Google Search Appliance

Google for Work Glossary

Google Search Appliance software version 7.2 and later

Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
www.google.com

GSA-GLOS_200.01
March 2015

Google for Work Glossary

This glossary provides definitions to search appliance terms.

A

Admin Console

The web-based user interface that enables administrators to configure a Google Search Appliance. Administrators use the Admin Console to specify or change the settings for crawling, serving, traversing, and monitoring.

advanced search reporting

An Admin Console feature that enables administrators to see what types of links users choose on a search results page, and to track all actions that a user performs such as clicking navigational links.

alert

Email updates that users can receive that provide the latest relevant search results based on a user's topic of interest.

API

An application programming interface.

authentication

The process of verifying a user's identity, using one of several available software mechanisms.

Automatic Self-Learning Scorer

A search appliance feature that automatically analyzes user behavior and the specific links that users click on for specific queries to fine tune relevance and scoring.

authorization

The process of determining whether an authenticated user has the rights to view a particular search result.

B

batch authorization requests

A feature that enables the search appliance to cache *SAML* authorization requests for users. For each user who performs a search query that involves secure content, the search appliance first determines the relevant URLs and then determines whether the user has access to the content. The search appliance makes an authorization request to the appropriate web servers and then stores the authorization data. The search appliance uses the cached authorization information for subsequent searches, making those searches faster.

blacklist

A list of words that cannot be expanded during *query expansion*, but which a search appliance can index and search.

blacklist file

A file of *blacklist* words.

Boolean search

Search queries that include Boolean operators such as: AND, OR, and NOT.

C

cached result

As part of its core technology, Google indexes all the content on a page, rather than just a portion of the content or just meta tags. Each indexed page can be served in a cached HTML format (up to 4 million bytes of each document before HTML conversion). When a user views a cached document, each query term is highlighted in a different color, making the query terms easy to see. Cached pages are always available for view, even if the server where the live content is stored is slow or not responding.

canonical host

In situations where a host is a mirrored server or a host has multiple aliases, one host can be designated as the standard or “canonical” host.

change interval

An estimate of the duration between changes to a URL. A search appliance uses the change interval of a URL to determine when to recrawl the URL.

CMS

See *content management system*.

collection

A segment of a search index. Administrators can divide a search index into collections to show different results to different users; for example, by geography, product, or job function.

collection biasing

A feature that enables search appliance administrators to influence the order of documents in search results based on the documents' memberships in collections.

composite collections

A set of collections from various Google Search Appliances which the primary search appliance in a unified environment can query. Formerly known as "remote composite collections."

connector

Software that provides connectivity between a search appliance and a content management system. A connector enables a search appliance to authenticate, authorize, traverse, and index content from a content management system. Developers can create connectors as Java applications that use the Spring framework (http://www.springsource.org/) for configuration and application parameters.

connector framework

A Google product that consists of the connector manager software, the service provider interface (*SPI*), documentation, and Google support for the connector manager.

connector instance

A programmatic instantiation of a connector for a specific content management system.

connector manager

An open source software package that Google provides that manages creation, instantiation, scheduling, and monitoring of connectors. The connector manager calls the *SPI* methods at stated management system. The connector manager software is provided as open source.

connector.properties file

A file that the connector manager creates and uses to store data from configuration form values. Spring Framework updates the `<property>` tag values in `connectorInstance.xml` file from the `.properties` file.

connector type

Identifies a connector to the connector manager, generates the configuration form that appears in the Admin Console of a search appliance.

content feed

A feed source from a content management system that provides documents, metadata, and a URL to each document's location in the content management system. A content feed requires that a connector traverse the content management system documents, and provide user authentication and authorization services (unless all documents are world readable or a single-sign on system is in place).

content management system

A software system that stores and manages documents and provides document source control services such as securing controlled-access documents and archiving. A content management system consists of a web client, server, management software, and storage of documents. A content management system is also known as a CMS (content management system) or an ECM (enterprise content management system).

content URL

The URL that retrieves content; not necessarily the same as URL that search results display. See also *display URL*.

continuous crawl

A crawl mode in the search appliance that sets the crawler to automatically locate and index content whenever content is updated. See *crawl schedule*.

