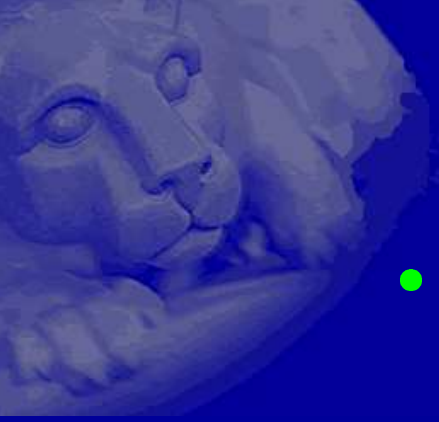
Using the Penn State Search Engine

and latest updates, 2006

Jeffrey D'Angelo, Jeremy Hill and James Leous
search@psu.edu

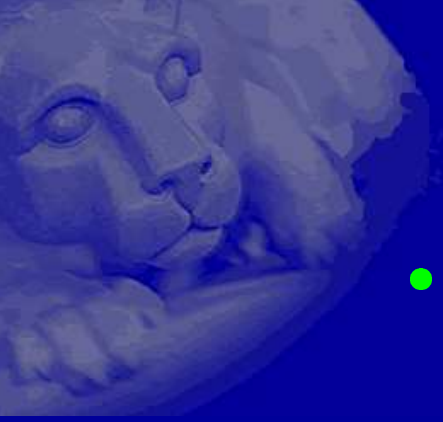
http://aset.its.psu.edu/

ITS / Academic Services and Emerging Technologies



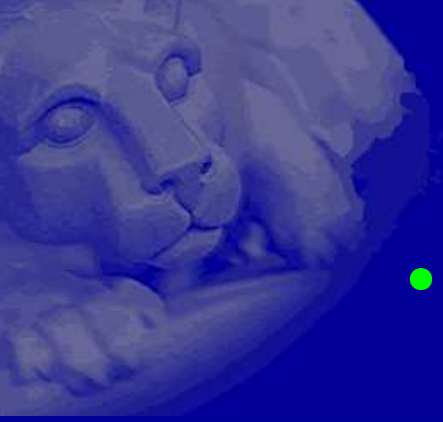
Overview

- New Upgrade Changes Affecting Existing Websites (For those already using search)
- How Search Works (For those new to search)
- Major Differences Between the Versions
- By The Numbers
- Search Integration



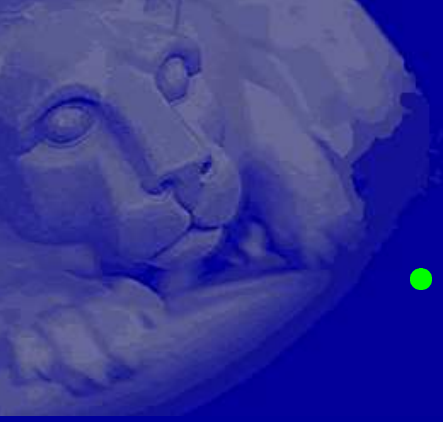
Changes Affecting Websites

- Upgrade: July 11, 2006



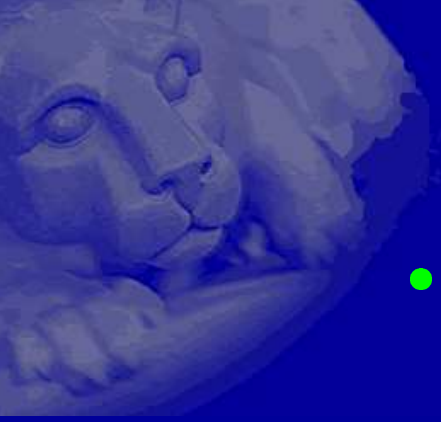
Changes Affecting Websites

- Upgrade: July 11, 2006
- What will work without changes
  - Basic search forms, as_sitesearch, sitesearch



Changes Affecting Websites

- Upgrade: July 11, 2006
- What will work without changes
  - Basic search forms, as_sitesearch, sitesearch
- What will *not* work without changes
  - search forms that use restrict and subcollections
  - search forms that use proxystylesheet to use a custom XSLT file



Changes Affecting Websites

- Upgrade: July 11, 2006
- What will work without changes
  - Basic search forms, as_sitesearch, sitesearch
- What will *not* work without changes
  - search forms that use restrict and subcollections
  - search forms that use proxystylesheet to use a custom XSLT file
- What *may* require some attention or changes
  - XML formatted search results & applications that use them

Change Details

Details: <http://aset.its.psu.edu/googledocs/2006_upgrade/>

How Does Search Work?

