On Unifying Diverse DNS Data Sources

Alfred Arouna
SimulaMet and OsloMet
Oslo, Norway
alfred@simula.no

Mattijs Jonker
University of Twente
Enschede, the Netherlands
m.jonker@utwente.nl

Ioana Livadariu
SimulaMet
Oslo, Norway
ioana@simula.no

ABSTRACT

The DNS maps human-readable identifiers to computer-friendly identifiers and relies on a reverse tree architecture to achieve this mapping. Backed by economic incentives, the DNS has become increasingly complex with data being shared among multiple autonomous stakeholders. The diversity of autonomous stakeholders limits data collection, access and sharing to researchers. For instance, each of stakeholder controls limited parts of the DNS space, thereby limiting analysis of real-world DNS behaviour. We aim to design and develop a software framework to unify diverse and large-scale public DNS data sources. The platform will facilitate the access to public DNS data by providing an efficient way of processing and analyzing large amounts of distributed data regardless of the DNS data format. Thus, the framework will help enable reproducibility in DNS studies.

CCS CONCEPTS

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

KEYWORDS

DNS, measurements, software framework, datasets

ACM Reference Format:
Alfred Arouna, Mattijs Jonker, and Ioana Livadariu. 2022. On Unifying Diverse DNS Data Sources. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22), October 25–27, 2022, Nice, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3517745.3563022

1 MEASUREMENTS INFRASTRUCTURES

The Domain Name System (DNS) maps easy-to-remember names to a computer-usable IP address and relies on a reversed tree architecture to achieve this mapping. Despite being introduced in 1987, most of the DNS building blocks have remained the same. Nowadays, and as Kurt Kayser stated during RIPE84, the DNS is considered as *the motor oil that keeps the Internet-engine running*[1].

To understand the critical role of the DNS; numerous studies have been conducted in several areas; including but not limited to resiliency, misconfiguration, privacy, and security. Figure 1 shows the use of DNS datasets or measurement infrastructure by researchers, based on proceedings of well-known conferences and or journals on Internet measurement over the past 10 years[2]. DNS data sources rely on a variety of data formats including but not limited to ASCII, PCAP, JSON, Apache Avro and ISC-dnsqr. However, more than 85% of the used formats are not suitable for large-scale and long-term analysis. We observe the rise of new active DNS measurement infrastructures i.e., OpenINTEL[3] and RIPE Atlas[4] as well as the reduction of PlanetLab[5] usage. However, DNSDB[6][7] and ICANN's datasets[8] are the oldest (since 2012) used long-term DNS datasets. Although Day in the Life of the Internet (DITL)[9] data collection is momentary, the collected data is massive making DITL the oldest used dataset of its kind. For instance, the 2017 KSK rollover data covered 72 hours of traffic from all root servers (except G-root). Note that DNS-Census[10] is used in a limited number of papers while registry data has been widely used due to the increase of DNS abuse studies. Interestingly, DNS-Coffee, which has been collecting changes from more than 1620 DNS zones over the last 10 years, is the oldest long-term, passive, and large-scale DNS zones dataset[11].

While some of the aforementioned data sources frequent papers, authors may choose to rely on self-instrumented, one-off measurements. As a result, many publications use active self-collected and/or passive self-instrumented datasets. In some works, researchers combine different datasets to synergize coverage. For instance, to get a better view of resolvers usage, DITL helps to get the perspective of the root servers while RIPE Atlas focuses on the client viewpoint [3]. Moreover, the combination of longitudinal zone data (DNS-Coffee) and active querying (OpenINTEL) has shown that lame delegations are common and expose hundreds of thousands of domains to adversarial takeover [1]. Without active measurements, evaluating DNS practices from zone files is challenging, i.e., zone files may contain dangling records. Therefore, combining diverse DNS datasets helps to characterize real-world DNS behaviour.

2 DNS COMPLEXITY

Backed by economic incentives, the DNS ecosystem has become increasingly complex with data shared among multiple autonomous stakeholders. Figure 2 shows a simplified view of the DNS ecosystem, which is composed of a myriad of actors like Authoritative Nameserver (AuthNS), resolvers, registries, registrars, registrants, policy makers, CDNs, and research centres/universities. DNS data can be collected: a) *on the wire*; b) *at rest*; or c) *sent onwards* [2].

[1] https://twitter.com/gjherbiet/status/1527583072920674307

[2] We selected 95 papers from 2011 to 2021 according to their a) impact from Scopus and b) usage of ongoing and or long-term datasets.
[3] https://data.openintel.nl/data/
[4] https://atlas.ripe.net
[5] https://planetlab.cs.princeton.edu
[6] https://www.farsightsecurity.com/solutions/dnsdb/
[7] DNSDB is based on ISC Passive DNS
[8] Domain Name Zone Alert (DNZA) and Centralized Zone Data Service (CZDS)
[9] https://www.dns-oarc.net/oarc/data/ditl
[10] https://dnscensus2013.neocities.org
[11] https://dns.coffee

