Amassing Country-Code Top-Level Domains from Public Data

Raffaele Sommese, Roland van Rijswijk-Deij, Mattijs Jonker

University of Twente

RIPE87 – Rome – 27 Nov – 1 Dec

UNIVERSITY OF TWENTE.

DACS
Design and Analysis of Communication Systems

Open INTEL

Before starting

A special thanks to the

RIPE Academic Cooperation Initiative!

Introduction

- The Web, as a vast and intricate network, has been the subject of extensive academic study from diverse **technical, economical, and policy aspects**.
- These studies often start with lists of **domain names**, which serve as a fundamental **building block** of the Web's addressing system.
- Domain lists can include top-ranked domain names or names extracted from **zone files** of top-level domains (TLDs).

Our (OpenINTEL) journey to domain names

- OpenINTEL: A research-oriented, large-scale DNS measurement platform scanning **~65%** of all registered domain names every day.
- **252 million domains** daily, **9.1 trillion data point** collected **since 2015**, more than **60 papers** made possible by our data.
- Goal: be the **long-term** memory of the DNS.
- We rely on domain name lists (or zones) to seed our (forward DNS) measurements!

Source 1: Public Top Lists

- Alexa top 1-Million (RIP)
- Cisco Umbrella top 1-Million
- Tranco top 1-Million
- Cloudflare Radar top 1-Million

Only popular domain names : **Biased, but public!**

Source 2: The open ccTLDs

- Switzerland (.ch), Estonia (.ee), Lichtenstein (.li), Niue (.nu) and Sweden (.se) via Zone Transfer **(intentional AXFR!)**
- France (.fr) and Slovakia (.sk) via OpenDATA.

Fully representative : **Public!**

Source: Super Straho - Unsplash

Hic sunt dracones

- The boundaries of our public data sharing!

https://data.openintel.nl/

Source: Wikipedia

Source 3: ICANN CZDS

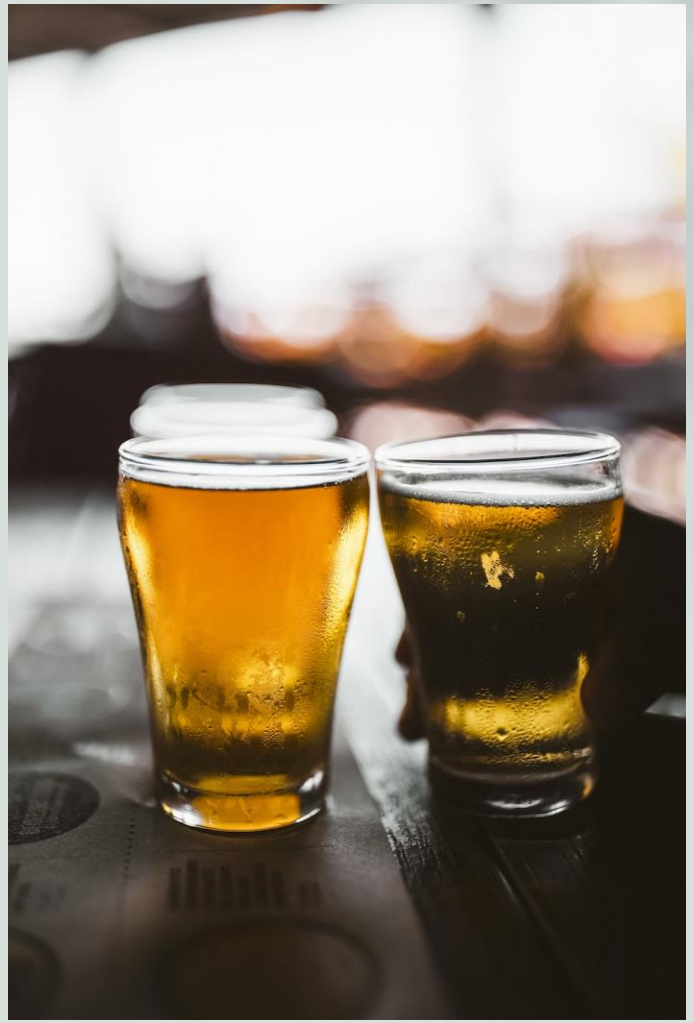
- With the expansion of new generic TLDs (gTLDs), ICANN **mandated** zone files to be accessible through a simplified and centralized process: ICANN's **Centralized Zone Data Service**.
- Most of the gTLDs approved by ICANN!

Easy and quick process, only for gTLDs :
Not re-shareable publicly (sort of)!