controlled-access content

Information that a search appliance must not display unless the user who requests the content has provided proper authentication credentials and who has authorization to view the information.

crawl

To search a web site or server for documents and pages to index.

crawl diagnostics

Shows the status of each URL that the search appliance crawled or attempted to crawl.

crawl mode

Whether a search appliance continuously checks its crawl URLs for changed content or crawls the URLs at a scheduled time (known as "scheduled crawl mode").

crawl queue

A list of URLs that the crawler has queued for crawling.

crawl schedule

The times that an administrator designates for a search appliance to crawl URLs for indexing. Administrators can select either continuous crawl, where a crawl occurs after users update content, or scheduled crawl, where a crawl occurs for a fixed time and duration.

credential group

A set of authentication mechanisms that share a username and password. Credential groups enable the search appliance to gather user credentials by using the *Universal Login Form*.

D

date biasing

Enables the search appliance to weigh document dates more heavily when it evaluates the order in which search results appear, and to prefer documents with newer dates to documents with older dates.

date range search

A search that an administrator restricts to return only documents that contain dates that fall within a time frame, or before or after a specified date.

display URL

A URL that appears in search results; not necessarily the same URL that the search appliance uses to retrieve the content. See also *content URL*.

distributed crawl and index replication

See *GSA^n*.

documents

Any content acquired by traversing or crawling. Content can include images, text files, binary files, or other file types. For a complete list of the files that can be indexed by a Google Search Appliance, see the *Indexable File Formats* document.

duplicate host

A web server that replicates the content of another web server. The administrator can create a list of these hosts, because their content does not need to be crawled.

DTD

Document Type Definition. The purpose of a DTD is to define the legal building blocks of an XML document. It defines the XML document structure with a list of legal elements.

dynamic navigation

A feature that helps users refine search results by using metadata. When a user clicks on an dynamic navigation attribute value, the search results are filtered to contain results from the original search query that also have that specific attribute value. The options are refreshed with the attribute values that are applicable to the new result set.

dynamic result cluster

A feature that narrows searches by providing dynamically formed subcategories ("dynamic result clusters") based on the results of each search query. Each subcategory groups similar documents together. Instead of reading through results to understand the results, end users can browse a subcategory.

dynamic scalability

See *GSA Unification*.

E

ECM

See *content management system*.

encoding scheme

Each language has an official encoding scheme which is used to represent all of the language's characters in an 8-bit data stream format. Google search uses encoding schemes to determine how to translate incoming and outgoing search requests.

Enterprise PageRank

Enterprise PageRank is a link analysis algorithm that assigns a numerical weighting to each element of the hyperlinked set of documents in the content for an enterprise, with the purpose of measuring a document's relative importance within the set.

Enterprise PageRank threshold

In the crawl queue, the lowest Enterprise PageRank of a URL that is within the license limit.

entity biasing

Increases or decreases a document's search result score when it contains an entity that matches a specified name:content pair.

entity recognition

A feature that discovers interesting entities in documents with missing or poor metadata and stores these entities in the search index.

excluded URL

A URL that represents a document that is specifically exempt from the crawl. The exclusion can be caused by a robots.txt file, a *URL pattern*.

expert search

A feature that helps your users find experts in your organization. When the user searches on a term, a list of experts for the search term appears in a sidebar next to the search results. The list might include photos, names, and phone numbers. There might also be a more detailed list of experts on a separate page that is linked to the search results page.

external metadata

Document properties originating in or stored in an external source such as a database.

external metadata indexing

Indexing document properties that originate in or are stored in an external source such as a database.

F

federation

See *GSA Unification*.

feed

An XML file that provides a search appliance with sources of data for its search index. A feed file can be either a list of URLs that the appliance searches and periodically recrawls, or a list of URLs and content that the appliance crawls once after the feed file is made available for access.

feeding

The process by how you direct content to the Google Search Appliance instead of having the search appliance locate content. Feeding is a push process, in which the content files are pushed to the Google Search Appliance.

feed client

An application that pushes a feed XML file to a Google Search Appliance.

forms authentication

An authentication rule for controlled-access content sites that the search appliance indexes through a single login form, typically used with a single sign-on (SSO) system. Content accessed through forms authentication can be served as public or secure content. You can only define one forms authentication rule for a search appliance.

freshness tuning

A setting that lets you fine-tune the frequency of crawling for specified URLs. An administrator can set a search appliance to crawl a set of URL patterns more or less frequently. Administrators set the frequency of the crawls based on how often users update content (active content versus archived content).

front end

A user interface for search users. Administrators can change the look and feel of the search and the search result pages. Administrators can customize one or more front ends to display different colors, fonts, and designs. If a company has multiple collections (see *collections*), an administrator can make each front end appear in a different format with its own configuration options.

