



Express Inclusion Partner Interface

This document describes how URLs can be added and modified in Entireweb's Express Inclusion system. The API handles only URLs and their indexing status – all matters relating to user/client management, payments, renewals etc should be handled by you, the partner.

The general API for handling URLs is the following:

https://services.entireweb.com/express_inclusion/auto/?id=ID&action=ACTION&e=EXTERNID&test=TEST&...

- **ACTION** describes what action to take, and may be ADD, MODIFY or QUERY. These are each explained in sections below.
- **ID**: Your private ID tag that identifies your Express Inclusion account. This tag will be provided by Entireweb, and should be used for all submissions.
- **EXTERNID**: An 32-bit signed integer of your choosing, which will be used for your future referencing of URLs. This value could for instance correspond to an internal SQL table id number at your end. This parameter may NOT be omitted, may NOT be zero, and MUST be unique. The interpretation and assignments of these numbers are completely up to you.
- **TEST**: Run in test mode if set to 1. In test mode, URLs will still be inserted and indexed, but no payments will be issued. Use this mode to test the API during development. Prior to going live, Entireweb will erase all URLs in your account and you will not be able to use the test parameter any more, which means that payments will be logged for all submissions you make.

In addition to these, certain actions will have additional parameters, described below.

All calls and all parameters may be sent with either GET or POST.

Data Types

The basic data type in this API is the URL. A URL comes in one of three different types:

- **Type 1**: A URL that has been added for indexing through the API, but from which no links will ever be followed. This is a legacy type and is not used for new URLs. It was used when Express Inclusion offered only indexing of distinct pages, not entire sites. This type is of interest only for partners that have type 1 URLs already and wish to upgrade these. See information on upgrading URLs below.
- **Type 2**: A URL that has been added for indexing through the API, with full link following capabilities. This is the type that new URLs will be inserted as.
- **Type 3**: A special URL type which cannot be directly added through the API, but instead is inserted by Entireweb for every URL that is found while indexing a site (I.e. a type 2 URL). Thus every type 2 URL has an associated list of type 3 URLs that refer to the actual pages found while crawling that site. Referred to as SUBIDs. This list may be changed by Entireweb's backend systems at any time, to reflect newly added/deleted/updated URLs within a site. Thus caching of these by an API partner should be limited. The list of SUBIDs can be retrieved with the QUERY action, described below.

Every type-1 and type-2 URL can be associated with an arbitrary EXTERNID, for your reference. Type 3 URLs are referenced by an Entireweb-provided ID number (SUBID) and the EXTERNID of their parent type-2 URL. Thus type-3 URLs do NOT have EXTERNIDs of their own.

Adding new URLs: The add action

To add new URLs, use the following call:

[https://services.entireweb.com/express_inclusion/auto/?
action=add&id=ID&e=EXTERNID&url=URL&ta=TIME&p=PARTNER](https://services.entireweb.com/express_inclusion/auto/?action=add&id=ID&e=EXTERNID&url=URL&ta=TIME&p=PARTNER)

URL is the URL to add, properly escaped. PARTNER is the canonical hostname of the partner through which the sale was performed, eg if sold through MegaPartner Inc., this may be p=www.megapartner.com. Omit or leave empty if the sale was not performed through a partner of yours (I.e. don't pass this parameter if you made the sale yourself).

The URL will be inserted as a type 2 URL. URL is the URL to add, and need not be unique (ie it is OK if multiple clients submit the same URL). TIME is the epoch time that should be used as the time of adding. The URL will expire 365*86400 seconds after this time. If omitted, it will be set to the current time. The existence of this parameter enables retroactive adding of URLs, by passing an add time that is older than the current time. The time cannot be in the future.

The response to this request will be a RESULT tag (see below) with one of the following codes: OK, BAD_REQUEST, BAD_URL, EXISTING_EXTERNID.

Modifying existing URLs: The modify action

To modify parameters for a URL, use the following call:

[https://services.entireweb.com/express_inclusion/auto/?
action=modify&id=ID&e=EXTERNID&url=URL&ta=TIME&volume=VOLUME&follow_links=
FOLLOW_LINKS&subdomain_lock=SUBDOMAIN_LOCK&seed=SEED&exclude=EXCLUDE](https://services.entireweb.com/express_inclusion/auto/?action=modify&id=ID&e=EXTERNID&url=URL&ta=TIME&volume=VOLUME&follow_links=FOLLOW_LINKS&subdomain_lock=SUBDOMAIN_LOCK&seed=SEED&exclude=EXCLUDE)

With this call, the URL, TIME and other parameters associated with an EXTERNID can be modified. This is useful eg for altering URLs or for doing renewals (by passing an updated ta). Any parameter omitted will not be modified. A URL can be effectively expired and prevented from being crawled by passing a TIME older than time()-365*86400.

