Poster: Through the ccTLD Looking Glass: Mining CT Logs for Fun, Profit and Domain Names

Raffaele Sommese
University of Twente
The Netherlands

Mattijs Jonker
University of Twente
The Netherlands

CCS CONCEPTS

• **Networks** → **Naming and addressing**.

KEYWORDS

Domain Names, Country Zones, ccTLDs, Public Data, Coverage

ACM Reference Format:
Raffaele Sommese and Mattijs Jonker. 2023. Poster: Through the ccTLD Looking Glass: Mining CT Logs for Fun, Profit and Domain Names. In *Proceedings of the 2023 ACM Internet Measurement Conference (IMC '23), October 24–26, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3618257.3624994

Introduction and Background

The Domain Name System (DNS) is a vital component of the Internet. Almost every interaction that we perform on the Internet, whether it is for communication or to access information, relies on the DNS. Given this critical role, the research community has developed a significant interest in understanding and analyzing the DNS ecosystem over the years. This has resulted in numerous studies, focusing on different aspects such as security [1, 3], resilience [2, 6], and abuse [4]. Researchers primarily rely on passive DNS and active DNS measurement for study. The prior involves collecting DNS traces from one to many vantage points across the Internet, which provides a view on user behavior and the outcomes of DNS resolutions at specific points in time. The latter, in turn, involves systematically and comprehensively scanning the DNS namespace, aiming to collect DNS state in a controlled and uniform manner.

Active DNS measurement requires a set of names to measure. Public lists such as the (discontinued) Alexa Top 1M, Cisco Umbrella, and (newer) Tranco Top 1M are often used to this end, providing a small view on the namespace, skewed towards popular names. A much more extensive set of names can be obtained from Top-Level Domain (TLD) zone files. While a small number of TLD registries publish their zone files openly, zone file access usually involves signing an agreement. The traditional approach involves outreach from requestor to registry. Access to newer generic TLDs was made easier by the Centralized Zone Data Service (CZDS), which ICANN introduced to promote transparency in the DNS ecosystem. Every new gTLD registry is required to provide zone file access to approved requestors.

IMC '23, October 24–26, 2023, Montreal, QC, Canada

ACM ISBN 979-8-4007-0382-9/23/10.
https://doi.org/10.1145/3618257.3624994

Despite ICANN's efforts, obtaining domain names under country code TLDs remains challenging. There are over 300 ccTLDs and their registries have autonomy over decisions related to sharing. National legislation and regulations affect how registries handle data sharing requests. Privacy regulations such as the General Data Protection Regulation (GDPR) have further complicated matters. In some cases, domain names are classified as personally identifiable information (PII), as they can contain personal names. Given these challenges, many ccTLD registries are cautious to avoid legal burdens and potential privacy violations. Consequently, they tend to deny access to third parties and researchers seeking access to ccTLD zone files. Notwithstanding, some ccTLD registries are both willing and able to share. The agreements involved, however, always limit the extent to which measurement data can be used and shared with the research community. While obtaining access to ccTLD zone files is onerous, ccTLD domain names hold significant importance for studying and comparing the behaviors of individual countries in terms of hygiene, resilience, and security within the DNS ecosystem. Being able to relate ccTLD domain name information to geopolitical, economic, and social events in different countries is a crucial step toward gaining a better understanding of the global evolution of the Internet.

Recognizing the value of this data, we investigate alternative public data sources to obtain comprehensive sets of domain names under ccTLDs. Specifically, we focus on extracting sets of ccTLD domain names from certificate transparency (CT) logs. While previous works explored CT logs from the perspective of information leakage [5], we examine the extent to which CT logs account for, i.e., cover, ccTLD zones. To this end, we leverage, as baseline, 18 ccTLD zone files to which we are fortunate to have access, combined with a sizable CT logs dataset. Our goal is to study the extent of ccTLD coverage in CT logs and analyze the longitudinal evolution thereof. Based on the coverage we reveal in this one of several large public sources of domain names, we form an argument to petition ccTLD registries for more transparency and zone file sharing with the research community. We also discuss possible further steps to enhance our analysis.

Methodology and Data

The two data sources that we leverage for this study are OpenINTEL and CT logs.

CT logs serve as public, append-only logs that store digital certificates issued by certificate authorities, enabling transparency and facilitating the detection of potentially malicious or misissued certificates. We established a collection pipeline several years ago, which continually scrapes and stores certificates from CT logs in active operation. The dataset for this study contains 5.6 trillion certificates extending back to 2017, spread across 38 distinct logs of

operators such as Cloudflare, Digicert, Google, Let's Encrypt and Sectigo. From these certificates, we extracted 390 million unique registered domain names (i.e., TLD+1) from the `Common Name` and `Subject Alternative Name` certificate attributes.

OpenINTEL is a large-scale active DNS measurement platform [7] that covers around 65% of the global DNS namespace. Its measurement is primarily seeded using zone files. OpenINTEL covers 18 of the over 300 ccTLDs in existence, which is possible through zone file access agreements with the related registries. Out of the (global) Top 10 ccTLDs by size, OpenINTEL covers two (`.nl` and `.ru`). We leverage the 18 ccTLDs as baseline to compare domain names learned from CT logs against. There are 33 million registered ccTLD names in the DNS data (on June 1, 2023), spread across the 18 ccTLDs. The starting year of our longitudinal analysis is 2017. Some ccTLDs were added to OpenINTEL later; for these ccTLDs we present the comparison as early as possible.

