# Unleashing the value of records management

Implementing the Sitemap protocol

# Sitemaps.org

An open, industry standard for web search engine crawling



#### What are Sitemaps?

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

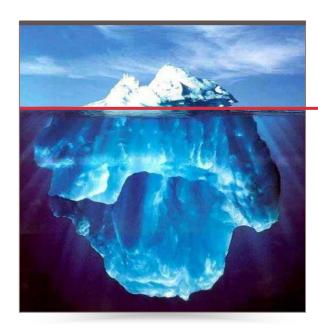
Web crawlers usually discover pages from links within the site and from other sites. Sitemaps supplement this data to allow crawlers that support Sitemaps to pick up all URLs in the Sitemap and learn about those URLs using the associated metadata. Using the Sitemap protocol does not guarantee that web pages are included in search engines, but provides hints for web crawlers to do a better job of crawling your site.

Sitemap 0.90 is offered under the terms of the Attribution-ShareAlike Creative Commons License and has wide adoption, including support from Google, Yahoo!, and Microsoft.



# But they cannot easily find upwards of 40% of public federal agency information

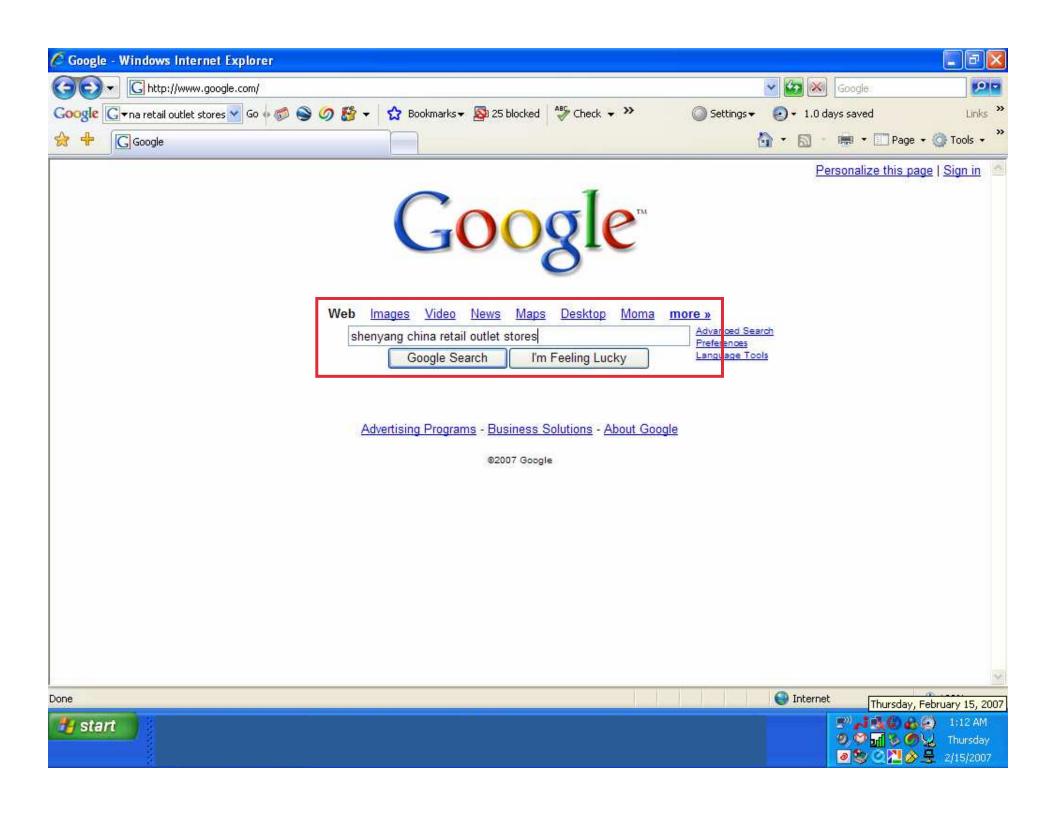
- Federal government web services invisible to search engine users:
  - GovBenefits.gov,
     Regulations.gov, USAJOBS.gov
     and other E-Gov sites are an
     island unto themselves
  - NOAA, NIH, EPA and other major agencies as well as nearly all departments are operating dozens, even hundreds, of invisible services
  - Less than 1% of web pages/records on NARA sites are visible

