WHOIS Right? An Analysis of WHOIS and RDAP Consistency

Simon Fernandez, Olivier Hureau, Andrzej Duda, Maciej Korczynski

WHOIS Right? An Analysis of WHOIS and RDAP Consistency

Simon Fernandez[1], Olivier Hureau[1], Andrzej Duda[12], and Maciej Korczyński[12]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
{first.last}@univ-grenoble-alpes.fr
[2] KOR Labs Cybersecurity, France
{firstname.lastname}@korlabs.io

Abstract. Public registration information on domain names, such as the accredited registrar, the domain name expiration date, or the abuse contact is crucial for many security tasks, from automated abuse notifications to botnet or phishing detection and classification systems. Various domain registration data is usually accessible through the WHOIS or RDAP protocols—*a priori* they provide the same data but use distinct formats and communication protocols. While WHOIS aims to provide human-readable data, RDAP uses a machine-readable format. Therefore, deciding which protocol to use is generally considered a straightforward technical choice, depending on the use case and the required automation and security level. In this paper, we examine the core assumption that WHOIS and RDAP offer the same data and that users can query them interchangeably. By collecting, processing, and comparing 164 million WHOIS and RDAP records for a sample of 55 million domain names, we reveal that while the data obtained through WHOIS and RDAP is generally consistent, 7.6% of the observed domains still present inconsistent data on important fields like IANA ID, creation date, or nameservers. Such variances should receive careful consideration from security stakeholders reliant on the accuracy of these fields.

Keywords: WHOIS · RDAP · DNS · domain names · registration data · measurements

1 Introduction

Malicious activities such as phishing scams, botnet operations, or malware distribution often involve the use of domain names. To investigate these activities and mitigate their impact, it is crucial to have access to specific information about domain registration. Essential information for investigating malicious activities related to domain names encompasses details such as the domain creation date, the registrant name, the sponsoring registrar, the domain status, the expiration date, email addresses designated for reporting domain name abuse, and other relevant data. However, in compliance with the European General Data Protection Regulation (GDPR) [37] and the Temporary Specification of the Internet Corporation for Assigned Names and Numbers (ICANN) for generic Top-Level

Domain (gTLD) registration data [20], personal information pertaining to registrants is typically obscured or hidden.

Different entities involved in the domain registration process typically provide registration information through two protocols: WHOIS [6] and RDAP (Registration Data Access Protocol) [15]. Despite the historical reasons for the coexistence of two protocols, each having its own specific format, and theoretically providing access to the same data, numerous studies [25,10,29,30] raised valid concern about the effectiveness and drawbacks of both protocols.

While both protocols were designed to provide registration information, there are no formal requirements mandating consistent results across different data sources. In practice, the registration data may vary between TLD registries, and registrars, as well as between the responses obtained from WHOIS and RDAP. This variability introduces an element of unpredictability with respect to the consistency and accuracy of the provided information.

Furthermore, studies that use registration data tend to favor one protocol over the other without providing explicit justification, and they base their preference on factors such as data retrieval speed, parsing capabilities, the presence of WHOIS and RDAP records for each domain, and other convenience-related considerations. Hence, an important issue emerges: to what degree do both protocols offer consistent information? Addressing this question requires a thorough and comprehensive analysis of how the data provided by the WHOIS and RDAP protocols align with each other.

To our knowledge, no previous research examined the assumption that information provided by WHOIS and RDAP is consistent. Nevertheless, many articles put forth classification algorithms, conducted studies on the domain behavior, or initiated abuse and vulnerability notification campaigns relying on data obtained through these protocols. In doing so, they implicitly depend on the accuracy and consistency of the information provided by WHOIS and RDAP.

Our paper makes the following contributions:

- We provide an overview of the disparities between WHOIS and RDAP, shedding light on the rationale behind the coexistence of multiple servers and protocols for accessing registration data. Delving into the historical and technical aspects, we highlight the intricate choices that have led to the current state of uncertainty surrounding the assurance of data consistency.
- We undertake a comprehensive data collection encompassing WHOIS and RDAP records for more than 55 million domains. Our focus is on parsing the fields commonly used in security and privacy studies. We will contribute all the collected registration data to the research community.
- We perform a thorough analysis of the parsed fields evaluating their consistency and deliberating over potential factors contributing to content variations. By doing so, we aim to raise awareness within the community about the importance of exercising caution with trust in registration data as 7.6% of the observed domains presented inconsistencies in fields used by security and privacy studies.

– We conduct a comprehensive analysis of the nameservers field, cross referencing the gathered data with the results obtained from active DNS measurements. Our aim is to determine which data source, whether WHOIS or RDAP, is more likely to provide accurate and trustful information.

2 Background

We begin by providing background information on the administration of domain names and the collaborative processes within the DNS ecosystem. Delving into the history of WHOIS and RDAP, we explore the reasons for their coexistence. Furthermore, we explain how to access registration data through both protocols, providing a clear outline of their respective procedures. Lastly, we elaborate on diverse approaches and challenges related to parsing WHOIS and RDAP.

2.1 The Ecosystem of Domain Management and Registration

The administration of a domain name entails the collaboration of multiple actors who collectively ensure the provision of all the necessary technical and administrative records vital for its operational use. At the top of the Domain Name System (DNS), the Internet Assigned Numbers Authority (IANA) manages the root nameservers and delegates the management of each top-level domain (TLD) to different registries. Country-code top-level domains (ccTLDs) such as `.uk` and `.fr` are managed by country-specific organizations (registries) like Nominet (for `.uk`) or AFNIC (for `.fr`). In contrast, generic top-level domains (gTLDs) such as `.com` and `.business` can be managed by any organization that meets the necessary requirements [19] and obtains authorization from the Internet Corporation for Assigned Names and Numbers (ICANN), like VeriSign Inc. (for `.com`) or Identity Digital (for `.business`). Registries are responsible for managing their top-level domain zones and have the authority to create new domains under their TLD. Each registry delegates the task of registering new domains to registrars, responsible for selling domains to users, referred to as registrants. When contacted by users, registrars collect and centralize user information, and communicate with the registry. In the interaction between registrars and registries, a variety of protocols may be used with the Extensible Provisioning Protocol (EPP) [12] commonly used for seamless communication. The registry then generates the required records such as DNS ones and administrative details to create the domain. For gTLDs under the ICANN agreement [19] and the majority of ccTLDs, both the registry and the registrar make the registration information available to the public. This information is typically accessible through the WHOIS and/or RDAP protocols.

2.2 Why Two Different Systems?

The existing WHOIS protocol as defined in RFC 3912 [6] published in 2004 formalized a practice in use since 1982 [21]. RFC 3912 established the guidelines on

how a server could offer the information about various Internet entities, including users, servers, domains, and IP addresses with a straightforward query/response protocol. However, it recognized that the WHOIS protocol had certain deficiencies in terms of crucial design goals like internationalization and robust security, typically expected of IETF protocols. RFC 3912 explicitly stated that it did not address these shortcomings and only required the content to be presented in a human-readable format. The decision to retain the original design flaws in the WHOIS protocol can be attributed to historical reasons. The original WHOIS system in use since the early 80s was already implemented on numerous servers. To maintain backward compatibility and prevent disruption to existing systems and practices, the IETF chose to accept the original design flaws rather than mandating widespread changes. This approach aimed to mitigate the risk of a new protocol facing low adoption rates, similar to what occurred with the `SPF` DNS record [24].

After several years, the IETF initiated efforts to design a new protocol aimed at providing domain registration information while addressing the limitations of WHOIS. This endeavor culminated in 2015 with the publication of RFC 7482 [33] that specified RDAP. RFC 7482 [33], along with subsequent extensions [34,16,15,3,28], specifies the protocol emphasizing the provision of machine-readable data in the JSON format. It defines data types, keys, and encoding to ensure structured information. Despite the introduction of RDAP, the WHOIS protocol has not been replaced, and both protocols continue to coexist, offering comparable data.

