Early Homograph Threat Detection: A DNS Study of IDNs and Native-Language Characters

WhoisXMLAPI
The Who Behind Domain, IP & Cyber Threat Intelligence

www.whoisxmlapi.com

TABLE OF CONTENTS

Executive Summary

Internationalized domain names (IDNs) have emerged as a powerful tool for organizations seeking to expand their global reach. They allow the use of native-language characters in domain names, enabling businesses to connect to non-English-speaking audiences and enhance their brand recognition in diverse markets.

The adoption of such domains has significantly grown, with 41% of generic top-level domains (gTLDs) and 85% of country code TLDs (ccTLDs) supporting IDNs. However, since the DNS was not initially designed to accept Unicode or native-language characters, TLD operators typically use Punycode to convert them into ASCII.

To date, there are more than 4 million second-level IDNs worldwide, but not all are used for legitimate purposes. One notable bad apple is the Nitrogen malware campaign, where threat actors exploited Punycode to create deceptive domains. They used wìnscp[.]net (Punycode translation: xn--wnscp-tsa[.]net) to lure targets into downloading a fake WinSCP installer. This and other similar scenarios show that IDNs also present cybersecurity risks.

As the use of IDNs and non-Latin scripts continues to grow, it's essential to understand more about them. As such, this white paper explores the TLD distribution, IP resolution, and WHOIS registration data of a sample of fully qualified domain names (FQDNs) containing native-language characters. Among our key findings are:

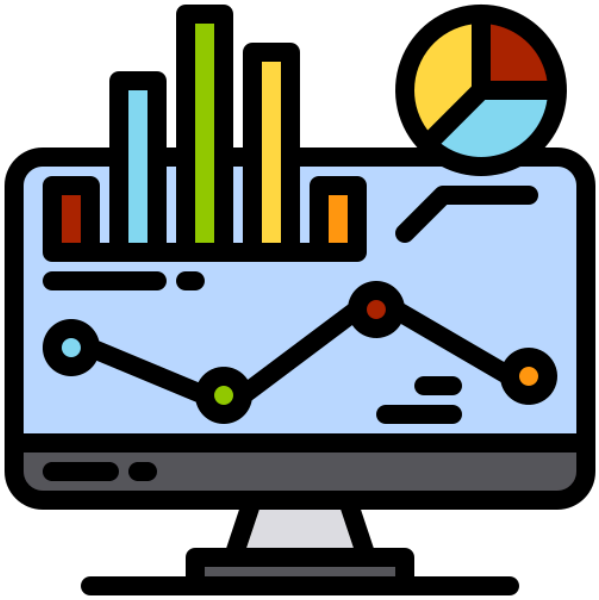
- The top 2 most used TLDs among the FQDNs were Chinese-looking TLDs, which jointly accounted for 52.75% of the total number.
- Fifty-eight of the top 100 IP addresses resolving the FQDNs were geolocated in the U.S.
- Amazon administered 36 of the top 100 IP addresses.
- Out of 34,437 root domains, duckdns[.]org hosted the most number of FQDNs with native-language characters.
- The top 100 root domains were distributed across 51 registrars, with MarkMonitor, Inc. leading the list.
- A total of 90% of the top 100 root domains had privacy-protected WHOIS information.

Research Data and Methodology

The study gleaned data from the Premium DNS Database file dated 6 June 2024, after we extracted a sample of 63,105 unique FQDNs containing native-language characters from more than 58 billion rows of data.

We then analyzed their IP resolutions by running the FQDNs' corresponding IP addresses on Bulk IP Geolocation API. This provided information about the ISP administration and geographic location of the servers hosting these domains.

Next, we obtained the root domains' WHOIS information using Bulk WHOIS Lookup and analyzed the registrar, registrant country, and WHOIS redaction of the top 100 root domains.