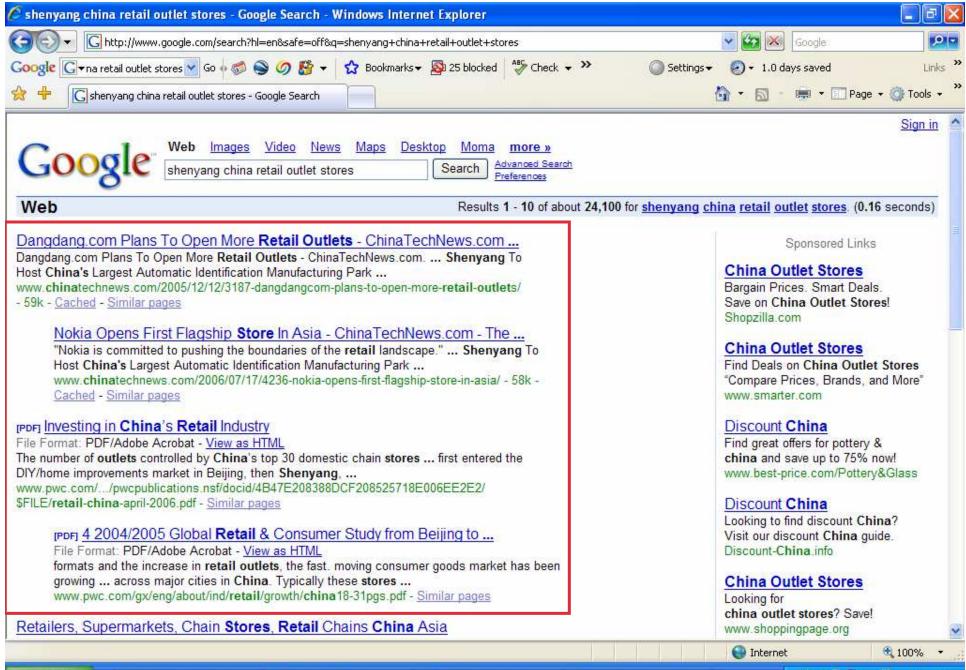


**WEB**Searchable

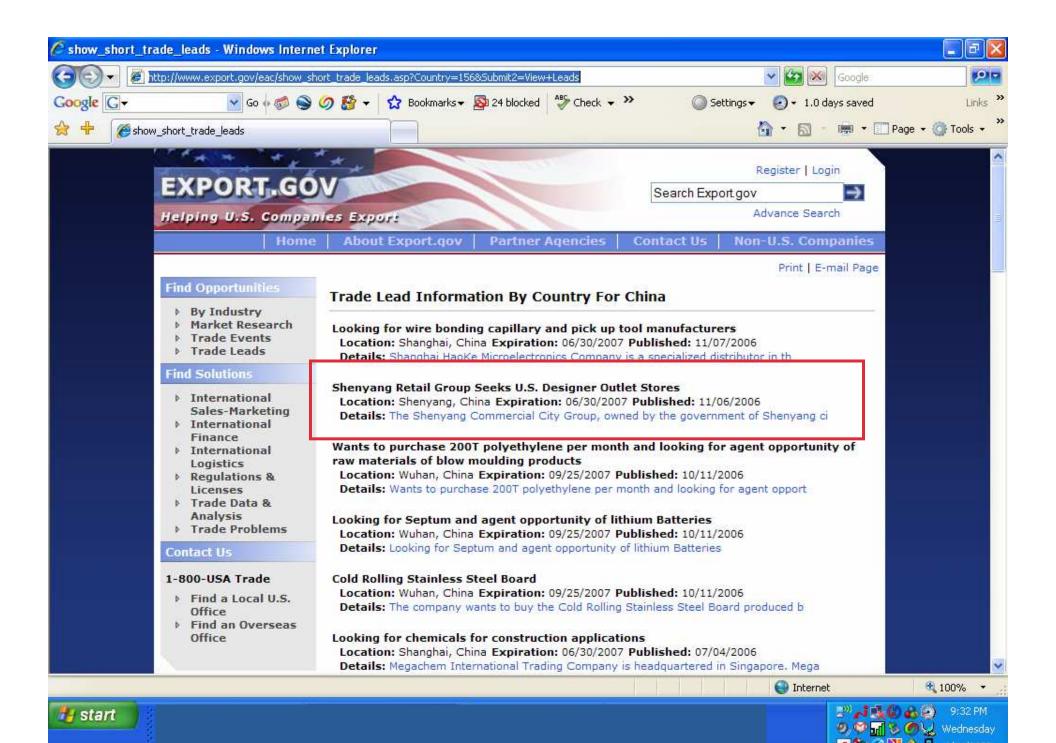
**DEEP WEB**Not searchable

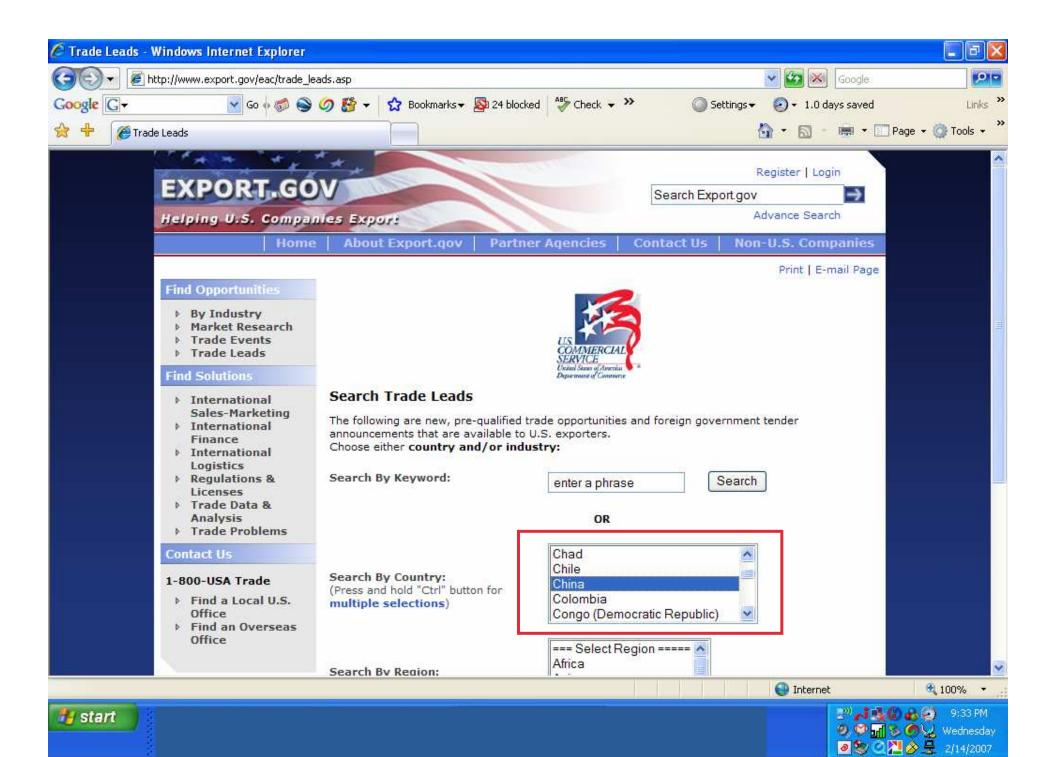


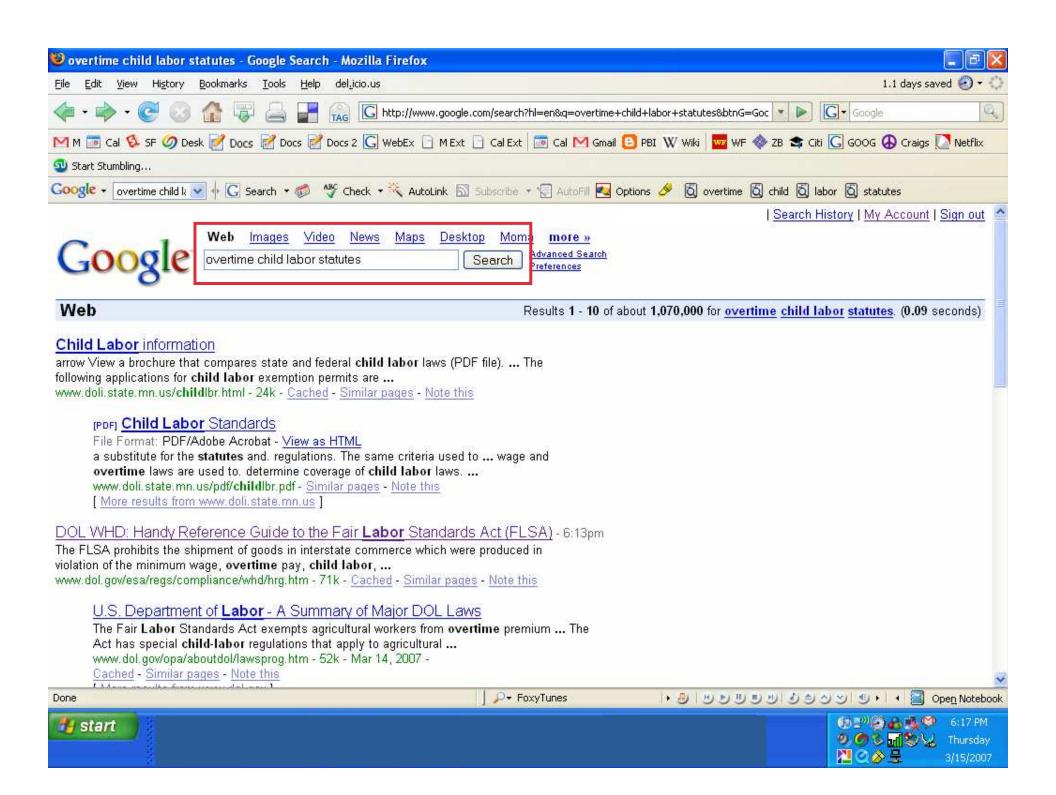


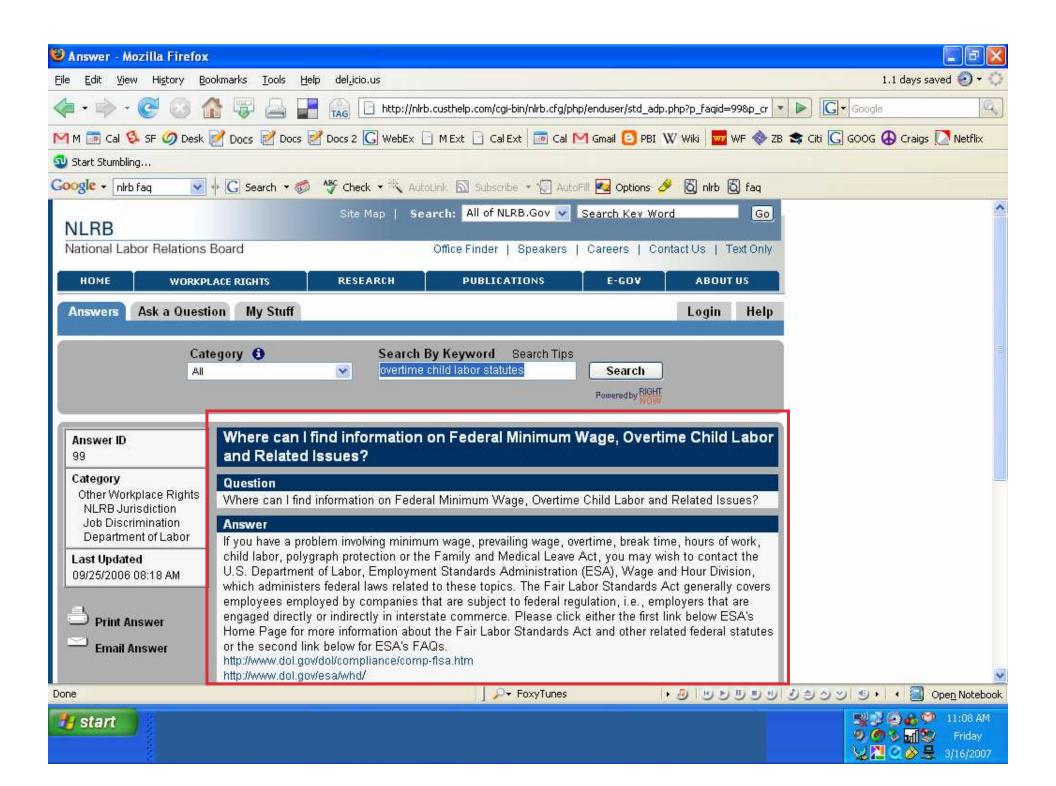


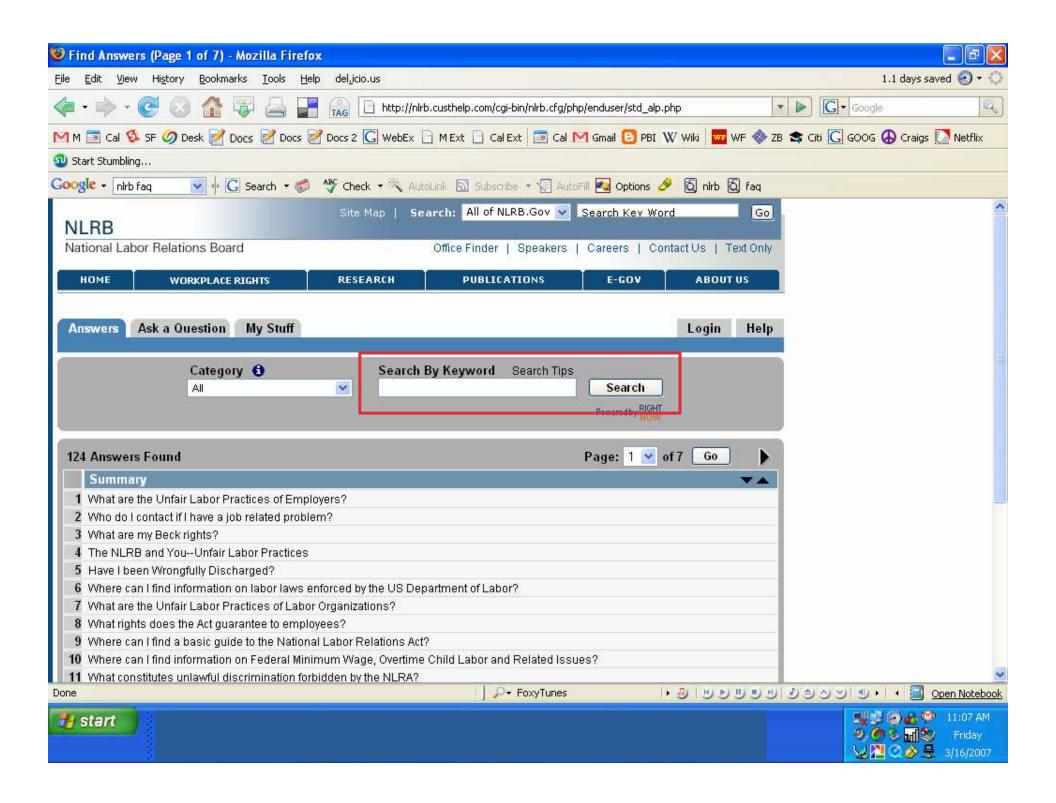


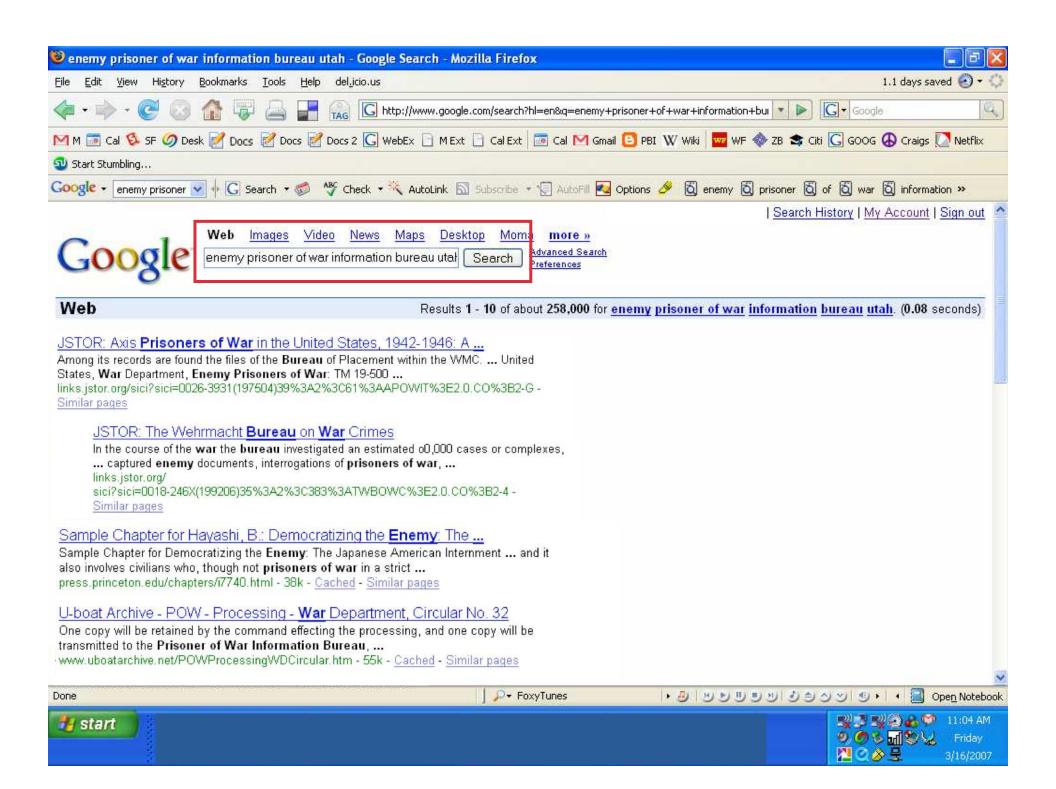


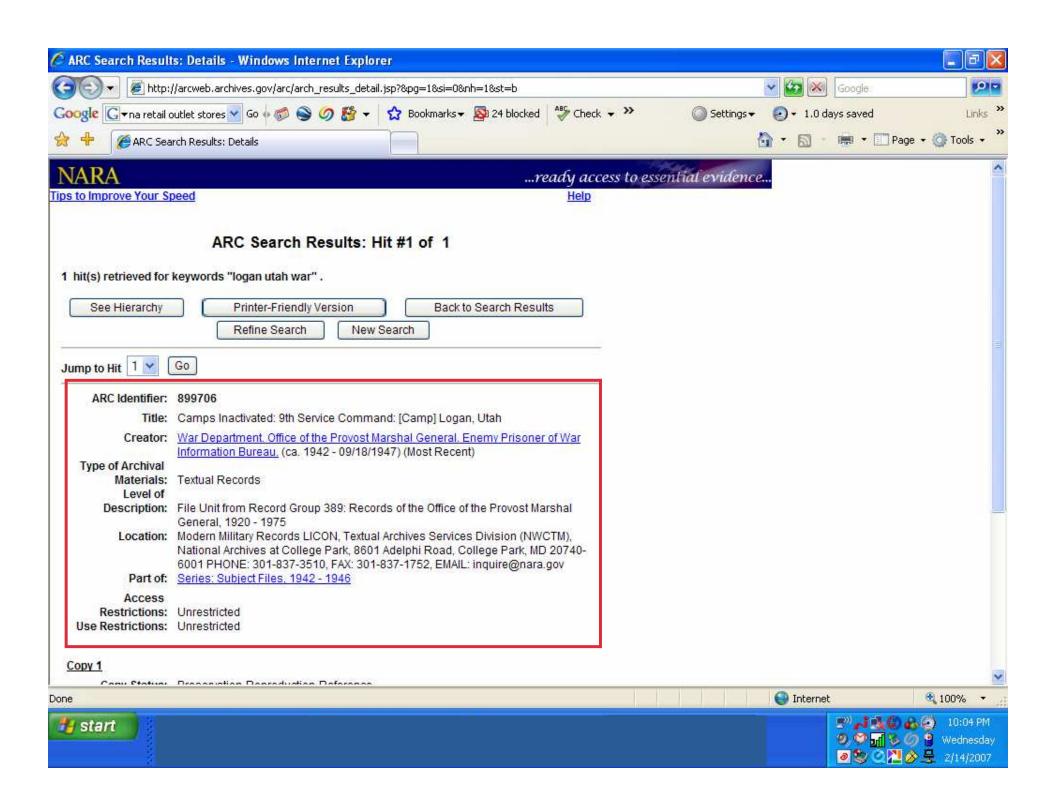


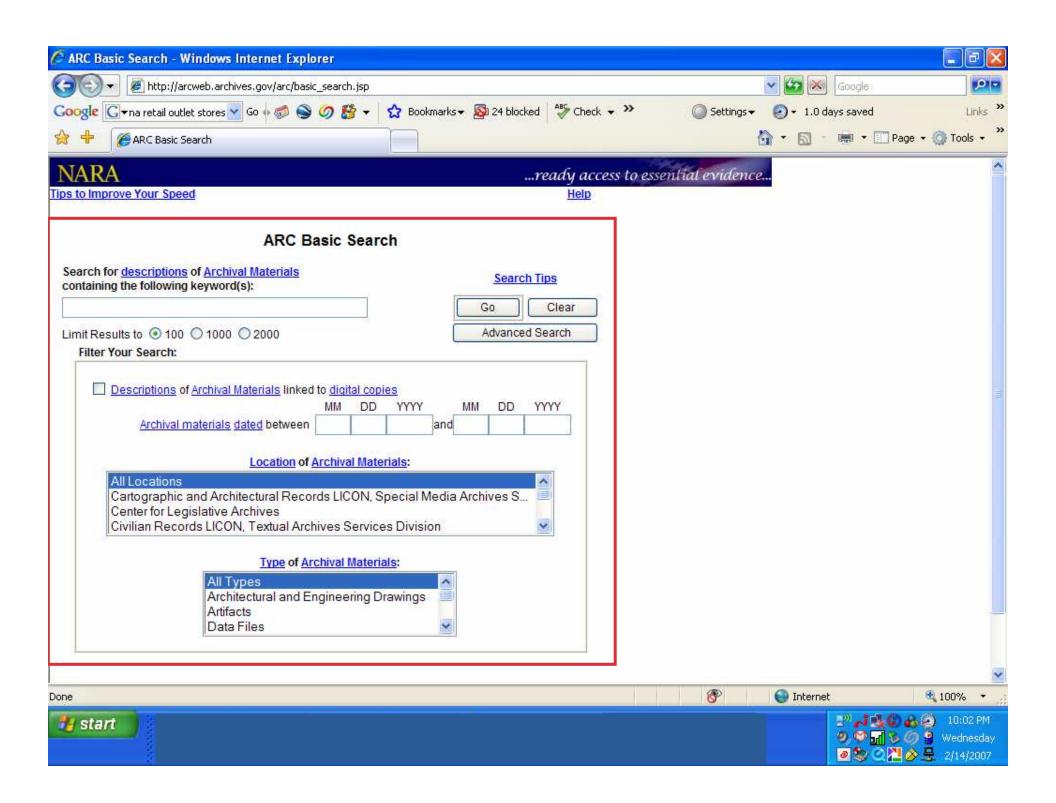












## Sitemaps for search engines

- A comprehensive list of a site's URLs, in simple text or rendered in XML