If the TIME parameter specifies a time later than the time currently assigned to the URL, the modify operation is considered a renewal. A payment will thus be implied.

The rest of the parameters are related to the way type-2 URLs are handled, and are only valid for type-2 URLs.

- VOLUME: The number of type-3 pages to store for a particular type-2 URL. Legal range is 1 to 1000. Default is 1000.
- FOLLOW_LINKS: Pass 0 to only index the particular URL and thus emulate the behaviour of an old-style type-1 page. Pass 1 to index the complete site. Default is 1.
- SUBDOMAIN_LOCK: If 1, only subdomains matching the type-2 URL or any of its seed URLs will be indexed. If 0, all URLs from the complete domain will be indexed. **Example:** If the URL of a type-2 URL is myblog.blogspot.com and SUBDOMAIN_LOCK=0, links to every page in blogspot.com will be followed, for example to someotherblog.blogspot.com. This is probably not what is desired. Passing SUBDOMAIN_LOCK=1 restricts indexing to myblog.blogspot.com. Default is 1.

- STORE_FULL. Pass 1 to store the complete information found on pages. This can later be retrieved with the query action using the ic=1 parameter. Default is 0.
- SEED: A newline-separated list of starting URLs from which to initiate the crawling of the site. This may be omitted, in which case crawling starts from the type-2 URL itself.
- INCLUDE: A newline-separated list specifying what URLs (or URL prefixes) will be included. No URLs not matching this list will be included. Leave empty to include everything.
- EXCLUDE: A newline-separated list of URLs (or URL prefixes) to exclude from crawling. Leave empty to exclude nothing.

SEED, INCLUDE and EXCLUDE are recommended to pass by POST.

The response to this request will be a RESULT tag (see below) with one of the following codes: OK, BAD_REQUEST, BAD_URL, BAD_EXTERNID.

Querying information about URLs: The query action

To retrieve info about URLs, use the following call:

https://services.entireweb.com/express_inclusion/auto/?action=query&id=ID&e=EXTERNID1,EXTERNID2,...&s=SUBID1,SUBID2,...&ic=<0,1>

In this case, e can be a comma-separated list of EXTERNIDs, or alternatively, e can be a single EXTERNID and s can be a comma-separated list of subids of type-3 pages for the type-2 page specified in e. If successful, this returns an XML document with one ENTRY block for each EXTERNID/SUBID, describing the URL belonging to that EXTERNID/SUBID. Up to 10 subids may be sent in a single list.

The resulting XML takes on slightly different forms depending on the type of URL returned. Specifically, for type-2 URLs, a tag called SUBIDS is included, which gives a comma-separated list of all subids (type-3 URLs) that is associated with the type-2 URL at the moment. You can use these subids to query information about the individual type-3 pages.

Here is an example, for which e=4284 and s= 16455326,16403861.

```
<QUERY>
<ENTRY EXTERNID="4284">
  <URL>http://www.test.com/</URL>
  <STATUS>1</STATUS>
  <TYPE>2</TYPE>
  <TIME_ADDED>1280620800</TIME_ADDED>
  <TIME_INDEXED>1280834593</TIME_INDEXED>
  <RE>4e051aaaaa51ada17923edd22ebc3421</RE>
  <IDXSTATUS>100</IDXSTATUS>
  <SUBIDS>16455326,16403861,16413573,16418236,16422936,16426903,16430953,1643517
6,16435729,16439677,16446376,16449216,16452343,16453289,16454792,16392023,163920
62,16502114,16406776,16422556,16424222,16432591,16436030,16488083,16437708,16440
214,16446810,16448131,16412871,16416142,16505614,16428432,16506920,16507272,1650
7667,16499194,16394826,16502416,16411446,16411447,16412161,16505690,16421642,164
24083,16506716,16430637,16507121,16437013,16438327,16442840,16448328,16507956,16
453015,16508159,16508184,16508185,16508211,16508225,16365968,16368653,16378497,1
6381976,16394330,16397488,16403111,16407007,16409883,16410921,16417494,16419082,
16420814,16424087,16425695,16428194,16435670,16436041,16440377,16441097,16441467
,16444806,16446235,16455648,16363160,16374424</SUBIDS>
</ENTRY>
<ENTRY EXTERNID="4284" SUBID="16455326">
```

```

<URL>http://www.test.com/</URL>
<STATUS>1</STATUS>
<TYPE>3</TYPE>
<TIME_ADDED>1280620800</TIME_ADDED>
<TIME_INDEXED>1280834466</TIME_INDEXED>
<RE>4e051aaaaa51ada17923edd22ebc3421</RE>
<IDXSTATUS>100</IDXSTATUS>
<TITLE>Title of page</TITLE>
<DESCRIPTION>Meta Description tag</DESCRIPTION>
<META>Author=John Doe, Keywords=test</META>
<SIZE>31259</SIZE>
<CONTENT>Content of page</CONTENT>
</ENTRY>
<ENTRY EXTERNID="4284" SUBID="16403861">
<URL>http://www.test.com/search.asp</URL>
<STATUS>1</STATUS>
<TYPE>3</TYPE>
<TIME_ADDED>1280620800</TIME_ADDED>
<TIME_INDEXED>1280834466</TIME_INDEXED>
<RE>bc53994daa50c16c1ea08577f8e73f05</RE>
<IDXSTATUS>100</IDXSTATUS>
<TITLE></TITLE>
<DESCRIPTION></DESCRIPTION>
<META></META>
<SIZE></SIZE>
<LASTMOD></LASTMOD>
<CONTENT></CONTENT>
</ENTRY>
</QUERY>