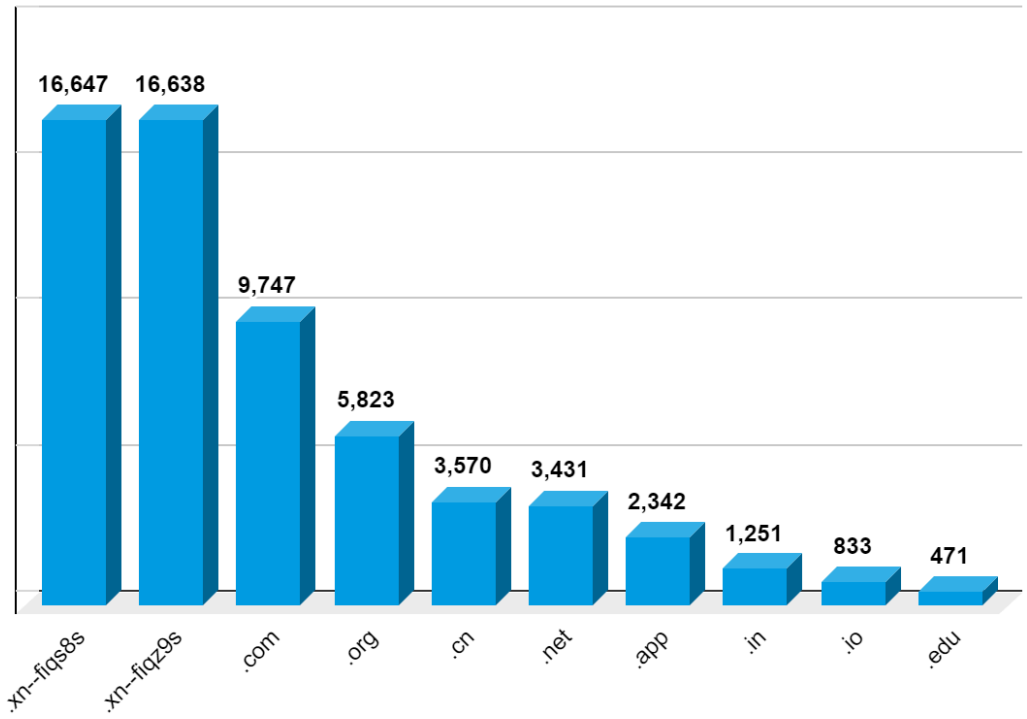


Research Findings

Most Used TLDs of the FQDNs

Two internationalized TLDs were among the most found in our passive DNS database, namely, .xn--fiqs8s or .中国 (16,647 FQDNs) and .xn--fiqz9s or .中國 (16,638 FQDNs). The rest of the top 10 were regular TLDs, comprising .com (9,747 FQDNs), .org (5,823 FQDNs), .cn (3,570 FQDNs), .net (3,431 FQDNs), .app (2,342 FQDNs), .in (1,251 FQDNs), .io (833 FQDNs), and .edu (471 FQDNs).

TOP 10 TLDs OF THE FQDNs



The top 10 TLDs accounted for 96.27% of the FQDNs, while the remaining domains were distributed across 104 other TLDs.