G

getfields

A parameter sent in the HTTP search request. The `getfields` parameter specifies one or more HTML tags whose content should be returned in the results. (These tags are typically included at the top of a document, providing information about the content in the document.)

Google Apps

Hosted web applications that organizations can use for communication, productivity, and collaboration. Google Apps include Gmail, Google Calendar, Google Sites, and Google Docs.

Google regular expression

Google regular expressions are similar to GNU regular expressions, except that a case insensitive expression starts with the `regexpIgnoreCase:` prefix and a case sensitive expression does not require a prefix, but you can use the `regexpCase:` and `regexp:` prefixes to specify case sensitivity. Google regular expressions also require that you escape special characters with a double backslash (\\).

googleoff/googleon

Special tags that you code into an HTML comment tag that stop and resume the indexing of text on a page. The `googleoff` tag stops a crawler from indexing and the `googleon` tag restarts indexing. For example, `fish <!--googleoff: index-->shark <!--googleon: index-->mackerel`

GSA^n

Combines multiple Google Search Appliances to increase document capacity and to enable single-node replication. GSA^n offers both distributed crawling and mirroring capabilities. With distributed crawling, several search appliances are configured to act as though they are a single search appliance, which greatly increases the number of documents that can be crawled. With mirroring, the search appliance can automatically clone itself. Mirrored search appliances can be used to handle additional query load or can be used as hot backup units that can take over at any point. The mirrored search appliance receives index updates in real time from the primary search appliance, ensuring that the search appliances are always in sync, and that crawling only needs to occur once. GSA^n was formerly known as "distributed crawl and index replication (multibox)."

GSA Unification

A configuration in which a search appliance, known as the primary search appliance, distributes queries to other search appliances, known as the secondary search appliances. The primary search appliance aggregates the results from all of the search appliances in the configuration and serves them to a search user. GSA Unification was formerly known as "dynamic scalability (federation)."

H

head requestor deny rule

A rule that identifies a URL where the content server denies users access with codes other than HTTP code 401 and the access-denied responses that the search appliance expects from the content server.

host load

Specifies the maximum number of concurrent connections open on every web server for crawling. Also known as web server host load.

I

index

To extract information from documents and create an index of terms found in the documents. Index can also mean a list of subjects or words and their locations in a body of text.

J

Jar file

(Java ARchive) A compressed file that contains compiled Java code and other files such as XML files.

K

KeyMatch

Administrator-defined keywords that promote specific web pages on a site. These keywords are associated with targeted URLs, so when search users type the keyword in the search box, they see the targeted URL displays above the main set of search results.

L

language bundle

A collection of resource files that the Google Search Appliance uses for query expansion and spelling in several languages.

M

metacharacter

A special character or special character combination that you can use in a regular expression to match a specific portion of a pattern. See also *regular expression*.

metadata biasing

Influences the display of search results depending on the metadata that is supplied with the documents listed in the search results.

metadata-and-URL feed

A feed source from a content management system that provides metadata and a URL for each document in the content management system.

meta tag

HTML tags that can be specified within an HTML document and that are not displayed to the end user, but which may contain information about the document. Google search uses some meta tags to enhance and filter search results when requested.

MIME

Multipurpose Internet Mail Extensions. The MIME type of a web document (or search result) identifies the format of the document it is associated with. Some sample MIME types include “text/html” for HTML documents, and “application/ms-word” for Microsoft Word documents.

multibox

See *GSA*^{n}.

N

network preparation form

A checklist of values that an administrator provides to configure a Google Search Appliance. The values include subnet mask, IP address, and other values.

number range search

A search that you restrict to only return documents that contain numbers within a specified range. For example, you can specify a range of weights, dimensions, or currencies.

