Typosquatting daily data feed reference manual

Table of Contents

1. About

The present document describes the "Typosquatting data feed" product, a service of daily data file downloads by WhoisXML API, Inc.

Document version: 3.0 dated 2020-08-25.

2. Data definition

The "Typosquatting data feed" captures at least 3-member groups DNS domains so that:

- Each domain in the group appeared on the same day in the DNS zone file;
- The domain names within the group are similar to each other.

Appearance on the same day in the zone file normally means that the domains were registered on the same day, close to the given date. By "similar to each other" we mean similarity with respect to a suitably chosen algorithmically calculable mathematical characterization.

The coverage of the search for such a domain by the data generator is the set of domains in the top-level domains covered by the following daily data feeds:

- `domain_names_new`

(http://www.domainwhoisdatabase.com/docs/Daily_GTLD_WHOIS_Updates_Reference_Manual/README_01_document.html#sec15),

- `ngtlds_domain_names_new`

(http://www.domainwhoisdatabase.com/docs/Daily_GTLD_WHOIS_Updates_Reference_Manual/README_01_document.html#sec32),

- and `cctld_registered_domain_names_new`

(http://www.domainwhoisdatabase.com/docs/Daily_CCTLD_WHOIS_Updates_Reference_Manual/README_02_document.html#sec15). This set covers almost all generic top-level domains and a number of country-code top-level domains.

3. Reasons for a domain to appear listed in the feed

Important disclaimer: a domain listed in this feed is not necessarily malicious or related to any dangerous activity. Nevertheless, it is a fact that if a domain is listed:

- There exist at least two additional domains which have a name similar to the given domain, so there is an increased risk of getting to another domain because of mistyping or misunderstanding the domain name;
- The domain has been registered together with similarly named domains on the same day.

The possible reasons for a malicious domain to appear in this feed include:

- Being registered in a burst for phishing purposes, typically to resemble the name of a brand.
- Being registered in bulk in a group of machine-generated domain names to be used by malware, e.g. as a potential command-and-control server.
- Being registered for typosquatting pages related to known other pages or brands (for ad money collection, or sometimes for phishing).

The possible reasons for a benign domain appearing in the feed include:

- Resulting from brand protection by the brand owner of a similar domain name;
- Being an element of a set of algorithmically generated domain names used e.g. for load balancing or assigning domain names to entities in bulk.

4. Data format

4.1. Simple data files

The data are provided in one file for each day, in the web directory

https://typosquatting.whoisxmlapi.com/datafeeds

The files are named according to the convention:

In case of a daily subscription

`typosquatting.YYYY-MM-DD.daily.basic.csv`

In case of a weekly subscription

`typosquatting.YYYY-MM-DD.weekly.basic.csv`, where the date corresponds to a Sunday which is the last day whose data are included in the file; thus a week is considered to start with the previous Monday, end with the Sunday in the file name, and the file is published the next Monday after the date in the file name.

In case of a monthly subscription

`typosquatting.YYYY-MM-DD.monthly.basic.csv`, where the date corresponds to the first day of the next month, thus e.g. data for July 2020 are in the file `typosquatting.2020-08-01.monthly.basic.csv`.

Note that the weekly and monthly data are derived from the concatenation of the respective daily data and the addition of the first field, the date.

The files are comma-separated value-files without text delimiters. The files use DOS/Windows-style line terminators (CR+LF). The first line is a header line with the field names. Each line has four or five fields depending on the subscription type:

`date`

The day when the group was detected.

group_number

Ordinal number of the group within the given day (in case of daily subscription, within the file).

group_member_number

Ordinal number of the domain within the group.

total_no_of_grp_members

Number of group members within the group.

domain

Domain name

domain_utf

Domain name transcribed to Unicode; only for domains with national (non-English) characters.