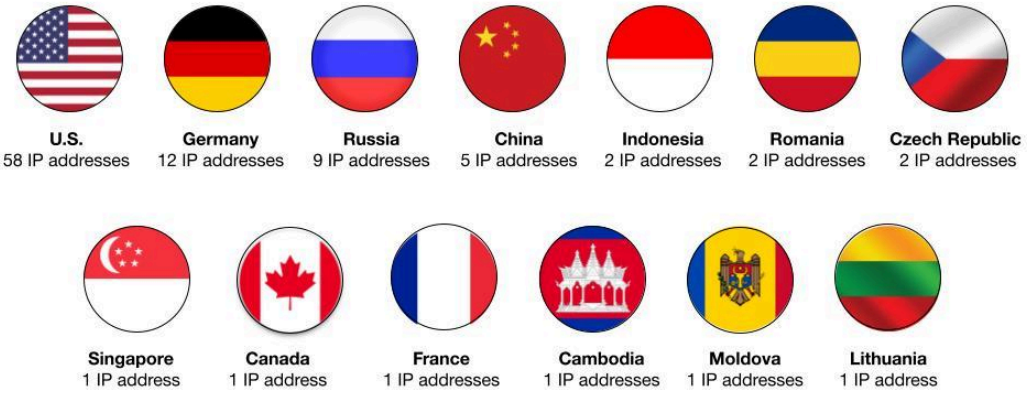
IP Resolutions of the FQDNs

From the Premium DNS Database file, we extracted the top 100 IP addresses to which the highest number of FQDNs resolved. We then obtained their IP geolocation data, which revealed that the top IP address (218[.]241[.]105[.]10), which accounted for a massive 49.06% of resolutions, was geolocated in China.

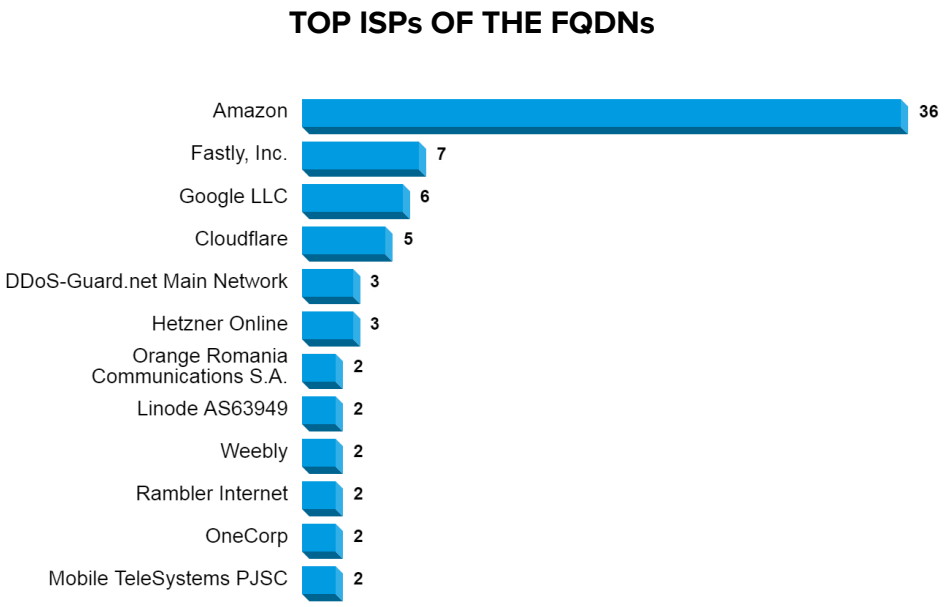
Overall, the geolocation information of the top 100 IP addresses showed that:

➡ A majority, 58 to be exact, were geolocated in the U.S. Twelve could be traced to Germany; nine to Russia; five to China; two each to Indonesia, Romania, and the Czech Republic; and one each to Singapore, Canada, France, Cambodia, Moldova, and Lithuania. Four IP addresses had unspecified geolocation countries.