O

OneBox

A search appliance feature that displays application content at the top of search results.

OneBox module

A unit of configuration that is defined in the Admin Console to configure the relationship between a search appliance and a OneBox provider. A OneBox module defines a search type, an optional keyword that invokes the search, and the way that a search appliance obtains and returns information after a user invokes a search.

OneBox provider

Either a collection in a search appliance (internal provider) or an external application that makes data available to a search appliance (external provider).

OneBox results template

See *results template*.

P

PageRank

See *Enterprise PageRank*.

people search

A deprecated feature that helps your users find people in your organization. When a user submits a search query, the search appliance searches any people search source collection that you specify, as well as the search index, and displays people search profile information in a sidebar element next to ranked search results.

per-URL ACL

A per-URL ACL is an access control list that has only a single URL associated with it.

policy ACL

(policy access control list). Enables administrators to specify serve result authorization rules for which users or groups can access which URLs in serve results. A policy ACL rule overrides all other search appliance authorization features.

protected content

See *controlled-access content*.

provider

See *OneBox provider*.

Q

query

Also known as search query. A string of one or more query terms that is submitted to Google search. The results returned satisfy all the query terms by default.

query expansion

A feature that causes search queries to auto-complete and query suggestions to appear when a user types a query in the search box.

query log

See *search log*.

query suggestion

Information that appears at the start of search results to suggest key words to help users refine a search query.

query term

A single term in a query. A single query term cannot contain any spaces or punctuation.

R

ranking framework

A feature that enables search appliance administrators to influence results of rankings programmatically for an unlimited number of URL prefixes.

regular expression

See *Google regular expression*.

related queries

Formerly called “synonyms.” Administrators for the search appliances can use related queries to associate alternative words or phrases with specified search terms. When a user enters the specified search term, the alternative appears as a suggestion.

remote composite collection

See *composite collections*.

remove URLs

A URL that represents a document that is specifically removed from search results by a front end. See also *excluded URL*.

repository

The storage component in a content management system.

result biasing

Influences how a search appliance ranks documents as relevant to a user's search query by tuning how results are scored and displayed.

results page

A page that appears after a search concludes. A results page contains *display URL*s and text from the link. A search results page may also contain a *OneBox module*.

results template

(OneBox) XSL code that specifies how search results, which are returned in XML, are displayed to the user in HTML.

return URL parameter

A parameter that gives a redirect URL's server information about the quickest path back to the search appliance after authentication.

S

SAML

Security Assertion Markup Language (SAML). An access control infrastructure with which the SAML Authentication and Authorization Service Provider Interfaces (SPIs) on a Google Search Appliance communicates.

SAML batch authorization requests

See *batch authorization requests*.

scheduled crawl mode

A feature that enables administrators to specify when a crawl takes place.

search log

A log file that an administrator can create in the Admin Console that lists the IP address of a user that conducts a search, along with a URL that the search appliance creates for the search.

search request

An HTTP GET command issued to the search appliance that includes parameters describing the query and returns the results of the search.

service provider interface

See *SPI*.

servlet container

Informally a web server, but more specifically describes the Java servlet API, which enables use of dynamic documents on a web server.

shard

A search appliance that participates in a *GSA*^{n} configuration. Shards are numbered starting with zero.

SMB URL pattern

A server message block URL pattern that begins with the `smb:` protocol; for example, `smb://fileserver/myshare/mydir/mydoc.txt/`. See also *URL pattern*.

snippets

Small section of text summarizing a search result. Snippets are key phrases that contain query terms in matching documents.

source biasing

Increases or decreases a document's search result score when a document's URL matches a specified pattern.