E.g. a two adjacent groups, No. 1058 and 1059, with 3 and 5 members, respectively, appear in the file as:

```
1058,1,3,slut.bar,
1058,2,3,slut.events,
1058,3,3,slut.red,
1059,1,5,worldthinkcreativity.online,
1059,2,5,worldthinkcreativity.org,
1059,3,5,worldthinkcreativity.com,
1059,4,5,worldthinkcreativity.info,
1059,5,5,xn--wrkdthinkcreativity-g5c.net,wırkdthinkcreativity.net
```

The last domain in the list has a non-English character ("i" without a dot) as a second letter, as seen in the non-empty last field. In a weekly or monthly file, the lines of a group will look like

```
2020-08-17,3,1,9,app1e1d05.com,
2020-08-17,3,2,9,app1e1d09.com,
2020-08-17,3,3,9,app1e1d03.com,
2020-08-17,3,4,9,app1e1d04.com,
2020-08-17,3,5,9,app1e1d02.com,
2020-08-17,3,6,9,app1e1d01.com,
2020-08-17,3,7,9,app1e1d07.com,
2020-08-17,3,8,9,app1e1d08.com,
2020-08-17,3,9,9,app1e1d06.com,
```

Note that it is the date and the ordinal number of the group (the first two fields) which identifies the group uniquely in these files.

4.2. Enriched data files

The enriched data files are comma-separated value-files with quotation marks (") as text delimiters. Numeric fields are not delimited. The files use DOS/Windows-style line terminators (CR+LF). The first line is a header line with the field names.

The file naming conventions are the same as the case of the basic files, except for having the word "enriched" instead of "basic" in the file names.

The fields are:

`date`

Only in weekly and monthly files; the detection date.

`group_number`

As in the case of the "simple" files

`group_member_number`

As in the case of the "simple" files

`total_no_of_grp_members`

As in the case of the "simple" files

`domain`

As in the case of the "simple" files

`domain_utf`

As in the case of the "simple" files

`registrant_name`

From the WHOIS record

`registrant_organization`

From the WHOIS record

`registrant_country`

From the WHOIS record

`registrant_state`

From the WHOIS record

registrant_city

From the WHOIS record

registrant_email

From the WHOIS record

registrarName

From the WHOIS record

registrarIANAID

From the WHOIS record

whoisServer

The server the WHOIS record was obtained from

nameServers

From the WHOIS record

status

From the WHOIS record; EPP status codes of the domain

createdDate

Registration datetime from the WHOIS record. Note: the date in the WHOIS record may differ from the file date. The file date corresponds to the day when the domain appears as new in the DNS, i.e. starts to resolve. It may not be the same day as the date administered in the WHOIS record, which has no technical implication.

updatedDate

Last update datetime from the WHOIS record

expiresDate

Expiry datetime from the WHOIS record

standardRegCreatedDate

Registration datetime in standard format. (See the comment at "createdDate")

standardRegUpdatedDate

Update datetime in standard format

`standardRegExpiresDate`

Expiry datetime in standard format

`IPs`

A space-separated list of IPv4 addresses assigned to the domain according to a DNS lookup on the day or the day after the file date.

5. Data availability

5.1. Daily subscriptions

"simple" data

are available by 6:00 p.m. UTC on most days. On some, as the generation input data coming from other daily feeds may take more time, the data are generated 8 hours later.

"enriched" data

are available normally at 8 p.m. the next day. Depending on the domain registration activity, up to 8 hours of delay is possible.

5.2. Weekly subscriptions

Weeks start on Sunday. The data files for the last week become available on Monday 8 p.m. UTC every week.

5.3. Monthly subscriptions

The data files for the last month are available on the second day of the month at 8 p.m. UTC.

6. Recommended applications

The data feeds hold domains which are, though possible benign, prone to typosquatting or phishing attacks and malware-related activity. Therefore it is recommended to doubly check them when used for any purpose (e.g. opening in a web browser) to maintain cybersecurity.

The listing can also be useful in studying manually or algorithmically with various purposes:

- proactive IT security investigations to reveal a future typosquatting or phishing attack. Correlating them with malware blacklists can extend the set of compromised domains in a list. Checking WHOIS and other technical data of the domains is also recommended.
- legal investigations related to past cybersecurity incidents
- brand protection to reveal or prevent misuse of brand names and domain names
- research of domain name registration activity trends, etc.

Consult our blogs at the product webpage for further details.