How Does Search Work?

In general search engines deploy a piece of software called a “crawler” which begins searching at a given URL or URLs and follows the links contained in one page to others.

There may be boundaries on this search ($e.g.$ only sites in this domain or only sites N levels deep in the current site).

Crawlers collect these pages and pass all or parts of it to a “catalog” or “index” of the site(s).

The index is passed to the actual search engine software where pages are ranked according to a numerical algorithm. Google’s numerical algorithm is called PageRank.

The usefulness of a crawler based search engine depends on all three of these parts.

Why Search?

- **Navigational** — The immediate intent is to reach a particular site.

Why Search?

- **Navigational** — The immediate intent is to reach a particular site.
- **Informational** — The intent is to acquire some information assumed to be present on one or more pages.

Why Search?

- **Navigational** — The immediate intent is to reach a particular site.
- **Informational** — The intent is to acquire some information assumed to be present on one or more pages.
- **Transactional** — The intent is to perform some Web–mediated activity.

Does Google Engine Scale?

Look at the numbers:

- www.Google.com $\rightarrow$ $\sim$4.3 Billion pages (or more)
- search.psu.edu $\rightarrow$ $\sim$1.0 Million pages

Look at the connections:

- www.Google.com $\rightarrow$ relies heavily on connections of "important" pages to other "important" pages
- search.psu.edu $\rightarrow$ with the exception of some central pages, very few departmental pages actually link across academic or administrative boundaries.

What Really Matters?

Although Google's PageRank algorithm is a secret, the following is known:

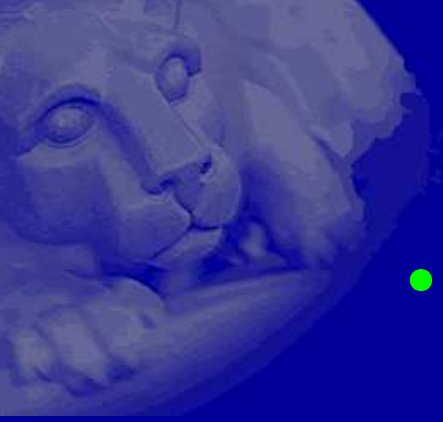
- Links from other pages are heavily weighted
- <TITLE> content is very important
- Keyword density within a page is important
- Keyword–laden links vs. "click here"
- Meta keyword tags are not that important

Are Tags Irrelevant?

Under our current system:

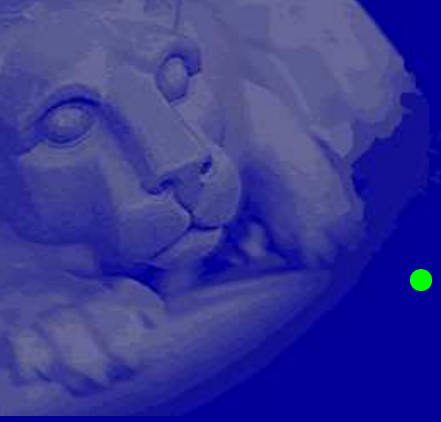
- Meta keywords are not that heavily weighted
- Penn State sample differs from Google sample with regard to interconnections
- Use keywords in <TITLE> tags of page as well as “keyword laden” links to other pages within and outside of your site
- Remember Google is current state of the art; the future is probably a more federated search mechanism

Major Version Differences



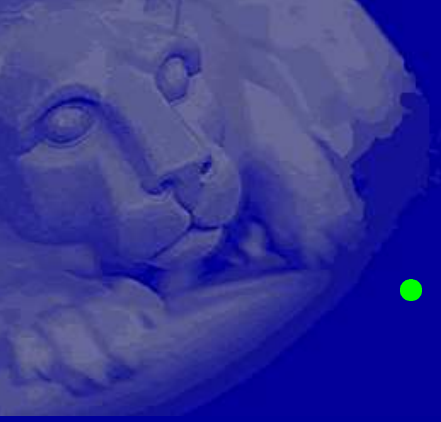
Major Version Differences

- Crawl/index method: all–at–once vs. incremental
  - faster/more efficient crawl rates
  - “How often does a page get crawled?” gets more complicated:
    <https://aset.its.psu.edu/googledocs/crawl_policy.html>



