



Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

# Advanced use of the Google Search Appliance

Sebastian Rahtz

OUCS

July 16th 2008



# Summary

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

- What is a Google Search Appliance?
- How do we use it?
- Configuration
- Giving control to webmasters
- Beyond the safe zone
  - teaching the GSA about keywords and phrases
  - changing the XSL stylesheet which formats the results
  - consuming the raw XML results directly
  - developing addon modules which integrate the GSA with other searches
  - giving the GSA access to protected resources



# What is a Google Search Appliance?

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The GSA is a server in the OUCS machine room. It:

- reads any web page it can reach by starting at `http://www.ox.ac.uk`
- accepts search requests and delivers answers in the manner of big brother Google
- sits outside the Oxford domain
- is a nice yellow sealed box box running Linux to which we have no access except a web-based console
- is open for any Oxford web site to query using their local search form



## Why

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The GSA was requested by the Web Strategy Group to provide a replacement for using the Oxford subset of big brother Google, because:

- we had insufficient control over the appearance
- we could not guarantee removal or addition of pages at short notice
- we had no contract with Google to say that the service would remain free and available
- we could not provide sophisticated sub-site searches

The WSG recognized that the public search interface to Oxford is a vital communication and publicity tool.

Our GSA is on a 2 year licence.



## Things our GSA is not

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

- It does not interact with Big Brother Google to determine hit rating
- It is not indexing Oxford-only IP-restricted sites
- It does not make an archive of Oxford web sites
- It does not have an infinite capacity. We have only paid for 1,000,000 documents



# Concepts and nomenclature

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The GSA manages an unlimited number of:

- collections** (also called **sites**): subsets of the overall index which match a set of URL patterns
- front ends** (also called **clients**): specifications for delivery of results
- stylesheets** (also called **proxystylesheets**) XSLT transformations to present the XML delivered by the system
- users** people who can log in and examine configuration or change settings



# An input form

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

```
<form
  method="get"
  action="http://googlesearch.oucs.ox.ac.uk/search">
  <fieldset>
    <legend>Search</legend>
    <input type="hidden" name="site" value="default_collection">
    <input type="hidden" name="client" value="oxford"/>
    <input type="hidden" name="proxystylesheet" value="oxford">
    <input type="hidden" name="output" value="xml_no_dtd"/>
    <div class="input">
      <input name="q" id="input-
search" value="" type="text"/>
      <input name="Go" value="Go!" type="submit"/>
      <br/>
    </div>
  </fieldset>
</form>
```



# The result URL

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

We type in 'cats dogs' and get sent to:

```
http://googlesearch.oucs.ox.ac.uk/search  
?site=default_collection  
&client=oxford  
&proxystylesheet=oxford  
&output=xml_no_dtd  
&q=cats+dogs
```





# Result page

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



search  
UNIVERSITY OF OXFORD



## University of Oxford Search — TEST

[Advanced Search](#)  
[Search Tips](#)

Advanced Search Results 1 - 10 of about 122 for **cats dogs**. Search took 0.03 seconds.

[Next >](#)

[Sort by date](#) / [Sort by relevance](#)

[PDF] [Specimen 1999](#)

... Section A [30 marks] 1. In English, most nouns form their plurals by the addition of -s : eg **cats, dogs**, cows, horses. This plural ...

[www.classics.ox.ac.uk/admissions/langapt\\_questions.pdf](http://www.classics.ox.ac.uk/admissions/langapt_questions.pdf) - 2005-11-09 - [Text Version](#)

[RTF] [Specimen 1999](#)

... Section A [30 marks]. 1. In English, most nouns form their plurals by the addition of -s : eg **cats, dogs**, cows, horses. This plural ...

[www.classics.ox.ac.uk/admissions/langapt\\_questions.rtf](http://www.classics.ox.ac.uk/admissions/langapt_questions.rtf) - 2005-11-09 - [Text Version](#)

[PDF] [Microsoft PowerPoint - loE morpheme day.ppt](#)

... sounds (but the same morpheme) are spelled the same: wanted called kissed **cats dogs**  
heal health tambor tamborilar martelo martelar Page 6. Page 7. Page 8. ...

[www.edstud.ox.ac.uk/.../general%20literacy%20presentations/theplaceofmorphologyindevelopment2005.pdf](http://www.edstud.ox.ac.uk/.../general%20literacy%20presentations/theplaceofmorphologyindevelopment2005.pdf) - 2006-03-20 - [Text Version](#)

[PDF] [Br leff ng](#)

... There are no plans for any farm animals to be housed in the new building, and the



## The XML returned (1)

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

```
<GSP VER="3.2">
  <TM>0.009117</TM>
  <Q>food</Q>
  <PARAM name="filter" value="1" original_value="1"/>
  <PARAM name="access" value="p" original_value="p"/>
  <PARAM name="entqr" value="0" original_value="0"/>
  <PARAM name="Go" value="Go!" original_value="Go!"/>
  <PARAM name="domains" value="ox.ac.uk" original_value="ox.ac.uk"/>
  <PARAM name="output" value="xml_no_dtd" original_value="xml_no_dtd"/>
  <PARAM name="sort" value="date:D:L:d1" original_value="date%3AD%3AL%3Ad1"/>
  <PARAM name="site" value="oucs" original_value="oucs"/>
  <PARAM name="ie" value="UTF-8" original_value="UTF-8"/>
  <PARAM name="client" value="oxford" original_value="oxford"/>
  <PARAM name="q" value="food" original_value="food"/>
  <PARAM name="ip" value="129.67.100.16" original_value="129.67.100.16"/>
  <RES SN="1" EN="10">
    <M>54</M>
    <FI/>
    <NB>
```