Source 4: The 'NDA' ccTLDs

- Obtaining domain lists for 'non-open' ccTLDs is a **complex** and often **arduous** process, involving **negotiations** and **restrictive contracts** with ccTLD governing bodies.
- And sometimes beers :)
- We managed to get access to 13 ccTLDs.

Under NDAs : **Not re-shareable**



Source: Teo Do Rio - Unsplash

Source 5:
The 'no way' ccTLDs

- Germany (.de): Nein, Nein, Nein! Privacy!
- Italy (.it): "We should ask to **all** the Italian domains owners if they agree to share the data with you"

Source: Wikipedia

Source 6: The 'we should share' ccTLDs

REGULATION (EU) 2020/857

"The Registry shall cooperate with competent authorities involved in the fight against cybercrime. It shall also cooperate with competent authorities and **public and private bodies** involved in the fight against speculative and abusive registrations, **in cybersecurity and information security**, in consumer protection, and in the protection of fundamental rights. It shall provide access to data to competent authorities and public bodies in line with Union or national law that complies with Union law, including with orders by courts or competent authorities vested with relevant powers."

Guess who is not sharing with public bodies (Universities) who operate in the cyber and information security spectrum?

Yes, looking at .eu!

To recap: The Problem

- Barriers to data-sharing **hurts** the research community!
- The Web extends beyond gTLDs: country-code top-level domains (**ccTLDs**) represent a significant portion of the **local online landscape**.
- This lack of transparency leads to an **underrepresentation** of ccTLDs in research, limiting our understanding of the global and regional Web ecosystems.

ccTLDs concerns

- Contrary to gTLDs, ccTLD governance is not under ICANN's purview, but instead a matter of **local policy**.
- Local policies often severely limit the data sharing possibility due to **privacy** and **liability** concerns.
- GDPR complicated the matter -- some entities consider **domain names personally identifiable information** (PII).

Source: Star wars
Movie

Another hope?

- We turn to alternative sources and rely on public data sources containing **substantial** numbers of domain names
- We explore two: **Certificate Transparency** (CT) logs and **Common Crawl** data.
- But how **representative** are these public sources of the ccTLDs landscape?

Our Dataset

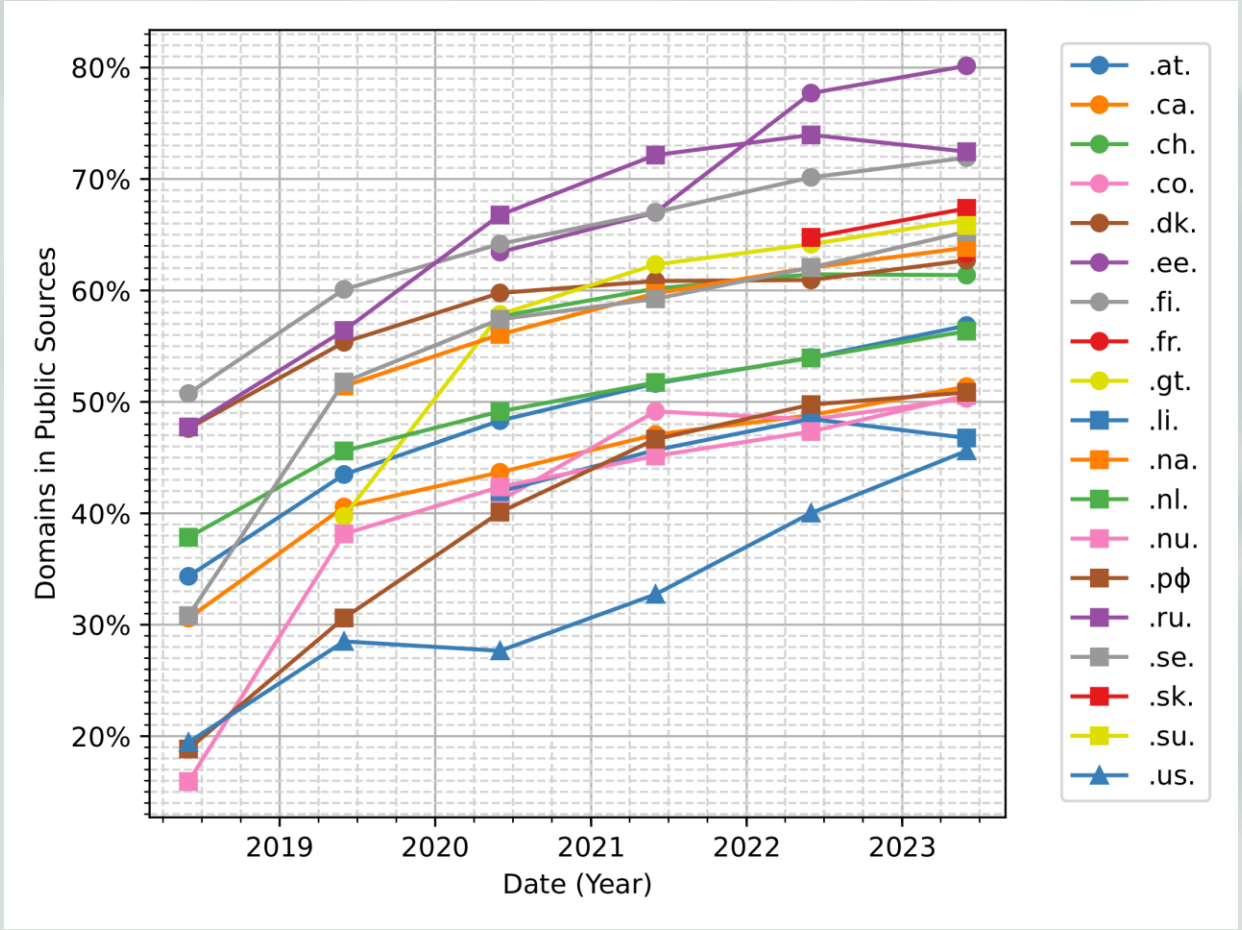
- We collect data from two sizeable and sustained sources: Certificate Transparency (CT) logs and Common Crawl data.
- CT logs track (as required by popular browsers) issuance of TLS certificates.
- Common Crawl provides a vast repository of Web crawl data (~**bi-monthl**y).
- We amass domain names into a **consolidated dataset** for analysis.
- Our methodology involves cross-referencing this dataset with a **ground-truth** of ccTLD zones to study coverage and timeliness.

