Data Sources against Malicious Domain Names: a Framework and Systematic Evaluation

Thomas Daniels[1,2,3], Jesse Davis[2,3], Maarten Bosteels[1], Rowena Schoo[4], and Pieter Robberechts[2,3]

[1]DNS Belgium
[2]Department of Computer Science, KU Leuven
[3]Leuven.AI
[4]NetBeacon Institute

June 25, 2026

Abstract

Domain names play a critical role in the operation of the internet, but are also exploited by malicious actors. Top-level domain registries, law enforcement, and cybersecurity researchers invest in efforts to combat such abuse, that span multiple goals: measurement, detection, and prevention. A high number of relevant data sources is available as input towards these goals: abuse feeds, domain registration data, domain crawls, passively collected DNS data, and others. This paper constructs a framework to systematically analyze evaluate those data sources across multiple dimensions: technical, impact, and legal/ethical. Each type of data source is then discussed following this framework. The analysis reveals that there is a high degree of variety among data sources of the same type, and combining multiple types of data sources is powerful.

1 Introduction

Domain names and the Domain Name System (DNS) play a critical role in the operation of the internet. Unfortunately, domain names are also exploited by malicious actors, for phishing, fraudulent web shops, Command-and-Control servers, spam, and other types of abuse. Such activities lead to serious financial losses for victims [100, 35, 8, 73]. For example, in 2024, Belgians lost 49 million euros to phishing alone [35].

Top-level domain (TLD) registries, law enforcement, and cybersecurity researchers actively invest in efforts to combat such abuse. These anti-abuse efforts span multiple goals. First, *measurement* is the study of volumes, trends, and characteristics of detected abuse [78, 70, 62, 67, 83]. Second, *detection* covers methods to find previously undiscovered abuse and is used as input for measurement, mitigation or even prevention [102, 107, 86, 9]. It should be noted that domain name registries and registrars usually do not host content, and mitigation at the DNS layer—takedown of the entire domain name—is very blunt,

so it may not be the most appropriate layer for all kinds of abuse [75]. Third, *prevention* aims to impede malicious activities by cybercriminals [40, 94, 21].

As input towards these goals, plenty of relevant data sources are available. These data sources span a wide variety of topics (feeds of malicious domains, web crawls, registration data, ...) and differ in their properties (accessibility, volume, completeness, ...). While overviews exist for abuse feeds [60, 34], to our knowledge there is no overview that centrally lists and discusses a wider range of (types of) data sources. Specific data sources are generally referred to in academic or industry works that use or collect that data, so the information is scattered throughout the literature and split between disciplines and venues. This means that there is no obvious starting point for registries or researchers that want to measure or tackle abuse in a data-driven manner.

To fill this gap, we aim to analyze a wide range of these data sources and highlight how they can be used against malicious domains, operationally or in research. Specifically, this work aims to fulfill the following objectives:

- To design a framework that enables uniform analysis of data sources on several dimensions: technical, impact, legal and ethical.
- To provide an overview of data sources that are useful for combating malicious activities facilitated by domain names. We cover data sources of all availabilities: publicly for free, on request, purchasable, or internal to the party fighting the abuse (registry, registrar, ISP, ...).
- To analyze and evaluate these data sources according to the framework.

Section 2 lists and describes the different types of malicious activities that can be facilitated by domain names. Section 3 describes the framework that is used to evaluate the data sources. Sections 4 to 8 discuss the different data sources. Section 9 concludes this work.

2 Types of maliciousness

There is a large variety of malicious activities that can be facilitated by domain names, of which this section gives an overview.

The established ICANN definition of DNS Abuse [43] covers botnets, malware, pharming[1], phishing, and spam (if used as delivery mechanism for the four other types). This definition is highly relevant because it captures a common understanding of technical harms that are appropriate to mitigate at DNS level for maliciously registered domains [23, 47], and the contracts for generic TLD's include obligations to mitigate evidenced DNS Abuse as per this definition.

However, many operators will go beyond this definition depending on their local policies, regulations, risk tolerance, etc. Therefore, this work also covers other malicious activities that have been subject to research and that can affect the trust in a top-level domain, even if the mitigation may in some cases be more appropriate at the hosting layer.

[1] Pharming is a form of fraud where the attacker alters the IP address of a reputable domain name, by hijacking the registration or poisoning DNS caches [44].

Malware and botnets make use of domain names in multiple fashions. First, for malware distribution: distribution URLs are a mixture of URLs with a domain name or an IP address as hostname (about 40% of URLs use a domain name, based on the recent data from URLhaus [99] at the time of writing). Second, to connect with Command-and-Control (C&C) servers: the usage of Domain Generation Algorithms (DGAs) enables attackers to register new domain names that can be found by the malware, rather than relying on a static domain [10, 83]. Third, the DNS protocol can also be used as a communication channel for C&C [22] or data exfiltration [108].

Phishing aims to deceive victims into sharing confidential information such as passwords or credit card details, by mimicking a trusted entity. Phishing websites can be hosted on maliciously registered domain names (about 80% of phishing domains globally [74]) or on an existing, legitimate website that was compromised. In the context of country-code TLDs, the approach largely depends on the target audience of the phisher: locally targeted phishing uses maliciously registered domains, internationally targeted phishing uses compromised websites [70].

Scams other than phishing also widely use domain names. Fraudulent web shops look like benign shopping websites, but orders are not delivered, or the sold products are counterfeit [102, 9, 53]. Investment scams lure victims by promising unrealistically high returns, sometimes in the context of cryptocurrency [72, 59]. Many other types of online scams exist, including technical support scams [66, 95], survey scams (in which the victim is falsely promised a reward for completion of a survey) [51], and pet scams (in which a website claims to sell pets and exploits the victim's emotional attachment to fictitious pets) [84].

Unlawful content may be made available through domain names, though how this is acted upon will vary hugely by operating jurisdiction. A recent, specific example in Belgium is the suspension of .be domain names used for illegal gambling websites through a cooperation between DNS Belgium and the national gambling authority [39].

Spam emails can use domain names in two ways. First, spammers link to domains that exhibit one of the malicious activities described above [41]. Second, domain names are used for the sending of emails, and these domains can be registered [20, 36], compromised, or spoofed [63] by a spammer.

3 Framework

To analyze and evaluate the different data sources in a structured and uniform way, we design a framework that spans several characteristics of the data sources. Due to the large number of data sources and providers, each data source will not be examined individually, but the general type of data source will be discussed. However, it would be possible to do such individual analysis according to this framework. This also makes the framework conceptually similar to existing

Listing 1: Definition of technical characteristics in the framework

Accessibility How broadly is the data available? How is access provided to the data?

Volume How large is the data source, in terms of samples and disk usage?

Time horizon Does the data source contain historical or short-term data?

Update frequency How often do the contents of the data source change?

frameworks to evaluate the quality of sports technology [87] or to measure the composition of machine learning datasets [68].

The framework consists of three dimensions. *Technical* characteristics affect the practical usage of the data source and can be derived purely from the data source or provider documentation itself. *Impact* characteristics view the data source in the context of combating malicious domain names. *Legal and ethical* characteristics discuss the usage conditions of the data and presence of personal data. The goal of these different dimensions is to give a clear picture of the data sources and help understand what is possible to achieve with them.

A one-page overview of the framework is also available in Appendix A.

3.1 Technical

Listing 1 defines the four considered technical characteristics, which are relevant because:

Accessibility affects whether or not a source is suitable. This could entail the difficulty of acquiring it which may be due to a low number of providers or the technicalities associated with collecting it. More practically, each use case may require the data to be delivered differently, e.g., it may need to be fully downloadable meaning an API-only provider would not suffice.