```

The parameter `ic` (include content) can take the values 0 or 1 (default 0). If 1, data found when indexing the URL (specifically, the tags `TITLE` and below) are included for each entry. This considerably expands the size of the XML response, so pass `ic=1` only if necessary.

`IDXSTATUS` gives the indexing status of the page. 100 means the page was correctly indexed, and the following gives a list of possible errors that can occur while processing the page:

- **Error 101.** *DNS error.* The host name of your page could not be resolved to a valid IP address. Make sure you've entered the correct host name.
- **Error 102.** *Forbidden IP number.* Your host name resolves to an IP number that is not allowed. Make sure you've entered the correct host name and that your host resolves to a public IP number.
- **Error 103.** *Excluded page.* We could not crawl your page since you have blocked access to it via `robots.txt` or meta tag exclusion.
- **Error 104.** *HTTP 404.* Your web server responded with error code 404 (file not found on server).
- **Error 105.** *HTTP 403.* Your web server responded with error code 403 (access disallowed).
- **Error 106.** *HTTP 4xx.* Your web server responded with an error code from the 400 series (other than 403 and 404), meaning your web server gives an error condition upon access.
- **Error 107.** *HTTP 5xx.* Your web server responded with an error code from the 500 series, signifying an internal problem with your server.
- **Error 108.** *HTTP error.* Your web server denied access and responded with an unhandled HTTP response code not from the 400 or 500 series.
- **Error 109.** *Bad redirection.* The URL you've provided redirects in an unhandled way.
- **Error 110.** *Cyclic redirection.* Your URL redirects to itself. Therefore, the crawler could not obtain an indexable resource from it.
- **Error 111.** *Site unreachable.* We failed to contact your page because of an undefined condition. Causes may include network timeouts, DNS problems or misconfigured web servers.

IDXSTATUS=127 means the page has not yet been indexed.

If an error arises, the result is a RESULT tag describing the error. If a particular EXTERNID is not found, the STATUS tag will have value 0. Otherwise STATUS=1 means the URL is active, and STATUS=2 means it is expired. The URL tag is the url currently associated with the EXTERNID. TIME_ADDED is the epoch adding time of the URL. Please use the query request only minimally for generation of account details in your system, since this will significantly slow down your interface. This call is mainly intended for administration and trouble-shooting purposes. If the request cannot be fulfilled, there will only be a RESULT tag with an error code. Codes can be OK, BAD_REQUEST, BAD_EXTERNID, BAD_SUBID.

Upgrading Legacy Type-1 URLs to Type-2

It is very simple to upgrade your old type-1 URLs to type-2, and get all the benefits of whole-site indexing. Just send a modify action for each EXTERNID to upgrade, with follow_links=1. This will automatically transform the URL to type-2. There is no cost for upgrading.

Typical Usage

The recommended basic usage of this API is to send an add action in response to a purchase made by one of your clients. Doing this will automatically register a payment in Entireweb's systems, for which we will send you an invoice in accordance with the partner agreement that you have signed with Entireweb. When sending the add action, you should assign it a unique EXTERNID at your end, and register this id in e.g. a database for later reference.

You are responsible for keeping track of the inclusion time and renewing subscriptions after one year. This is done with the modify action, as described above.

To be able to display information about the URL and its indexing status to the client, use the query action with the previously assigned EXTERNID.

Error Codes

The following response codes are currently defined:

- OK: The URL was successfully added.
- BAD_REQUEST: Bad parameters were given, typically a malformed id tag.
- ALREADY_ADDED: This URL has already been added to your queued list of URLs.
- BAD_URL: URL was malformed and could not be added.
- BAD_EXTERNID: Externid was malformed or not found.
- BAD_SUBID: Subid was malformed or not found.
- EXISTING_EXTERNID: Externid already mapped to a URL.