2.3 Data Access and Availability

RFC 3912 [6] and RFC 8521 [14] define the WHOIS and RDAP data access protocols, respectively. The RDAP protocol operates over HTTP(s) using the REST paradigm and returns data in JSON format, while a WHOIS user needs to connect to a server over TCP on port 43 and receive a plain text response.

The registration data may be incomplete, and some registries may only offer minimal information—in this case, they are called "thin", in opposition to "thick" registries that directly provide the full registration data. This difference in the completeness of registration data remains valid for both WHOIS and RDAP. For instance, the `.com` registry provides minimal information and does not include the registrant organization data. To obtain complete information (with respect to GDPR), the user of both protocols may need to follow referrals to one or several servers (see Figure 1): they first need to locate the registry server (①), then submit a query to the registry to obtain the registration information (②), and optionally, retrieve more detailed data from the registrar (③).

For WHOIS queries, users can rely on command line tools provided by their system to bundle most steps and referrals, like the Debian *whois* package. On the contrary, there is no widely deployed command line tool to query RDAP databases.

The user needs to follow the steps below to retrieve registration information of `google.com` using RDAP:

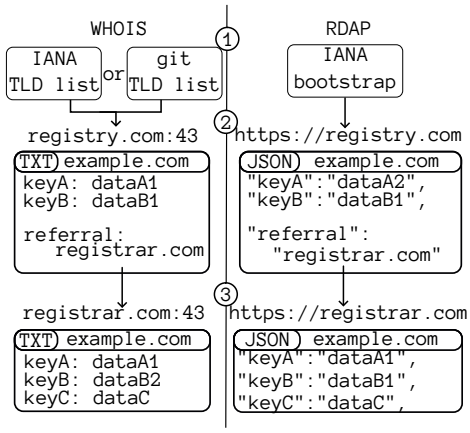

Fig. 1: Referral system to obtain complete registration data

① The user begins by retrieving the bootstrap configuration file from IANA,[3] as specified in RFC 9224. From this file, they obtain the URI of the `.com` RDAP server.
② The user appends the string `domain/google.com` to the server URI obtained in step ①, and forms the query to retrieve the registry RDAP answer at `https://rdap.verisign.com/com/v1/domain/google.com`. (an illustration of the result can be found in Appendix A, Figure 7)
③ The returned JSON object contains a referral to the registrar server (in this example, MarkMonitor, Inc). The user can access this information at `https://rdap.mark monitor.com/rdap/domain/google.com`.

For WHOIS, RFC 3912 [6] does not provide a bootstrap file for step ①. Instead, users can query the IANA WHOIS server at `whois.iana.org` to retrieve TLD-related information. The response includes the details about the TLD registry, in particular, the domain name of the WHOIS server for that zone. As an example, let us examine the procedure involved in retrieving the registration information for the domain `google.com` using the WHOIS protocol:

① The user proceeds by querying the IANA WHOIS server for the `.com` TLD and locates the record `whois: whois.verisign-grs.com`. This information directs them to the VeriSign server.
② Next, the user queries this server that provides registry WHOIS information for the domain `google.com` (the result is presented in Appendix A, Figure 6).
③ Within this record, there is a referral to the registrar server `WHOIS Server: whois.markmonitor.com`. The user can retrieve the most detailed registration data by querying this registrar WHOIS server.

Nevertheless, users may encounter problems when following this approach:

[3] `https://data.iana.org/rdap/dns.json`

Table 1: Number of active TLDs providing RDAP and WHOIS servers

	RDAP	WHOIS	
Source	Boostrap	IANA	GitHub
ccTLD (309)	27 (9%)	222 (72%)	231 (75%)
gTLD (1152)	1152 (100%)	999 (86%)	1147 (99%)

- Certain WHOIS servers may require specific query flags. For example, the WHOIS server for the `.de` TLD expects the flags `"-T dn,ace"`.
- The IANA database may not always be up to date, resulting in inaccurate information about certain TLDs. For example, it does not provide a WHOIS server for the `.cm` TLD.
- In some cases, the TLD registry may not handle the registration information for domain names associated with public suffixes. For instance, the registry server `whois.nic.uk` for the `.uk` TLD does not manage the `.ac.uk` TLD, managed instead by `whois.nic.ac.uk`.

For these reasons, the Debian *whois* package[4] adopts a different approach. It uses a dedicated database that specifies servers responsible for the public suffixes and the corresponding flags to be used. The source code for this package is accessible in a collaborative GitHub repository.[5] While the repository allows anyone to propose modifications, it has been mainly maintained by Marco d'Itri since 1999. This repository serves as a valuable alternative to the IANA WHOIS server, acting as a reliable starting point for retrieving WHOIS information (referred to as the `git TLD list` in Figure 1, step ①).

We have retrieved the information from the RDAP bootstrap file, the GitHub repository of the *whois* package, and queried the server `whois.iana.org` for all active gTLD and ccTLD listed on the IANA website. Table 1 shows that the GitHub repository provides 148 additional WHOIS servers compared to the IANA list. For instance, it includes a WHOIS server for the `.cm` TLD, not available on `whois.iana.org`. The table also highlights the proportion of active gTLDs and ccTLDs that offer WHOIS and RDAP services. It is important to highlight that ccTLDs provide relatively less access to registration data than gTLDs. In particular, the adoption of the RDAP protocol among ccTLDs is significantly low, accounting for only 9%. We can attribute the disparity between ccTLDs and gTLDs to the agreement established between gTLDs and ICANN [19]. As per this agreement, registries have to offer access to registration data through the RDAP protocol. However, it does not require gTLDs to maintain WHOIS servers, and it does not apply to ccTLDs. Contrarily, the deployment of RDAP by ccTLD registries is influenced by various factors such as voluntary adoption, local regulations, and technical considerations.

[4] `https://tracker.debian.org/pkg/whois`
[5] `https://github.com/rfc1036/whois`

2.4 Parsing Registration Data

One of the primary motivations behind the design of RDAP is to address the inherent limitations of the WHOIS system, in particular, its vague and loosely defined "human-readable" format for data. By incorporating the JSON-structured response format and well-defined data element features, among others, RDAP provides a more standardized, machine-readable approach to accessing registration data. This enhancement significantly improves the efficiency and reliability of parsing and extracting information from RDAP responses when compared to the traditional WHOIS system.

WHOIS data has been presented in various formats, undergone frequent changes, and may even be expressed in the local language of the registrar or TLD registry (e.g., the Bolivian ccTLD `.bo` WHOIS records are written in Spanish). The absence of normalization or implicit conventions raises a significant challenge when parsing WHOIS records, as highlighted in the studies that use WHOIS data [11,30,38,27,29,25].

We can categorize traditional algorithms for parsing WHOIS data into two distinct approaches: templates and rules. The template-based approaches, such as `Net::Whois`[6] (Perl), `whoisrb`[7] (Ruby), and `PHPWhois`[8] (PHP), offer regular expression templates specifically tailored to each registry or registrar. When using this approach, the user obtains WHOIS data from the registry, parses it using the relevant template for the TLD and registry, extracts any potential referral link to a registrar WHOIS server, and then retrieves and parses the registrar WHOIS data using the corresponding template. This approach is effective when the templates are available and regularly maintained. However, it becomes challenging when no template is available for a specific entity or if the format undergoes changes. Therefore, its success heavily relies on the quantity and quality of the templates, necessitating manual updates for each template.

Rule-based approaches such as `python-whois`[9] use a collection of predefined rules, regular expressions, and Natural Language Processing techniques to identify prevalent formats found in WHOIS records such as `Key: Value`, and extract as many fields as feasible. This approach is versatile and can be applied to any registrar without the need for dedicated templates. It may also accommodate format changes over time. However, it is generally less efficient compared to the use of custom-made templates [27].

Previous work explored existing parsers to train machine-learning algorithms based on Natural Language Processing or used techniques like Conditional Random Field [31] to automatically deduce the data structure and enhance the accuracy of field extraction. This approach demonstrated improved capabilities in extracting various fields from data.