## The XML returned (2)

Advanced use  
of the Google  
Search  
Appliance


Sebastian  
Rahtz

```
<R N="5">
  <U>http://www.oucs.ox.ac.uk/ltg/projects/jtap/rose/letters.
  <UE>http://www.oucs.ox.ac.uk/ltg/projects/jtap/rose/letters
  <T>Rosenberg&#39;s Letters</T>
  <RK>7</RK>
  <FS NAME="date" VALUE="2005-07-22"/>
  <S>  <b>...</b> Except that the <b>food</b> is
  unspeakable, and perhaps luckily, scanty, the rest<br>
  is pretty tolerable. I have <b>food</b> sent up from
  home and <b>...</b>  </S>
  <LANG>en</LANG>
  <HAS>
    <L/>
    <C SZ="23k" CID="cE1498LlUfWJ" ENC="ISO-8859-1"/>
  </HAS>
</R>
<R N="6">
  <U>http://www.oucs.ox.ac.uk/email/oxford/index.xml.ID=body.
  <UE>
  http://www.oucs.ox.ac.uk/email/oxford/index.xml.ID%3Dbody.
  </UE>
  <T>[oucs] Oxford Email Addresses: 14. History -
  Long-form Addresses</T>
  <RK>7</RK>
```

# default stylesheet

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

 [Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#) [Appliance](#)

Search:   [Advanced Search](#)  
[Search Tips](#)

Search:  public content  public and secure content

---

**Search** Results 1 - 10 of about 54 for **food**. Search took 0.01 seconds.

[Next >](#) [Sort by date](#) / Sort by relevance

[MS WORD] [Job Description](#)  
... other items: **Food** handling - the safe handling and disposal of **food** when required;  
Liaising with contract cleaners; Holiday / sickness ...  
[www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc - 2008-07-03 - Text Version](#)

[PDF] [Job Description](#)  
... Cleaning kitchen equipment, maintaining fridges, coffee vending machines, microwave  
and other items • **Food** handling - the safe handling and disposal of ...  
[www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf - 2008-07-03 - Text Version](#)

[oucs] [Further Webmail: 8. Changing Your Sending Options](#)  
... So from the example of Mary at **Food** Studies and Balliol, she could choose that others  
see the Balliol address in preference to her official **food-studies** address ...  
[www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body\\_1\\_div.8 - 15k - 2008-06-27 - Cached](#)

[MS EXCEL] [OUCS RESTAURANT - A proposal](#)  
... 1, **OUCS** RESTAURANT - A proposal, 2, 3, Input Area, 4, Fixed Costs /Week, 5, 6, Rent,  
**Food** Cost, 7, Rates, Meal Price, 8, Bills, 9, Wages, No of Weeks, 10, 11, Output ...  
[www.oucs.ox.ac.uk/coursematerials/excel/sy\\_xlt - 2005-07-22 - Text Version](#)

[Rosenberg's Letters](#)  
... Except that the **food** is unspeakable, and perhaps luckily, scanty, the rest  
is pretty tolerable. I have **food** sent up from home and ...  
[www.oucs.ox.ac.uk/itg/projects/jtap/rose/letters.html - 23k - 2005-07-22 - Cached](#)

[oucs] [Oxford Email Addresses: 14. History - Long-form Addresses](#)  
... For example, Mary Brown could have four addresses: mary.brown@**food-studies**.oxford.  
ac.uk (long form) mary.brown@**food**.ox.ac.uk (short form) mary.brown@enstone ...  
[www.oucs.ox.ac.uk/email/oxford/index.xml.ID=body\\_1\\_div.14 - 14k - 2008-06-27 - Cached](#)

[oucs] [Oxford Email Addresses: 3. Email Addresses in Oxford](#)  
... addressed. For example, Mary Brown is a member of staff at the Department  
of **Food** Studies and is a fellow of Enstone College. She ...  
[www.oucs.ox.ac.uk/email/oxford/index.xml.ID=body\\_1\\_div.3 - 14k - 2008-06-27 - Cached](#)  
[ [More results from www.oucs.ox.ac.uk/email/oxford](#) ]

[oucs] [Guide for Experienced Email Users: 3. Oxford Email ...](#)  
... chris.jones@chch.ox.ac.uk Graduate and staff users also get an email address relating  
to their departmental affiliation, for example chris.jones@**food**.ox.ac.uk ...  
[www.oucs.ox.ac.uk/email/quickstart/index.xml.ID=body\\_1\\_div.3 - 12k - 2008-06-28 - Cached](#)



# admin stylesheet

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



## Results for University of Oxford

Results 1 - 10 of about 54 for **food**. Search took 0.01 seconds.

### Internal search

Search **Oxford Only** restricted pages

[Sort by date](#), Sort by relevance.

#### [MS WORD] Job Description

... other items; **Food** handling - the safe handling and disposal of **food** when required;  
Liaising with contract cleaners; Holiday / sickness ...

[www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc) - 2008-07-03 - Text Version

#### [PDF] Job Description

... Cleaning kitchen equipment, maintaining fridges, coffee vending machines, microwave  
and other items • **Food** handling - the safe handling and disposal of ...

[www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf) - 2008-07-03 - Text Version

#### [oucs] Further Webmail: 8. Changing Your Sending Options

... So from the example of Mary at **Food** Studies and Balliol, she could choose that others  
see the Balliol address in preference to her official **food**-studies address ...