Volume The data sizes impacts technical requirements regarding storage as well as how quickly particular records can be found and retrieved.

Time horizon Sources vary in their time scope: sometimes long horizons are needed (e.g., a longitudinal study based on 10 years of phishing [70]) whereas in others only recent data is needed (e.g., 2.5 months of DNS traffic data to analyze behavior of malicious domains [7]).

Update frequency impacts requirements regarding how frequently new data can or should be fetched and downstream analyses be run. Update frequency is a technical property of the data source but has synergy with the impact characteristics Time-variance and Freshness.

3.2 Impact

Listing 2 defines the five considered impact characteristics, which are relevant because:

Listing 2: Definition of impact characteristics in the framework

Completeness Does the data source contain complete or partial coverage of what it aims to collect?

Generality How broadly applicable is the data source, is it rather specific (and thus only relevant for one or a few tasks) or rather general?

Time-variance Does it capture behavior that is time-variant or generally consistent over time?

Freshness Do the time-variance and update frequency align well, or is the latest available data likely to get stale?

Overlap If multiple providers offer similar data sources, how much overlap is there between them?

Listing 3: Definition of legal/ethical characteristics in the framework

License or usage conditions Under what conditions is the data available?

Privacy Does the data source include personal data?

Completeness may indicate the expected utility of a data source for a given task: it is typically desirable to have coverage as complete as possible.

Generality captures for which and how many tasks the data source is useful, and could indicate which data sources are a higher priority to collect.

Time-variance can indicate if patterns observed in a snapshot of the data remain relevant for a long time, or if frequent re-analysis is needed.

Freshness provides insights into the risk of working with stale data.

Overlap conveys how useful it is to collect data from multiple providers.

3.3 Legal/ethical

Listing 3 defines the two considered legal/ethical characteristics, which are relevant because:

License or usage conditions impact when—or if—a given data source can be used.

Privacy may affect data availability and impacts how the data should be handled, since some sources include personal or sensitive data.

4 Abuse feeds

Abuse feeds, also known as blocklists, are lists of online threats. The entries of these lists are usually URLs, domain names, or IP addresses. Abuse feeds have several use cases: blocking malicious sources at browser-level or network-level; notifications to registrants, registrars and registries about threats on domain names they manage; or training data for machine learning models to distinguish benign from malicious activities.

Typically, such feeds are not designed for the purpose of registry/registrar-level mitigation, but for network protection. They may therefore include items that are less relevant for the DNS industry, such as IP addresses. Additionally, they may have a higher tolerance for false positives than required for registry/registrar-level mitigation, and an entry on an abuse feed does not equate an evidenced report. They also do not provide a distinction between maliciously registered and compromised domains. Nevertheless, they are still a useful resource for anti-abuse research and actions in the context of domain names.

A framework to evaluate abuse feeds has been developed by ICANN [60]; this study goes into more detail and is more specifically tailored towards abuse feeds than our framework. Abuse feeds have also been extensively studied and compared by others [5, 34, 98, 1].

4.1 Technical

Accessibility. Abuse feeds are widely available from several providers. OpenPhish[2], PhishTank[3], and APWG[4] offer abuse feeds for phishing. URLHaus[5] provides an abuse feed for malware. SURBL[6], Netcraft[7], and Google Safe Browsing[8] offer feeds for multiple types of malicious activities. This overview is not exhaustive: Feal et al. studied 2,093 open feeds from 69 providers [34].

Generally, abuse feed providers offer downloadable files with a list of malicious entries. Some providers also offer APIs to check if a single entry is known to be malicious.

Volume. The number of entries strongly varies between abuse feeds [5, 34]. Across the ones analyzed by ICANN, the number of unique domain names ranged between a few thousand to hundreds of thousands per month [60].

Each entry in an abuse feed contains a limited number of textual fields, so abuse feeds with downloadable files do not demand an enormous amount of disk space.

Time horizon. Some abuse feeds only offer snapshots of the currently active malicious domains/URLs/IP addresses. Others provide a longer, historical view that contains all malicious entries added within a given time period. For

[2] https://openphish.com/
[3] https://phishtank.org/
[4] https://apwg.org/
[5] https://urlhaus.abuse.ch/
[6] https://surbl.org/
[7] https://www.netcraft.com/
[8] https://safebrowsing.google.com/

example, OpenPhish offers both: they provide feeds with active threats that refresh frequently [81], and databases with historical threats from up to 180 days ago [80].

Update frequency. The update frequencies differ a lot between abuse feeds, with only 7% of the feeds studied by Feal et al. changing at least daily [34]. Of the popular, actively updated feeds, the update frequencies generally range from the order of seconds to the order of hours. For example, SURBL's MULTI feed updates every 30-40 seconds on average [96], and OpenPhish's Community feed updates every 12 hours [81].

4.2 Impact

Completeness. Even when combined, abuse feeds cannot guarantee a complete coverage of malicious activity. A study on the .eu top-level domain found that almost 20% of registrations associated with malicious campaigns do not appear on abuse feeds [101]. Note that the opposite is also true, entries on abuse feeds may be false positives.

Generality. Some abuse feeds focus on a specific type of abuse (e.g., phishing, malware, spam). Others are more general and include multiple abuse types. Abuse feeds are generally combined with other data sources in the context of research; see the next sections.

Time-variance. Abuse feeds are highly time-variant: new threats pop up continuously, and reported threats can be taken offline. For example, the majority of phishing websites in the .nl and .ie top-level domains are mitigated within a day [70].

Freshness. The freshness of an abuse feed can be captured through two metrics proposed by ICANN: *timeliness* and *churn* [60]. Timeliness refers to the rapid addition of new threats. If abuse feeds overlap, it often happens that an entry appears on one feed hours before it does on another [60, 5]. It is not always the same feed that contains an entry first, which suggests that combining multiple feeds improves timeliness.

Churn refers to the removal of obsolete entries. Note that some abuse feeds do not remove entries, in which case they represent lists of everything that has been flagged as malicious over an extended period.

Overlap. The overlap between different abuse feeds was studied in detail by ICANN [60]. They found that the majority of overlaps are small (less than 5%), while some are much larger. This happens when openly available abuse feeds are incorporated in others.

4.3 Legal/ethical

License or usage conditions. Because of the large number of abuse feed providers, there is also a large variety in licenses and usage conditions. Some abuse feeds are openly available for free, others are available commercially.

Privacy. In principle, abuse feeds do not contain personal data.

5 (Historical) TLD registration database

The registration database of a top-level domain contains all records about current and past domain name registrations, and details about the transactions related to those registrations (creates, deletes, updates, ...).

To optimally support the registry operations, the structure of the registration database may be quite complex. Such a structure may not be ideal for analytical purposes. Therefore, we consider the *historical* registration database as a flattened timeline through all transactions on registrations. It is thus a derived database from the registration database, and its simpler structure enables more straightforward analysis of registrations over time.

A TLD registry can broadly work according to two different models: in the "thin" model, the registry only possesses the technical information associated with the registration, while in the "thick" model, the registry also possesses a copy of the contact details of the registrant [46].

5.1 Technical

Accessibility. This complete (historical) registration database is only accessible to the TLD registry operator. However, (very) limited parts of the registration database may be publicly available.

For generic top-level domains (gTLDs), access to the zone files of participating gTLDs can be requested through ICANN's Centralized Zone Data Service [45]. These files provide a mapping of registered domain names to their name servers, and DNSSEC-related data.