IMC '22, October 25–27, 2022, Nice, France

ACM ISBN 978-1-4503-9259-4/22/10.
https://doi.org/10.1145/3517745.3563022

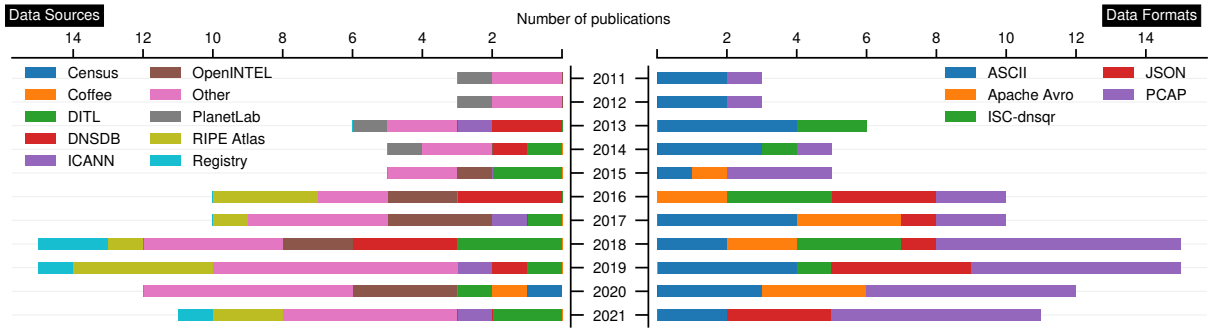

Figure 1: DNS data source usage over the last 10 years. JSON, ISC-dnsqr and Apache Avro formats are related to the use of long-term DNS datasets such as RIPE Atlas, DNSDB and OpenINTEL. Although the increase of publications can be correlated with the rise of long-term datasets, DNS researchers relied in majority on one-time snapshot of the state of (parts of) the DNS.

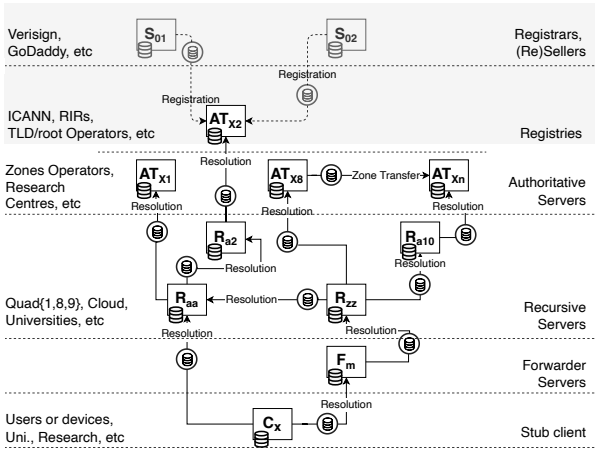

Figure 2: DNS ecosystem overview. Business relationships (grey section) allow registration data sharing. DNS data can be collected at-rest, on-the-fly or send onwards.

Therefore, to access to DNS data, researchers can either: 1) create new datasets; 2) collect partial or total existing datasets from DNS data providers; or 3) combine datasets. Active and/or passive measurements approaches can be used separately or combined to collect data. As seen in Figure 1, one common way of collecting DNS data is to create a one-off snapshot of the state of (parts of) the DNS depending on the researcher needs and privacy/confidentiality requirements. Thus, when it comes to DNS data, researchers face many challenges including but not limited to: distributed architecture, high diversity of actors, datasets, and measurement approach artefacts in addition to privacy, confidentiality, coverage, frequency, complexity, and availability [4]. Consequently, data sharing and/or combination is limited. Thus, restricting the characterization of real-world and global DNS behaviour.

3 UNIFYING DIVERSE DNS DATASETS

Our goal is to design and develop a framework that unifies diverse and large-scale public DNS data sources and eases processing of this data. To this end, we aim to build our platform using the following research questions:

***RQ1**: How can we characterize DNS data sources?*

***RQ2**: To what extent can we unify various DNS data sources? Can we also support non publicly available datasets, like self-instrumented DNS data?*

***RQ3**: How can we best design a software architecture to unify DNS data and facilitate easy access?*

To answer these questions, we started with a survey introduced in section 1. One of our preliminary results is the characterization of seven large scale measurements infrastructures: RIPE Atlas, BGPStream[12], OpenINTEL, Censored Planet[13], M-Lab[14], DNS Coffee and ISC Passive DNS.

Challenges. This characterization is based on challenges including but not limited to: *1)* a variety of data providers; *2)* time ordered data stream; *3)* framework scalability; *4)* data storage considerations; *5)* data volume and its impact on the global Internet infrastructure; *6)* diverse data formats; and *7)* the development of an active community around the project. This characterization helps to provide insights in term of design, implementation, and daily operation of successful software frameworks.

Acknowledgements. Alfred and Ioana's work is internally funded by SimulaMet. Mattijs work was supported in part by the EU H2020 CONCORDIA project (830927).

REFERENCES

[1] Gautam Akiwate, Mattijs Jonker, Raffaele Sommese, Ian Foster, Geoffrey M Voelker, Stefan Savage, and KC Claffy. 2020. Unresolved Issues: Prevalence, Persistence, and Perils of Lame Delegations. In *Proceedings of the ACM Internet Measurement Conference*. 281–294.

[2] Sara Dickinson, Benno Overeinder, Roland van Rijswijk-Deij, and Allison Mankin. 2020. Recommendations for DNS Privacy Service Operators. RFC 8932. https://doi.org/10.17487/RFC8932

[3] Moritz Müller, Giovane CM Moura, Ricardo de O. Schmidt, and John Heidemann. 2017. Recursives in the wild: Engineering authoritative DNS servers. In *Proceedings of the 2017 Internet Measurement Conference*. 489–495.

[4] Olivier van der Toorn, Moritz Müller, Sara Dickinson, Cristian Hesselman, Anna Sperotto, and Roland van Rijswijk-Deij. 2022. Addressing the challenges of modern DNS a comprehensive tutorial. *Computer Science Review* 45 (2022), 100469.

[12] https://bgpstream.caida.org/
[13] https://censoredplanet.org/
[14] https://www.measurementlab.net/