



Google and the Search for Federal Government Information

by Bonnie Klein (Defense Technical Information Center) <BKlein@dtic.mil>

Why Can't I Find It?

As a librarian working at a federal government information center, I agree with **Senator Joseph Lieberman** (I-Conn) that the public frequently cannot

find information and services placed on government Websites specifically for their benefit.¹ It is true that information and services on many government sites, through practice or policy, are simply inaccessible to commercial search engines. A bill introduced by the **Senate Homeland Security and Governmental Affairs Committee** chaired by **Senator Lieberman** seeks to remedy the situation by requiring federal agencies to review, report, and test search accessibility capabilities. The **E-Government Reauthorization Act of 2007 (S.2321)**² includes a provision for government agencies to employ standards such as **Google's sitemap protocol**³ to make government information more easily indexed by commercial search engines and discoverable by citizens.

But, it takes two to tango. Commercial search engines are under no obligation in their practice or policy to give ranking preference to information from a government source. The **Defense Technical Information Center (DTIC)**,⁴ the organization I work for, and other government information centers that have exposed their data to commercial search engines often find our products are not listed or highly ranked in search results and are, therefore, still invisible. The proposed legislation will not fix that.

Like the earlier **E-Government Act of 2002 (P.L. 107-347)**⁵, the new bill assigns the responsibility for policy, guidance and oversight to the **Office of Management and Budget (OMB)**, **Office of Information and Regulatory Affairs (OIRA)**.⁶ In my opinion, the current policy in **OMB Circular A-130 "Management of Federal Government Information"**⁷ already covers the search capabilities provision by directing agencies to "use electronic media and formats, including public networks, as appropriate and within budgetary constraints, in order to make government information more easily accessible and useful to the public."

At the December 11, 2007 Senate Committee hearing on "**E-Government 2.0: Improving Innovation, Collaboration, and Access**,"⁸ **Karen Evans**, Administrator of the Office of Electronic Government and Information Technology at **OMB**, reported on the progress the government has made in getting services and information online and available to citizens. One avenue is **USA.gov**,⁹ the official U.S. Government Internet portal and centralized point of entry for locating government information, benefits, and services. In FY 2007, **USA.gov** received approximately 97 million visits during the year or 1.87 million visits per week.

At the same hearing, **John Lewis Needham**, **Google's** Manager for Public Sector Content Partnerships, testified that: "The government produces a lot of information and these databases cannot be navigated by Web crawlers." **Needham** correctly stated that the most prevalent technical barriers to search engine access to "deep Web" government information are: (1) agency use of dynamic query-based databases, (2) Robots.txt. files that prevent crawling and (3) outdated links.

Needham also opined that "Agencies are concerned more about how information is presented than if users are finding it." The fact is that agencies are concerned about both. To meet reporting requirements and scorecards, Government agencies want the searching public to readily discover, recognize, and choose the agency as their preferred trusted and authoritative information provider.

Seek and Ye Shall Find?

The premise of the proposed legislation is that if agencies make their data searchable, it will be indexed and discoverable. Hear ye, citizens,

seek and ye shall find. Well, maybe. It depends on where you search, what you are searching for, and how you are searching.

Most search engine users expect and accept that they must sort through a large amount of material, much of it irrelevant to their purpose. To aid users in narrowing results, **Google** and other search engines offer options that limit a search to material types such as images, video, maps, news, and books or by specific interest groups such as Scholar and Finance. **Google** also offers a **U.S. Government option**¹⁰ that searches U.S. federal, state and local government domains and sites; but this option resides under "Special Searches" and is not readily apparent to most **Google** users.

If agencies do apply sitemap or another indexing standard, will search engines rank the federal government information higher in search results? The answer is "No." **Google** states in its **Public Sector Frequently Asked Questions (FAQ)** that "implementing sitemaps does not affect the ranking of a Webpage in search results."¹¹

The answer to the FAQ "What pages will **Google** index? Will they appear in **Google.com** or **Google's US Government Search**?" "is both a disclaimer and business policy. **Google** "can't guarantee that we'll include all pages that we crawl on your agency's Website in our index. However, we'll include all pages we believe are relevant to our users, so that they appear in search results of **Google.com** and **Google's US Government Search**, as well as other **Google** services."