TOP GEOLOCATION COUNTRIES IP ADDRESSES RESOLVING THE FQDNs

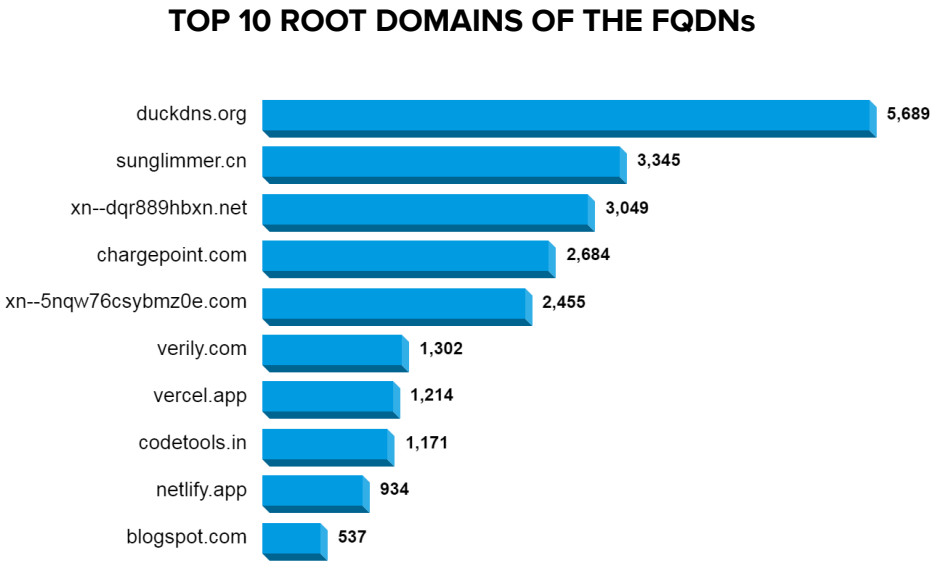


➡ Amazon administered 36 of the IP addresses, followed by Fastly, Inc. (seven IP addresses); Google LLC (six IP addresses); Cloudflare, Inc. (five IP addresses); DDoS-Guard.net Main Network (three IP addresses); and Hetzner Online (three IP addresses). Orange Romania Communications S.A., Linode AS63949, Weebly, Rambler Internet, OneCorp, and Mobile TeleSystems PJSC managed two IP addresses each. Twenty other ISPs administered one IP address each, while eight did not have current ISP data.



Top Root Domains of the FQDNs

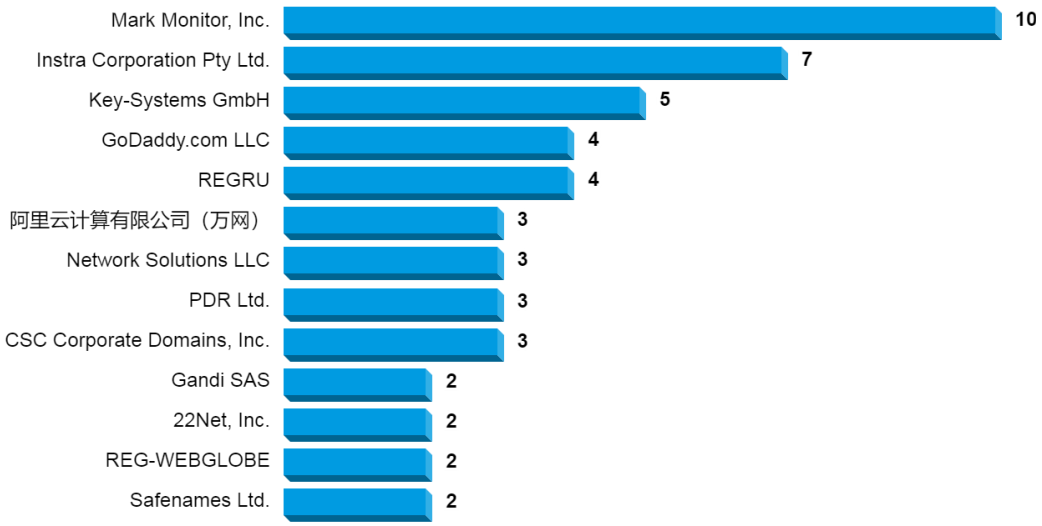
Since most of the internet properties containing native-language characters were subdomains, we decided to extract their root domains and found 34,437. The root domain duckdns[.]org hosted the most number of FQDNs, 5,689 to be exact. The rest of the top 10 root domains were sunglimmer[.]cn, xn--dqr889hbxn[.]net, chargepoint[.]com, xn--5nqw76csybmz0e[.]com, verily[.]com, vercel[.]app, codetools[.]in, netlify[.]app, and blogspot[.]com. These root domains accounted for 35.46% of the FQDNs.