We applied the following methodology. For the available ccTLDs (namely `.at`, `.ca`, `.ch`, `.co`, `.dk`, `.ee`, `.fi`, `.gt`, `.li`, `.na`, `.nl`, `.nu`, `.ru`, `.se`, `.sk`, `.su`, `.us`, and `.pϕ`), we join the sets of ccTLD domain names from OpenINTEL with the sets learned from CT data. We compare both sources on the first of June of each year, starting in 2017 and ending with 2023. We do not have a reliable temporal indication of when precisely certificates were appended to CT logs. Therefore, for our longitudinal comparison, we take the start of validity of the certificate as an indicator. This means that, for example, when we compare domain names in the DNS data with CT data for our once-yearly datapoint, we consider domain names on certificates with valid from date of up to 2017-06-01 in the CT data. This should provide a lower bound on CT coverage.

Preliminary Results

Figure 1 shows the results of our longitudinal coverage analysis. Our preliminary analysis immediately reveals two clear outcomes. First, for 2023, the sets of ccTLD domain names extracted from CT logs account for significant parts of the complete ccTLD zone files (i.e., our baseline), ranging from the highest coverage of 72% for `.ee`, to the lowest coverage of 42% for `.us`. The average coverage for the 18 ccTLDs under consideration is 52%. Second, we observe that coverage has increased over the years. We argue that this result is a natural consequence of the surging adoption of certificates for securing web communication, which has skyrocketed since the introduction of free TLS certificates. We also investigated the number of valid certificates during the analysis period (i.e., not expired as of 2023-06-01). We found that considering only these certificates would result in an average reduction of 20% in the number of obtainable domain names.

Conclusions and Future Directions

Our results establish a promising outlook for researchers interested in studying the ccTLD domain name ecosystem. The sets of ccTLD names that can be obtained from CT logs constitute significant parts of the respective ccTLDs zones, which are challenging to access due to legal considerations. Interestingly, one of the best-covered ccTLD zones is the Estonian `.ee` ccTLD, which is one of a handful ccTLDs that is published openly. Additionally, there are other intriguing cases such as the Russian Federation `.ru`. This is currently one

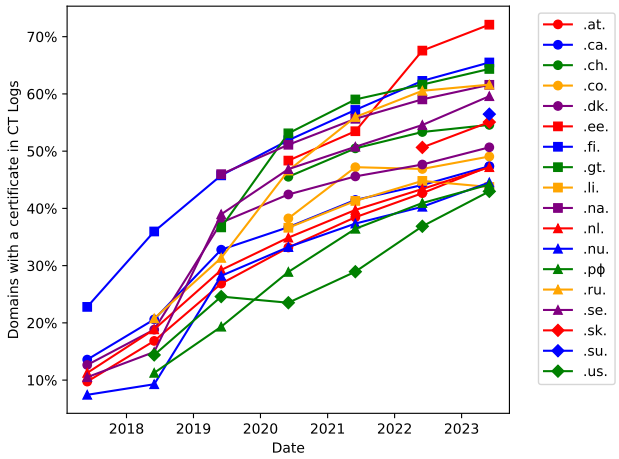

Figure 1: Domain names learnable from CT Logs

of the most difficult zones to access, but 60% is covered by CT logs. We hope that, having revealed that significant parts of the ccTLD domain name populations can already be inevitably learned from public sources, it becomes easier for ccTLD registries to favor sharing decisions towards the legitimate interest of the research community – by making zones either public or lowering barriers for vetted access.

As future work, we will explore other public data sources to harvest names from (e.g., Common Crawl). Further, we will analyze `A` and `AAAA` records and hosted websites to identify if their presence tends to increase the coverage. We also plan to investigate the publication delay (i.e., the time between domain name publication in its zone file and in CT logs) to better understand their representativeness of the current state of the DNS ecosystem. Finally, we may expand our analysis beyond registered domain names and consider how many subdomains we can learn under zones.

REFERENCES

[1] Gautam Akiwate et al. 2022. Retroactive Identification of Targeted DNS Infrastructure Hijacking *(IMC '22)*.
[2] Mark Allman. 2018. Comments on DNS Robustness *(IMC '18)*.
[3] T. Chung et al. 2017. A Longitudinal, End-to-End View of the DNSSEC Ecosystem. In *USENIX Security 17*.
[4] Maciej Korczynski et al. 2018. Cybercrime After the Sunrise: A Statistical Analysis of DNS Abuse in New GTLDs *(ASIACCS '18)*.
[5] Quirin Scheitle et al. 2018. The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem *(IMC '18)*.
[6] Raffaele Sommese. 2023. *Everything in Its Right Place: Improving DNS resilience.* Ph. D. Dissertation.
[7] Roland van Rijswijk-Deij et al. 2016. A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements. *IEEE Journal on Selected Areas in Communications* (2016).