Google

Google Search Appliance Review Guide

Introduction

Thank you for reviewing the Google Search Appliance. This guide is intended to help reviewers experience and understand the key features and benefits of the Google Search Appliance. The four main topics covered in this document include:

- **Getting Started**

- **Setting Up a Crawl**

- **Testing Search Results**

- **Customizing Search Results**

This guide focuses on the Google Search Appliance GB-1001, for departments, mid-sized companies, and universities. Google offers a range of corporate search products, all of which share the same code-base, including the GB-5005 and the GB-8008.

Google offers extensive documentation to evaluators and customers, so this document will refer you to the *Google Search Appliance Model GB-1001 Installation Guide* for more detailed instructions on how to perform specific tasks. There is also extensive product documentation available via the product's online Help Center that is available on the upper right corner of the Admin Console.

Background

The Google Search Appliance is an integrated hardware/software appliance for corporate intranets and websites.

Google began development of the Google Search Appliance more than three years ago, in response to demand for a product that could be deployed securely behind the firewall. The product was beta tested in early 2001 and launched commercially in February 2002.

The latest version of the Google Search Appliance, version 4.0, includes many new features including a continuous crawler, the ability to create an unlimited number of collections and support for forms-based single sign-on security systems.

Featured Google Search Appliance customers

- FindLaw
- Kaiser Permanente
- Medtronic
- National Semiconductor

- Nextel
- NIC
- PBS
- Sur La Table

- University of Florida
- USATODAY.com
- World Bank
- Xerox

Benefits

The Google Search Appliance is designed to provide customers and employees with relevant search results while minimizing the demands on an IT organization.

Simplicity: Google's philosophy is to make search simple, and we designed the Google Search Appliance to reflect these values. The Google Search Appliance is a ready-to-run system that comes with all the software and hardware needed to power an intranet and/or website search service. Customers are often able to install, configure and deploy the GB-1001 in just a few short hours. The complex tasks of indexing and searching corporate content are automated, and the need to manually configure the search engine is eliminated.

> The Google Search Appliance is built on the same technology used at Google.com and has been optimized for corporate search environments.

Relevancy: The Google Search Appliance utilizes the same search algorithms that power Google.com. As a result, you will find the product delivers highly accurate search results that are easy to understand and reduce the amount of time required to connect users with the information they're looking for. In addition, Google Search Appliance customers benefit from the continued enhancement to Google search technology, as improvements to Google web search technology are regularly added to the appliance.

Cost-Effectiveness: The Google Search Appliance is an affordable corporate search product that is offered with a straightforward cost structure that makes budget planning easy and simple. Google's pricing is structured as a flat rate that includes hardware, software, updates, and all technical support for a two-year period. Prices for the various Google Search Appliance models are based on the number of documents. Total cost of ownership is an important benefit of the Google Search Appliance because the product is significantly easier to configure and maintain than competing products. In fact, once the Google Search Appliance is configured, there is minimal maintenance required to keep it up and running.

Products

The three Google Search Appliance models are:

GB-1001 (reviewer's model)

- Two-unit (2U) server
- Searches up to 1.5 million documents
- Supports up to 300 queries per minute
- Dimensions: 2U Height 3.5" x Width 19.0" x Depth 25.13"
- Other: 56K external data/fax modem for remote diagnostics
- Power: Rating switch full range 90~250V, 50~60Hz, 6A@120V / 3A@240V

GB-5005

- Five-server cluster with built-in load balancing and failover
- Searches up to 3 million documents
- Supports up to 300 queries per minute
- Dimensions: Height 36" x Width 24" x Depth 43"
- Other: 56K external data/fax modem for remote diagnostics
- Power: Rating switch full range 90-250V, 47-63Hz, 250 watts

GB-8008

- Eight-server cluster with built-in load balancing and failover
- Searches up to 15 million documents
- Supports up to 1,000 queries per minute
- Dimensions: Height 36" x Width 24" x Depth 43"
- Other: Two APC switches, network switch, load balancer and 56K modem
- Power: 1,200 watts average, 1,500 watts peak

Reviewing the Google Search Appliance

Getting Started

The Google Search Appliance is unique among corporate search products because administrators aren't required to install software or interact with operating systems, service packs, CD-ROM drives or bulky download files. Plug it in to an electrical outlet, connect it to the network and it turns on automatically. After a minute or two, the Google Search Appliance will play a tune when it's ready to be configured.

Install and configure the Google Search Appliance using a laptop and a browser, as detailed on pages 4 through 18 of the Installation Guide. Product instructions are also available via the Google Search Appliance on-line link in the Admin Console.