A WHOIS analysis of the 100 most used root domains revealed that:

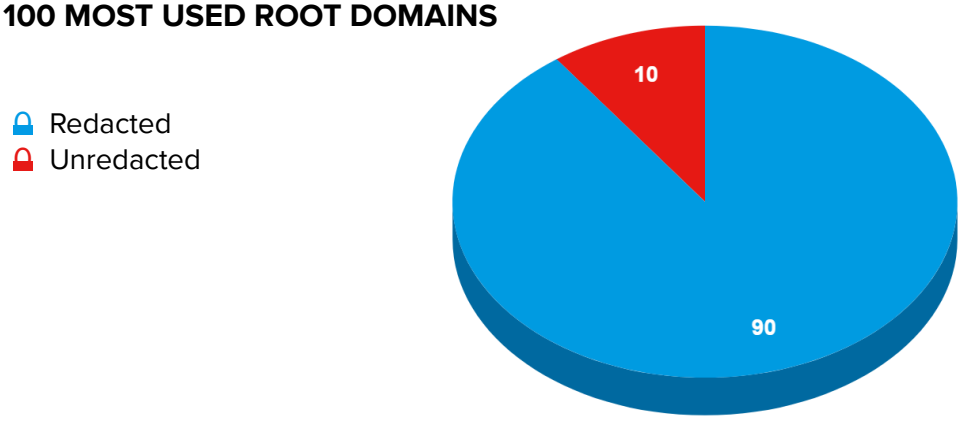
➡ They were distributed across 51 registrars. The registrars that administered more than one domain included MarkMonitor, Inc. (10 domains); Instra Corporation Pty. Ltd. (seven domains); Key-Systems GmbH (five domains); GoDaddy.com LLC (four domains); REGRU (four domains); 阿里云计算有限公司（万网 (three domains); Network Solutions LLC (three domains); PDR Ltd. (three domains); CSC Corporate Domains, Inc. (three domains); Gandi SAS (two domains); 22Net, Inc. (two domains); REG-WEBGLOBE (two domains); and Safenames Ltd. (two domains).

TOP REGISTRARS OF THE TOP 100 MOST USED ROOT DOMAINS



➡ Based on their registrant name fields, 90% of the root domains had privacy-protected WHOIS details, while only 10% had public WHOIS records.

NUMBER OF REDACTED AND UNREDACTED WHOIS RECORDS OF THE TOP 100 MOST USED ROOT DOMAINS



➡ Most of the root domains, 30 to be exact, were registered in the U.S. The U.K. (12 domains), France (six domains), China (five domains), the Czech Republic (three domains), Canada (three domains), Russia (two domains), Seychelles (two domains), Bulgaria (two domains), and Iceland (two domains) completed the top 10.

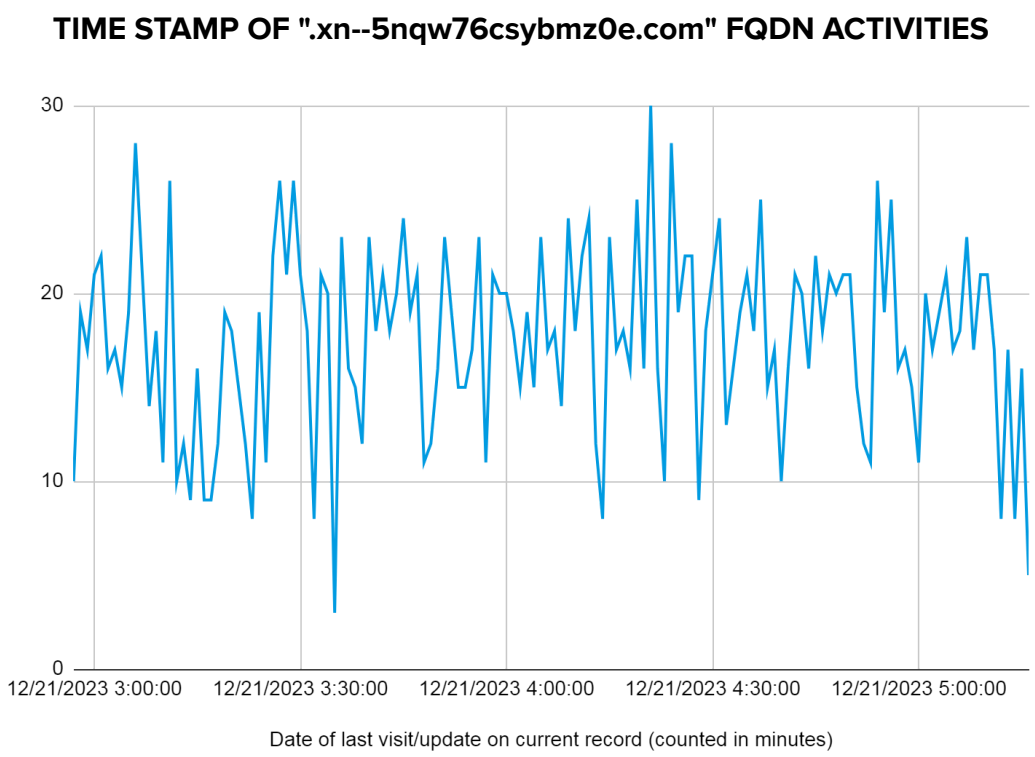
TOP 10 REGISTRANT COUNTRIES OF IDN ROOT DOMAINS