Instead **Google** assesses relevancy based on its **PageRank** technology. **Donna Bogatin** in her January 26th, 2007 ZDNet post "**Google search PageRank excludes relevant Websites**" observes that "By requiring that Web pages have inbound links from third-party Web sites, the **PageRank** based algorithm may result in automatic exclusion of the most relevant pages for a given query simply because no other Websites have linked to them."¹² You'll have to take it on faith, but there is a lot of esoteric and eclectic government information that only a few, if any, would seek or need to find.

We also need to keep in mind that **Google** and other search engines are commercial enterprises, not public utilities. **Barbara Frist's** description of the search engine business model is: "**Google** gets content for free, gives it away for free, and makes its money by being an enormous distribution channel for everything from physics research to 19th century scanned books to the latest **YouTube** video."¹³ Content is a means to an end. In 2007, **Google** had 57% of the market share and reported 4th quarter revenue of \$4.83 billion, a 50% increase over 2006. **AdSense** revenue increased 30%, amounting to \$1.45 billion of the total. Business operations and revenue-generating advertising partnerships, not altruism, factor into page ranking.

As I said earlier, when federal agencies have taken the initiative to open deep Web databases, commercial search engines do not always rank the government-source content above commercial or for-fee suppliers. The page rank depends on what, where and how one searches. I offer the experience of my agency as an example.

The DTIC Experience

Since 1945, the **Defense Technical Information Center** and its predecessor agencies have served as the **Department of Defense (DoD)** institutional repository and secondary disseminator of scientific, technical, research and development information. Note the term "secondary disseminator." **DTIC** is an aggregator and not the originator, owner or publisher of the information in our collection. It is possible, actually highly likely, that our reports are available from other sources such as the **DoD** office that sponsored the research or from the contractor or grantee that produced the report.

Starting in 1995, **DTIC** provided public online access to searchable bibliographic citations for **DoD Public Release Technical Reports**

continued on page 32

via its **Scientific and Technical Information Network (STINET)**.¹⁴ Internet technology quickly evolved from “gophers” and **Wide Area Information Service (WAIS)** to **World Wide Web (WWW)** browsers and increasingly sophisticated database search engines, computer applications and tools. By 1998, **DTIC** was linking the bibliographic records to full-text. **STINET** content was part of the “deep Web” until **DTIC** implemented the **Open Archives Initiative (OAI) protocol**¹⁵ in early 2006. **OAI** allows third party harvesters easy access to **DTIC**’s content in a variety of formats such as **COSATI**, **MARC**, **Dublin Core (DC)** and **HTML** using **XML** technology with links to the digital content using **DTIC**’s **Handle Service**.¹⁶ Today **DTIC** offers free online access to more than 343,000 full-text documents and 1,109,000 citations. This number grows as **DTIC** adds new documents and digitizes its legacy collection.

DTIC was motivated to expose its content to search engines to provide citizens with free open access to the full-text of **DoD** public release research reports. In 2002, a techno-savvy entrepreneur saw a money-making opportunity to exploit the **DTIC** collection by harvesting the citations, making them searchable via **WWW** search engines and providing the full-text downloaded from **DTIC** for a fee. Now that the **DTIC** collection is **OAI** compliant, the commercial supplier still frequently ranks above **DTIC**. And sometimes the **DTIC** citation does not make the list at all.

At this writing, my **Google Web** search for the **DTIC** technical report “A Wavelet Analysis of Mining Explosions” ranks the commercial supplier first and a **Department of Energy Office of Scientific and Technical Information version (DOE OSTI)** second. The **DTIC** source citation is not listed nor does it appear when searching **Google Books** or **Google Scholar**. It does, however, rank first in **Google’s US Government Search**.

In another example, the results for a **Google Web** search for the **DTIC** title “Planetary Defense: Eliminating the Giggle Factor” authored by a **National Defense University** student, ranks a **US Air Force** source first and the commercial supplier second. Once again **DTIC** is not listed. **Google Scholar**, however, ranks **DTIC** first above the commercial supplier, but does not list the **US Air Force** version. In **Google’s US Government Search**, **DTIC** ranks second after the **US Air Force**.

Access vs. Use – What About Copyright?