- If XML, then tagged with location, last modification date, change frequency and priority

```
<?xml version="1.0" encoding="UTF-8"?>
<url><urlset xmlns="http://www.google.com/schemas/sitemap/0.84"></ur>
   <ur1>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </ur1>
   <ur1>
      <loc>http://www.example.com/catalog?item=12&amp;desc=vacation hawaii</loc>
      <changefreq>weekly</changefreq>
   </ur1>
   <<u>ur1</u>>
      <loc>http://www.example.com/catalog?item=73&amp;desc=vacation new zealand</loc>
      <lastmod>2004-12-23</lastmod>
      <changefreq>weekly</changefreq>
   </ur1>
   <ur1>
      <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>
   </ur1>
      <loc>http://www.example.com/catalog?item=83&amp;desc=vacation usa</loc>
      <lastmod>2004-11-23</lastmod>
   </ur1>
</urlset>
```



## Relevant legislation and OMB policy

- The Sitemap protocol supports the E-Government Act of 2002 requirements to:
  - "Organize and categorize information intended for public access and ensure it is searchable across agencies...[using] sophisticated Internet search functions (including their crawl and index mechanisms)..."
  - "...publish your information directly to the Internet...expos[ing] information to freely available and other search functions [that] adequately [organize] and [categorize] your information."
  - "...[When] disseminating significant information dissemination products, advance preparation, such as using formal information