Some country code top-level domain (ccTLD) registries make their zone files publicly available as well, for example .se and .nu [48], .ch and .li [97], .ee [37], and .sk [77]. Additionally, .fr publishes a monthly export of various public data about their domain names [2]. When zone files or domain name lists are not public, partial lists of domain names can be constructed from public sources: OpenINTEL processes Certificate Transparency logs and Common Crawl to collect domain names [90] and publishes lists of found domains for 307 ccTLDs [79]. They measured the coverage for 19 ccTLDs and found that their data covers 43–80% of these TLDs [90].

To obtain more information about the registration of an individual domain name, one can query WHOIS [19] and/or RDAP [42] services offered by domain name registries. However, these services are not conceived for bulk requests and may not provide all details of a registration (e.g., personal data of the registrant).

Volume. The volume of the registration database varies between registries, since it is directly dependent on the number of registrations. The number of registered domains per TLD ranges from very few to millions [11]. When considering historical data, the number of transactions within a registry is expected to be an order of magnitude larger than the number of active registrations.

Time horizon. The (historical) registration database goes back as long as the TLD registry has kept transaction details. However, registries will likely remove personal information tied to old, deleted registrations after a given amount of time (see Privacy).

Update frequency. The registration database itself is updated immediately when a transaction on a registration happens. A derived historical database can be generated at a frequency chosen by the registry. The same is true when a registry publishes a subset of data, and the update frequency of those strongly varies in practice (e.g., hourly for the .se zone file, monthly for the .fr open data).

5.2 Impact

Completeness. The registration database is complete because it contains the details of all transactions (with the exception of obsolete personal data, see Privacy). The completeness of a derived historical database depends on the implementation, but is controlled by the registry. A public data export is only a subset of the registration database, the scope of which is decided by the registry.

Generality. When combined with other data sources, registration data is helpful for a variety of tasks. Together with abuse feeds (Section 4), this data can be leveraged to characterize malicious domains [70, 41, 78] or train a machine learning model to proactively detect malicious registrations [40, 94, 21]. Combined with domain crawling (Section 6), the registration database is useful to detect fraudulent web shops [102].

Time-variance. The registration database is rather time-variant, because domain names get registered and deleted every day. For example, .be sees about 15-20k new registrations per month and a similar number of deletions [25], and .nl sees tens of thousands of new registrations and deletions per month [88].

A large part of the registered domains are years old and are less time-variant, though these domains can undergo changes as well, e.g.: changes in name servers, changes in registrant contact details, transfers.

Freshness. Since the registration database itself is updated immediately on a transaction, it is fresh in the sense that it contains the latest known information. However, while the list of registered domains is always up-to-date, the underlying registration data may be outdated, e.g., due to a registrant moving but not changing their address, or a domain not being transferred while it does in practice belong to someone else. This is difficult to measure, since the actual information is unknown.

Derived databases and exports can get stale if not updated frequently, especially if new registrations and deletions are of interest. Daily zone file snapshots are found to miss at least 1% of short-lived domains, which are often registered with likely malicious intent [91].

Overlap. The (historical) registration database of a TLD does not overlap with other sources. However, overlap is possible between sources that aim to reconstruct zone files or lists of domain names based on public data.

5.3 Legal/ethical

License or usage conditions. The complete registration database is only available to the registry and its usage is governed by the registry's terms of service and privacy policy. Public exports of data subsets have their own license and usage conditions, generally covering topics such as purpose of use or limitations on download frequency.

Privacy. If the registry possesses thick registration data, then their (historical) registration database contains personal data, i.e., the contact details of the registrants, which makes this a sensitive data source. For registrations that have been deleted, registries are likely to delete the personal information after a given amount of time, compliant with their privacy policies and relevant regulations.

6 Domain crawls

Domain crawling refers to the process of automatically requesting data associated with a domain name, such as fetching the website associated with the domain or querying the domain's DNS records. Crawl data is useful both to detect malicious domain names and to gain more insights related to known malicious and benign domains.

When the subject of the crawling is specifically DNS records, it is commonly referred to as *active DNS* (where the "active" refers to the means of collection, as opposed to passively collected DNS data; Section 7).

6.1 Technical

Accessibility. Over 300 billion crawled web pages are freely available through Common Crawl [14], with billions of new pages added each month. Common Crawl publishes information about the requests, and the full HTTP response (headers and content).

A large collection of active DNS data is available for researchers through the Active DNS Project [85, 54]. This project has been collecting DNS records for large volumes of domain names since 2015.

Several domain name registries perform their own crawling on their zones [24, 104, 18, 32]. These crawls may go beyond just web data (e.g., crawling of DNS records) and can be tailored towards the specific needs of a registry, while Common Crawl only collects web data. A registry can also make sure that all domain names in their zone are crawled. The output of crawls performed by a registry is usually not public, but an internal data source.

Cybersecurity researchers may also perform crawls specific to their research subject. The domain names or URLs for crawling can be collected from public sources such as Certificate Transparency logs [59, 6], Reddit comments [9], abuse feeds [64], or zone files [67]. In some cases, the datasets gathered by these crawls are made publicly available.

Volume. The volume of the Common Crawl dataset is very large, containing petabytes of data in total. The crawl of March 2026, containing 1.97 billion pages, has a compressed size of almost 100 TiB [15].

Crawls of registries and researchers are smaller, though the exact size depends on the number of crawled domains and the scope of the crawl.

Time horizon. Common Crawl and the Active DNS Project go back to 2008 and 2015 respectively, and are still being updated at the time of writing. For crawls by registries and researchers, the time horizon depends on when collection started and, if applicable, ended. However, several registries have started crawling their zones years ago and still do so today.

Update frequency. Common Crawl is updated every month, the Active DNS Project daily. The update frequency of crawls from registries may vary but is often done monthly. Crawls from researchers collected for a specific publication are generally frozen after collection is completed.

6.2 Impact

Completeness. Considering the large scale of the web, it is impossible to crawl it in its entirety. Within the scope of a top-level domain, crawls performed by registries are complete in the sense that every domain name in the zone can be probed, but such crawls are unlikely to encompass all subpages on the websites associated with the domain names. Malicious subpages, such as for phishing, also do not often occur on the home page and are not findable starting from the home page.

Generality. Domain crawling is useful for a large variety of research, in many cases combined with abuse feeds. Examples of use cases are detecting or gaining insight in phishing domains [86, 70, 6, 71], measurement of abuse and its uptime [76, 62], detecting fraudulent web shops [102, 9] or other scams [59], classifying compromised and maliciously registered domains [64], and understanding domain drop catching [67].

Time-variance. Domain crawls are rather time-variant because domains and web pages come and go every day. However, for established web pages, a large portion may remain constant between two crawls: of the nearly 1.3 million websites[9] with a .be domain that have been successfully crawled in both February and March 2026, the textual content of almost 80% was unchanged between the two crawls.

Freshness. The freshness of a regular (say, monthly) crawl is not uniform: some websites change quickly (and thus the crawl gets stale for those), others don't. To improve freshness for crawls performed by registries, newly registered domains can be crawled shortly after their registration, rather than waiting for the next normally scheduled crawl.

[9] Specifically, the home page of the website.

Overlap. Crawls performed by different registries are generally disjoint collections, since the crawled domains belong to different top-level domains. However, a portion of the web pages collected by those crawls are also present in Common Crawl: statistics on the representation of top-level domains in Common Crawl are publicly available [16].

6.3 Legal/ethical

License or usage conditions. The usage of Common Crawl data is governed by their Terms of Use [17]. Data from the Active DNS Project is available to researchers on request. The output of crawls performed by registries is generally not publicly available.

Privacy. Domain crawls for cybersecurity purposes are not performed with the intent to collect personal data, however when performing website crawls, it is likely that personal data incidentally appears on a number of pages. This makes it less straightforward to share the collected data across organizational boundaries.