Reviewer Notes: When you are configuring your network settings, the Google Search Appliance does a quick network check to make sure your settings are valid. Even the best system administrators can mistype an IP address, so this check helps identify problems easily.
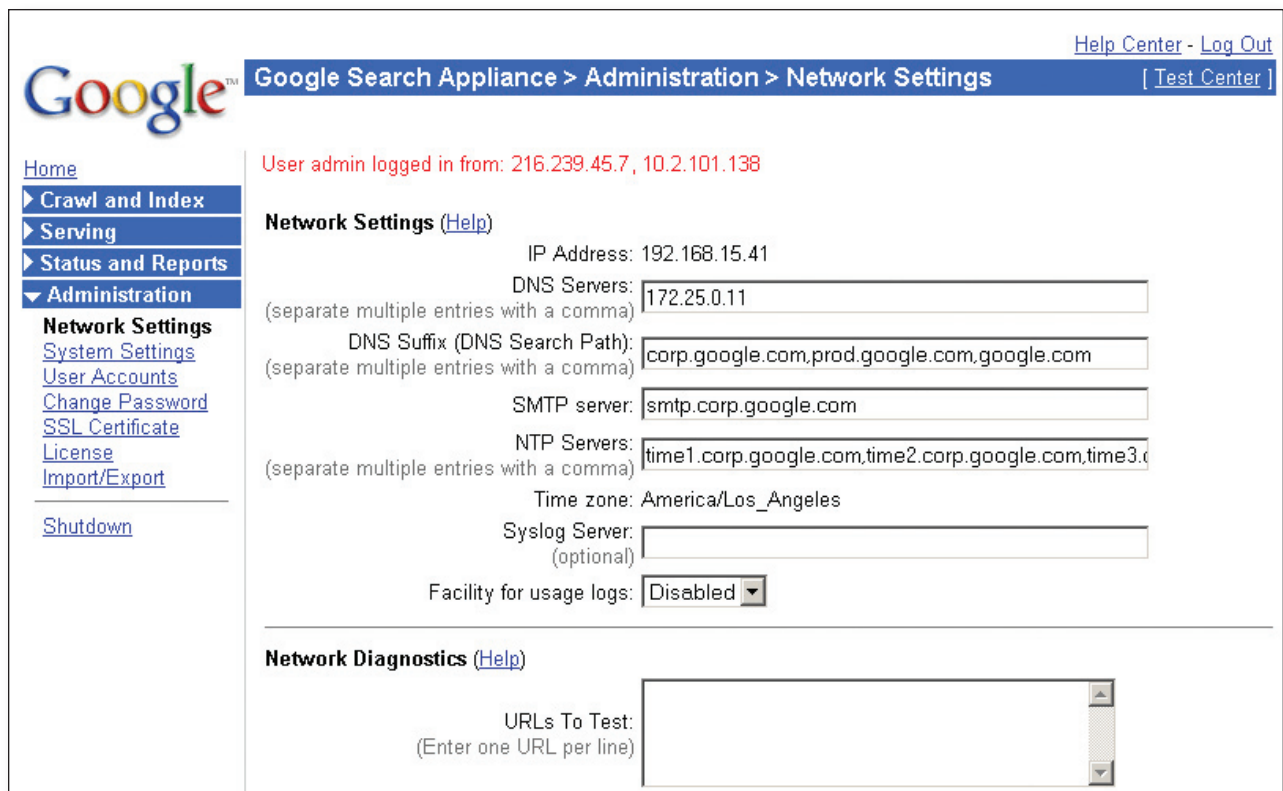
This is an image of the Google installation wizard that will guide you through the set-up of the Google Search Appliance.

Getting Around the Google Search Appliance

The Google Search Appliance is controlled entirely from a browser-based Admin Console. The Google Search Appliance Admin Console enables you to control all aspects of search configuration such as creating collections, managing crawls, customizing search layout pages and running test queries.

Reviewer Notes: You will find fewer controls and options in the Google Admin Console than you might expect. We have simplified the administrative environment and shifted much of the complexity to be automatically handled by the software.

More details regarding the Admin Console are on pages 19 through 26 of the Installation Guide.

