agency Web sites can ensure that their resources are indexed by search engines and are available to the American public through USA.gov and most commercial search engines.

7. How Sitemaps work

The sitemap is merely a file that lists URLs, and simple information about the pages in XML. No new development of the Web site itself is necessary, nor is the development of a site map an onerous task. In the development of one federal Web site (http://www.plainlanguage.gov), it took only eight hours for the site's web manager to fully implement the Sitemap protocol.

The sitemap can communicate information about the location, importance, frequency of change, and last modification of a page. Using a sitemap, a search engine can optimize its indexing of any site in order to more effectively aggregate the content on the site; thus, the site can be accurately represented in search engine results.

Tools have been developed to ensure that the creation of a sitemap is relatively simple. Google and Yahoo both offer specific tools to help manage a Web site, and there are also open source scripts to help Webmasters create sitemaps.

For more information about the technical implementation of the Sitemap protocol, please see http://www.sitemaps.org/protocol.php.

8. Five Websites on the Right Track

Plain Language, www.plainlanguage.gov
This site promotes the use of 'plain English' to help citizens better understand the workings of the government. After installing a sitemap, their examples of 'plain' and 'obfuscated' language are now high results on Google, and the site is the top result for the "Plain Language" search. The site has seen an increase in traffic, and as new content is added, the sitemap is regenerated in order to reflect the new content.

Energy Department's Office of Science and Technology Information, www.osti.gov
OSTI makes available the research of the Department of Energy and cites sharing this
information with the American people as central to its mission. When OSTI implemented the
Sitemap protocol several years ago, the increase in traffic directed to the site was immediate.
"The first day that Yahoo offered up our material for search, our traffic increased so much that
we could not keep up with it," OSTI Director Walt Warnick said.

Education Department's National Center for Education Statistics, http://nces.ed.gov/
The NCES provides statistical information on educational facilities. The Webmaster created a
sitemap for five previously uncrawlable databases. Search engines are now indexing NCES and
sending their traffic to the latest statistical data, and users now find the original source of the
information.

Library of Congress American Memory project, http://memory.loc.gov/
The Library of Congress' American Memory project is a vast collection of American historical
sources and objects. Before implementing the Sitemap protocol, this powerful resource was not
available to people using third-party search engines. Today, much of the collection is fully
indexed and reachable via search engine queries.

State websites partnering with Google
The governments of Arizona, California, Utah, and Virginia have partnered with Google to
make searching for their materials easier online using the Sitemap protocol.[4] This has made a
great deal of state government content available to the American public.

Many other sites are working toward making their content more accessible to search engines,
and thus the general public. For example, USAJobs.gov just made available a feed of vacancy
announcements,[5] giving major search engines access to this previously uncrawlable
information. In this ongoing process, we applaud those agencies making their information
available.

9. Policy on Access of Government Information
Adoption of information policies that would promote searchability is supported by the goals of
government regulations and legislation, including the E-Government Act of 2002, the
Paperwork Reduction Act, Electronic FOIA, and other federal materials regarding the
management of public informational resources.

Paperwork Reduction Act/Circular A-130, 2000
Circular A-130 was published by OMB to establish policy and guide the management of the
informational resources of federal agencies. This circular reinforces the importance of efficient
management of information resources, including the "free flow of information" and the
effective dissemination of government information to citizens.

---

[5] "USAJOBS Connects with Major Search Engines," OPM News Release, December 5 2007. —
http://www.opm.gov/news/usajobs-connects-with-major-search-engines,1346.aspx

Circular A-130 indicates that agencies should use techniques that reduce the burden on the public to access agency materials. Agencies are required to "[d]isseminate information in a manner that achieves the best balance between the goals of maximizing the usefulness of the information and minimizing the cost to the government and the public." Since the information is already distributed via agency Web sites, it would take very little effort to ensure that the information is widely accessible to the public via search engines.

While Circular A-130 delves deeply into the specifics of how to manage information resources, it is clear that at a higher level, it is a document that mandates and guides agencies in making government resources easily available. This includes making the agency resources and information available to the largest possible audience.

E-Government Act, Section 207, 2002
Section 207 of the 2002 E-Government Act seeks to improve the organization and accessibility of government information. The E-Government Act directed OMB to require that agencies use information technology and Internet-based technologies to improve citizens' ability to access government information and services.

Section 207 of the E-Government Act specifically mandates that each agency director be responsible for creating guidelines for their agency's Web site, with two of the goals being to speed the retrieval of search results and to improve the relevance of those results.

As President Bush said in his signing statement for the E-Government Act, "[t]he Act will also assist in expanding the use of the Internet and computer resources in order to deliver Government services, [...] for a citizen-centered, results-oriented, and market-based Government."

Electronic FOIA, 1996
The Freedom of Information Act (FOIA) was signed into law in 1966; a recent amendment in 1996 broadened FOIA to cover electronic records. The Act was created to "ensure an informed citizenry, vital to the functioning of a democratic society, needed to check against corruption and to hold the governors accountable to the governed." FOIA affirmed the public's right to know about the business of government as a central principle of our open society.

The 1996 amendments to FOIA were intended to simplify and expedite access to federal government records through the use of electronic communications media. The 1996 amendments received widespread bipartisan support.

Making this information available electronically is a step in the right direction. However, if this information is made available only through agency Web sites, many users searching for these resources will not find them. Now that so many resources have been made available in electronic form, it is relatively simple to ensure that they are easily accessible by using the Sitemap protocol.

10. How Policy Impacts Search

Agency policies can strongly affect the availability of information to ordinary users. If agencies take the step of making their information available online, then they should put in place policies to ensure the widest possible availability of that information. One simple way to do this is to use the Sitemap tools to ensure that users can find the information through their preferred search engine.

By allowing users their choice of search engines, rather than limiting them to using a single tool or Web site, agencies can ensure the broadest possible audience for their valuable information and resources.

Policy and legislation clearly outline the priorities for making government resources easy to find and use. They require agency Web sites create effective methods of sharing information with Internet users. The Sitemap protocol can help to make federal agency Web sites more accessible to search engine users with a minimal investment of resources.

11. Recommendations to Help Agencies Ensure Their Content Is Accessible to Search

- Congress should pass the E-Government reauthorization act, which would require OMB to create best practices to encourage searchability of federal Web sites.

- OMB should officially recognize the importance of commercial search to Internet users and work with the CIO Council to adopt policies to help users find information.

- Agencies should adopt an information policy that makes public accessibility of online content and resources a priority.

- Agencies should create Sitemaps of content on their sites, with special attention given to materials stored in databases and accessible only through drop-down menus. For example, many agencies have FAQ databases that are not accessible to search crawlers but contain very succinct and useful answers to common questions.

- Agencies should review their use of robots.txt files in order to ensure they are used in the least restrictive way possible. Every effort should be made to include, rather than exclude, materials from the website, whether materials were excluded purposefully or accidentally in the past.