While the template-based and rule-based approaches offer some potential for obtaining registration data through WHOIS, they require regular mainte-

[6] `https://metacpan.org/pod/Net::Whois`
[7] `https://whoisrb.org/`
[8] `https://github.com/SimpleUpdates/phpwhois`
[9] `https://pypi.org/project/python-whois/`

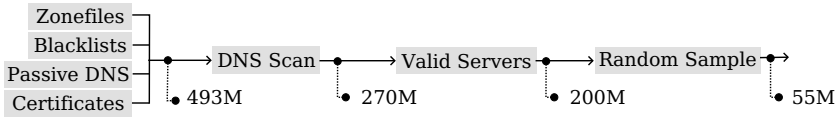

Fig. 2: The stages of domain selection with the number of domains at each step

nance and may be less efficient than RDAP. The introduction of RDAP offers a promising alternative for enhanced parsing efficiency and accuracy.

3 Methodology

In this section, we outline our methodology for collecting and parsing WHOIS and RDAP records. Considering the significant volume of data, we have meticulously designed our scheme to efficiently collect and parse registration data for a large number of domains within a reasonable time frame. All this is achieved while ensuring that WHOIS and RDAP servers experience minimal strain. We begin by explaining the process of domain selection, as illustrated in Figure 2, followed by a comprehensive description of the WHOIS and RDAP parsing process. Lastly, we provide an overview of how we have identified and analyzed discrepancies among the records.

3.1 Domain Data Collection and Filtering

Compilation of registered domain names. First, we gathered an extensive list of domains by consolidating multiple data sources:

- gTLD zone files obtained from the ICANN Centralized Zone Data Service (CZDS),[10]
- ccTLD zone files accessible via `AXFR` zone transfers (`.se`, `.nu`, `.li`, `.ch`),
- Passive DNS feed from SIE Europe,[11]
- Domain blacklists including SpamHaus,[12] APWG,[13] OpenPhish,[14] URLHaus,[15] ThreatFox,[16] and SURBL,[17]
- Google Certificate Transparency Logs,[18] which we continuously monitored to identify newly issued Transport Layer Security (TLS) certificates and extract the corresponding domain names.

[10] `https://czds.icann.org`
[11] `http://sie-europe.net`
[12] `https://www.spamhaus.org`
[13] `https://apwg.org`
[14] `https://openphish.com`
[15] `https://urlhaus.abuse.ch`
[16] `https://threatfox.abuse.ch`
[17] `https://surbl.org`
[18] `https://certstream.calidog.io`

All the collected domains are aggregated and deduplicated, resulting in a list of 493 million unique domain names. To guarantee the inclusion of only registered domains, we performed an active DNS scan on each domain, querying for `A` resource records using `zdns` [23], and exclude those for which the response is `NXDOMAIN` (non-existent domain).

Filtering domains with valid WHOIS and RDAP servers. To study the inconsistencies between WHOIS and RDAP records, we carefully filtered out domains that lacked a recognized WHOIS or RDAP server. This filtering process involved cross-referencing the official IANA list [17] and the GitHub repository, as detailed in Section 2.3. After this filtering step, our dataset comprised 200 million domain names.

Scanning all 200M domains would be a time-consuming process spanning several months, along with significant storage challenges. To address this, we opted to work with a representative subset of domains. This subset was randomly chosen from the pool of 200 million domains, with a sample size of 55 million domains carefully determined to facilitate the collection and parsing of WHOIS and RDAP records within a one-month time frame.

3.2 Gathering and Parsing Resgistration Data

Data collection. After identifying WHOIS and RDAP servers for the sampled domain names, we proceeded with the collection of the corresponding records. We gathered the registration data of the selected domains between December 6th and December 31st, 2022. During the collection process, we parsed each record to determine if it belonged to a "thin" registry that delegated a part of the data to a referral server, and follow the eventual referral. This step was iteratively repeated to ensure we obtained all versions of the registration data, following all referrals. At the end, we successfully collected a total of 164 million unique records, covering information from over 55 million distinct domains.

To ensure accurate comparisons, we collected WHOIS and RDAP records of each domain within a narrow time window, typically under 1 minute. This prevents the comparison of records collected at different times and reduces discrepancies resulting from domain updates during the scanning process. Moreover, some registrars impose query limits on IP addresses and enforce timeouts or blacklist IP addresses that exceed these limits. To ensure compliance and prevent any disruptions, we adjusted our data collection speed accordingly.

After the collection process, we carefully examined the gathered WHOIS and RDAP records. Any malformed responses (like invalid HTTP packets or JSON objects for RDAP) or timeouts were discarded, while valid responses underwent parsing for further analysis.

Parsing WHOIS. Parsing WHOIS data and extracting all pertinent fields presents a challenge, as detailed in Section 2. Consequently, this study focuses

Table 2: Fields extracted from WHOIS and RDAP records

Field	Data type	Missing rate		Domain inconsistency	Used by
		Records	Domains		
`Nameservers`	Text	3.2%	6.6%	573,790 (1%)	[5,9,13]
`IANA ID`	Integer	5.9%	13.7%	106,813 (0.2%)	[1,5,26,8]
`Creation date`	Date	0.8%	2.2%	3,138,024 (5.7%)	[11,1,26,8]
`Expiration date`	Date	1.0%	2.7%	2,424,951 (4.4%)	[25,26]
`Emails`	Email	7.9%	14.8%	18,958,821 (34.5%)	[8,4,29,38] [11,5,26,30]

on specific fields used in previous research (see Table 2), using custom templates designed to accurately parse various formats. We developed 242 custom templates comprising regular expressions that outline the extraction process for selected fields from WHOIS records across numerous registrars. The templates are designed to handle multiple languages and formats, maximizing the comparability of records.

Parsing RDAP. Contrasted with WHOIS, parsing RDAP records is typically more straightforward, primarily due to the JSON format. Nevertheless, despite the data format being defined in RFC 9083 [15], there might be ambiguity regarding the correct placement of information within the data structure. Consequently, different registries and registrars may have varying interpretations of where specific information should be located.

We gathered the designated fields from all locations allowed by the RFC. We considered malformed fields, those containing incorrect data types, or located in the wrong place within the data structure as missing. For instance, there are two primary representations of domain names in RDAP: as a string object (e.g., `ns.example.com`) or as an array of labels (e.g., `[ns, example, com]`). However, according to RFC 9083 [15], when listing domain nameservers, they must be in the string format. Therefore, if we encountered a nameserver in the array format instead of the expected string format, we considered it as missing. This decision was based on the assumption that most automated systems would adhere to the RFC and disregard the field due to its invalid type.

Field selection. To compare different data sources, it is important to note that not all registration data records share the same set of fields. As a result, we selected a limited number of fields, which have been commonly used in previous security studies and are consistently present in both WHOIS and RDAP records, whether at the registry or registrar levels. Table 2 presents the selected fields, along with the type of data they hold and the articles that have used them. For this research, we have chosen the following fields:

– `Nameservers`: this field indicates the name servers that have the authority over a particular domain.

- `IANA ID` and `Registrar`: the sponsoring registrar responsible for managing the domain is captured in the `Registrar` text field. Additionally, the `IANA ID` is an integer field that typically represents the unique identifier assigned by IANA [18] to each ICANN-accredited registrar (if applicable).
- `Creation date` and `Expiration date`: these fields denote the date of the initial registration for the domain and the subsequent expiration date. Once the registration expires, the domain becomes available for purchase again unless the owner renews it.
- `Emails:` This field contains a range of contact email addresses that can be used, for instance, for reporting domain-related abuse.

We deliberately omitted selecting fields associated with a registrant, despite their use in several studies, due to their absence in many registries. Furthermore, the implementation of the European General Data Protection Regulation (GDPR) resulted in the removal or redaction of the field content by most servers. The impact of GDPR on the content of these fields falls outside the scope of this paper and has already been analyzed in prior research [29].

When a field is absent from a record, or the content could not be parsed, the data is marked as missing. Table 2 shows the proportion of records missing each field. The record missing rate indicates the proportion of records with missing data, whereas the domain missing rate represents the percentage of domains that have at least one record with missing data, considering that each domain has multiple records (i.e., WHOIS and RDAP, including records collected by following referrals).