The adage “consider the source” applies when seeking government information. There are and always have been resellers and repackagers of government information who have profited by knowing where and how to get it and then supplying it to others. This is perfectly legal and fills a need. What is not, is when the supplier does not credit the source or misrepresents themselves as the copyright

against the grain people profile

Technical Reports Team Lead/Copyright Specialist
Defense Technical Information Center
8725 John J. Kingman Road, Ft. Belvoir, VA 22060-6218
Phone: (703) 767-8037 • Fax: (703) 767-9244
<bklein@dtic.mil> • www.dtic.mil

Bonnie Klein

BORN & LIVED: Born in Chicago IL. Lived in DesPlaines IL, (University of Illinois) Urbana IL, (Indiana University) Bloomington, IN, (WIU) Macomb IL, Uijongbu S. Korea, Hohenfels Germany, Heidelberg Germany, Springfield Virginia.

EARLY LIFE: Unexceptional.

PROFESSIONAL CAREER AND ACTIVITIES: Always the librarian, true to my calling.

PET PEEVES/WHAT MAKES ME MAD: Copyright ambiguity.

PHILOSOPHY: Anything goes.

MOST MEMORABLE CAREER ACHIEVEMENT: *CENDI Frequently Asked Questions About Copyright: Issues Affecting the U.S. Government* <http://www.cendi.gov/publications/04-8copyright.html>.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Getting an icon, tag, and machine-actionable metadata package to identify works of the U.S. Government not subject to copyright in the U.S.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: Oh, boy! This is a hard one to answer. I have no idea, but expect that anything goes. 🐼

owner and imposes restrictive terms and conditions of use. Even **Google Books** sometimes adds a copyright watermark to post-1923 public domain government works provided to it by third parties.¹⁷

No matter how or where one finds government information, once found we need to know what uses we can make of it. E-Government initiatives have overlooked the importance of administrative copyright management metadata in building the Government digital infrastructure. I believe this is attributable to a common misconception that all government information is in the public domain and may be used by anyone, anywhere, anytime without permission, license or royalty payment. The reality is that government information products include a variety of copyrighted and public domain materials. Only government works prepared by officers and employees of the U.S. Government as part of their official duties are not protected by copyright in the U.S. (**17 USC §105**).¹⁸ Contractors and grantees are not considered Government employees and may hold copyright in works they produce for the Government. The Government also publishes and distributes other third-party copyrighted materials with permission or under license.

Adding to the confusion is another generally-held misconception that a work is in the public domain if it does not have a copyright notice. Although once true, the **U.S. Copyright Law** was amended in 1989 to automatically grant copyright protection to original works of authorship once fixed in a discernable format (**17 USC §102**).¹⁹ No formality, registration, or effort on the part of an author is required for a work to be protected. Use of a copyright

notice is voluntary. Absent a notice, the burden is on the user to investigate the copyright status of the work.

Typically U.S. Government works have no statement that clearly identifies them as such. The lack of notice creates an element of uncertainty. It may factor into why the **Google Books** digitization program errs on the side of caution by adding a copyright watermark to U.S. Government works published after 1923 (Note: Works published before 1923 are in the public domain — an easy math computation!). Social networks such as **Wikipedia** that operate in an open intellectual property environment also struggle with copyright/copyleft management and have developed **tags**²⁰ to document their decisions. As diligent as they are, it’s no surprise that Wiki editors and contributors do not always get it right in assessing the copyright of U.S. Government information.

Conclusion

Rather than legislating search capabilities, citizens might be better served if the Government would mandate a system-neutral method to unambiguously identify government information and its copyright status. Visual icons and machine-readable tags would tell users (1) that the information is from a government source and (2) if there are any intellectual property considerations or use constraints. The identifiers could be applied to all materials in all formats (paper, physical media, digital, datasets, software, etc.), across domains and no matter the dissemination channel. In the digital environment, search engines and successor technologies could factor in the tags to


continued on page 34

elevate the government information ranking or as a criteria to narrow a search by usage rights ala **Creative Commons**.²¹

Although the intent is different, the **Government Printing Office (GPO)** has a pilot program underway to identify, mark and certify the integrity of government information it disseminates. The system uses digital signature technology and adds a visible icon or "Seal of Authenticity" to assure users that the content is authoritative. The icon graphic is an eagle next to the words "Authenticated U.S. Government Information."²²

We could all benefit if Government agencies would mark the copyright status of their information products at the time of creation or acquisition. As **Clifford Lynch** points

out: "There's a difference between viewing the presence of tags as conclusive positive information and being able to count on the absence of a tag as negative information."²³

Models, methods, technologies and tools exist to implement a marking system. What we need is the mandate to do it. 