7 Passively collected DNS (traffic) data

DNS (traffic) data can be passively collected from various vantage points, such as resolvers or authoritative name servers. These locations process DNS queries and can therefore log them. Each vantage point provides a different perspective on DNS traffic; for example, the view from an authoritative TLD name server is less granular and detailed than from an ISP resolver.

The most common technique referred to as Passive DNS [103] logs DNS queries and responses in an aggregated, privacy-preserving manner. This data is collected by *sensors* deployed in networks that process DNS traffic, rather than active querying (Section 6). Passive DNS data provides the contents of DNS records associated with a domain, the time range during which a record was observed, and the number of queries in that range.

A concrete example of this kind of data, in the Passive DNS Common Output Format [31], is given in the documentation of CIRCL [13], a provider of Passive DNS:

```
1 {"rrtype": "A", "rrname": "185.194.93.14", "rdata":
    "circl.lu", "count": "19", "time_first": "1696798385",
    "time_last": "1697890824"}
2 {"rrtype": "AAAA", "rrname": "2a00:5980:93::14", "rdata":
    "circl.lu", "count": "18", "time_first": "1696798385",
    "time_last": "1697890824"}
3 {"rrtype": "MX", "rrname": "10 cppy.circl.lu", "rdata":
    "circl.lu", "count": "149", "time_first": "1696786636",
    "time_last": "1697897232"}
4 {"rrtype": "NS", "rrname": "ns1.eurodns.com", "rdata":
    "circl.lu", "count": "5", "time_first": "1696798385",
    "time_last": "1697701116"}
5 {"rrtype": "NS", "rrname": "ns2.eurodns.com", "rdata":
    "circl.lu", "count": "5", "time_first": "1696798385",
    "time_last": "1697701116"}
```

```
6 {"rrtype": "NS", "rrname": "ns3.eurodns.com", "rdata":
    "circl.lu", "count": "5", "time_first": "1696798385",
    "time_last": "1697701116"}
7 {"rrtype": "NS", "rrname": "ns4.eurodns.com", "rdata":
    "circl.lu", "count": "5", "time_first": "1696798385",
    "time_last": "1697701116"}
8 {"rrtype": "SOA", "rrname": "ns1.eurodns.com
    hostmaster.eurodns.com 2023091306 43200 7200 1209600
    86400", "rdata": "circl.lu", "count": "260",
    "time_first": "1696780845", "time_last": "1697183586"}
```

For example, the first line of this data tells that 19 `A` requests to `circl.lu` have been observed between 2023-10-08 20:53:05 and 2023-10-21 14:20:24 UTC with as response `185.194.93.14`.

Another passively collected data source consists of the query logs of an authoritative TLD name server. Such query logs do provide details about each individual query, as opposed to the aggregated Passive DNS data. However, these query logs are more coarse-grained, because resolvers heavily make use of caching, so not all DNS queries will induce a query to the authoritative TLD name servers. Moreover, the query response is less informative, because an authoritative TLD name server's response is limited to the name server of the queried domain (host name and/or IP address) and DNSSEC-related data.

Individual queries can be logged in a similar manner at DNS resolvers, such as public resolvers or private resolvers at an ISP, university network, or enterprise network [61, 82]. The scope of these logs consists of all DNS queries within the monitored network. This is more granular and detailed than aggregated Passive DNS logs and than logs from an authoritative TLD name server.

7.1 Technical

Accessibility. Different providers offer Passive DNS data, though this is not as widespread as abuse feeds. DomainTools[10] [30], Spamhaus [93], and WhoisXML API [3] provide Passive DNS data commercially. CIRCL [13] offers Passive DNS access to trusted partners. Mnemonic [69] publicly provides limited Passive DNS data. Generally these services are accessed by querying their API, though some also offer a full database export [3, 28].

TLD query logs are not offered publicly, and are internal data sources to the registries that collect them. Registries can use ENTRADA [105] to process DNS query logs from PCAP files and transform them into Parquet files for more convenient analysis.

Query logs from public resolvers or ISPs are typically not accessible to researchers, often due to privacy concerns [107]. A relevant project in this context is DNS TAPIR [26], offering a platform that enables collection of DNS traffic data from ISP resolvers in a privacy-preserving manner [27]. At the time of writing, this project is in its test phase.

Volume. The volume of available Passive DNS depends on the provider, but can grow to tens or hundreds of billions of records [3, 28]. DomainTools mentions processing 200,000 observations per second [28], and Spamhaus 200 million per

[10] Also known as the Farsight database. DomainTools acquired Farsight Security in 2021.

hour [33]. The DomainTools DNSDB export requires at least tens of terabytes of storage [28].

For TLD query logs, the volume mainly depends on the size and popularity of the TLD. The authoritative .be name servers receive over 2 billion queries per day, the logs of which use tens of gigabytes of space.

The volume of query log data from resolvers depends on how much logging is in place, and also on the number of users of the resolver and how much traffic these users generate.

Time horizon. Passive DNS sources provide historical data, and it depends on the provider how far back this goes. DomainTools has data since 2010 [29], WhoisXML API since 2008 [4], and Spamhaus has archived data since 2014 but their API only provides access to one year's worth of data [92].

The longevity of TLD or resolver query logs or depends on how long the operator chooses to keep the logs.

Update frequency. Passively collected DNS data is generally updated in near-real-time.

7.2 Impact

Completeness. The completeness of passively collected DNS data is dependent on the vantage point. For Passive DNS, completeness could be expressed in terms of how much of global traffic is captured, or how many (sub)domains or DNS records are discovered. Passive DNS data is collected from a number of sensors deployed globally, but where those sensors are located is generally not publicly shared. None of these completeness metrics seem trivial to estimate.

TLD or resolver query logs are complete in the sense that all queries within their scope can be logged if desired. However, this scope may be limited: TLD query logs only provide a coarse view, public resolvers are limited to the set of users that query them, and private resolver query logs are limited to the network in which they reside.

Generality. Passive DNS has been used for a variety of tasks: detecting spam domains [36], identifying botnets and malware campaigns [12, 106], or detecting malicious domains in general without focusing on one specific type [50, 7]. An extensive survey on the detection of malicious domains through DNS data (not limited to Passive DNS) is given by Zhauniarovich et al. [107].

TLD query logs appear in fewer works than Passive DNS, though have also been successfully used to identify suspicious domains [105, 89] and botnets [105]. Query logs from enterprise network resolvers have been applied to the detection of malware [61, 82].

Time-variance. Passively collected DNS data is very time-variant due to new domains, a continuous stream of traffic, and—when responses are logged—changing DNS records.

Freshness. Because passively collected DNS data is updated in near-real-time, the data will be fresh.

Overlap. Given the large quantities of data collected by Passive DNS providers, it is inevitable that there is overlap in their observed traffic or DNS records. However, this cannot be quantified without having access to the complete archives of all providers.

TLD query logs do not have much overlap with Passive DNS sources or resolver query logs, given their different contents and points of view. Naturally, TLD query logs from different registries also do not overlap.

7.3 Legal/ethical

License or usage conditions. Each Passive DNS provider has their own license and conditions for the usage of their data. TLD or resolver query log data is generally not offered publicly.

Privacy. Passive DNS data is aggregated and does not contain any identification of the querying clients, so this preserves privacy. TLD query log data does include information about individual queries, including the client. Therefore, even though this client will identify a resolver rather than an individual end user for the vast majority of queries, TLD query log data is more sensitive than Passive DNS data. Individual query log data from resolvers is the most sensitive, because the client does in this case identify an individual. Even under changing client IP addresses, different sessions of a user could be linked through patterns in their DNS traffic [52].

8 Miscellaneous sources

This section shortly discusses a few data sources that do not fit under the previous sections but are still relevant for the detection of malicious domains. These sources are also used in several of the cited works.