The missing rates for all fields, except for the IANA ID and Emails fields, are relatively low. This result was expected since the IANA ID solely pertains to domains under generic TLDs and ICANN-accredited registrars. Furthermore, each field presented its own set of parsing challenges, particularly in the case of WHOIS records, but also for RDAP. In RDAP, certain records, such as email contact addresses, can be located in different parts of the JSON structure as defined by RFC 7483 [34].

3.3 Analyzing Data Consistency

After collecting, parsing, and cleaning the registration data for all studied domains, we analyzed the consistency among various WHOIS and RDAP records.

For a given domain, if we were able to collect registration data from multiple sources and if these records have common fields, we evaluated the consistency of the data. If the formatted data in same fields is identical, we considered them to be matching fields. On the other hand, if there is a discrepancy between the data, it results in a mismatch. We consider two types of mismatches: the first one involves two records from the same protocol, such as the registry WHOIS not aligning with the registrar WHOIS. The second type involves two records from different protocols, for instance, the registrar WHOIS not corresponding to the registrar RDAP.

3.4 Ethical Consideration

We adhered to the best practices recommended by the measurement community to ensure reliable results with minimal disruption to the servers [35,7]. When gathering various data sources, including WHOIS, RDAP, and DNS records, we meticulously adhered to server rate limits [23]. Additionally, upon visiting the scanner's source IP address, users are presented with a webpage that provides information about our identity, work, and instructions for adding a scanned server to our opt-out lists, allowing them to cease receiving requests from us. Throughout the study, we did not receive any opt-out requests via email.

The raw data we collected may include information about registrants. However, after the implementation of GDPR, most registrars provide options for their customers to choose which fields are visible or automatically redact personal information. In practice, most fields that could potentially contain personal data were redacted by default.

4 Results

In this section, we present the analysis of inconsistencies and explore the root causes of the disparities observed in specific fields. Table 2 provides a breakdown, field by field, indicating the count of records where the field was missing, the number of domains in which at least one mismatch was identified, or if the field was entirely absent from the records. Excluding the `emails` field, which raises its unique challenges discussed in Section 4.3, we observed that 7.6% of all examined domains exhibited at least one inconsistency in the remaining fields.

4.1 Nameservers

The typical method to obtain a list of authoritative nameservers for a given domain involves sending recursive queries within the DNS tree, starting from the root zone and progressing toward the registry nameserver, which then provides the relevant information [8]. However, in certain prior studies that had a primary focus on detecting malicious domains [5,9,13], the nameserver information used in the analysis was obtained from WHOIS.

The primary purpose of the nameserver fields was either to cluster domains with identical nameservers [5,9] or to conduct further analysis on the nameserver itself. For instance, investigations could involve verifying whether the nameserver is self-hosted, such as `ns.example.com` being authoritative for `example.com`, determining if it is managed by well-kown DNS service operators, or identifying if the apex domain of the nameserver is newly registered [13].

In the subsequent part of this section, we begin by examining the various types of nameserver mismatches and their frequency. Then, we use DNS as a reference point to ascertain the accuracy of the data sources involved in cases of mismatches.

Table 3: Number of records and domains with mismatching nameservers

Case	Records	Domains
All	1,044,268	576,204
Inclusion	314,633 (30.1%)	224,833 (39.1%)
Intersection	48,693 (4.6%)	23,934 (4.1%)
Disjoint	680,942 (65.2%)	343,994 (60.0%)

Mismatch Types. We identified a total of 1,044,268 mismatches between two registration records of the same domain, encompassing 576,204 unique domain names. This accounts for approximately 1% of the overall collected domains; hence 99% of the measured domains did not have mismatching nameservers records.

When the nameservers of two records (referred to as A and B) are found to be inconsistent, three potential scenarios may arise:

Inclusion. $A \subset B$ or $A \supset B$: one set is a subset of the other one.
Intersection. No inclusion but $A \cap B \neq \emptyset$: A and B do not match but they have at least one common server.
Disjoint. $A \cap B = \emptyset$: A and B do not have common nameservers.

Table 3 presents the number of mismatches detected in each scenario. As described in Section 3.3, a given domain may have multiple records for each protocol, as each registration record may contain a referral field. As a result, each domain can exhibit multiple types of mismatches. For example, the nameservers extracted from the registrar's WHOIS record could be included in the list of nameservers found in the registrar's RDAP record, and additionally, the nameservers listed in the registry's WHOIS record could entirely differ from the servers in the registry's RDAP record. In such cases, a domain would be counted in both the inclusion and disjoint categories. Consequently, the values in the Domains column may exceed 100%.

When using DNS to fetch the domain's resource records, if the client (e.g., a recursive resolver) has multiple nameservers to choose from, it can use any of them interchangeably or query all of them and process the first received answer [22]. This means that the inclusion and intersection cases may be less concerning, as both records share at least one nameserver, potentially indicating that all nameservers serve the same data. Conversely, the disjoint case, in which both records do not have common servers, is concerning as it raises suspicion that the nameservers may not serve the same data or be authoritative for the domain name. This situation concerns 65% of the studied mismatches and 60% of the domains with mismatching records. The mismatch often involves records from different protocols. We have observed that 67.6% of the nameserver mismatches were between a WHOIS record and an RDAP record, whereas 17% were between two RDAP records (registry RDAP and registrar RDAP) and 15.4% were between two WHOIS records of the same domain.

In summary, while affecting only 1% of domains, nameserver mismatches, especially the 67.6% involving disparities between WHOIS and RDAP, raise concerns. In 60% of such cases, both sources lack any common nameservers, making the choice between WHOIS and RDAP for gathering nameserver information non-neutral and yielding incompatible results.

Who is Right? To successfully collect any DNS record for a domain it is essential to have an `NS` record in the parent zone file, specifying the authoritative nameserver for the domain. To gather the nameserver information, we actively queried the DNS infrastructure and performed a comparison with the nameservers listed in the WHOIS and RDAP records.

Methodology. To find the `example.com` nameservers, the client (e.g., a recursive resolver) first sends an `NS` query to the DNS root servers and receives the name of the servers that have authority over the `.com` zone. The client then sends another `NS` query to one of these servers and receives the `NS` record of `example.com`. This last answer comes from the registry in charge of the `.com` zone. The client can then either return the result because it retrieved the `NS` record of `example.com` from the authoritative nameservers of the parent (nameserver of `.com`) or perform additional `NS` queries to the nameservers received at the previous step and get the nameservers configured by the administrator of the domain. RFC 1034 [32] states that the nameservers returned by the registry and the nameservers configured by the administrator must be identical, but previous study [36] revealed that around 10% of the domains in the `.com`, `.org` and `.net` zones had differences between the nameservers provided by the parent registry servers and the nameservers provided by the child domain servers. If a domain is active, it must have an `NS` record at the registry level, as it is a part of the resolution chain. On the contrary, some domain owners do not put `NS` records in the child nameservers. To maximize the number of collected domains, we queried the `NS` resource records for each domain at the registry level.

Scans. To determine the consistency between registration data sources and DNS data, we used `zdns` [23] to retrieve the `NS` resource records of each domain where a mismatch was detected. Additionally, we collected their WHOIS and RDAP records for a second time, specifically between January 24th and January 27th, 2023, which ensured that all three data sources (DNS, WHOIS, and RDAP) were collected simultaneously, eliminating cases in which domain configurations were altered during our scans.

While some domains had expired between our initial scan and this supplementary analysis, approximately 90% of the domains remained active and returned a `NOERROR` DNS response with non-empty results during the scan.

Results. The second data collection unveiled 365,521 distinct domains exhibiting nameserver mismatches.

After the collection of the new registration data and the `NS` records from the authoritative DNS servers, the resulting data falls into two categories: the

mismatch can be between two records from the same protocol (two WHOIS records or two RDAP records), or between two records of different protocols.