models, may be necessary to ensure effective interchange or dissemination. This procedure is needed when freely available and other search functions do not adequately organize and categorize your information."
- The Sitemap protocol also supports the Federal Enterprise Architecture's Data Reference Model 2.0 requirements to:
  - "Identify how information and data are created, maintained, accessed, and used...[and] Define data and describe relationships among data elements used in the agency's information systems."
  - \* Source: Federal Sitemaps (http://colab.cim3.net/cgi-bin/wiki.pl?FederalSitemaps), Semantic Interoperability (SICoP) and XML (xmlCoP) Community of Practice.



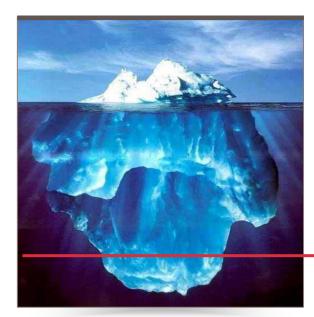
#### Relevant ERM metrics

- Relevant metric: "ERA-8 Percentage of Electronic Records Holdings Open and Accessible Online"
- Proposed metric: Percentage of electronic records holdings accessible through commercial search engines:
  - For NARA, effective accessibility is currently less than 1%
  - However, NARA's AAD and ARC services are now being sitemapped, which will open 100 million records to search engine users



#### Join in

- Many government agencies are now implementing sitemaps
- How many of your agency's web and ERM services are invisible to citizen users? What can you do about it?
- Find out:
  - Query John Lewis (JL)
     Needham,
     jlneedham@google.com
  - Track the Federal Sitemaps effort: http://tinyurl.com/3byhy7



**WEB** Searchable

**DEEP WEB**Searchable

