



Entireweb Search API

Entireweb Sweden AB
Last revision: 2010-04-28

This document describes how to obtain search results from Entireweb via an XML or JSON feed.

Invoking Entireweb

Search results from Entireweb are obtained by invoking the `/xmlquery` URL on the web server www.entireweb.com by means of an HTTP 1.1 GET request. The parameters governing the data to return shall be encoded in the URL parameters. The response to this request will be a well-formed XML document (MIME type `text/xml`) or a JavaScript array in JSON format (MIME type `application/json`), both suitable for parsing and rendering into a search result page.

The following is an example URL for invoking Entireweb to search for the query *entireweb search engine*.

```
http://www.entireweb.com/xmlquery?  
pz=01234567012345670123456701234567&ip=1.2.3.4&n=10&of=0&sc=9&  
format=xml&q=entireweb+search+engine
```

Note especially that proper URL encoding must be employed (see the *q* parameter above). All parameters used, and many more, will be explained below. At the bare minimum, *pz*, *ip* and *q* must always be supplied.

The XML Response

For the search above, the following result is typical:

```
<?xml version="1.0" encoding="UTF-8" standalone='yes'?>  
<ENTIREWEB VERSION="3.0">  
  <QUERY Q="entireweb" TYPECLASS="1" N="15" FIRST="1" THISPAGE="1" TOTPAGES="50"  
  SEARCHTIME="0.002" ESTIMATE="2740000">  
    <HIT INDEX="1">  
      <TITLE><![CDATA[Entireweb Search Engine]]></TITLE>  
      <SNIPPET><![CDATA[Entireweb makes searching easy with a user friendly search engine. ...  
Express Inclusion lists your site throughout the Entireweb Network within 48 hours. The  
network receives over 100 million searches every month.]]></SNIPPET>  
      <URL><![CDATA[http://www.entireweb.com/]]></URL>  
      <LINK><![CDATA[http://www.entireweb.com/]]></LINK>  
      <DISPLAYURL><![CDATA[entireweb.com/]]></DISPLAYURL>  
      <LANG><![CDATA[English]]></LANG>  
      <REGION><![CDATA[com]]></REGION>  
      <CLUSTERING><![CDATA[1]]></CLUSTERING>  
      <ID>3501874746803392</ID>  
      <LOCATION><![CDATA[se]]></LOCATION>  
      <LASTMOD><![CDATA[26 Apr 2010]]></LASTMOD>  
      <LASTCRAWL><![CDATA[26 Apr 2010]]></LASTCRAWL>  
      <DISCOVERED><![CDATA[1 Jan 2000]]></DISCOVERED>  
      <SIZE><![CDATA[11287]]></SIZE>  
      <INDEX>1</INDEX>  
      <IMP>0</IMP>  
    </HIT>  
    ... More hits follow...  
  </QUERY>  
</ENTIREWEB>
```

The JSON Response

Passing *format=json* (as will be detailed below) returns a JSON response rather than XML. For the query above, this could look like:

```
{
  "version":"2.0",
  "q":"entireweb",
  "typeclass":"1",
  "n":"15",
  "first":"1",
  "last":"15",
  "total":"1939",
  "thispage":"1",
  "totpages":"50",
  "searchtime":"0.004",
  "estimate":"2740000",
  "hits": [
    {
      "title":"Entireweb Search Engine",
      "snippet":"Entireweb makes searching easy with a user friendly search engine. ... Express Inclusion lists your site throughout the Entireweb Network within 48 hours. The network receives over 100 million searches every month.",
      "url":"http://www.entireweb.com/",
      "link":"http://www.entireweb.com/",
      "displayurl":"entireweb.com/",
      "lang":"English",
      "region":"com",
      "clustering":"1",
      "id":"3501874746803392",
      "location":"se",
      "lastcrawl":"26 Apr 2010",
      "lastmod":"26 Apr 2010",
      "discovered":"1 Jan 2010",
      "size":"11287",
      "imp":"0",
      "index":"1"
    },
    ... more hits follow...
  ]
}
```

Response Details

Both the XML and JSON response formats use identical tags to identify data fields. Thus we explain the tags only for the XML case. The examples above show only one search result, normally there would be more. From the top, the tags have the following meanings:

ENTIREWEB	Always opens XML feeds from Entireweb.
QUERY	The start of the query response section.
QUERY\Q	The query string passed to xmlquery
QUERY\TYPECLASS	Internal characterization of the query. Not of interest.
QUERY\N	The number of results per page. Useful for pagination.
QUERY\FIRST	The index of the first result returned. Useful for pagination.
QUERY\LAST	The index of the last result returned. Useful for pagination.