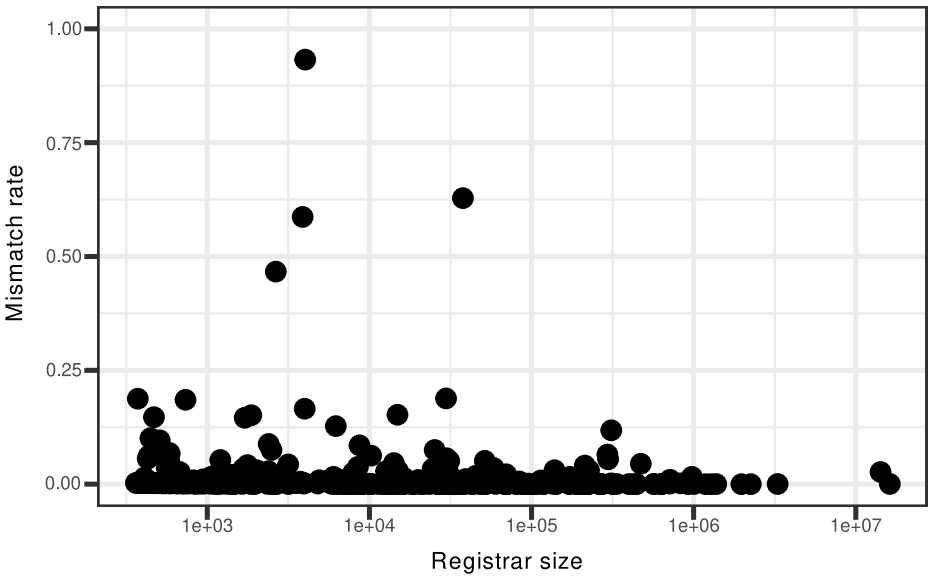

Fig. 3: Nameserver mismatch rate per registrar

WHOIS-RDAP mismatches. In 74.9% of the identified mismatch cases, the disparity exists between a record gathered through WHOIS and a record collected through RDAP. As previously described, the nameservers obtained from DNS may constitute a subset, superset, or have a non-empty intersection with each record. Upon examining all possible scenarios, we found that in 99.5% of cases, the DNS record corresponded to either the WHOIS or RDAP record. The remaining 0.5% involved intermediate situations in which the DNS result only partially matched one of the records. Due to the limited number of domains affected by this situation, we opted for concentrating our analysis on cases in which the DNS matched one of the records.

In 78.5% of cases, the DNS data corresponded to the nameservers provided by the RDAP record. This underscores the fact that, although nameservers obtained from DNS typically align with data from RDAP, there are still 21% of mismatch instances in which the DNS results match the WHOIS record. Interestingly, Figure 3 highlights that a few registrars exhibit a notably high mismatch rate compared to others. We observed that only four registrars have a mismatch rate exceeding 25%, while the largest registrars, representing the majority of domains, maintain a very low mismatch rate.

Registry-Registrar mismatches. The remaining 25.1% of cases represent the situations in which the mismatch is between two records from the same protocol but collected from different servers. In this case, the collector queried the registry server, got a referral to another server, and recursively called it, gathering an additional record. If two records are inconsistent, we checked if the nameservers provided by the DNS matched the records collected at the registry server or at the referral servers. In 99.2% of the cases, the DNS data matched the registry record, and in the remaining 0.8% of the cases, it did not match either records. The DNS data matched the registrar record in only 0.008% of the cases.

As described in Section 4.1, we decided to collect the `NS` records at the DNS authoritative nameservers of the registry. Consequently, we expected the record provided by the registry to be consistent with the DNS data from the same registry. Hence, the mismatches between two records from the same protocol almost always come from invalid data from the referral server.

The main takeaway is that when both sets of nameservers do not have common elements, and the discrepancy lies between an RDAP and a WHOIS record, the RDAP record is accurate and aligns with the `NS` records from DNS in 78% of the cases.

4.2 IANA ID, Creation and Expiration Dates

When it comes to obtaining the IANA ID, creation date or registrar name of a domain, research primarily relies on the WHOIS and RDAP protocols. Unlike nameservers, which can also be retrieved from DNS, there is no third-party service that offers direct access to this data. Consequently, when two sources diverge in these fields, there is no simple method to determine which record contains the accurate information.

In this section, we outline the types of mismatches identified in IANA ID, creation and expiration dates, and highlight a few cases in which we can ascertain the correct record.

Creation and Expiration Dates. The creation date represents the domain's initial registration instant, providing insight into its age. In domain-related research, the domain age is a pivotal factor as older domains, active for multiple years, are generally deemed more trustworthy than newly registered ones. The extensive analyses of the domain registration behavior [1,13] have shown that malicious domains tend to have shorter lifespans and are used in attacks shortly after registration. Other studies [8,9] have used the creation date to detect bulk registrations of malicious domains.

The domain age is also frequently combined with other parameters to distinguish between benign and malicious domains [11,26]. While some approaches [1] attempt to estimate the domain activity period by monitoring its appearance and disappearance in publicly accessible zone files, this method is contingent on zone file accessibility and the availability of historical data for the domain. Consequently, most studies depend on WHOIS or RDAP to acquire the creation date.

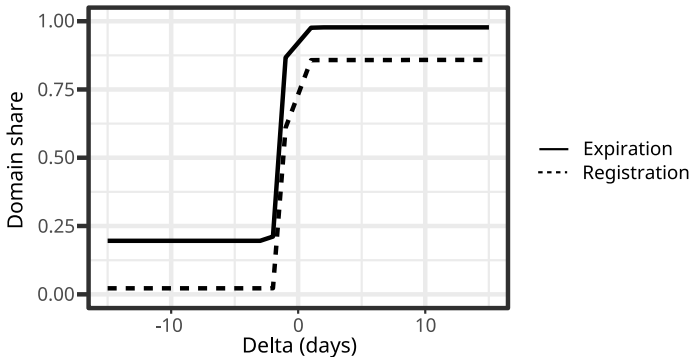

Fig. 4: Cumulative distribution of creation and expiration date mismatches

The expiration date also provides insights into the domain behavior and can shed light on various scenarios. For instance, if a domain is removed from its zone file before its expiration date, it may suggest actions taken by the registrar or seizure by authorities [1]. Additionally, parking and drop-catching entities use the expiration date to identify when a domain will become available for re-registration [38].

Both creation and expiration dates are usually found in the majority of WHOIS and RDAP records. However, in the case of WHOIS, they may be listed under various names, such as `Creation Date`, `Registration Date`, `Created at`, `Valid until`, and more.

After filtering out the dates that were not possible to parse and the dates lower or equal to the UNIX Epoch (which may indicate a default value or a configuration error), we observed that 5.7% (for creation dates) and 4.4% (for expiration dates) of the domains exhibited inconsistencies across their records. Figure 4 illustrates the distribution of time differences between these records.

We can observe that in 84% of the cases for creation dates and 78% of the cases for expiration dates, the differences are less than 2 days. These discrepancies have minimal impact on the analyses relying on creation dates to gauge the domain age [13] or on the speed of domain re-registration after expiration [1].

Previous studies [25] highlighted common misunderstanding of the different expiration steps before the deletion of a domain and pointed out that these steps can account for a mismatch of up to 30 days, as a confusion could be made between the expiration date, the deletion date and how the grace and redemption periods should be accounted for, but the collected data shows no specific mismatch proportion at 30 days. However, our analysis points out that