[www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body\\_1\\_div.8](http://www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body_1_div.8) - 15k - 2008-06-27 - Cached

#### [MS EXCEL] OUCS RESTAURANT - A proposal

... 1. OUCS RESTAURANT - A proposal, 2, 3, Input Area, 4. Fixed Costs /Week, 5, 6, Rent,  
**Food** Cost, 7, Rates, Meal Price, 8, Bills, 9, Wages, No of Weeks, 10, 11, Output ...

[www.oucs.ox.ac.uk/coursematerials/excel/xy.xls](http://www.oucs.ox.ac.uk/coursematerials/excel/xy.xls) - 2005-07-22 - Text Version

#### Rosenberg's Letters

... Except that the **food** is unspeakable, and perhaps luckily, scanty, the rest  
is pretty tolerable. I have **food** sent up from home and ...

[www.oucs.ox.ac.uk/ftg/projects/jtap/rose/letters.html](http://www.oucs.ox.ac.uk/ftg/projects/jtap/rose/letters.html) - 23k - 2005-07-22 - Cached

#### [oucs] Oxford Email Addresses: 14. History - Long-form Addresses

... For example, Mary Brown could have four addresses: mary.brown@**food**-studies.oxford.  
ac.uk (long form) mary.brown@**food**.ox.ac.uk (short form) mary.brown@enstone ...

[www.oucs.ox.ac.uk/email/oxford/index.xml?ID=body\\_1\\_div.14](http://www.oucs.ox.ac.uk/email/oxford/index.xml?ID=body_1_div.14) - 14k - 2008-06-27 - Cached

Insert Oxford Email Addresses - 2. Email Addresses in Oxford

# oucs stylesheet

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



## University of Oxford Search — OUCS

[Advanced Search](#)  
[Search Tips](#)

Advanced Search Results 1 - 10 of about 54 for **food**. Search took 0.36 seconds.

[Next](#)

[Sort by date](#) / [Sort by relevance](#)

### [MS WORD] [Job Description](#)

... other items; **Food** handling - the safe handling and disposal of **food** when required, Liaising with contract cleaners; Holiday / sickness ...

[www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc) - 2008-07-03 - [Text Version](#)

### [PDF] [Job Description](#)

... Cleaning kitchen equipment, maintaining fridges, coffee vending machines, microwave and other items • **Food** handling - the safe handling and disposal of ...

[www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf) - 2008-07-03 - [Text Version](#)

### [oucs] [Further Webmail: 8. Changing Your Sending Options](#)

... So from the example of Mary at **Food** Studies and Balliol, she could choose that others see the Balliol address in preference to her official **food**-studies address ...

[www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body\\_1\\_div\\_8](http://www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body_1_div_8) - 15k - 2008-06-27 - [Cached](#)

### [MS EXCEL] [OUCS RESTAURANT - A proposal](#)

... 1, OUCS RESTAURANT - A proposal, 2, 3, Input Area, 4, Fixed Costs /Week, 5, 6, Rent, **Food** Cost, 7, Rates, Meal Price, 8, Bills, 9, Wages, No of Weeks, 10, 11, Output ...

[www.oucs.ox.ac.uk/coursematerials/excel/ky.xls](http://www.oucs.ox.ac.uk/coursematerials/excel/ky.xls) - 2005-07-22 - [Text Version](#)

### [Rosenberg's Letters](#)

... Except that the **food** is unspeakable, and perhaps luckily, scanty, the rest

# o um stylesheet



## Oxford University Museum of Natural History



Home

Visiting Us Collections & Research Teaching & Learning Services News

### Navigate

Contact us  
Accessibility  
Sponsors  
Copyright  
Site map



### Site search

### Search

Results 1 - 10 of about 54 for **food**. Search took 0.31 seconds.

#### [MS WORD] Job Description

... other items: **Food** handling - the safe handling and disposal of **food** when required; Liaising with contract cleaners; Holiday / sickness ...  
[www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc) - 2008-07-03 - **Text Version**

#### [PDF] Job Description

... Cleaning kitchen equipment, maintaining fridges, coffee vending machines, microwave and other items • **Food** handling - the safe handling and disposal of ...  
[www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf) - 2008-07-03 - **Text Version**

#### [oucs] Further Webmail: 8. Changing Your Sending Options

... So from the example of Mary at **Food** Studies and Balliol, she could choose that others see the Balliol address in preference to her official **food**-studies.oxford.ac.uk ...  
[www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body\\_1\\_div.8-15k](http://www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body_1_div.8-15k) - 2008-06-27 - **Cached**

#### [MS EXCEL] OUCS RESTAURANT - A proposal

... 1, OUCS RESTAURANT - A proposal, 2, 3, Input Area, 4, Fixed Costs /Week, 5, 6, Rent, **Food** Cost, 7, Rates, Meal Price, 8, Bills, 9, Wages, No of Weeks, 10, 11, Output ...  
[www.oucs.ox.ac.uk/coursematerials/excel/xy.xls](http://www.oucs.ox.ac.uk/coursematerials/excel/xy.xls) - 2005-07-22 - **Text Version**

#### Rosenberg's Letters

... Except that the **food** is unspeakable, and perhaps luckily, scanty, the rest is pretty tolerable. I have **food** sent up from home and ...  
[www.oucs.ox.ac.uk/itg/projects/jtap/rosa/letters.html](http://www.oucs.ox.ac.uk/itg/projects/jtap/rosa/letters.html) - 23k - 2005-07-22 - **Cached**