Setting Up a Crawl
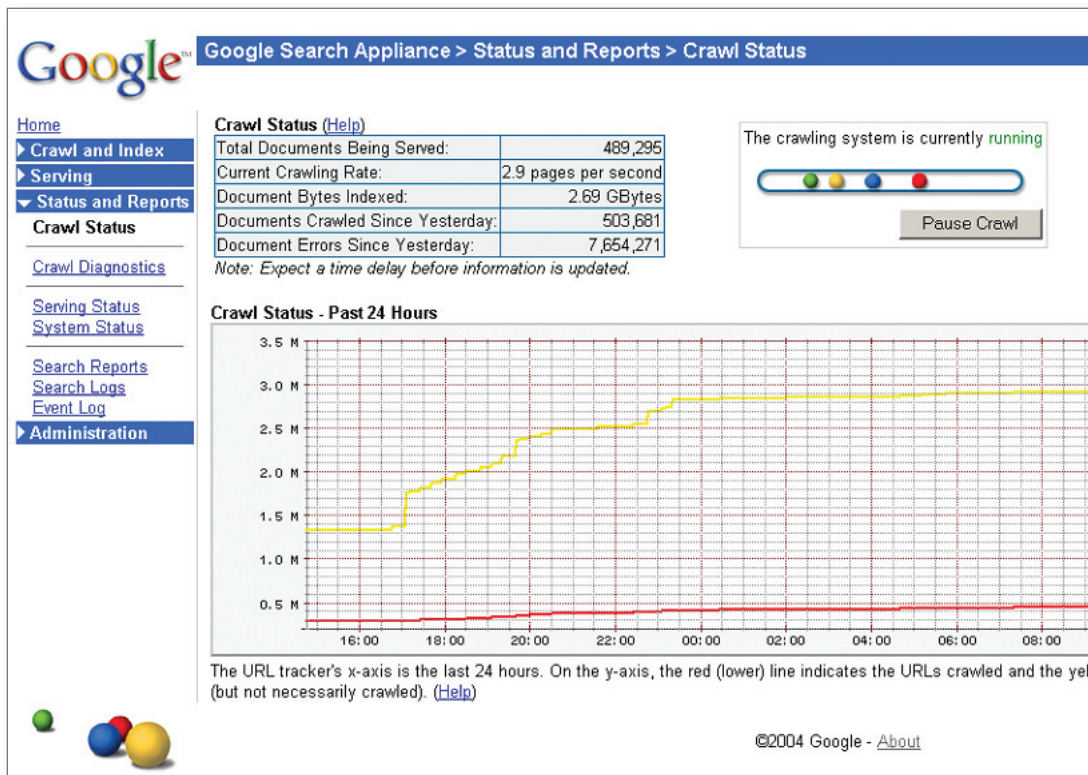
Crawls are set up and managed via the Admin Console. This enables you to manage the configuration of your search index and include which URLs to crawl or which URL patterns, servers or filetypes to exclude from the search index.

The Google Search Appliance crawler automatically reads and indexes a wide variety of document formats, including HTML, PDF, Word and Excel – more than 200 file formats in total. In addition to web servers, the crawler recognizes a variety of document repositories, including a number of corporate knowledge management, content management and application server systems. This includes systems that embed session IDs in URLs, common to many J2EE implementations. The crawler also works with HTTPS and the most prevalent security domains, including NTLM and Basic HTTP Authentication as well as forms-based single sign-on systems.

Reviewer Notes: For your first crawl, we recommend testing a small site or part of a site. This will help you understand the crawling process and will enable you to perform a quick search.

Go to the Configure Crawl tab and enter a start page such as www.example.com/about_us/index.html and set the Follow and Crawl Only URLs patterns to www.example.com/about_us/ (replace "example" with your website name and "about_us" with your path).

Click the Save URLs To Crawl button and then click the Start Crawling button to view the progress as the crawler reads each start page, gathers URLs, etc.

Detailed crawling instructions are found on pages 19 through 27 of the Installation Guide.

Reviewer Notes: For more extensive testing, Google recommends you crawl as much content as possible (preferably 50,000 documents or more). Search result quality improves when users search a broader set of content.

If you are not sure whether a specific page or directory will be included in the index, try the "Test These Patterns" link from the URLs To Crawl page. This is a handy tool that helps administrators test their URL patterns and regular expressions when setting up the crawl.

Testing Search Results

When the index building and replication status is substantially underway, you can perform the first test search. Click the Test Center link in the upper right corner of the Admin Console and a new window will open with the default search form. Although this looks like the standard Google interface, it applies only to the index you have created.

Reviewer Notes: If you have indexed your own publication or website, try looking for your name or a recent article. If you are testing another website, try a related topic or word to gauge how Google Search Appliance results are presented.

You can search for any number of words, and the Google Search Appliance will quickly find every document that matches all those terms.

Refer to pages 25 and 26 of the Installation Guide for more details.

Special Features

In addition to standard search results, the Google Search Appliance offers several features that improve search results for specific situations. Administrators can access the Google Search Appliance menu of special features via the Crawl and Index and Serving sections of the Admin Console. The Google Search Appliance special features are:

- **Automatic Spellchecker:** Google's self-learning spellchecker automatically detects misspellings and suggests corrections – even for names and company-specific terms.

- **Keymatch:** This enables administrators to designate the top search result for a given query.

- **Synonyms:** Most sites have specific terms and abbreviations, and the Google Search Appliance enables administrators to enter synonyms that will be presented to users as alternatives. An example might be "doctor" and "physician," or "PDA" and "handheld."