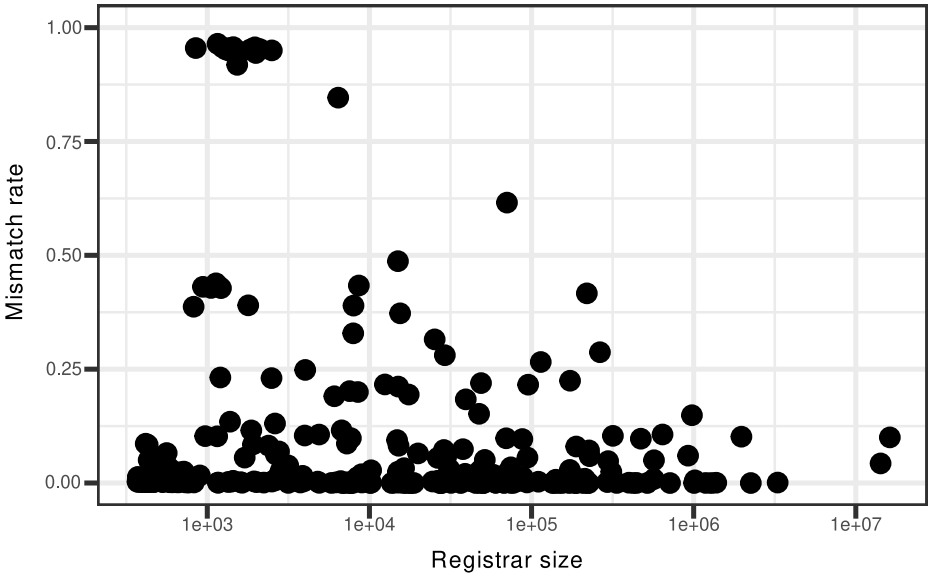

Fig. 5: Creation date mismatch rate per registrar

several records present an expiration date difference of exactly one year, which corresponds to the minimal duration of a registration, so the difference could come from the fact that the renewal of the domain was taken into account in one of the records and not in the other. Then, 98% of expiration date mismatches are either under 2 days or exactly 1 year, leaving only a few domains with unexplained expiration date mismatches.

Approximately 16% of the creation date mismatches extend beyond 2 days. In contrast to expiration date mismatches, creation date mismatches are more evenly distributed. One possible explanation for these discrepancies is that different entities may have distinct definitions of the `Creation Date`. While RFC 9083 [15] clearly defines keywords to describe creation events in RDAP, such as `registration`, `reregistration`, `reinstantiation`, and `transfer`, WHOIS lacks such precision. Consequently, the `Creation Date` recorded in the WHOIS record may not correspond to the same events in the domain life cycle as the `registration` event in the RDAP record.

The `Creation Date` mismatch rate for each registrar, as shown in Figure 5, highlights that while many registrars have over 10% of their domains with creation date mismatches, a few registrars exhibit nearly 100% of their domains with mismatched creation dates. This observation supports our hypothesis that some of these mismatches may result from registrar misinterpretations, custom registration processes, or systematic configuration errors. For example, the vast majority of domains presenting a `Creation Date` mismatch of 30 or 31 days are under the `.com` TLD and share the same registrar, `FastDomain Inc`. For these domains, the registrar record `Creation Date` is always one month earlier that

the one in the registry record. After investigation, we found that this registrar allows their customers to cancel their domain order up to 30 days after payment, while the ICANN Agreement [19] only imposes a 5-day refund window. Consequently, we can hypothesize that the creation of the registry record was delayed until the end of the 30-days period, while the registrar record was created when the customer first ordered the domain.

Table 4: Number of records and domains with mismatching emails

Case	Records	Domains
All	50.1M	19.0M
Inclusion	37.1M (74%)	15.1M (79.8%)
Intersection	0.59M (1.2%)	0.56M (2.9%)
Disjoint	12.4M (24.8%)	4.9M (26%)

IANA ID and Registrars. ICANN-accredited registrars play a crucial role in domain registration and management. The IANA ID associated with each registrar is a unique identifier, often found in WHOIS and RDAP records, helping to trace domain ownership and authority.

The content of the `Registrar` field in WHOIS and RDAP may differ from the name listed in the IANA registry. For example, 2.4% of domains with IANA ID 146 (`GoDaddy.com, LLC`) have different `Registrar` entries, including `GoDaddy LLC`, `GoDaddy.com, Inc.`, `GODADDY` or `Go Daddy, LLC`. Therefore, parsing the `Registrar` field to identify registrars can be challenging, and users often rely on the `IANA ID` for accuracy. However, in certain ccTLDs, registrars receive local accreditation, and the corresponding IANA IDs are not assigned or displayed in the public WHOIS and RDAP. In these cases, extracting registrar information solely relies on the `Registrar` field.

Our analysis uncovered that a mere 0.2% of domain names had records with inconsistent `IANA ID`. The analysis of IANA IDs reveals that the majority of mismatches occur between specific pairs of IDs. Approximately 91% of these detected mismatches involve a record with IANA ID 1556 (`Chengdu West Dimension Digital Technology Co., Ltd.`) and another record with IANA ID 1915 (`West263 International Limited`). Additionally, 4% of the mismatches involve IANA ID 3951 (`Webempresa Europa, S.L.`) and ID 5555555, which is an invalid ID. This pattern may suggest misconfiguration issues by particular entities, resulting in consistent mismatches across all the domains they manage.

In the second case, we confirmed the issue by registering a domain name with the registrar `Webempresa Europa, S.L.` and examining its records. While the registry WHOIS record correctly indicated the valid IANA ID 3951, the registrar WHOIS record contained an IANA ID field with the value 5555555, which does

not correspond to any valid registrar number. The registrar's WHOIS record also displayed placeholder values for various fields, including the abuse contact phone number and the reseller name. We verified that all domains registered with this registrar had inconsistent records. We reported the issue to the registrar, and over several months, we noticed that all the domains they managed were updated with correct registration data, resolving the inconsistencies. We suspect that the mismatches between ID 1556 and ID 1915 share the same origin. However, we were unable to test this hypothesis, as both registrars exclusively serve users in China and Hong Kong.

4.3 Email Addresses

Various types of email addresses are included in registration data, serving different purposes. These addresses are associated with the registry, registrar, or registrant, as well as for technical, administrative, and abuse-related functions. RFC 9083 [15] provides specific keywords in RDAP for describing the role of each email address, such as `administrative`, `abuse`, `billing`, or `technical`, which allows for easy identification of the address role, a capability that WHOIS lacks. For these reasons, we chose to collect all addresses in each record without distinguishing their roles. We then compared the records based on the sets of addresses they contain. Mismatches can occur due to protocol-specific contact addresses; for instance, the technical contact email for RDAP records may differ from that in WHOIS records if a registrar delegates technical administration to a third party. However, we anticipate that some addresses will be common across multiple records for the same domain, such as the abuse contact email for reporting domain-related abuse.

To analyze email mismatches, we applied the techniques described in Section 4.1. Initially, email addresses were parsed and duplicates removed. Subsequently, we compared the various possible inclusion and intersection cases. Table 4 presents the results of this analysis.

We identified 50 million mismatches for 19 million unique domains, encompassing 34.5% of the domains in this study. Among them, 74% of mismatches and 79.8% of domains featured one set of email addresses included in the other. About 75.2% of mismatches were either inclusions or intersections, potentially arising from shared addresses (e.g., abuse or registrant emails) while the addition of server or protocol-specific addresses by different entities (e.g., contact addresses for WHOIS or RDAP servers) may result in differences. However, nearly 5 million domains (8.8% of all analyzed domains) had a pair of records with no common email addresses.

The disjoint cases may be attributed to the GDPR implementation. Previous research [29] explored the impact of GDPR on the availability of personal information fields before and after its enactment. Following the GDPR implementation, many registrars and registries replaced the registrants' personal details like the name, the phone number, and the email address in WHOIS and RDAP records with entries such as 'REDACTED FOR PRIVACY', effectively

concealing this information. However, some entities introduced proxy email addresses to safeguard the registrants' actual addresses. These proxy servers mediate communication between proxy addresses and registrant emails. For example, in an RDAP record under the `registrant` role, one might encounter the address `b4ebaf9bfeba@withheld forprivacy.com`. While this conceals the registrants' personal data from the public, a valid contact address remains accessible. Protecting user privacy by redacting or using proxy email addresses can create discrepancies between WHOIS and RDAP records, as the registrant's address, which should be consistent in all records, may be redacted or hidden behind proxies.

Email mismatches can also occur when registrars or registries use distinct addresses for WHOIS and RDAP, even though both email addresses are administered by the same organization, such as `abuse.whois@registrar.com` and `abuse.rdap@registrar.com`.

To address these discrepancies, we conducted a new analysis by extracting and comparing only the domain names from email addresses, discarding the local parts. This approach considered email addresses within the same domain as consistent. The results are presented in Table 5. We found that this approach resolved 18.6% of the mismatches and reduced the rate of disjoint email addresses from 24.8% to 9.7%, which suggests that in many cases in which email addresses appeared disjoint, they actually originated from records with different email addresses hosted under the same domain.