Ground Truth

- Our ground truth data is derived from the OpenINTEL project, which measures **19 ccTLDs** (out of ~300 in existence).
- **12 ccTLDs** zone files were obtained through NDA agreements; **7** are public.
- Our dataset spans from **2018 to 2023.**

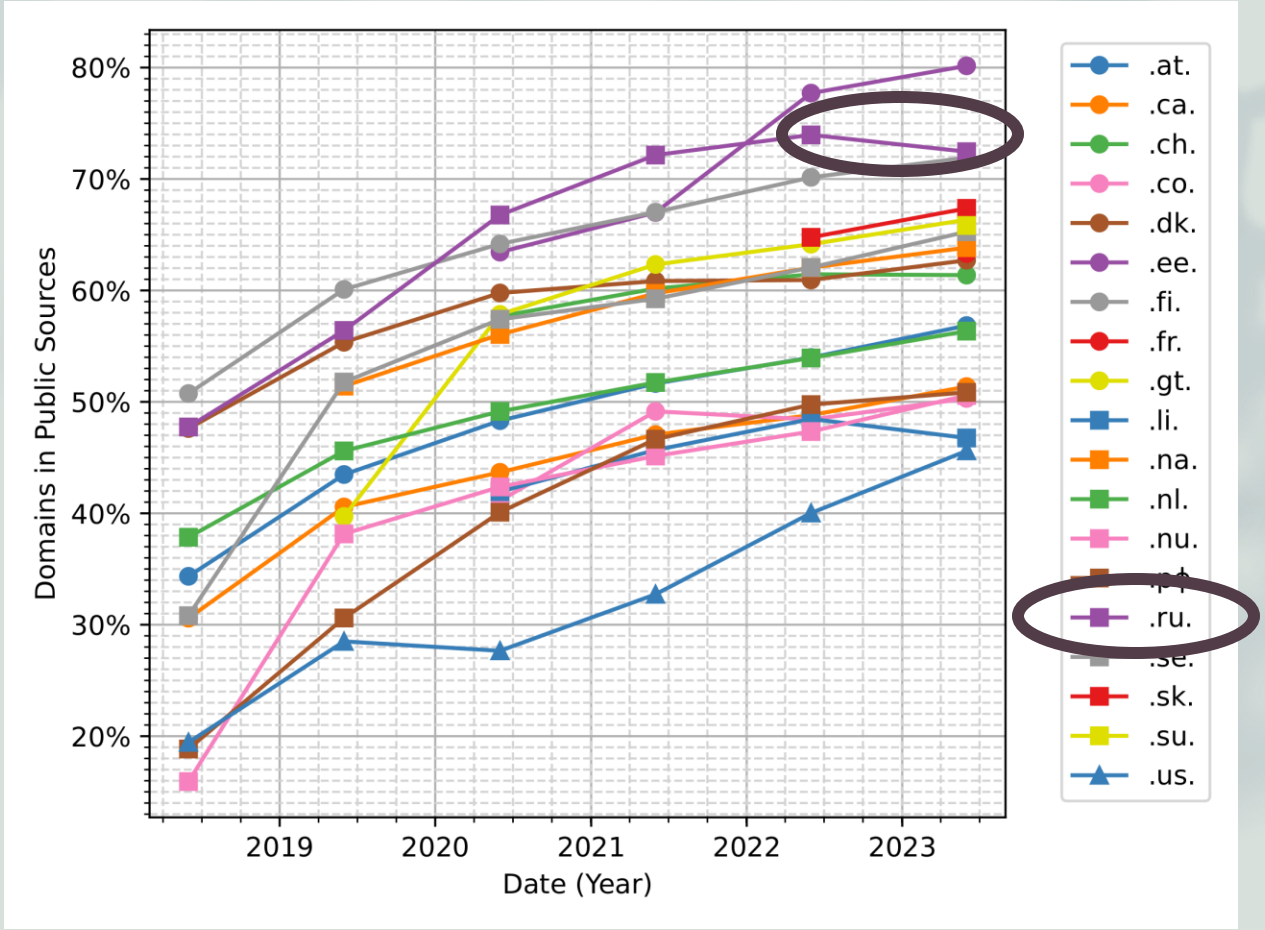
Half of the zones are already public!