#### [oucs] Oxford Email Addresses: 14. History - Long-form Addresses

... For example, Mary Brown could have four addresses: mary.brown@**food**-studies.oxford.ac.uk (long form) mary.brown@**food**.ox.ac.uk (short form) mary.brown@enstone ...  
[www.oucs.ox.ac.uk/email/oxford/index.xml?ID=body\\_1\\_div.14-14k](http://www.oucs.ox.ac.uk/email/oxford/index.xml?ID=body_1_div.14-14k) - 2008-06-27 - **Cached**

#### [oucs] Oxford Email Addresses: 3. Email Addresses in Oxford

... addressed. For example, Mary Brown is a member of staff at the Department of **Food** Studies and is a Fellow of Estates Policies. She



Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz




# oum-learning stylesheet

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

Oxford University Museum of Natural History

## The Learning Zone



Home Animals Fossils Insects Minerals Rocks Funstuff

Search the learning zone

food

**Search** Results 1 - 10 of about 54 for **food**. Search took 0.02 seconds.

[Next >](#)

[MS WORD] [Job Description](#)  
... other items; **Food** handling - the safe handling and disposal of **food** when required; Liaising with contract cleaners; Holiday / sickness ...  
[www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.doc) - 2008-07-03 - [Text Version](#)

[PDF] [Job Description](#)  
... Cleaning kitchen equipment, maintaining fridges, coffee vending machines, microwave and other items - **Food** handling - the safe handling and disposal of ...  
[www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf](http://www.oucs.ox.ac.uk/jobs/CleaningAssistant.pdf) - 2008-07-03 - [Text Version](#)

[JOURNALS] [Further Webmail: 8. Changing Your Sending Options](#)  
... So from the example of Mary at **Food** Studies and Balliol, she could choose that others see the Balliol address in preference to her official **food-studies** address ...  
[www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body\\_1\\_dir.8](http://www.oucs.ox.ac.uk/email/webmail/manual.xml?ID=body_1_dir.8) - 15k - 2008-06-27 - [Cached](#)

[MS EXCEL] [OUCS RESTAURANT - A proposal](#)  
... 1, **OUCS RESTAURANT** - A proposal, 2, 3, Input Area, 4, Fixed Costs /Week, 5, 6, Rent, **Food** Cost, 7, Rates, Meal Price, 8, Bills, 9, Wages, No of Weeks, 10, 11, Output ...  
[www.oucs.ox.ac.uk/coursematerials/excel/xy.xls](http://www.oucs.ox.ac.uk/coursematerials/excel/xy.xls) - 2005-07-22 - [Text Version](#)

[Rosenberg's Letters](#)  
... Except that the **food** is unspeakable, and perhaps luckily, scanty, the rest is pretty tolerable. I have **food** sent up from home and ...  
[www.oucs.ox.ac.uk/itg/projects/jtap/rose/letters.html](http://www.oucs.ox.ac.uk/itg/projects/jtap/rose/letters.html) - 23k - 2005-07-22 - [Cached](#)

[JOURNALS] [Oxford Email Addresses: 14. History - Long-form Addresses](#)  
... For example, Mary Brown could have four addresses: mary.brown@**food-studies**.oxford.ac.uk (long form) mary.brown@**food**.ox.ac.uk (short form) mary.brown@enstone ...  
[www.oucs.ox.ac.uk/itg/projects/jtap/rose/addresses.html](http://www.oucs.ox.ac.uk/itg/projects/jtap/rose/addresses.html) - 4.8k - 2005-07-22 - [Cached](#)





# Simple XSL (1)

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

```
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:template match="/GSP">
    <html>
      <head>
        <title>Google Search Appliance results</title>
      </head>
      <body>
        <h2>Results from Google Search</h2>
        <ul>
          <li>Query:<xsl:value-of se-
            lect="PARAM[@name='q']/@original_value"/>
          </li>
          <li>Site:<xsl:value-of
            select="PARAM[@name='site']/@original_value"/>
          </li>
          <li>Clien:<xsl:value-of
            se-
            lect="PARAM[@name='client']/@original_value"/>
          </li>
        </ul>....</body>
      </html>
    </xsl:template>
  </xsl:stylesheet>
```



# Simple XSL

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

```
<table
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <tr>
    <td>Title</td>
    <td>Context</td>
    <td>URL</td>
    <td>Crawl date</td>
  </tr>
  <xsl:for-each select="RES/R">
    <tr>
      <td>
        <xsl:value-of select="T" disable-output-
          escaping="yes"/>
      </td>
      <td>
        <xsl:value-of select="S" disable-output-
          escaping="yes"/>
      </td>
      <td>
        <a href="UE">
          <xsl:value-of select="U"/>
        </a>
      </td>
      <td>
```