Major Version Differences

- Crawl/index method: all–at–once vs. incremental
  - faster/more efficient crawl rates
  - “How often does a page get crawled?” gets more complicated:
    <https://aset.its.psu.edu/googledocs/crawl_policy.html>
- More capacity
  - 3,000,000 $\rightarrow$ 5,000,000 max docs
    - old engine licensed for 1.5M, new engine 2.0M
  - dual 1.0 GHz Pentium3 CPUs $\rightarrow$ dual 2.6 Pentium4
  - 3 $\times$ 80 GB $\rightarrow$ 5 $\times$ 250 GB disk drives
  - 2 GB $\rightarrow$ 12 GB system memory (RAM)



New Features Ready

- googleon/googleoff tags
- Server-side stored XSLT (FrontEnd)

<http://aset.its.psu.edu/googledocs/instructions.html#features>

New Features Possible

New Features Possible

- Databases

New Features Possible

- Databases
- Feeds

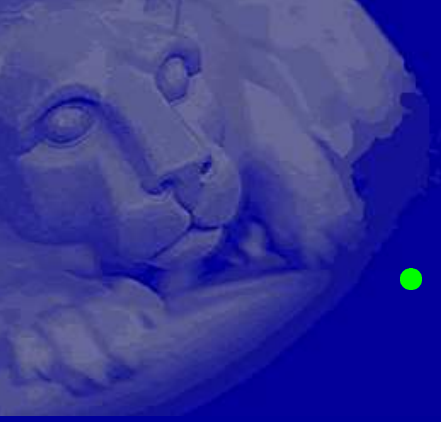
New Features Possible

- Databases
- Feeds
- Protected sites
  - Also in old engine: HTTP Basic, NTLM
  - New: form-based, cookie-based, Google AuthZ API (SAML, X509, LDAP)
  - New: SSO (Oblix, Netegrity and Cams ... CoSign?)

By the Numbers

From the last crawl (Sunday evening, May 14, 2006) older search engine:

- Time:
  - took 9 hours 25 min 54 sec to query the Web servers,
  - 2 hours 55 min 27 sec to build the index
  - and another 40 min and 19 sec to replicate and test the index
  - Start to finish: 11 hr 20 min 38 sec to crawl/index Penn State



By the Numbers (cont'd)

- Amount:
  - 993,631 total URLs crawled/indexed
  - 848,551 URLs found that we excluded (non .psu.edu, containing ?, etc)
  - 89 GB index size
- Queries:
  - Spring Semester 2006: 2,756,628
  - Average for Spring 2006: 23,561/day
  - 1 every 3.7 seconds
  - over 1,000 in the time to give this seminar

By the Numbers (cont'd)

Newer Search Engine

- Time:
  - 26.5 hours to crawl all of Penn State the first time
  - changes can be noticed in as little as 15 minutes for some pages (up to 90 days)
- Amount:
  - 1,343,604 total indexed (7pm 2006/05/15)

Search Integration

Integrating the Penn State Search Engine Into Your Site

Search Integration

Integrating the Penn State Search Engine Into Your Site

- Invoking

Search Integration

Integrating the Penn State Search Engine Into Your Site

- Invoking
- Restricting Search Results

Search Integration

Integrating the Penn State Search Engine Into Your Site

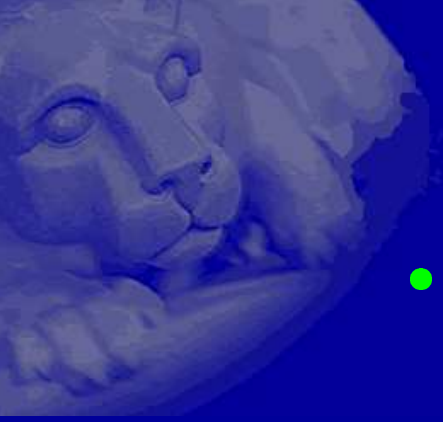
- Invoking
- Restricting Search Results
- Customizing Search Results Style