- Varying coverage for 19 ccTLDs, ranging from 43% to **80%** in 2023.
- Coverage has steadily increased over the years, from avg 37% in 2018 to **59%** in 2023.
- This already negates the **privacy concerns** of ccTLDs!



Half of the zones are already public!

- Varying coverage for 19 ccTLDs, ranging from 43% to **80%** in 2023.
- Coverage has steadily increased over the years, from avg 37% in 2018 to **59%** in 2023.
- This already negates the **privacy concerns** of ccTLDs!



Coverage Contributions

In 2023, CT logs and Common Crawl together contribute to a coverage of **59% for ccTLDs.**

CT logs account for most of this coverage, with **24%** due to co-appearing names, **28%** solely thanks to CT, and **7%** exclusively from Common Crawl.

Common Crawl adds value by capturing additional domain names.

The supplemental coverage provided by Common Crawl is **decreasing over time**, possibly due to the increased adoption of TLS.

Delay in Publication

How long does it take for a ccTLD domain to appear in public sources after registration?

- Nearly **60% of newly registered domains** appeared in CT logs on the same day they were added to the zone, and **80%** within five days.
- CT logs can provide timely data about newly registered domains!

Source: Alice in wonderland Movie

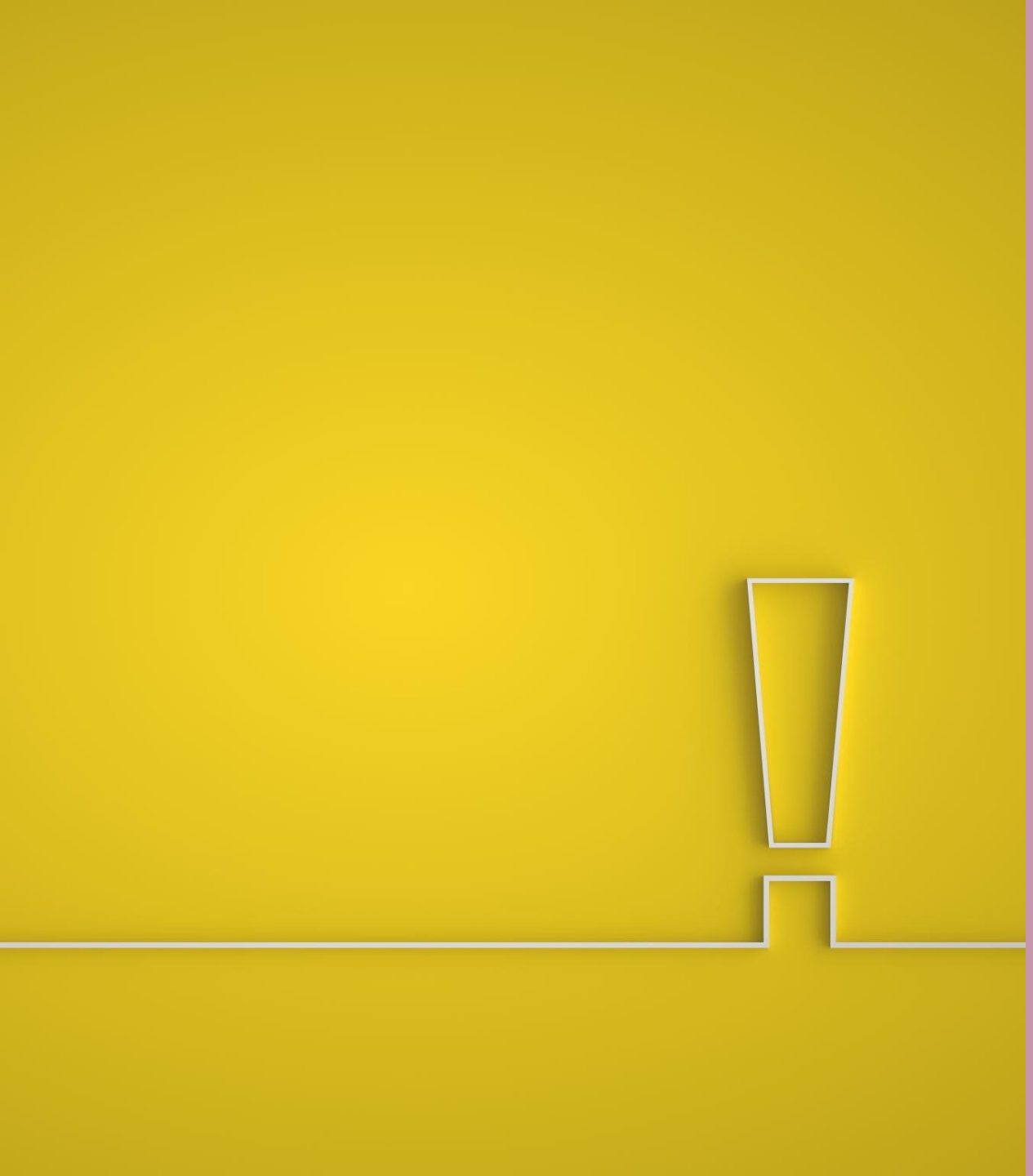
Can we generalize our results?

- Our sample of ccTLD covers only 19 of **~300** ccTLDs in existence.
- To generalize our results, we assessed coverage in 2023 for **1153** gTLDs!
- Coverage across gTLDs varies, but generally falls within the **range of 38% to 80%**.
- Larger gTLDs tend to have higher coverage rates.
- Our findings suggest that public sources can provide **substantial coverage across different TLDs**.