# Simple stylesheet output

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

## Results from Google Search

- Query: food
- Site: default\_collection
- Clien: oucs-test

| Title                                     | Context                                                                                                                                                    | URL                                                                                                             | Crawl date  |
|-------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|-------------|
| 1 club : <b>food</b> & drink              | Club : <b>food</b> & drink. ... lunch and afternoon. café-bar we offer hot <b>food</b> from the café during the week and from the bar at the weekends: ... | <a href="http://www.club.ox.ac.uk/catering/">http://www.club.ox.ac.uk/catering/</a>                             |             |
| 2 magdalen > college life > <b>food</b>   | ... available. Prices are set just above basic <b>food</b> costs. Cooked breakfast, lunch and dinner are available, and no booking is required. ...        | <a href="http://www.magd.ox.ac.uk/college_life/food.shtml">http://www.magd.ox.ac.uk/college_life/food.shtml</a> |             |
| 3 <b>Food</b> and drink - Wolfson College | Wolfson College. Home   Catering. <b>Food</b> and drink. Lunch and Dinner. Informal                                                                        | <a href="http://www.wolfson.ox.ac.uk/catering/">http://www.wolfson.ox.ac.uk/catering/</a>                       | 15 Jul 2008 |

---



# A collection

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

## This specifies

- URL patterns which should be matched (could be anything in Oxford)
- URL patterns which should be excluded



# What are URL patterns?

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

| Valid URL Patterns                                                                                | Examples                                                                             | Explanation                                                                                                      |
|---------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| Any <b>substring</b> of a URL that includes the host/path separating slash                        | <code>http://www.ox.ac.uk/</code>                                                    | Any page on <code>www.ox.ac.uk</code> using the HTTP protocol.                                                   |
| Any <b>suffix</b> of a string. You specify the suffix with the \$ at the end of the string.       | <code>home.html\$</code>                                                             | All pages ending with <code>home.html</code> .                                                                   |
| Any <b>prefix</b> of a string. You specify the prefix with the ^ at the beginning of the string.  | <code>^https://</code>                                                               | Any page using the HTTPS protocol.                                                                               |
| An <b>arbitrary substring</b> of a URL. These patterns are specified using the prefix "contains". | <code>contains:coffee</code>                                                         | Any URL that contains "coffee."                                                                                  |
| <b>Exceptions</b> denoted by - (minus) sign.                                                      | <code>cheese.ox.ac.uk/</code><br><code>-</code><br><code>www.cheese.ox.ac.uk/</code> | Means that " <code>cheese.ox.ac.uk</code> " is a match, but " <code>www.cheese.ox.ac.uk</code> " is not a match. |



# What are URL patterns? (more)

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

**Regular expressions** from the GNU Regular Expression library.

**Comments**

```
#this is a comment
```

See the GNU Regular Expression library.

Empty lines and comments starting with # are permissible. These comments are removed from the URL pattern and ignored.

```
# Law School PHP is trusted
-regex:^http://denning.law.ox.ac.uk.*php ?.*
# mysource matrix cms - exclude 'str1?str2=str3'
regexIgnoreCase:^http://www.chinacentre.ox.ac.uk/
[-a-z0-9_/.]+?[-a-z0-9_]+=
```



# Definition of a collection

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

[Back to List of All Collections](#)

Edit Collection:

Note: The default `_collection` is by default defined with the pattern `"/*"`, which will display search results for the entire Search Appliance index. For administration purposes it is helpful to have at least one collection with this pattern; it will allow the administrator to see all indexed URLs in the Crawl Diagnostics under [Status and Reports > Crawl Diagnostics](#) by selecting this collection.

**Include Content Matching the Following Patterns:** ([Help](#) - [Test these patterns](#))

```
medsci.ox.ac.uk/  
nda.ox.ac.uk/  
cardiov.ox.ac.uk/  
ndcls.ox.ac.uk/  
ndm.ox.ac.uk/  
cineuro.ox.ac.uk/  
clinpharm.ox.ac.uk/
```

*example: <http://www.mycompany.com/engineering/>*

---

**Do Not Include Content Matching the Following Patterns:** ([Help](#) - [Test these patterns](#))

```
contains:acl_users  
.js$
```



## Configuration: starting points

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The GSA has been told to start at <http://www.ox.ac.uk>  
and follow links as far as it can, within the following domains:

---

[ox.ac.uk/](http://ox.ac.uk/)  
[malariagen.net/](http://malariagen.net/)  
[oss-watch.ac.uk/](http://oss-watch.ac.uk/)  
[www.cricketintheparks.org.uk/](http://www.cricketintheparks.org.uk/)  
[www.ethox.org.uk/](http://www.ethox.org.uk/)  
[www.gmap.net/oxford/](http://www.gmap.net/oxford/)  
[www.gprg.org/](http://www.gprg.org/)  
[www.isis-innovation.com/](http://www.isis-innovation.com/)  
[www.ntrac.org.uk/](http://www.ntrac.org.uk/)  
[www.octo-oxford.org.uk/](http://www.octo-oxford.org.uk/)  
[www.oushop.com/](http://www.oushop.com/)  
[www.oxfordlimited.co.uk/](http://www.oxfordlimited.co.uk/)  
[www.ww1lit.com/](http://www.ww1lit.com/)  
[www.conference-oxford.com/](http://www.conference-oxford.com/)  
[oxforduniversity.newcomersclub.googlepages.com/](http://oxforduniversity.newcomersclub.googlepages.com/)

More can be added





## Configuration: searching

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

By default the GSA indexes every document it can find, including binary documents such as PDF, Word and Powerpoint.