Zooming in on the FQDN Dynamics

After analyzing the ownership and attributable aspects of the FQDNs, where we determined their most used TLDs, ISPs, geolocations, and WHOIS registration preferences, we took a closer look at some of their related activities and groupings.

A Possible Software-Related Domain Got Busy

The root domain containing native-language characters xn--5nqw76csybmz0e[.]com (Punycode: 核新软件[.]com), which appears to use a software-related term, caught our attention. In addition to having as many as 2,455 FQDNs, it seems to have logged several activities over a short period. In fact, the FQDNs tallied record visits and updates over only a few hours, as seen in the chart below.



The activity of this particular domain could be indicative of a highly dynamic operation. Recording the timestamps in minutes allowed us to see the rapid updates, giving us a clear picture of how quickly these internet properties made it into the DNS or were modified.

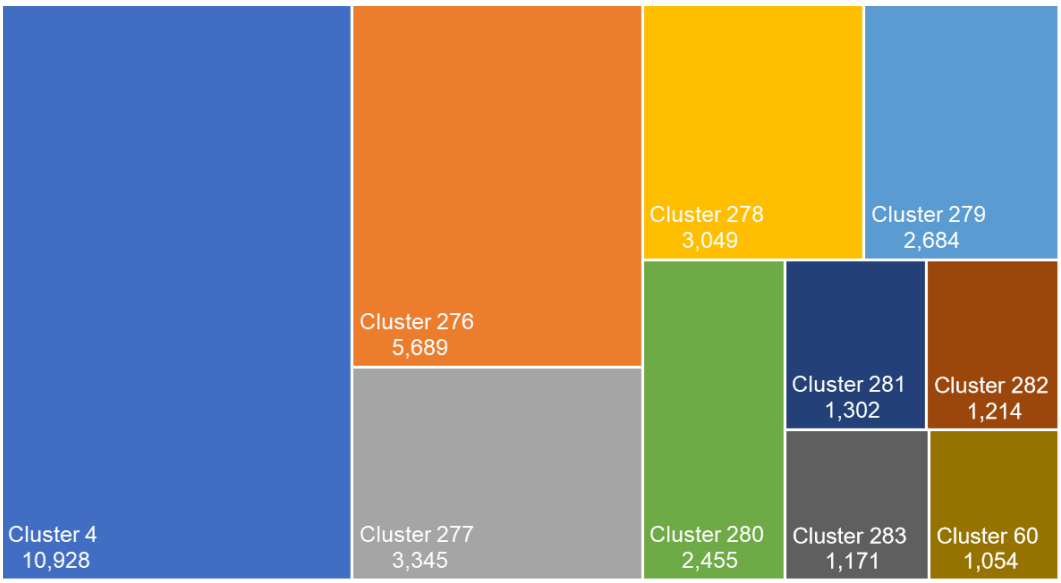
Homograph Clusters

The 63,105 FQDNs in the study can be grouped into 1,046 clusters, identified by the high degree of string similarity or the close resemblance among the members.