Certificate Transparency logs [55, 56] are public, append-only logs of the issuance of TLS certificates. In the context of malicious domain detection, they are often used to discover the existence of (sub)domains. In 2023, the Certificate Transparency logs covered 52% of domains in 19 ccTLD zones [90].

DGArchive [83, 38] is a database of DGA-generated domain names (commonly used in malware; Section 2). This database contains pre-computed domain names from reverse-engineered DGAs, of which only a subset has ever been registered. This distinguishes DGArchive from abuse feeds (Section 4).

Tranco [58, 57] is a ranked list of popular websites, robust against manipulation by adversaries. In the context of malicious domain detection, this list is useful (albeit biased) to sample benign domains and websites from.

IP data accessible through providers such as MaxMind [65] and IPinfo [49] provides contextual data for IP addresses, such as their Autonomous System or approximate location.

9 Conclusion

Domain names are exploited by cybercriminals to facilitate malicious activities. This work provided an overview of relevant data sources for TLD registries or cybersecurity researchers to combat such abuse, systematically evaluated through a framework we constructed. At a high level, we analyzed abuse feeds, TLD registration databases, domain crawls, passively collected DNS (traffic) data, and miscellaneous sources across multiple dimensions.

The analysis reveals that **there is a lot of variety between data sources of the same type.** Data sources typically have multiple providers or collection methods, which results in differences across all dimensions and characteristics. The presence of multiple options makes it more likely that a good fit exists for a specific task, but also adds a lot of complexity.

Another important conclusion is that **combining multiple types of data sources is powerful.** Abuse feeds play a role in a lot of works, but also other types of data sources have been combined, such as domain registration data and crawling [70, 102], or crawling and passively collected DNS data [64, 36]. This highlights the value in using multiple types of data sources, and it is plausible that some combinations offer undiscovered potential.

Acknowledgements. TD received funding from VLAIO (Flemish Innovation & Entrepreneurship) through the Baekeland PhD mandate [HBC.2023.0718]. JD and PR acknowledge support from the Flemish government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

[1] Antonia Affinito, Alessio Botta, and Anna Sperotto. *Unveiling Domain Blocklist Performance: An Analysis over Four Years*. https://ripe89.ripe.net/wp-content/uploads/presentations/68-Affinito-Presentation.pdf. RIPE89 presentation. 2024. (Visited on Mar. 28, 2026).

[2] Afnic. *Shared data: reuse .fr data*. https://www.afnic.fr/en/products-and-services/fr-and-associated-services/shared-data-reuse-fr-data/. (Visited on Apr. 14, 2026).

[3] WhoisXML API. *DNS History for Enhanced Cybersecurity*. https://dns-history.whoisxmlapi.com/. (Visited on Apr. 23, 2026).

[4] WhoisXML API. *Passive DNS: A Complete Primer*. https://dns-history.whoisxmlapi.com/blog/passive-dns. 2024. (Visited on Apr. 29, 2026).

[5] Simon Bell and Peter Komisarczuk. "An Analysis of Phishing Blacklists: Google Safe Browsing, OpenPhish, and PhishTank". In: *Proceedings of the Australasian Computer Science Week, ACSW 2020, Melbourne, VIC, Australia, February 3-7, 2020*. Ed. by Prem Prakash Jayaraman et al. ACM, 2020, 3:1–3:11. DOI: 10.1145/3373017.3373020.

[6] Hugo L. J. Bijmans et al. “Catching Phishers By Their Bait: Investigating the Dutch Phishing Landscape through Phishing Kit Detection”. In: *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*. Ed. by Michael D. Bailey and Rachel Greenstadt. USENIX Association, 2021, pp. 3757–3774. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/bijmans.

[7] Leyla Bilge et al. “EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis”. In: *Proceedings of the Network and Distributed System Security Symposium, NDSS 2011, San Diego, California, USA, 6th February - 9th February 2011*. The Internet Society, 2011. URL: https://www.ndss-symposium.org/ndss2011/exposure-finding-malicious-domains-using-passive-dns-analysis.

[8] BioCatch. *French lose €7.6 billion to scams in 12 months*. https://www.biocatch.com/press-release/french-lose-billions-to-scams-in-12-months. 2025. (Visited on May 29, 2026).

[9] Marzieh Bitaab et al. “Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale”. In: *2023 IEEE Symposium on Security and Privacy (SP)*. 2023 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, May 2023, pp. 2566–2583. ISBN: 978-1-6654-9336-9. DOI: 10.1109/SP46215.2023.10179461. URL: https://ieeexplore.ieee.org/document/10179461/ (visited on Jan. 12, 2024).

[10] Bogdan Constantin Cebere et al. “Down to earth! Guidelines for DGA-based Malware Detection”. In: *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*. RAID ’24. Padua, Italy: Association for Computing Machinery, 2024, pp. 147–165. ISBN: 9798400709593. DOI: 10.1145/3678890.3678913.

[11] CENTR. *Registration Statistics: Total Domains by TLD*. https://stats.centr.org/public/registrations. (Visited on June 18, 2026).

[12] Hyunsang Choi and Heejo Lee. “Identifying botnets by capturing group activities in DNS traffic”. In: *Computer Networks* 56.1 (2012), pp. 20–33. ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2011.07.018. URL: https://www.sciencedirect.com/science/article/pii/S1389128611002787.

[13] CIRCL. *Passive DNS*. https://www.circl.lu/services/passive-dns/. (Visited on Apr. 23, 2026).

[14] Common Crawl. *Common Crawl - Open Repository of Web Crawl Data*. https://commoncrawl.org/. 2007. (Visited on Apr. 17, 2026).

[15] Common Crawl. *Common Crawl March 2026 Crawl Archive (CC-MAIN-2026-12)*. https://data.commoncrawl.org/crawl-data/CC-MAIN-2026-12/index.html. 2026. (Visited on Apr. 17, 2026).

[16] Common Crawl. *Statistics of Common Crawl Monthly Archives - Top-Level Domains*. https://commoncrawl.github.io/cc-crawl-statistics/plots/tlds. (Visited on Apr. 17, 2026).

[17] Common Crawl. *Terms of Use*. https://commoncrawl.org/terms-of-use. 2024. (Visited on Apr. 17, 2026).

[18] CZ.NIC. *Launching DNS Crawler*. https://en.blog.nic.cz/2020/06/01/launching-dns-crawler/. 2020. (Visited on Apr. 17, 2026).

[19] Leslie Daigle. *WHOIS Protocol Specification*. RFC 3912. Sept. 2004. DOI: 10.17487/RFC3912. URL: https://www.rfc-editor.org/info/rfc3912.

[20] Kenya Dan et al. "Spam Domain Detection Method Using Active DNS Data and E-Mail Reception Log". In: *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 1. 2019, pp. 896–899. DOI: 10.1109/COMPSAC.2019.00133.

[21] Thomas Daniels et al. "RegCheck: A Real-Time Approach for Flagging Potentially Malicious Domain Name Registrations". In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. KDD '25: The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Toronto ON Canada: ACM, Aug. 3, 2025, pp. 4375–4386. ISBN: 979-8-4007-1454-2. DOI: 10.1145/3711896.3737260. URL: https://dl.acm.org/doi/10.1145/3711896.3737260.

[22] Christian J. Dietrich et al. "On Botnets That Use DNS for Command and Control". In: *2011 Seventh European Conference on Computer Network Defense*. 2011, pp. 9–16. DOI: 10.1109/EC2ND.2011.16.

[23] DNS Abuse Framework. *Framework to Address Abuse*. https://dnsabuseframework.org/media/files/2020-05-29_DNSAbuseFramework.pdf. 2020. (Visited on June 11, 2026).