The exceptions are:

- 1 all graphic, music, and font formats
- 2 all executable programs and library files
- 3 software distributions and other archives
- 4 pages clearly personal (pictures of cats)
- 5 dynamically-generated calendars
- 6 dynamically-generated search templates with no content
- 7 personal pages on users.ox.ac.uk
- 8 endless queries which seem unlikely to be of use, eg those monitoring network activity



# Where does our love go?

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The following table lists the top sites as of 2008-07-15, in descending order of size.

|                          |       |
|--------------------------|-------|
| people.maths.ox.ac.uk    | 44364 |
| www.ashmus.ox.ac.uk      | 33356 |
| fenix.ouls.ox.ac.uk      | 21873 |
| www-pnp.physics.ox.ac.uk | 21571 |
| web.comlab.ox.ac.uk      | 20276 |
| griffith.ashmus.ox.ac.uk | 16974 |
| www.griffith.ox.ac.uk    | 14471 |
| www.ox.ac.uk             | 14379 |
| www.mansfield.ox.ac.uk   | 13311 |
| www.maths.ox.ac.uk       | 13115 |
| dps.plants.ox.ac.uk      | 12614 |
| www.comlab.ox.ac.uk      | 12115 |
| ptcl.chem.ox.ac.uk       | 11541 |
| www.oucs.ox.ac.uk        | 10525 |
| www.fmrrib.ox.ac.uk      | 8910  |
| web2.comlab.ox.ac.uk     | 8684  |
| www.chem.ox.ac.uk        | 8450  |
| www.stats.ox.ac.uk       | 7646  |



# Examination of details


Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The box's admin interface allows the administrators to examine the details of these and all other sites, down to the file level. For example:

[All hosts](#) > <http://www.fmrib.ox.ac.uk/fs/faq/scripting.html>

#### More information about this page

- [Link to this page](#)
- [Cached version](#)
- PageRank: 
- Last successful crawl:
  - Time: 09 Mar 3:22 PM
  - Authentication method: None
- Number of links on this page to crawled pages: 3
- [View list of public crawled pages that link to this page](#)
- [View list of all crawled pages that link to this page: 7 pages.](#)
- This page is in the following collections:
  - default\_collection

The GSA has its own algorithm to decide how often to revisit a page, looking at how often it changes. Pages are typically looked at once every day or two, but this can be speeded up or slowed down.

# Excluded patterns (1): default setup

Advanced use  
of the Google

Search  
Appliance

Sebastian  
Rahtz

```
#: Images      .mov$          /?S=A$        contains:\006  .htm/$
.gif$         .mpg$          /?S=D$        contains:\007  .phtml/$
.jpg$         .mpeg$        /?D=A$        contains:\010  .ghtml/$
.jpeg$        .mp3$         /?D=D$        contains:\011  .asp/$
.png$        .ogg$         /?M=A$        contains:\012  .jsp/$
.jpe$         .dat$         /?M=D$        contains:\013  .shtml/$
.pcx$         .dta$         /?N=A$        contains:\014  !/
.tif$         .log$         /?N=D$        contains:\015  "/
.tiff$        .lst$         /?C=N&O=A$    contains:\016  $/
.bmp$         .bz2$         /?C=M&O=A$    contains:\017  %/
.dll$         .jar$         /?C=S&O=A$    contains:\020  &/
.exe$         .arj$         /?C=D&O=A$    contains:\021  '/
.a$           .cab$         /?C=N&O=D$    contains:\022  (</
.o$           .rar$         /?C=M&O=D$    contains:\023  )/
.so$          .rpm$         /?C=S&O=D$    contains:\024  +/
.bin$         .tar$         /?C=D&O=D$    contains:\025  ./
.class$      .zip$         /?C=N;O=A$    contains:\026  ./
.ttf$        .tar.gz$      /?C=M;O=A$    contains:\027  &lt;/
.pfb$        .upp$         /?C=S;O=A$    contains:\030  =/
.pfm$        .tgz$         /?C=D;O=A$    contains:\031  &gt;/
.afm$        .sdd$         /?C=N;O=D$    contains:\032  {/
.hqx$        .hdr$         /?C=M;O=D$    contains:\033  |/
.sea$        .iso$         /?C=S;O=D$    contains:\034  }/
.eps$        .img$         /?C=D;O=D$    contains:\035  ~/
.ai$         .gpg$         contains:\001  contains:\036  [/
.ram$        .gbk$         contains:\002  contains:\037  \\  
.wav$        .fac$         contains:\003  contains:\040  ]/  
.avi$        .ghg$         contains:\004  contains:\177  ^/  
.mid$        .mdic$        contains:\005  .html/$
```



## Excluded patterns (2): added locally

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

---

```
# don't index personal pages or CGI
users.ox.ac.uk/~
users.ox.ac.uk/cgi-bin/
# assorted database accesses which go on for ever
www.chem.ox.ac.uk/timetableweek.asp
cms.ouls.ox.ac.uk/law/e-resources_and_guides/databases/
etcsl.orinst.ox.ac.uk/cgi-bin
external.materials.ox.ac.uk/private/
foodweb.hertford.ox.ac.uk/main/
herbaria.plants.ox.ac.uk/vfh/image/
library.ox.ac.uk/find?
linacre.ox.ac.uk/forum/
manageserver.physics.ox.ac.uk/cgi-bin/
mhs.ox.ac.uk/epact/
ora.ouls.ox.ac.uk/access/
poinikastas.csad.ox.ac.uk/4DLink3/
portal.imm.ox.ac.uk/booking
scm2005.chem.ox.ac.uk/gallery2/
vindolanda.csad.ox.ac.uk/4DLink2/
www.ashmus.ox.ac.uk/ash/cis/Searches/searches/
www.lincoln.ox.ac.uk/component/option,com_events/
www.oppf.ox.ac.uk/pn/?POSTNUKE
```