QUERY\THISPAGE	The index of this result page = $(FIRST+N-1)/N$. Useful for pagination.
QUERY\TOTPAGES	The number of pages that can be obtained for this query, by increasing the <i>of</i> parameter, with this setting for <i>n</i> . Useful for pagination.
QUERY\SEARCHTIME	The number of seconds the system spent working on answering this query.
QUERY\ESTIMATE	Estimated total number of search results for the query. This value is typically displayed at the top of a search result page to indicate the total number of global results for a search query.
HIT	A result hit.
HIT\INDEX	The index, counting from <i>of</i> , i.e. from QUERY\FIRST-1
TITLE	The title of the document.
LINK	The link that should be used when a user clicks this result.
DISPLAYURL	Cleaned-up version of the URL suitable for display in the search results.
LANG	Language of the document (see the <i>lang</i> parameter).
REGION	Region of the document (see the <i>reg</i> parameter).
CLUSTERING	The order in the cluster set. 1,2,3 means the first, second and third results from a domain/subdomain. 4 means a link (when $sc \in \{7, 8, 9\}$). See the <i>sc</i> parameter.
LOCATION	Geographic location of the web server serving this page.
LASTCRAWL	Time of last crawling/updating in the index.
LASTMOD	Time of last update (if reported by the web server).
DISCOVERED	Time when the document was first found on the Web.
SIZE	Size of document.
IMP	A measure of how much the search query had to be broadened to include this particular hit. IMP takes values in the range $\{0,1,2,\dots\}$, in a monotonously increasing sequence. A lower value means less broadening took place. A value of 0 means the query corresponded completely to the result. This tag can normally be ignored.

Request details

There is a great number of parameters that can be sent to the */xmlquery* module to alter what results are returned and in what format the data should be formatted.

Parameters required for authentication and authorization

These parameters must always be provided.

pz	The Partner Identification tag. The value of this parameter must be a 32-byte hexadecimal string provided by Entireweb. It identifies you as a partner and is also used for regulating your allotted query volume.
ip	The IP number (dotted quad) of the client performing the search. This is NOT the IP number of the server invoking Entireweb, but the IP number of the client utilizing that server. This is used both for result biasing and authorization/abuse detection.

Parameters that control the search results returned

q	The query to search for. Example: <i>q=entireweb+search+engine</i>
format	The format to return the response in. Valid values are <i>xml</i> and <i>json</i> , with <i>xml</i> being the default.
n	The number of results to return per page. $n \in \{10, 15, 20, 25, 30, 40, 50\}$. Default is 10.
of	The offset into the result list. Example: To get to the third page of ten results, pass <i>n=10</i> and <i>of=20</i> . This would make the first result returned number $20 + 1 = 21$, the first result on the third page. Default is 0.
lang	Return only documents written in this language. Language is an ISO 2- or 3-character code. Omitting <i>lang</i> (or passing <i>world</i>) returns documents in any language. Example: <i>lang=en</i> or <i>lang=swe</i>
reg	Return only documents from this region or continent. Region is either an ISO region code or a continent name. Omitting <i>reg</i> (or passing <i>world</i>) returns documents from any region (default). The recognized continent names are {scan, euro, name, same, came, asia, afri, ocea} corresponding to all countries in Scandinavia, Europe, North America, South America, Central America, Asia, Africa or Oceania. Example: <i>reg=se</i> or <i>reg=euro</i>