[24] DNS Belgium. *Our open source crawler*. https://www.dnsbelgium.be/en/news/open-source-crawler. 2022. (Visited on Apr. 17, 2026).

[25] DNS Belgium. *Statistics*. https://www.dnsbelgium.be/en/statistics. (Visited on Apr. 15, 2026).

[26] DNS Tapir. *A privacy-preserving operational platform for DNS threat detection and shared situational awareness*. https://www.dnstapir.se/. (Visited on May 11, 2026).

[27] DNS Tapir. *Information Management*. https://www.dnstapir.se/docs/tapir-info-mgmt-en/. (Visited on May 11, 2026).

[28] DomainTools. *DNSDB Export Requirements*. https://docs.domaintools.com/api/dnsdb/export/requirements/. (Visited on Apr. 29, 2026).

[29] DomainTools. *DNSDB FAQ*. https://docs.domaintools.com/api/dnsdb/faq/. (Visited on Apr. 29, 2026).

[30] DomainTools. *Farsight DNSDB API*. https://docs.domaintools.com/api/dnsdb/. (Visited on Apr. 23, 2026).

[31] Alexandre Dulaunoy et al. *Passive DNS - Common Output Format*. Internet-Draft draft-dulaunoy-dnsop-passive-dns-cof-13. Work in Progress. Internet Engineering Task Force, Aug. 2024. 14 pp. URL: https://datatracker.ietf.org/doc/draft-dulaunoy-dnsop-passive-dns-cof/13/.

[32] DENIC eG. *DENIC Crawler*. https://www.denic.de/en/crawler/. (Visited on Apr. 17, 2026).

[33] Milly Fawcett. *What is Passive DNS? A beginner's guide*. https://www.spamhaus.com/resource-center/what-is-passive-dns-a-beginners-guide/. 2022. (Visited on Apr. 29, 2026).

[34] Alvaro Feal et al. "Blocklist Babel: On the Transparency and Dynamics of Open Source Blocklisting". In: *IEEE Transactions on Network and Service Management* 18.2 (June 2021), pp. 1334–1349. ISSN: 1932-4537, 2373-7379. DOI: 10.1109/TNSM.2021.3075552. URL: https://ieeexplore.ieee.org/document/9416274/.

[35] Febelfin. *Figures 2024: 'If it smells phishy, it probably is!'* https://febelfin.be/en/themes/fraud-security/numbers-and-trends/figures-2024-if-it-smells-phishy-it-probably-is. 2025. (Visited on May 29, 2026).

[36] Simon Fernandez, Maciej Korczyński, and Andrzej Duda. "Early Detection of Spam Domains with Passive DNS and SPF". In: *Passive and Active Measurement*. Ed. by Oliver Hohlfeld, Giovane Moura, and Cristel Pelsser. Cham: Springer International Publishing, 2022, pp. 30–49. ISBN: 978-3-030-98785-5. DOI: 10.1007/978-3-030-98785-5_2.

[37] Estonian Internet Foundation. *.ee zone file*. https://www.internet.ee/domains/ee-zone-file. (Visited on Apr. 14, 2026).

[38] Fraunhofer FKIE. *DGArchive*. https://dgarchive.caad.fkie.fraunhofer.de/. (Visited on May 5, 2026).

[39] Gaming Commission. *Samenwerkingsakkoord met DNS Belgium*. https://kansspelcommissie.be/en/samenwerkingsakkoord-met-dns-belgium. Dec. 15, 2025. (Visited on June 10, 2026).

[40] Shuang Hao et al. "PREDATOR: Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16: 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna Austria: ACM, Oct. 24, 2016, pp. 1568–1579. ISBN: 978-1-4503-4139-4. DOI: 10.1145/2976749.2978317. URL: https://dl.acm.org/doi/10.1145/2976749.2978317.

[41] Shuang Hao et al. "Understanding the domain registration behavior of spammers". In: *Proceedings of the 2013 conference on Internet measurement conference*. IMC '13. New York, NY, USA: Association for Computing Machinery, Oct. 23, 2013, pp. 63–76. ISBN: 978-1-4503-1953-9. DOI: 10.1145/2504730.2504753. URL: https://dl.acm.org/doi/10.1145/2504730.2504753.

[42] Scott Hollenbeck and Andy Newton. *Registration Data Access Protocol (RDAP) Query Format*. RFC 9082. June 2021. DOI: 10.17487/RFC9082. URL: https://www.rfc-editor.org/info/rfc9082.

[43] ICANN. *Acronyms and Terms. Domain Name System Abuse (DNS Abuse)*. https://www.icann.org/en/icann-acronyms-and-terms/domain-name-system-abuse-en. (Visited on June 19, 2026).

[44] ICANN. *Acronyms and Terms. Pharming*. https://www.icann.org/en/icann-acronyms-and-terms/pharming-en. (Visited on June 19, 2026).

[45] ICANN. *Centralized Zone Data Service*. https://czds.icann.org/home. (Visited on June 25, 2026).

[46] ICANN. *Thick WHOIS Transition Policy for .COM, .NET, and .JOBS*. https://www.icann.org/en/contracted-parties/consensus-policies/thick-registry-registration-data-directory-services-transition-policy/thick-whois-transition-policy-for-com-net-and-jobs-01-02-2017-en. 2017. (Visited on Apr. 14, 2026).

[47] Internet & Jurisdiction Policy Network. *Domains & Jurisdiction Program - Operational Approaches - Norms, Criteria, Mechanisms*. https://www.internetjurisdiction.net/uploads/pdfs/Papers/Domains-Jurisdiction-Program-Operational-Approaches.pdf. 2019. (Visited on June 11, 2026).

[48] Internetstiftelsen. *Zone Data*. https://internetstiftelsen.se/en/zone-data/. (Visited on Apr. 14, 2026).

[49] IPinfo. *IP Data Intelligence for Developers & Enterprises*. https://ipinfo.io/. (Visited on May 5, 2026).

[50] Issa Khalil, Ting Yu, and Bei Guan. "Discovering Malicious Domains through Passive DNS Data Graph Analysis". In: *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2016, Xi'an, China, May 30 - June 3, 2016*. Ed. by Xiaofeng Chen, XiaoFeng Wang, and Xinyi Huang. ACM, 2016, pp. 663–674. DOI: 10.1145/2897845.2897877.

[51] Amin Kharraz, William Robertson, and Engin Kirda. "Surveylance: Automatically Detecting Online Survey Scams". In: *2018 IEEE Symposium on Security and Privacy (SP)*. 2018, pp. 70–86. DOI: 10.1109/SP.2018.00044.

[52] Matthias Kirchler et al. "Tracked Without a Trace: Linking Sessions of Users by Unsupervised Learning of Patterns in Their DNS Traffic". In: *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*. AISec '16. Vienna, Austria: Association for Computing Machinery, 2016, pp. 23–34. ISBN: 9781450345736. DOI: 10.1145/2996758.2996770.

[53] Platon Kotzias et al. "Scamdog Millionaire: Detecting E-commerce Scams in the Wild". In: *Proceedings of the 39th Annual Computer Security Applications Conference*. ACSAC '23. New York, NY, USA: Association for Computing Machinery, Dec. 4, 2023, pp. 29–43. ISBN: 979-8-4007-0886-2. DOI: 10.1145/3627106.3627184. URL: https://dl.acm.org/doi/10.1145/3627106.3627184.

[54] Athanasios Kountouras et al. "Enabling Network Security Through Active DNS Datasets". In: *Research in Attacks, Intrusions, and Defenses*. Ed. by Fabian Monrose et al. Cham: Springer International Publishing, 2016, pp. 188–208. ISBN: 978-3-319-45719-2.