Table 5: Number of records and domains with email domain mismatches after removing the local part of the address, retaining only the base domain name

Case	Records	Domains
All	50.1M	19M
Equality	9.3M (18.6%)	4.0M (21.4%)
Inclusion	35.7M (71.3%)	14.5M (76.7%)
Intersection	0.24M (0.5%)	0.23M (1.2%)
Disjoint	4.8M (9.7%)	2M (10.6%)

In conclusion, this analysis underscores the need for caution when gathering email addresses, especially for notification campaigns [30]. The choice of data source significantly affects the collected email addresses for 34.5% of domains. Additionally, in 10% of cases for which email records mismatch, the domains hosting these addresses are unrelated, suggesting that email servers may be managed by different entities, potentially leading to varying effectiveness in notification campaigns.

5 Related Work

Table 2 provides an overview of prior research that used WHOIS and RDAP data for domain name registration information. Nevertheless, the accuracy of the collected data has not been thoroughly investigated. Some earlier studies [5,8,25,9] relied on WHOIS data prior to the introduction of RDAP. However, as discussed in Section 2, inconsistencies are also present in WHOIS data obtained from servers managed by registries and registrars.

Challenges in processing WHOIS records have been identified, particularly concerning the reliability of extracted data such as AS numbers for IP WHOIS [2] and domain status [25]. In a previous in-depth analysis of the `.com` zone [27], the authors developed a machine-learning algorithm to address the multiple formats used in WHOIS records, demonstrating the difficulties in consistently parsing relevant fields.

The performance analysis of WHOIS and RDAP [10] focused on the speed but lacked the examination of data consistency across different servers and protocols.

In our work, we observed that 7.6% of the scanned domains exhibited mismatching records, raising concerns about the reliability of security metrics relying on such data. Notably, metrics that use the `Creation Date` field [26] or the bulk registration status [1] may be impacted, especially for registrars with high mismatch rates as presented in Figure 5. Obtaining accurate creation dates for domains under these registrars may require alternative data sources.

The `Emails` field exhibited the highest mismatch rates, even with a conservative parsing approach. Previous studies on notification campaigns [30,4] reported difficulties in extracting valid email addresses from WHOIS records, with email bounce rates exceeding 50%. These findings raise concerns about the effectiveness of notification campaigns due to the challenges associated with obtaining consistent and valid abuse emails from different entities.

6 Conclusions

Registration data plays a crucial role in the development of detection systems and gaining insights into the domain name behavior and entity management. However, obtaining this information may require interacting with various servers (either registries or registrars) and protocols (either WHOIS or RDAP). Our extensive analysis of 164 million records from 55 million domains unveiled that the data obtained through WHOIS and RDAP is generally consistent. Nonetheless, 7.6% of the analyzed domains displayed discrepancies in one or more of the following fields: IANA ID, creation and expiration dates, or nameservers. In cases related to the nameserver field, we used active DNS measurements to determine the accurate record. When disparities involved RDAP and WHOIS records, our findings showed that RDAP records were correct in 78% of instances for which mismatches occurred.

The principal insight underscores the importance of studies that rely on dependable registration data to diversify their data sources by collecting it from

various servers and protocols. Although larger registrars generally display lower mismatch rates, this observation does not inherently guarantee the accuracy of the data. Smaller registrars present a wide range of outcomes, with some demonstrating minimal discrepancies, while others exhibit higher rates. The potential risk exists for malicious actors to exploit registrars with inconsistent data, allowing them to evade detection systems that rely on the availability and reliability of registration data. An analysis of the extent of malicious domains managed by such inconsistent registrars could offer valuable insights into evasion strategies.

To facilitate future research, all collected records[19] and the associated data analysis[20] are made public.

Acknowledgments

We thank KOR Labs and Sourena Maroofi (KOR Labs) for their valuable support in parsing the registrar's data. This work has been partially supported by the French Ministry of Research projects PERSYVAL-Lab under contract ANR-11-LABX-0025-01 and DiNS under contract ANR-19-CE25-0009-01.

References

1. Affinito, A., Sommese, R., Akiwate, G., Savage, S., kc claffy, Voelker, G.M., Botta, A., Jonker, M.: Domain name lifetimes: Baseline and threats. In: 6th Network Traffic Measurement and Analysis Conference, TMA 2022. IFIP (2022), https://dl.ifip.org/db/conf/tma/tma2022/tma2022-paper32.pdf
2. Bianzino, A.P., Pezzuolo, D., Mazzini, G.: Who is whois? An analysis of results consistence. In: 2014 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM). pp. 289–292. IEEE (Sep 2014). https://doi.org/10.1109/SOFTCOM.2014.7039137
3. Blanchet, M.: Finding the Authoritative Registration Data Access Protocol (RDAP) Service. Request for Comments RFC 9224, Internet Engineering Task Force (Mar 2022). https://doi.org/10.17487/RFC9224
4. Çetin, O., Hanif Jhaveri, M., Gañán, C., van Eeten, M., Moore, T.: Understanding the role of sender reputation in abuse reporting and cleanup. Journal of Cybersecurity **2**(1), 83–98 (Dec 2016). https://doi.org/10.1093/cybsec/tyw005
5. Christin, N., Yanagihara, S.S., Kamataki, K.: Dissecting one click frauds. In: Proceedings of the 17th ACM Conference on Computer and Communications Security - CCS '10. p. 15. ACM Press (2010). https://doi.org/10.1145/1866307.1866310
6. Daigle, L.: WHOIS Protocol Specification. Request for Comments RFC 3912, Internet Engineering Task Force (Sep 2004). https://doi.org/10.17487/RFC3912
7. Dittrich, D., Kenneally, E.: The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research (2012)
8. Du, K., Yang, H., Li, Z.: The Ever-changing Labyrinth: A Large-scale Analysis of Wildcard DNS Powered Blackhat SEO. USENIX Security 2016 p. 19 (2016)

[19] https://doi.org/10.57745/RJX9XH

[20] https://github.com/drakkar-lig/whois-right-dataset

9. Felegyhazi, M., Kreibich, C., Paxson, V.: On the Potential of Proactive Domain Blacklisting. LEET 2010 (2010)
10. Ganan, C.: WHOIS sunset? A primer in Registration Data Access Protocol (RDAP) performance. TMA p. 8 (2021)
11. Ghaleb, F.A., Alsaedi, M., Saeed, F., Ahmad, J., Alasli, M.: Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning. Sensors **22**(9), 3373 (Apr 2022). https://doi.org/10.3390/s22093373
12. Gould, J.: Extensible Provisioning Protocol (EPP) and Registration Data Access Protocol (RDAP) Status Mapping. Request for Comments RFC 8056, Internet Engineering Task Force (Jan 2017). https://doi.org/10.17487/RFC8056
13. Hao, S., Kantchelian, A., Miller, B., Paxson, V., Feamster, N.: 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. pp. 1568–1579. ACM (Oct 2016). https://doi.org/10.1145/2976749.2978317
14. Hollenbeck, S., Newton, A.: Registration data access protocol (rdap) object tagging. Request for Comments RFC 8521, Internet Engineering Task Force (Nov 2018). https://doi.org/10.17487/RFC8521
15. Hollenbeck, S., Newton, A.: JSON Responses for the Registration Data Access Protocol (RDAP). Request for Comments RFC 9083, Internet Engineering Task Force (Jun 2021). https://doi.org/10.17487/RFC9083
16. Hollenbeck, S., Newton, A.: Registration Data Access Protocol (RDAP) Query Format. Request for Comments RFC 9082, Internet Engineering Task Force (Jun 2021). https://doi.org/10.17487/RFC9082
17. IANA: List of tlds (2023), `https://www.iana.org/domains/root/db`
18. IANA: Registrar ids (2023), `https://www.iana.org/assignments/registrar-ids/registrar-ids.xhtml`
19. ICANN: Icann registrar agreement, `https://www.icann.org/resources/pages/registrars-0d-2012-02-25-en`
20. ICANN: Icann temporary agreement for gtlds to comply with gdpr, `https://www.icann.org/resources/pages/gtld-registration-data-specs-en`
21. ICANN: Icann whois history, `https://whois.icann.org/en/history-whois`
22. IETF: Domain names - implementation and specification. Request for Comments RFC 1035, Internet Engineering Task Force (Nov 1987). https://doi.org/10.17487/RFC1035
23. Izhikevich, L., Akiwate, G., Berger, B., Drakontaidis, S., Ascheman, A., Pearce, P., Adrian, D., Durumeric, Z.: ZDNS: A fast DNS toolkit for internet measurement. In: Proceedings of the 22nd ACM Internet Measurement Conference. pp. 33–43. ACM (Oct 2022). https://doi.org/10.1145/3517745.3561434
24. Kitterman, S.: Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1. Request for Comments RFC 7208, Internet Engineering Task Force (Apr 2014). https://doi.org/10.17487/RFC7208
25. Lauinger, T., Onarlioglu, K., Chaabane, A., Robertson, W., Kirda, E.: WHOIS Lost in Translation: (Mis)Understanding Domain Name Expiration and Re-Registration. In: Proceedings of the 2016 Internet Measurement Conference. pp. 247–253. ACM (Nov 2016). https://doi.org/10.1145/2987443.2987463
26. Le Pochat, V., Van hamme, T., Maroofi, S., Van Goethem, T., Preuveneers, D., Duda, A., Joosen, W., Korczynski, M.: A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints. In: Proceedings 2020 Network and Distributed System Security Symposium. Internet Society (2020). https://doi.org/10.14722/ndss.2020.24161