SPI

Service provider interface that consists of classes and methods that the connector manager calls at stated intervals to facilitate authentication, authorization, and traversal. A developer supplies the logic for each method. Google provides open source code (http://google-enterprise-connector-manager.googlecode.com) for the SPI.

start and follow URLs

Start and follow URLs control where the Google Search Appliance begins crawling content. Google Search Appliance administrators enter start and follow URLs in the **Start Crawling from the Following URLs** section on the **Content Sources > Web Crawl > Start and Block URLs** page in the Admin Console.

static route

An administrator-defined route to a host or network that is not on the default route the search appliance follows to the hosts that it crawls.

stop words

Common words, such as articles, prepositions, and pronouns that are not used in a search when entered in a query.

synonym file

A text file in UTF-8 encoding that contains phrases to use in *query expansion*. A phrase can replace text such as `product abc = product xyz`, which replaces references in a search request from "product abc" to "product xyz". A phrase can append text to a search query using the > operator, such as `xyz123 > Sales`, so that whenever a user searches for the xyz123 part number, Sales is appended to the end of the part number so that the part number can be routed to the correct department. A phrase can be a list of terms in brackets that expand a search to contain additional words. In the phrase `{phone, cell, mobile, telephone}`, if a user searches for phone, the search is expanded to include cell, mobile, and telephone.

T

Test Center

A feature in the Admin Console that you can use to test the output format and search results for a front end or collection. The Test Center displays a search page in a separate window with drop-down menus for the front ends and collections configured in the Admin Console. You can also enter text in the search box and view the results within the Test Center.

traverse

Acquire documents, URLs, and metadata from a content management system for indexing.

trigger

(OneBox) A keyword that, when entered in a search query, causes a search appliance to invoke OneBox results.

trust duration

Specifies how long an authentication mechanism's verification of user credentials will be trusted, in seconds.

trusted application

An application that the search appliance trusts to send pre-validated ids along with end-user's search requests. The search appliance returns secure results without requiring more validation of the user.

An example of a trusted application is a web-based enterprise portal that provides secure access to search using the Google Search Appliance as its engine. If the portal is a trusted application, the only time that end users need to supply credentials is when they log in to the portal.

To register a trusted application with the search appliance, use the **Search > Secure Search > Trusted Applications** page.

U

Universal Login

Universal Login centralizes serve-time authentication for the Google Search Appliance. See also *Universal Login Form* and *credential group*.

Universal Login Form

The primary way the Google Search Appliance gathers user credentials (usernames and passwords). The user's credentials are applied to all the systems in the credential groups for which the user supplies a username and password.

URL pattern

A URL that an administrator specifies as a pattern to match the URLs found by the crawler. URL patterns can be positive to include documents that match, or a negative to exclude documents that match.

URL rewrite rules

Rules that a search appliance follows to rewrite URLs that match a *URL pattern*.

URL status

The status of a URL in the crawl list for a search appliance, indicating whether the content to which a URL points was fetched, was excluded because of a rule, or returned an error.

user results

A feature that gives users the capability of creating moderated results that appear on the results page for specific keyword searches.

UTF-8

Unicode Transformation Format (8-bit). UTF-8 is a Unicode based encoding scheme for describing language data by representing the data using 8-bit codes. Google search uses UTF-8 to support multiple languages simultaneously.

V

version manager

The component used to update the search appliance from one software version to another.

W

war file

(Web Application aRchive). A compressed file that Apache Tomcat uncompresses to create folders and provide jar files. The connector manager is distributed in the `connector-manager.war` file. A war file can be renamed with the .zip file type and its contents and folders examined in the same way that you view a zipped file.

web client

Software that provides web access to:

- An API that a connector uses to acquire metadata, a URL for a document's location, and possibly documents from the content management system. Each document must have a unique document ID. Google recommends that each document also have a last-modified date and a MIME type that corresponds to the files in the *Indexable File Formats* document.
- End users to access the documents in a content management system.

web directory

Files on a web server stored in a directory.

web front end

Another name for the web client.

wildcard search

A feature that enables end users to search by entering a word pattern rather than the exact spelling of a term.

X

XML

eXtensible Markup Language. XML is a markup language, similar to HTML, which was designed to describe data. The tags used in XML are not pre-defined, and are described by a DTD or the data provider.

XSL

eXtensible Stylesheet Language. XSL is a language that is designed to describe how an XML document should be displayed. XSL is used to transform results from XML format into custom HTML output.

XSLT

XSL Transformation. XSLT describes the process of transforming an XML document into another format. The search administrator can use XSLT stylesheets to customize the look and feel of the search results pages.