## Excluded patterns (3): oddities

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

---

```
# pictures
http://www-pnp.physics.ox.ac.uk/~karagozm/pix/
# Pete Biggs says this can go
#!http://ptcl.chem.ox.ac.uk/~doye/jon/
# Law School PHP is trusted
-regex: ^http://denning.law.ox.ac.uk.*php\?.*
# recursive
sbc.bioch.ox.ac.uk/stansfeld.php
# more recursion, in Ashmolean
contains: ?q=printme
# huge never-ending
www4.bioch.ox.ac.uk/~oubs/ABTD
# duplicate
www4.bioch.ox.ac.uk/oubs/ABTD
www2.bioch.ox.ac.uk/~oubs/
# another calendar
http://www.philosophy.ox.ac.uk/calendar?SQ_CALENDAR_VIEW
# admissions not to be index
www.admissions.ox.ac.uk/
# sers018.sers.ox dev server duplicates www.ouls.ox
sers018.sers.ox.ac.uk/
# endless recursive
contains: SQ_DESIGN_NAME=print
```



# Control for web masters

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

A user can

- manage the definition of a collection
- edit the details of a front end and associated stylesheet
- see crawl status and diagnostics for a collection
- see serving and search logs for a collection

Note: search reports and logs are not dynamic, they have to be requested and generated



# What does a frontend comprise?

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

- XSL stylesheet (can be edited raw, or tweaked in simple ways with settings)
- KeyMatch: force results to the top of the page if a keyword is matched
- Related queries: teach GSA about synonyms
- Filters:
  - Domain - restrict searches to one or more domain names (not IP addresses)
  - File type - restrict searches to one or more file types, such as HTML, PDF, and so on
  - Query expansion - determine the extent to which queries are expanded with synonyms
  - Meta tags - filter searches by values and value types in meta tags
- Remove URLs: simply exclude certain patterns
- Onebox modules: merge in other searches



# Definition of a keymatch

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



The screenshot shows the Google Search Appliance administration interface. At the top, the Google logo is on the left, and navigation links for 'Help Center' and 'Log Out' are on the right. Below the logo is a breadcrumb trail: 'Google Search Appliance > Serving > Front Ends > KeyMatch'. A blue navigation bar contains 'Home', 'Crawl and Index', 'Serving', 'Front Ends', 'Status and Reports', and 'Administration'. The 'Serving' section is expanded to show 'KeyMatch' as the active tab, along with 'Format', 'Related Queries', 'Filters', 'Remove URLs', and 'OneBox Modules'. The 'KeyMatch' tab contains links for 'View Matches', 'Edit Matches', 'Add Matches', and 'Import/Export Matches'. Below these links is a 'Search for KeyMatches containing:' field with a 'Search' button. A table below shows a single match entry with columns for 'Delete', 'Search Terms', 'Terms Occur As', 'URL for Match', and 'Title for Match'. The entry shows 'where' as the search term, 'KeywordMatch' as the term type, and 'http://www.oucs.ox.ac.uk/about/travel.xml' as the URL. The title is 'Where is OUCS?'. A 'Save Changes' button is at the bottom right of the table. At the bottom of the page, the copyright notice '©2002-2006 Google - About' is visible.

Google™ Google Search Appliance > Serving > Front Ends > KeyMatch [Help Center](#) - [Log Out](#) | [Test Center](#)

[Home](#) [Back to List of All Front Ends](#) Edit Front End:

[Crawl and Index](#) **Format** **KeyMatch** [Related Queries](#) [Filters](#) [Remove URLs](#) [OneBox Modules](#)

[View Matches](#) - [Edit Matches](#) - [Add Matches](#) - [Import/Export Matches](#)

After editing, click the Save Changes button. ([Help](#))

Search for KeyMatches containing:

| Delete                   | Search Terms                       | Terms Occur As                            | URL for Match                                                          | Title for Match                             |
|--------------------------|------------------------------------|-------------------------------------------|------------------------------------------------------------------------|---------------------------------------------|
| <input type="checkbox"/> | <input type="text" value="where"/> | <input type="text" value="KeywordMatch"/> | <input type="text" value="http://www.oucs.ox.ac.uk/about/travel.xml"/> | <input type="text" value="Where is OUCS?"/> |

©2002-2006 Google - [About](#)



# Result of using a keymatch

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



## University of Oxford Search



[Advanced Search](#)  
[Search Tips](#)

**Advanced Search** Results 1 - 10 of about 110000 for **where**. Search took 0.4 seconds.

[Next](#)>

[Where is OUCS?](#)

<http://www.oucs.ox.ac.uk/about/travel.xml>

[Where](#) are we? — [The Nuffield Laboratory of Ophthalmology](#)

**Where** are we? The laboratory is located at the West Wing levels 5 6 John



# Definition of related queries

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

Google™ [Google Search Appliance > Serving > Front Ends > Related Queries](#) [Help Center](#) - [Log Out](#)  
[ [Test Center](#) ]

[Home](#)  
▶ [Crawl and Index](#)  
▼ [Serving](#)  
    **Front Ends**  
    [Query Expansion](#)  
    [Access Control](#)  
    [Forms Authentication](#)  
    [OneBox Modules](#)  
▶ [Status and Reports](#)  
▶ [Administration](#)

[Back to List of All Front Ends](#) Edit Front End:

**View Related Queries** - [Edit Related Queries](#) - [Add Related Queries](#) - [Import/Export Related Queries](#)

Browse existing Related Queries. ([Help](#))  
Search for Related Queries containing:

| Search Terms | Related Queries |
|--------------|-----------------|
| food         | grub            |