27. Liu, S., Foster, I., Savage, S., Voelker, G.M., Saul, L.K.: Who is .com?: Learning to Parse WHOIS Records. In: Proceedings of the 2015 Internet Measurement Conference. pp. 369–380. ACM (Oct 2015). https://doi.org/10.1145/2815675.2815693
28. Loffredo, M., Martinelli, M.: Registration Data Access Protocol (RDAP) Partial Response. Request for Comments RFC 8982, Internet Engineering Task Force (Feb 2021). https://doi.org/10.17487/RFC8982
29. Lu, C., Liu, B., Zhang, Y., Li, Z., Zhang, F., Duan, H., Liu, Y., Chen, J.Q., Liang, J., Zhang, Z., Hao, S., Yang, M.: From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy under the GDPR. In: Proceedings 2021 Network and Distributed System Security Symposium. Internet Society, Virtual (2021). https://doi.org/10.14722/ndss.2021.23134
30. Maass, M., Stöver, A., Pridöhl, H., Bretthauer, S., Herrmann, D., Hollick, M., Spiecker, I.: Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support. USENIX Security 2021 (2021). https://doi.org/10.48550/ARXIV.2011.06260
31. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 -. vol. 4, pp. 188–191. Association for Computational Linguistics, Edmonton, Canada (2003). https://doi.org/10.3115/1119176.1119206
32. Mockapetris: Domain names - concepts and facilities. Request for Comments RFC 1034, Internet Engineering Task Force (Nov 1987). https://doi.org/10.17487/RFC1034
33. Newton, A., Hollenbeck, S.: Registration Data Access Protocol (RDAP) Query Format. Request for Comments RFC 7482, Internet Engineering Task Force (Mar 2015). https://doi.org/10.17487/RFC7482
34. Newton, A., Hollenbeck, S.: JSON Responses for the Registration Data Access Protocol (RDAP). Request for Comments RFC 7483, Internet Engineering Task Force (Mar 2015). https://doi.org/10.17487/RFC7483
35. Partridge, C., Allman, M.: Ethical Considerations in Network Measurement Papers. Commun. ACM **59**(10), 58–64 (sep 2016)
36. Sommese, R., Moura, G.C.M., Jonker, M., van Rijswijk-Deij, R., Dainotti, A., Claffy, K.C., Sperotto, A.: When Parents and Children Disagree: Diving into DNS Delegation Inconsistency. In: Passive and Active Measurement, pp. 175–189. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-44081-7_11
37. Union, E.: General data protection regulation, `https://eur-lex.europa.eu/eli/reg/2016/679/oj`
38. Vissers, T., Joosen, W., Nikiforakis, N.: Parking Sensors: Analyzing and Detecting Parked Domains. In: Proceedings 2015 Network and Distributed System Security Symposium. Internet Society (2015). https://doi.org/10.14722/ndss.2015.23053

A Examples of records

```
Domain Name: GOOGLE.COM
Registry Domain ID: 2138514_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.markmonitor.com
Registrar URL: http://www.markmonitor.com
Updated Date: 2019-09-09T15:39:04Z
Creation Date: 1997-09-15T04:00:00Z
Registry Expiry Date: 2028-09-14T04:00:00Z
Registrar: MarkMonitor Inc.
Registrar IANA ID: 292
Registrar Abuse Contact Email: abusecomplaints@markmonitor.com
Registrar Abuse Contact Phone: +1.2086851750
Domain Status: clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited
Domain Status: clientTransferProhibited https://icann.org/epp#clientTransferProhibited
Domain Status: clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited
Domain Status: serverDeleteProhibited https://icann.org/epp#serverDeleteProhibited
Domain Status: serverTransferProhibited https://icann.org/epp#serverTransferProhibited
Domain Status: serverUpdateProhibited https://icann.org/epp#serverUpdateProhibited
Name Server: NS1.GOOGLE.COM
Name Server: NS2.GOOGLE.COM
Name Server: NS3.GOOGLE.COM
Name Server: NS4.GOOGLE.COM
DNSSEC: unsigned
URL of the ICANN Whois Inaccuracy Complaint Form: https://www.icann.org/wicf/
```

Fig. 6: Registry WHOIS record of `google.com` obtained from the VeriSign server

```
{
 "objectClassName": "domain",
 "ldhName": "GOOGLE.COM",
 "links": [{
     "value": "https://rdap.verisign.com/com/v1/domain/GOOGLE.COM",
     "rel": "self",
     "href": "https://rdap.verisign.com/com/v1/domain/GOOGLE.COM",
     "type": "application/rdap+json"
   },{
     "value": "https://rdap.markmonitor.com/rdap/domain/GOOGLE.COM",
     "rel": "related",
     "href": "https://rdap.markmonitor.com/rdap/domain/GOOGLE.COM",
     "type": "application/rdap+json"}],
 "entities": [{
     "objectClassName": "entity",
     "handle": "292",
     "roles": ["registrar"],
     "publicIds": [{"type": "IANA Registrar ID","identifier": "292"}],
     "vcardArray": [
       "vcard", [
         ["version",{},"text","4.0"],
         ["fn",{},"text","MarkMonitor Inc."]]],
     "entities": [{
         "objectClassName": "entity",
         "roles": ["abuse"],
         "vcardArray": ["vcard",[
             ["version",{},"text","4.0"],
             ["fn",{},"text",""],
             ["tel",{"type": "voice"},"uri","tel:+1.2086851750"],
             ["email",{},"text","abusecomplaints@markmonitor.com"]]]}]}],
 "events": [
   {"eventAction": "registration", "eventDate": "1997-09-15T04:00:00Z"},
   {"eventAction": "expiration", "eventDate": "2028-09-14T04:00:00Z"},
   {"eventAction": "last changed", "eventDate": "2019-09-09T15:39:04Z"},
   {"eventAction": "last update of RDAP database", "eventDate": "2023-05-26T13:57:10Z"}],
 "nameservers": [
   {"objectClassName": "nameserver","ldhName": "NS1.GOOGLE.COM"},
   {"objectClassName": "nameserver","ldhName": "NS2.GOOGLE.COM"},
   {"objectClassName": "nameserver","ldhName": "NS3.GOOGLE.COM"},
   {"objectClassName": "nameserver","ldhName": "NS4.GOOGLE.COM"}],
 }
```


Fig. 7: Part of the Registry RDAP record of `google.com` obtained from the VeriSign server