Invoking Search Engine

Invoking the Search Engine from Your Site

- Easy and convenient way for visitors to search from your site
- Directions:

  <http://aset.its.psu.edu/googledocs/instructions.html#invoke>

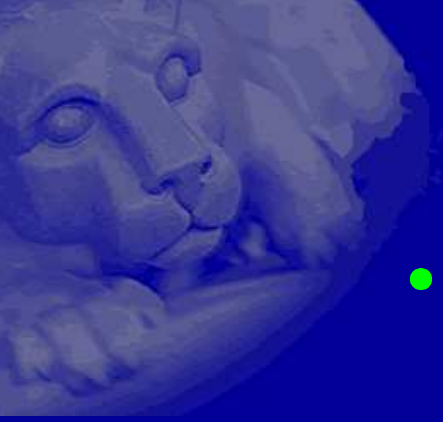


Restricting Search Results

- as_sitesearch

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>

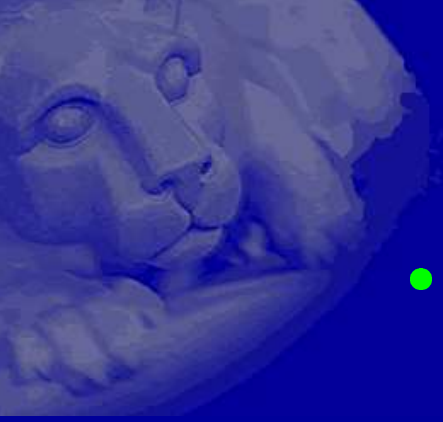


Restricting Search Results

- as_sitesearch — Restrict results to only URLs that begin with the value you set for this parameter. Adds “site:–your–url–” to the search query.
- sitesearch

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>

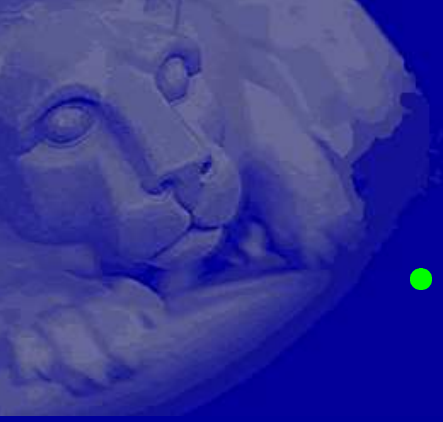


Restricting Search Results

- as_sitesearch — Restrict results to only URLs that begin with the value you set for this parameter. Adds “site:–your–url–” to the search query.
- sitesearch — Restrict results to only URLs that begin with the value you set for this parameter. Does not add “site:–your–url–”
- **[Changing feature]** restrict & site

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>



Restricting Search Results

- as_sitesearch — Restrict results to only URLs that begin with the value you set for this parameter. Adds “site:–your–url–” to the search query.
- sitesearch — Restrict results to only URLs that begin with the value you set for this parameter. Does not add “site:–your–url–”
- **[Changing feature]** restrict & site — Uses a (sub)collection which search administrators must set up in advance. Allows multiple sites to be grouped together; there will be some delay (days to hours).

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>

Customize Format/Style

Customize Search Results Format/Style

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#format>

Customize Format/Style

Customize Search Results Format/Style

- Two choices for alternate format: XML and XSLT.

Directions:
<http://aset.its.psu.edu/googledocs/instructions.html#format>

Customize Format/Style

Customize Search Results Format/Style

- Two choices for alternate format: XML and XSLT.

- XML formatted results — good for dynamic applications

Directions:
<http://aset.its.psu.edu/googledocs/instructions.html#format>

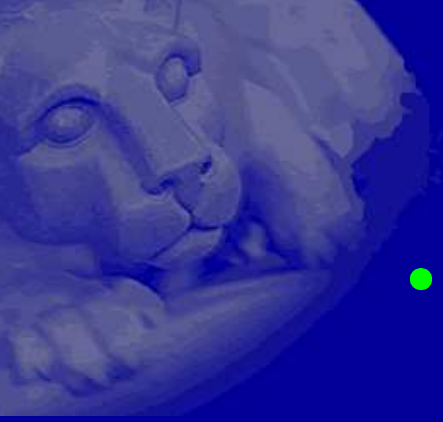
Customize Format/Style

Customize Search Results Format/Style

- Two choices for alternate format: XML and XSLT.
- XML formatted results — good for dynamic applications
- XSLT — translate HTML into your desired format — good for Web visitors