This characteristic translates to a ratio of about 60 FQDNs per cluster. However, 152 groups were outliers, containing more than the average number of FQDNs. How large were these homograph clusters?

The largest cluster, cluster 4, had 10,928 FQDNs or a 17.32% share. The rest of the top 10 included clusters 276 with a 9.02% share, 277 with 5.3%, 278 with 4.83%, 279 with 4.25%, 280 with 3.89%, 281 with 2.06%, 282 with 1.92%, 283 with 1.86%, and 60 with 1.67%. Overall, the top 10 largest groups accounted for 52.12% of the FQDNs.

TOP 10 LARGEST HOMOGRAPH CLUSTERS



Now, what do the domains in these groups look like? Certain clusters had varied IDN root domains and suspicious FQDNs. The table below shows some examples.

CLUSTER	SAMPLE MEMBERS	LIKELY IMITATED ENTITY
4	1stcâpital[.]xn--fiqs8s 1stcâpital[.]xn--fiqz9s 1stcépital[.]xn--fiqz9s	First Capital Bank
	1stcălgary[.]xn--fiqs8s 1stcălgăry[.]xn--fiqs8s 1stcălgăry[.]xn--fiqz9s	First Calgary Financial
60	ablearizȯna[.]xn--fiqs8s ablearizȯna[.]xn--fiqz9s ablearizona[.]xn--fiqs8s	Achieving a Better Life Experience (ABLE) Account (Arizona)
	ablevermont[.]xn--fiqz9s ablevermont[.]xn--fiqs8s ablevɐrmont[.]xn--fiqz9s	ABLE Account (Vermont)

Conclusion

The study examined the FQDNs containing native-language characters in our passive DNS database and revealed some of their most common ownership and infrastructure characteristics. We also found some suspicious usage patterns, highlighting the fact that, like any technological advancement, IDN and non-Latin script usage presents both opportunities and risks.

It is to be expected that attackers use look-alike IDNs to mimic legitimate domains and deceive users in phishing, brand impersonation, and other malicious campaigns. Therefore, organizations must be vigilant in monitoring and potentially blocking IDNs and FQDNs that closely resemble brand names or trademarks, whether their own or those of other entities.

About WhoisXML API

For over a decade now, WhoisXML API has been at the forefront of data aggregation, providing one of the most comprehensive collections of WHOIS, IP, and DNS information. Our cyber intelligence platform boasts of an impressive repository of 21+ billion historical WHOIS records, 50+ billion domains and subdomains, 10.4+ million IP netblocks, and a staggering 99.5% coverage of active IPv4 and IPv6 addresses.

Recognized for our exceptional growth and innovation, WhoisXML API has earned its place on the Inc. 5000 list and the Financial Times list of Top Fastest-Growing Companies for several years. Our solutions have garnered the trust of more than 52,000 users, including Fortune 500 companies, leading security firms, and organizations spanning diverse industries.