What do we not see?

We examined the presence of IPv4 address and open web ports.

- Of the **names in both CT logs and Common Crawl**, **91.5%** had **open Web ports.**
- Domains **not found in either source: 70.5%** had **open Web ports.**
- Domains in both CT logs and Common Crawl showed the highest rates of HTTPS deployment (87.7%).



Our Appeal to Registries

Registries with closed zones should consider **opening their zones** with minimal delay to address security concerns and benefit the community.

Legal framework should be rediscussed, even at the European Level.

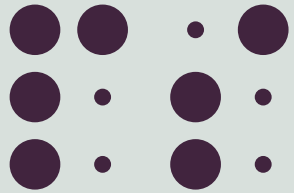
NIS2 ?!

Conclusion

- CT Logs and Common Crawl provide visibility on **more than 50%** of closed ccTLDs domains (on average).
- OpenINTEL data sharing efforts have been curbed in the past due to contractual agreements. We will make **more measurement data public soon.**
- We hope these results can spark a discussion with ccTLD operators for more **transparency** and **data sharing efforts.**

Questions?

r.sommese@utwente.nl



Let's discuss data sharing!

Amassing Country-Code Top-Level Domains from Public Data

Raffaele Sommese, Roland van Rijswijk-Deij, Mattijs Jonker

University of Twente

RIPE87 – Rome – 27 Nov – 1 Dec

UNIVERSITY OF TWENTE.

DACS
Design and Analysis of Communication Systems

Open INTEL

Backup Slides

Amassing Domain Names Challenges

- Considering other public data sources:
  - There are alternative sources such as the HTTP Archive and TLS scans that could complement domain name amassment efforts.
- Feasibility of amassing domain names:
  - Scraping CT logs and Common Crawl data at scale requires resources.
  - CT Logs are retired over time (especially temporally sharded logs).