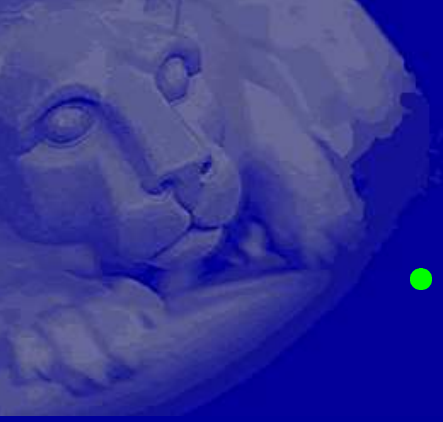
Directions:
<http://aset.its.psu.edu/googledocs/instructions.html#format>

Using XML Results



Using XML Results

- XML is the eXtensible Markup Language. It is a format containing data that can be easily passed between programs and systems.



Using XML Results

- XML is the eXtensible Markup Language. It is a format containing data that can be easily passed between programs and systems.
- You can have a dynamic application query the search engine directly.
  - This application can do offline processing.
  - This application can serve as an intermediary between the Web visitor and the search engine. This allows you to provide your own formatting controls. – Downside: it requires some programming work.

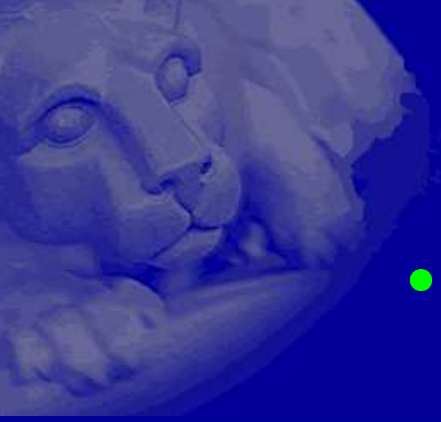
Using XML Results (cont'd)

Directions on receiving XML:

<http://aset.its.psu.edu/googledocs/instructions.html#xml>

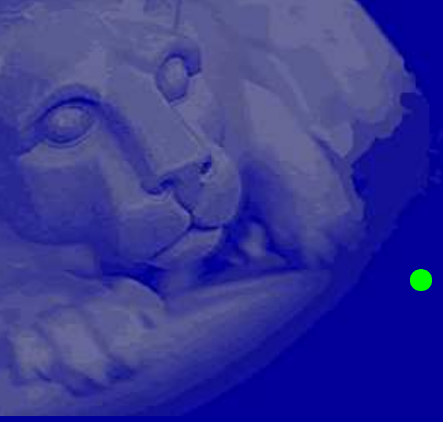
Note: Current users of this feature should compare test results on new appliance. The old and new formats are documented at above URL.

Customizing HTML results



Customizing HTML results

- All search results begin in XML format. The Google Search Appliance “translates” them into another format, such as HTML.



Customizing HTML results

- All search results begin in XML format. The Google Search Appliance “translates” them into another format, such as HTML.
- It uses Extensible Stylesheet Language Transformation (XSLT) documents to translate into a particular format or style. Use the proxystylesheet parameter to specify what XSLT you wish to use.
  - proxystylesheet=PennState means, use the XSLT for the “FrontEnd” called “PennState” (default). The new appliance allows multiple FrontEnds.
  - proxystylesheet=<a URL> means, attempt to read an XSLT file at the address provided (only on old)

Self Serve XSLT/FrontEnd

The self serve XSLT/FrontEnd custom search results tool.

- Basic knowledge of HTML required; knowledge of XSLT *not* required.
- You may design your site style in any program you wish (Dreamweaver, etc), and simply copy/paste the HTML into the Web form.
- XSLT is saved on the server side for you (new engine).

Generator URL with directions:

<http://aset.its.psu.edu/googledocs/custom_style.html>

Final Thoughts

Pictures:

<http://www.personal.psu.edu/jcd/useful/webcon/2006/pics/>

Documentation: <http://aset.its.psu.edu/googledocs/>

Help: search@psu.edu