Appendix

Sample FQDNs

- 0-17www[.]xn--h1aedf6aze[.]1000rub[.]workato-preview[.]app
- 0-1u[.]xn--5nqw76csybmz0e[.]com
- 0-discover[.]pli[.]edu[.]hopac[.]xn--5nqw76csybmz0e[.]com
- 0-jc[.]incites[.]thomsonreuters[.]com[.]library[.]xn--dqr889hbxn[.]net
- xn--0000-angeldavidsoto276606300-0qc[.]codeanyapp[.]com
- xn--0000-angeldavidsto276606300-45v[.]codeanyapp[.]com
- xn--0000-adavidsoto276606300-lhkz37h25dtxf[.]codeanyapp[.]com
- xn--0000-ngeldvidst276606300-ubcf40cb[.]codeanyapp[.]com
- 00000-okta-idp[.]xn--coinbasegatevy-nhb334l[.]kahoot[.]it
- 00000-okta-idp[.]xn--coinlisegateway-vibr[.]kahoot[.]it
- xn--00000-kt-dp-s8a8lrc[.]xn--fwd1-2oa[.]testing1[.]kahoot[.]it
- xn--00000-kta-idp-nmb[.]fwdc1[.]testing1[.]kahoot[.]it
- 00000aaaaa[.]xn--5nqw76csybmz0e[.]com
- 000024922[.]xn--dqr889hbxn[.]net
- 000031697[.]xn--5nqw76csybmz0e[.]com
- 000032131[.]xn--dqr889hbxn[.]net
- 000032178[.]xn--dqr889hbxn[.]net
- 000052924[.]xn--5nqw76csybmz0e[.]com
- 00010[.]xn--00000-kt-dp-s8a8lrc[.]xn--fwd1-2oa[.]testing1[.]kahoot[.]it
- 00010[.]xn--00000-okta-idp-ggb[.]coinbasegateway[.]kahoot[.]it
- 000109748[.]xn--5nqw76csybmz0e[.]com
- 000gman000[.]xn--5nqw76csybmz0e[.]com
- 000ko7d[.]xn--dqr889hbxn[.]net
- 000snapscan[.]xn--5nqw76csybmz0e[.]com
- xn--000-angeldavidsoto276606300-r5v[.]codeanyapp[.]com
- 001[.]xn--ehv9b625fm7t[.]com[.]ph
- 0012806031[.]xn--dqr889hbxn[.]net

Sample Root Domains

- duckdns[.]org
- sunglimmer[.]cn
- xn--dqr889hbxn[.]net
- chargepoint[.]com
- xn--5nqw76csybmz0e[.]com
- verily[.]com
- vercel[.]app
- codetools[.]in
- netlify[.]app
- blogspot[.]com
- github[.]io
- harvard[.]edu
- involve[.]cz
- eu[.]com
- myshopify[.]com
- weworkers[.]io
- nomufy[.]com
- weeblysite[.]com
- 10cek[.]ru
- br[.]com
- du4t[.]cn
- com[.]ph
- weebly[.]com
- us[.]org
- fandom[.]com
- us[.]com
- profihunter[.]ru
- com[.]de
- iddls[.]com
- cohuashop[.]com
- amazon[.]in
- gr[.]com
- myqcloud[.]com
- cn[.]com
- web[.]app
- winkdrop[.]net
- ngrok-free[.]app
- livejournal[.]com
- uffizzi[.]com
- aternos[.]me

- hotelatour[.]cn
- ctbss[.]net
- gravitational[.]io
- logisy[.]tech
- koesio[.]dev
- guest[.]net

Sample IP Addresses

- 218[.]241[.]105[.]10
- 139[.]180[.]143[.]10
- 8[.]140[.]249[.]154
- 104[.]18[.]21[.]81
- 104[.]18[.]20[.]81
- 36[.]86[.]63[.]182
- 86[.]35[.]3[.]193
- 86[.]35[.]3[.]192
- 100[.]127[.]132[.]229
- 104[.]237[.]2[.]98
- 54[.]153[.]56[.]183
- 185[.]199[.]111[.]153
- 185[.]199[.]109[.]153
- 185[.]199[.]108[.]153
- 185[.]199[.]110[.]153
- 134[.]174[.]153[.]11
- 76[.]76[.]21[.]22
- 76[.]76[.]21[.]241
- 76[.]76[.]21[.]61
- 76[.]76[.]21[.]9
- 104[.]237[.]2[.]92
- 76[.]76[.]21[.]164
- 76[.]76[.]21[.]123
- 77[.]93[.]208[.]80
- 76[.]76[.]21[.]142
- 45[.]79[.]222[.]138
- 76[.]76[.]21[.]98
- 35[.]156[.]224[.]161
- 54[.]185[.]169[.]147
- 44[.]228[.]118[.]237
- 52[.]58[.]254[.]253
- 18[.]192[.]231[.]252
- 76[.]76[.]21[.]93
- 3[.]70[.]101[.]28
- 18[.]192[.]94[.]96
- 3[.]72[.]140[.]173
- 142[.]251[.]116[.]132
- 23[.]227[.]38[.]74