ir	Return only documents from servers located in this region, by means of geo lookup of the server IP number. Region is either an ISO region code or a continent name. Omitting <i>reg</i> returns documents from any region (default). The recognized continent names are {scan, euro, name, same, came, asia, afri, ocea} corresponding to all countries in Scandinavia, Europe, North America, South America, Central America, Asia, Africa or Oceania. Example: <i>ir=se</i> or <i>ir=euro</i> .
int	Return only results that were discovered after a certain point in time. The value should be a the number of seconds into the past that we want to return results for. It can also be one of the special constants { <i>d</i> , <i>w</i> , <i>m</i> , <i>y</i> }, which means day, week, month and year. For example, passing <i>int=86400</i> or <i>int=d</i> will both return only results found in the last 24 hour period.
sf	SafeFilter, the adult content protection filter. $sf \in \{0, 1, 2\}$, Passing 0 turns SafeFilter off, which makes adult content available. Passing 1 enables SafeFilter completely, and all adult material will be filtered. Passing 2 (the default) enables adaptive SafeFilter: In this case, the filter will be engaged for non-adult queries.
df	Duplicate filter. $df \in \{0,1,2,3\}$. 0 disables duplicate filtering. 1 detects and removes binary identical documents. 2 and 3 both enable an advanced linguistics-based duplicate filtering system, with 2 (the default) being more permissive than 3.
sc	Site Clustering, i.e. reduction of the number of results returned from each domain. $sc \in \{0,1,2,3,4,5,6,7,8,9\}$. 0 turns clustering off. 1, 3, 5 returns one, two or three documents, respectively, from each subdomain (e.g., with $sc=3$, at most two results are returned from <i>sa.entireweb.com</i> and at most two results are returned from www.entireweb.com). 2, 4, 6 returns one, two or three documents, respectively, from each domain (e.g with $sc=2$, at most one result is retrieved from <i>entireweb.com</i> , regardless of subdomain). 7 and 8 returns one master result from each subdomain or domain, respectively, and a small number (typically at most 4) clustered results from the same subdomain/domain. This is useful for presenting a list of "related links" under a result. 9 is an adaptive mode, that automatically composes a suitable list of clustered and "related link" results. $Sc=9$ is the default.
ol	Origin language. If an ISO2/3 language code is specified here, it will be used as the origin language of the visitor, and bias results towards this language or related languages.
oi	Origin interface language. Same principle as <i>ol</i> , but with a much stronger effect. This is normally used for sending a user-selected site interface language to Entireweb, which is seen as strong evidence tha the user favors this language.
or	Origin region. If an ISO region code is specified here, it will be used as the origin region of the visitor, and bias results towards this region or related regions.

Parameters controlling the appearance of returned data

opt_hilitet opt_hilited opt_hilites opt_hiliteu	Hiliting of query terms in the data in the TITLE, DESC, SNIPPET and URL tags, respectively. opt_hiliteX $\in\{0,1,2,3\}$. Passing 0 hilites terms by surrounding them by a [[]] pair. These can then be replaced by e.g. a "" / "" pair, or anything else of the partner's choosing. Passing 1 disables hiliting. Passing 2 hilites with tags, and passing 3 hilites with <i></i> tags. Default is 1 (no hilite).
opt_sztxt	Appearance of the SIZE data. opt_sztxt $\in\{0,1\}$. Passing 0 (default) returns the size in bytes (e.g. "12345"), passing 1 returns the size in formatted text (e.g. "12.1k").
opt_lf	Language format. opt_lf $\in\{0,1,2\}$. Appearance of the LANG data. Passing 0 (default) returns the English name of the language (e.g. <i>Swedish</i>), passing 2 returns the ISO 3-character code (e.g. <i>swe</i>), passing 1 returns the ISO 2-character code if available and the 3-character code otherwise (for Swedish this would return <i>sv</i> , but Ancient Greek would be returned as <i>grc</i> since no 2-character code is assigned to this language).
opt_lmep opt_lcep opt_dep	Time format for the LASTMOD, LASTCRAWL and DISCOVERED tags, respectively. opt_Xep $\in\{0,1\}$. Passing 0 (default) returns the date formatted as <i>1 December 2000</i> . Passing 1 returns the epoch time, i.e. the number of seconds since January 1, 1970 (e.g. <i>975625200</i>).

Attribution

To use the Entireweb Search API, we require our partners to [provide attribution](#) to Entireweb by linking to Entireweb's main page and displaying one of the images presented on the [Implementation](#)'s page. The image should be clearly visible on each page that makes requests to, or gets results from, the Entireweb Search API.