[55] Ben Laurie, Adam Langley, and Emilia Kasper. *Certificate Transparency*. RFC 6962. June 2013. DOI: 10.17487/RFC6962. URL: https://www.rfc-editor.org/info/rfc6962.

[56] Ben Laurie, Eran Messeri, and Rob Stradling. *Certificate Transparency Version 2.0*. RFC 9162. Dec. 2021. DOI: 10.17487/RFC9162. URL: https://www.rfc-editor.org/info/rfc9162.

[57] Victor Le Pochat et al. *Tranco*. https://tranco-list.eu/. (Visited on May 5, 2026).

[58] Victor Le Pochat et al. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation". In: *Proceedings 2019 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium. San Diego, CA: Internet Society, 2019. ISBN: 978-1-891562-55-6. DOI: 10.14722/ndss.2019.23386. URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_01B-3_LePochat_paper.pdf.

[59] Xigao Li, Anurag Yepuri, and Nick Nikiforakis. "Double and Nothing: Understanding and Detecting Cryptocurrency Giveaway Scams". In: *Proceedings 2023 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium. San Diego, CA, USA: Internet Society, 2023. ISBN: 978-1-891562-83-9. DOI: 10.14722/ndss.2023.24584. URL: https://www.ndss-symposium.org/wp-content/uploads/2023/02/ndss2023_f584_paper.pdf (visited on Jan. 12, 2024).

[60] Siôn Lloyd, Carlos Hernández-Gañán, and Samaneh Tajalizadehkhoob. "RBL Evaluation Methodology". In: *ICANN Office of the Chief Technology Officer (OCTO) Publications* (2023). URL: https://www.icann.org/en/system/files/files/octo-037-11dec23-en.pdf (visited on June 25, 2026).

[61] Pratyusa K. Manadhata et al. "Detecting Malicious Domains via Graph Inference". In: *Computer Security - ESORICS 2014*. Vol. 8712. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 1–18. ISBN: 978-3-319-11203-9. DOI: 10.1007/978-3-319-11203-9_1. URL: http://link.springer.com/10.1007/978-3-319-11203-9_1 (visited on Nov. 3, 2025).

[62] Sourena Maroofi and Maciej Korczyński. *NetBeacon Measurement and Analytics Platform (MAP): Methodology*. https://netbeacon.org/wp-content/uploads/2026/02/NetBeaconMAP-Methodology.pdf. (Visited on May 12, 2026).

[63] Sourena Maroofi et al. "Adoption of Email Anti-Spoofing Schemes: A Large Scale Analysis". In: *IEEE Transactions on Network and Service Management* 18.3 (2021), pp. 3184–3196. DOI: 10.1109/TNSM.2021.3065422.

[64] Sourena Maroofi et al. "COMAR: Classification of Compromised versus Maliciously Registered Domains". In: *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2020 IEEE European Symposium on Security and Privacy (EuroS&P). Genoa, Italy: IEEE, Sept. 2020, pp. 607–623. ISBN: 978-1-7281-5087-1. DOI: 10.1109/EuroSP48549.2020.00045. URL: https://ieeexplore.ieee.org/document/9230367/ (visited on Apr. 17, 2026).

[65] MaxMind. *MaxMind GeoIP® Databases*. https://www.maxmind.com/en/geoip-databases. (Visited on May 5, 2026).

[66] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. "Dial One for Scam: A Large-Scale Analysis of Technical Support Scams". In: *Proceedings 2017 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium. San Diego, CA: Internet Society, 2017. ISBN: 978-1-891562-46-4. DOI: 10.14722/ndss.2017.23163. URL: https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/dial-one-scam-large-scale-analysis-technical-support-scams/.

[67] Najmeh Miramirkhani et al. "Panning for gold.com: Understanding the Dynamics of Domain Dropcatching". In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. the 2018 World Wide Web Conference. Lyon, France: ACM Press, 2018, pp. 257–266. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186092. (Visited on Jan. 12, 2024).

[68] Margaret Mitchell et al. *Measuring Data*. Feb. 13, 2023. DOI: 10.48550/arXiv.2212.05129. arXiv: 2212.05129[cs.AI]. URL: http://arxiv.org/abs/2212.05129.

[69] Mnemonic. *PassiveDNS Integration Guide*. https://docs.mnemonic.no/api/services/pdns/. (Visited on Apr. 23, 2026).

[70] Giovane C. M. Moura et al. "Characterizing and Mitigating Phishing Attacks at ccTLD Scale". In: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*. Ed. by Bo Luo et al. ACM, 2024, pp. 2147–2161. DOI: 10.1145/3658644.3690192.

[71] Morium Akter Munny et al. *Infrastructure Patterns in Toll Scam Domains: A Comprehensive Analysis of Cybercriminal Registration and Hosting Strategies*. Oct. 16, 2025. DOI: 10.48550/arXiv.2510.14198. arXiv: 2510.14198[cs]. URL: http://arxiv.org/abs/2510.14198.

[72] Muhammad Muzammil et al. "The Poorest Man in Babylon: A Longitudinal Study of Cryptocurrency Investment Scams". In: *Proceedings of the ACM on Web Conference 2025*. WWW '25. Sydney NSW, Australia: Association for Computing Machinery, 2025, pp. 1034–1045. ISBN: 9798400712746. DOI: 10.1145/3696410.3714588. URL: https://doi.org/10.1145/3696410.3714588.

[73] National Anti-Scam Centre (Australia). *National Anti-Scam Centre in Action - Quarterly update - April to June 2024*. https://www.nasc.gov.au/system/files/NASC-quarterly-update-Q4-2024.pdf. 2024. (Visited on June 11, 2026).

[74] NetBeacon. *Interactive charts*. https://netbeacon.org/map/interactive-charts/. (Visited on June 10, 2026).

[75] NetBeacon Institute. *DNS Abuse Definition: Attributes of Mitigation*. https://netbeacon.org/dns-abuse-definition-attributes-of-mitigation/. 2021. (Visited on June 11, 2026).

[76] NetBeacon Institute. *Measurement and Analytics Platform*. https://netbeacon.org/map/. (Visited on May 12, 2026).

[77] SK-NIC. *domains.txt - zoznam registrovanych domen v zivom stave / list of registered domains in live status*. https://sk-nic.sk/subory/domains.txt. (Visited on Apr. 14, 2026).

[78] Yevheniya Nosyk et al. "Exposing the Roots of DNS Abuse: A Data-Driven Analysis of Key Factors Behind Phishing Domain Registrations". In: *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*. CCS '25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 618–632. ISBN: 9798400715259. DOI: 10.1145/3719027.3744869.

[79] OpenINTEL. *Domain List - ccTLD Names*. https://openintel.nl/data/domain-lists/cctld-names/. (Visited on Apr. 14, 2026).

[80] OpenPhish. *OpenPhish Database*. https://openphish.com/phishing_database.html. Accessed on 2026-03-26.

[81] OpenPhish. *Phishing Feeds*. https://openphish.com/phishing_feeds.html. Accessed on 2026-01-15.

[82] Alina Oprea et al. "Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data". In: *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*. 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). Rio de Janeiro, Brazil: IEEE, June 2015, pp. 45–56. ISBN: 978-1-4799-8629-3. DOI: 10.1109/DSN.2015.14. URL: https://ieeexplore.ieee.org/document/7266837 (visited on Oct. 28, 2025).

[83] Daniel Plohmann et al. "A Comprehensive Measurement Study of Domain Generating Malware". In: *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, Aug. 2016, pp. 263–278. ISBN: 978-1-931971-32-4. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/plohmann.