> Keymatches and synonyms are configured as a feature of "Front Ends." This enables administrators to provide different keyword matches and synonyms to different user groups.
>
> Click the Serving tab and then choose to modify an individual "Front End." Choosing the appropriate tab enables an administrator to enter keymatches and synonyms directly or import them from a spreadsheet or database.
>
> To view keymatches and synonyms, use the Test Center (link in the upper right corner) and be sure that you have selected the relevant Front End from the drop-down menu.

- **Collections:** This feature enables administrators to divide their index into an unlimited number of specific collections that can be searched separately. Collections can be used to segment content in many different ways, including by job function, business unit, location or domain.

> To create a collection, click "Crawl and Index," Collections. On that page, an administrator can enter the location of content to be included in the index and also specify (by location or file type) any content to be excluded from the collection.
>
> **Note:** In creating collections, content is NOT being re-crawled by the Google Search Appliance; collections are simply a segment of the main index. Thus, administrators should ensure that all content that should be in a collection has already been included on the main crawl set-up page.

View the Search Results

As you look at the default results page, you'll see how the Google Search Appliance shows users just what they need to see:
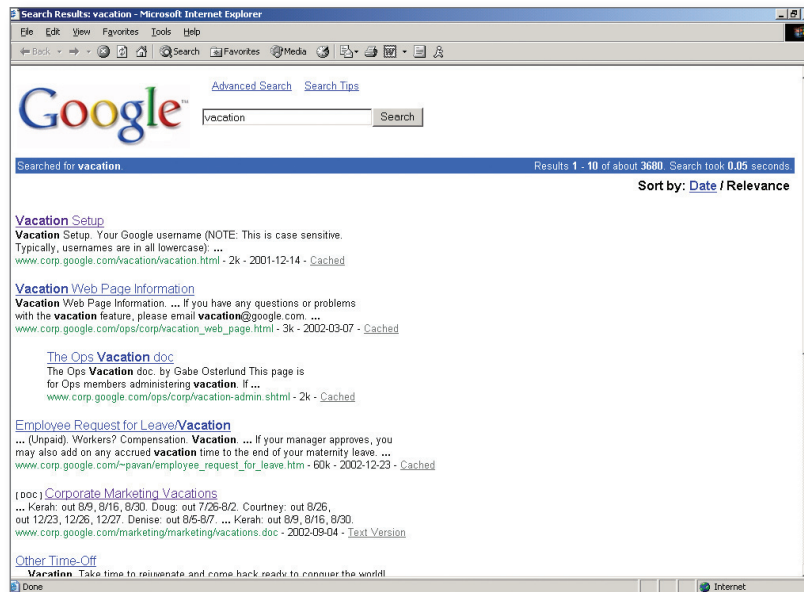
Among other elements, the search results page includes:

- **Relevance Ranking:** Shows the best pages first, based on word matches, links and other criteria.

- **Dynamic Page Summaries:** Dynamically generated snippets that show the match word in context from the document.

- **Sort by Date:** Enables administrators to extract the publication date from a page, even if the web server "Date Modified" is incorrect.

- **Results Grouping:** Clusters results from each subdirectory.

- **Cache Links:** Link to view the original page, with match words marked, even if the publishing server is unavailable.

- **HTML Version:** Link to view documents, without needing the original client application of the file format, using automatic reformatting of more than 220 file types into HTML.

- **Security Control:** Does not display result item if the user doesn't have authorized access.

Customizing Search Results

The default search results format is the standard Google layout, but the Google Search Appliance allows administrators to customize their interface to match the look and feel of their company site.

Reviewer Notes: To customize the Search Box and the Search Results page, click the Serving tab and then select a Front End (or create a new one) to edit. To make simple changes, such as replacing the Google logo with another logo, adding an HTML header and footer, and showing or hiding elements on the results page, use the Page Layout helper (Admin Console, Serving tab, Edit Selected Front End, Page Layout Helper link).

For many more extensive changes, the XSLT style sheet editor allows you to adjust every element. Just click the XSLT Styles link in the Output Format section. For details, see the Help Center Output Format section.

For even more flexibility and integration with local application servers, the Google Search Appliance can return an XML data stream. For more information, see the Help Center XML Protocol appendix.

Search Logs, Reports and Metrics

Search engine usage logs provide valuable information about a website as well as the search engine. The Google Search Appliance provides reports of the number of searches per day and hour, the top 100 searches and so on. In addition, raw search logs preserve information including the query, result, collection, style sheet used and time required. This can be exported for analysis in a spreadsheet, database or script.

Reviewer Notes: To view this information, wait until you have performed several tests. Then choose the Status and Reports tab and the Search Results and/or Search Logs link, and choose today or this month. After a few moments, you'll see the reports appear.

Contact Information

Nathan Tyler
Google Inc.
+1 650 623 4311
nate@google.com