 ©2002-2006 Google - [About](#)



# Result of using related queries

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



search  
UNIVERSITY OF OXFORD

## University of Oxford Search



[Advanced Search](#)  
[Search Tips](#)

**Advanced Search** Results 1 - 10 of about **14400** for **food**. Search took **0.07** seconds.

[Next](#)>

You could also try: [grub](#)

[club](#) : [food & drink](#)

Club : **food** & drink. ... lunch and afternoon. café-bar we offer hot **food** from



## Beyond the safe zone: Onebox modules

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

You can ask the GSA to pass the query to another system and merge the results back in. Caveats:

- Only 3 seconds is allowed for the external search to return
- Results must be returned in XML to a schema defined by Google
- Only the top 4 results will be shown
- Only administrators (not managers) can *create* Onebox modules



# What a Onebox module needs to know

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

- 1 Name
- 2 Trigger: one of
  - simple match with any query
  - keyword and query
  - regular expression
- 3 URL of provider. This must respond to queries of the form `www.example.com/answer?query=XXXX`
- 4 authentication details, if any



## Definition of a Onebox module

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz


```
<onebox id="contact" suppressDateTime="false" suppressI-  
PAddr="false" suppressKeyword="true" type="external">  
  <name>contact</name>  
  <security userAuth="none"/>  
  <description>Search database of Lexicon  
data</description>  
  <trigger triggerType="keyword">name</trigger>  
  <providerURL> http://clas-lgpn2.class.ox.ac.uk/cgi-  
bin/search.pl?searchBy=summary&style=onebox  
</providerURL>  
</onebox>
```



# Effect of Onebox


Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz




## search

UNIVERSITY OF OXFORD



### University of Oxford Search

---

[Advanced Search](#)  
[Search Tips](#)

**Advanced Search** Results 1 - 1 of about 1 for **name Ampelis**. Search took **0.13** seconds.

---

[Sort by date](#) / [Sort by relevance](#)

---

[Lexicon of Greek Personal Names results](#)  
[Αμπελις \(7 -0200 to 0310\)](#)

---

[Lexicon of Greek Personal Names - Names](#)  
... Krat-Ippos, Fil-Ippos, Plants: ampeloV 'vine' Ampelos, Ampelides, Ampellon, **Ampelis** ...  
of meaning individual Greek parents were when making a choice of **name**. ...  
[www.lgpn.ox.ac.uk/names/meaning.html](http://www.lgpn.ox.ac.uk/names/meaning.html) - 10k - 2005-08-30 - [Cached](#)

---

[Advanced Search](#)  
[Search Within Results](#) [Search Tips](#)





# XML returned to Onebox

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

```
<OneBoxResults>
  <Diagnostics>success</Diagnostics>
  <provider>Lexicon of Greek Personal Names</provider>
  <title>
    <urlText>Lexicon of Greek Personal Names
results</urlText>
    <urlLink>http://clas-
lgpn2.class.ox.ac.uk/LGPN/index.xml</urlLink>
  </title>
  <MODULE_RESULT>
    <U>http://clas-
lgpn2.class.ox.ac.uk/lexname/Bo1spwn</U>
    <Title>Βόσπων (4, -0269 to -0100)</Title>
  </MODULE_RESULT>
</OneBoxResults>
```



# Authorized access

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

The GSA has an important extra capability:

- allowing the box through secure systems and delivering the results to authenticated users only



# Typical simple authorization challenge

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz





# Setting up a username and password for a site

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz



Google Search Appliance > Crawl and Index > Crawler Access

[Home](#)

▼ [Crawl and Index](#)

[Crawl URLs](#)

[Databases](#)

[Feeds](#)

[Crawl Schedule](#)

**Crawler Access**

[Proxy Servers](#)

[Cookie Sites](#)

[Forms Authentication](#)

[HTTP Headers](#)

[Duplicate Hosts](#)

[Document Dates](#)

[Host Load Schedule](#)

[Index Rollback](#)

User admin logged in from: 192.76.8.123, 192.76.8.123

## Users and Passwords for Crawling: [\(Help\)](#)

To allow the appliance to crawl web servers protected by user authentication, add a username and password. Specify a domain only if needed (typically when crawling Microsoft IIS web servers).

| For URLs Matching Pattern, Use:                              | Username:                        | In Domain:           | Password:            |
|--------------------------------------------------------------|----------------------------------|----------------------|----------------------|
| <input type="text" value="http://tei.oucs.ox.ac.uk/Punch/"/> | <input type="text" value="gsa"/> | <input type="text"/> | <input type="text"/> |
| <input type="text"/>                                         | <input type="text"/>             | <input type="text"/> | <input type="text"/> |
| <input type="text"/>                                         | <input type="text"/>             | <input type="text"/> | <input type="text"/> |

\* Stored passwords are not displayed on these entries



## Other fun you can have

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

You may wish to:

- allow access to your SQL database for the GSA to range over
- feed (push) documents to the GSA from protected sites
- index Sharepoint
- index SSO authenticated resources



# What next?

Advanced use  
of the Google  
Search  
Appliance

Sebastian  
Rahtz

- Information for webmasters is at <http://www.oucs.ox.ac.uk/googlesearch/>
- Mail [webmaster@oucs.ox.ac.uk](mailto:webmaster@oucs.ox.ac.uk) if you need:
  - new username
  - new collection
  - new frontend
  - definition of a Onebox module
  - assistance with Web forms or XSLT