[84] Benjamin Price and Matthew Edwards. "Resource Networks of Pet Scam Websites". In: *2020 APWG Symposium on Electronic Crime Research (eCrime)*. 2020, pp. 1–10. DOI: 10.1109/eCrime51433.2020.9493253.

[85] Active DNS Project. https://activednsproject.org/. (Visited on Apr. 24, 2026).

[86] Routhu Srinivasa Rao and Alwyn Roshan Pais. "Detection of phishing websites using an efficient feature-based machine learning framework". In: *Neural Computing and Applications* 31 (2018), pp. 3851–3873.

[87] S. Robertson et al. "Development of a sports technology quality framework". In: *Journal of Sports Sciences* 41.22 (Nov. 17, 2023), pp. 1983–1993. ISSN: 0264-0414, 1466-447X. DOI: 10.1080/02640414.2024.2308435. URL: https://www.tandfonline.com/doi/full/10.1080/02640414.2024.2308435.

[88] SIDN Labs. *.nl statistics - domain names*. https://stats.sidnlabs.nl/en/registration.html. (Visited on Apr. 15, 2026).

[89] Marcos Rogério Silveira, Adriano Mauro Cansian, and Hugo Koji Kobayashi. "Semi-supervised approach for detecting malicious domains in TLDs in their first query". In: *International Journal of Information Security* 24.2 (Apr. 2025), p. 80. ISSN: 1615-5262, 1615-5270. DOI: 10.1007/s10207-025-00996-3. URL: https://link.springer.com/10.1007/s10207-025-00996-3.

[90] Raffaele Sommese, Roland van Rijswijk-Deij, and Mattijs Jonker. "This Is a Local Domain: On Amassing Country-Code Top-Level Domains from Public Data". In: *SIGCOMM Comput. Commun. Rev.* 54.2 (Aug. 2024), pp. 2–9. ISSN: 0146-4833. DOI: 10.1145/3687234.3687236. URL: https://doi.org/10.1145/3687234.3687236.

[91] Raffaele Sommese et al. "DarkDNS: Revisiting the Value of Rapid Zone Update". In: *Proceedings of the 2024 ACM on Internet Measurement Conference*. IMC '24. Madrid, Spain: Association for Computing Machinery, 2024, pp. 454–461. ISBN: 9798400705922. DOI: 10.1145/3646547.3689021. URL: https://doi.org/10.1145/3646547.3689021.

[92] Spamhaus. *FAQs — Passive DNS*. https://www.spamhaus.com/faqs/passive-dns/. (Visited on Apr. 29, 2026).

[93] Spamhaus. *Passive DNS API*. https://www.spamhaus.com/data-access/passive-dns-api/. (Visited on Apr. 23, 2026).

[94] Jan Spooren et al. "Premadoma: an operational solution for DNS registries to prevent malicious domain registrations". In: *Proceedings of the 35th Annual Computer Security Applications Conference*. ACSAC '19: 2019 Annual Computer Security Applications Conference. San Juan Puerto Rico USA: ACM, Dec. 9, 2019, pp. 557–567. ISBN: 978-1-4503-7628-0. DOI: 10.1145/3359789.3359836. URL: https://dl.acm.org/doi/10.1145/3359789.3359836.

[95] Bharat Srinivasan et al. "Exposing Search and Advertisement Abuse Tactics and Infrastructure of Technical Support Scammers". In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 319–328. ISBN: 9781450356398. DOI: 10.1145/3178876.3186098.

[96] SURBL. *Our Reputation Lists*. https://www.surbl.org/lists. Accessed on 2026-01-15.

[97] Switch. *Freely available data*. https://portal.switch.ch/pub/open-data/. (Visited on Apr. 14, 2026).

[98] Mitsuhiro Umizaki et al. "Understanding the Characteristics of Public Blocklist Providers". In: *IEEE Symposium on Computers and Communications, ISCC 2022, Rhodes, Greece, June 30 - July 3, 2022*. IEEE, 2022, pp. 1–7. DOI: 10.1109/ISCC55528.2022.9913009.

[99] URLhaus. *URLhaus Community API*. https://urlhaus.abuse.ch/api/. (Visited on June 23, 2026).

[100] US Federal Bureau of Investigation's Internet Crime Complaint Center. *Internet Crime Report 2025*. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf. 2026. (Visited on May 29, 2026).

[101] Thomas Vissers et al. “Exploring the Ecosystem of Malicious Domain Registrations in the .eu TLD”. In: *Research in Attacks, Intrusions, and Defenses*. Ed. by Marc Dacier et al. Cham: Springer International Publishing, 2017, pp. 472–493. ISBN: 978-3-319-66332-6. DOI: 10.1007/978-3-319-66332-6_21.

[102] Thymen Wabeke et al. “Counterfighting Counterfeit: Detecting and Taking down Fraudulent Webshops at a ccTLD”. In: *Passive and Active Measurement - 21st International Conference, PAM 2020, Eugene, Oregon, USA, March 30-31, 2020, Proceedings*. Ed. by Anna Sperotto, Alberto Dainotti, and Burkhard Stiller. Vol. 12048. Lecture Notes in Computer Science. Springer, 2020, pp. 158–174. DOI: 10.1007/978-3-030-44081-7_10. URL: https://doi.org/10.1007/978-3-030-44081-7%5C_10.

[103] Florian Weimer. “Passive DNS Replication”. In: 17th Annual FIRST Conference. 2005.

[104] Maarten Wullink, Giovane C. M. Moura, and Cristian Hesselman. “Dmap: Automating Domain Name Ecosystem Measurements and Applications”. In: *IFIP/IEEE Network Traffic Measurement and Analysis Conference (TMA 2018)*. June 2018.

[105] Maarten Wullink et al. “ENTRADA: Enabling DNS Big Data Applications”. In: *APWG Symposium on Electronic Crime Research (eCRIME 2016), Toronto, ON, Canada. June 1, 2, and 3, 2016*. 2016.

[106] Jialong Zhang et al. “Systematic Mining of Associated Server Herds for Malware Campaign Discovery”. In: *2015 IEEE 35th International Conference on Distributed Computing Systems*. 2015, pp. 630–641. DOI: 10.1109/ICDCS.2015.70.

[107] Yury Zhauniarovich et al. “A Survey on Malicious Domains Detection through DNS Data Analysis”. In: *ACM Computing Surveys* 51.4 (July 31, 2019), pp. 1–36. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3191329. URL: https://dl.acm.org/doi/10.1145/3191329.

[108] Kristijan Žiža, Predrag Tadić, and Pavle Vuletić. “DNS exfiltration detection in the presence of adversarial attacks and modified exfiltrator behaviour”. In: *International Journal of Information Security* 22.6 (Dec. 1, 2023), pp. 1865–1880. ISSN: 1615-5270. DOI: 10.1007/s10207-023-00723-w. URL: https://doi.org/10.1007/s10207-023-00723-w.

A Framework definitions

A.1 Technical

Accessibility How broadly is the data available? How is access provided to the data?

Volume How large is the data source, in terms of samples and disk usage?

Time horizon Does the data source contain historical or short-term data?

Update frequency How often do the contents of the data source change?

A.2 Impact

Completeness Does the data source contain complete or partial coverage of what it aims to collect?

Generality How broadly applicable is the data source, is it rather specific (and thus only relevant for one or a few tasks) or rather general?

Time-variance Does it capture behavior that is time-variant or generally consistent over time?

Freshness Do the time-variance and update frequency align well, or is the latest available data likely to get stale?

Overlap If multiple providers offer similar data sources, how much overlap is there between them?

A.3 Legal/ethical

License or usage conditions Under what conditions is the data available?

Privacy Does the data source include personal data?