USGovWork (17USC §105). Not subject to copyright. *This article is a United States Government work. The author is a U.S. Government employee. Copyright protection is not available for any work prepared by an officer or employee of the United States Government as part of that person's official duties. The views presented in this article are those of the author and do not reflect the official position of the Department of Defense or U.S. Government.*

Additional Information

E-Gov : The Official Website of the President's E-Government Initiative — <http://www.whitehouse.gov/omb/egov/>

OMB Policies for Federal Government Websites — http://www.usa.gov/webcontent/reqs_bestpractices/omb_policies.shtml

GSA Request for Information: Efficient and Effective Information Sharing and Retrieval (Sep 15 2005) — <http://www.fbo.gov/servlet/Documents/R/1282831>

OMB Memorandum M-06-02, "Improving Public Access to and Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model" — <http://www.whitehouse.gov/omb/memoranda/fy2006/m06-02.pdf>

CENDI Frequently Asked Questions About Copyright: Issues Affecting the U.S. Government — <http://www.cendi.gov/publications/04-8copyright.html>

Endnotes

1. **E-Government 2.0: Improving Innovation, Collaboration, and Access.** Member Statement: **Senator Joseph I. Lieberman.** http://hsgac.senate.gov/public/_files/121107JILOpen.pdf
2. **E-Government Reauthorization Act of 2007 (S.2321).** <http://thomas.loc.gov/cgi-bin/bdquery/z?d110:s.02321>
3. **Google Sitemap Protocol.** <https://www.google.com/webmasters/tools/docs/en/protocol.html>
4. **Defense Technical Information Center.** <http://www.dtic.mil>
5. **E-Government Act of 2002 (P.L. 107-347).** http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ347.107.pdf
6. **Office of Management and Budget (OMB), Office of Information and Regulatory Affairs (OIRA).** <http://www.whitehouse.gov/OMB/infoereg/infoportaltech.html>
7. **OMB Circular A-130 Management of Federal Information Resources.** <http://www.whitehouse.gov/omb/circulars/a130/a130trans4.html>
8. **E-Government 2.0: Improving Innovation, Collaboration, and Access.** Witness Testimony. <http://hsgac.senate.gov/public/index.cfm?Fuseaction=Hearings.Detail&HearingID=5830bc2b-275b-40fd-9309-ea0710e66516>
9. <http://www.usa.gov>
10. **Google US Government Search.** <http://www.google.com/ig/usgov>
11. **Google Public Sector Frequently Asked Questions (FAQ).** <http://www.google.com/publicsector/faq.html>
12. **Google Search Page Rank Excludes Relevant Websites by Donna Bogatin.** *ZDNet* January 26, 2007. <http://blogs.zdnet.com/micro-markets/?p=864>
13. **Face Value by Barbara Fister.** *insidehighered.com* February 18, 2008. <http://www.insidehighered.com/views/2008/02/18/fister>
14. **Scientific and Technical Information Network (STINET).** <http://stinet.dtic.mil/>
15. **Open Archives Initiative (OAI) protocol.** <http://www.dtic.mil/dtic/prodsrv/oai.html>
16. **DTIC's Handle Service.** <http://www.dtic.mil/dtic/handles/index.html>
17. **Google Book Search Treats Government Documents as Copyrighted Material.** *Prelinger Library Blog.* February 18, 2006. <http://prelingerlibrary.blogspot.com/2006/02/google-book-search-treats-government.html>
18. **U.S. Copyright Law, 17 USC §105.** <http://www.copyright.gov/title17/92chap1.html#105>
19. **U.S. Copyright Law, 17 USC §102.** <http://www.copyright.gov/title17/92chap1.html#102>
20. **Wikipedia Public Domain-US Government Image Tags.** http://en.wikipedia.org/wiki/Wikipedia:Image_copyright_tags/USA
21. **Creative Commons.** www.creativecommons.org
22. **Government Printing Office Authentication System.** <http://www.gpoaccess.gov/authentication>
23. **The Shape of the Scientific Article in The Developing Cyberinfrastructure by Clifford Lynch.** *CTWatch Quarterly*, Volume 3, Number 3, August 2007. <http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/>