Challenges on Working with DNS Data

Alfred Arouna
SimulaMet and OsloMet
alfred@simula.no

Mattijs Jonker
University of Twente
m.jonker@utwente.nl

Ioana Livadariu
SimulaMet
ioana@simula.no

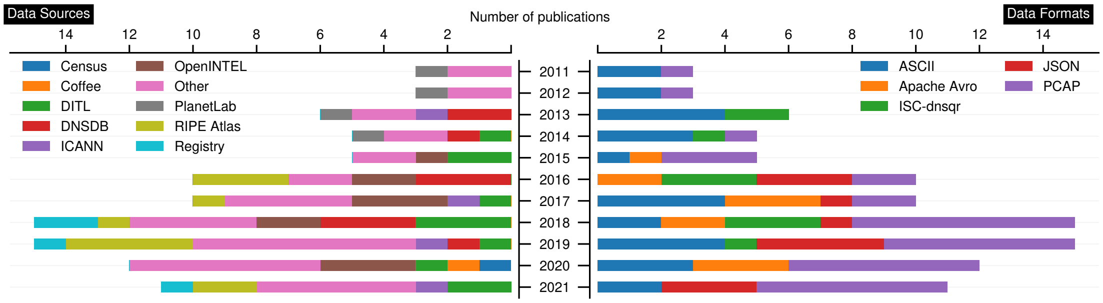

Figure 1: DNS datasets usage over the last 10 years. JSON, ISC-dnsqr and Apache Avro formats are related to the use of long-term DNS datasets such as RIPE Atlas, DNSDB and OpenINTEL. Although the increase of publications can be correlated with the rise of long-term datasets, DNS researchers relied in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

1 OVERVIEW OF DNS ECOSYSTEM

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. DNS data can be collected: a) *on the wire*; b) *at rest*; or c) *sent onwards* [1]. Therefore, researchers can either: 1) create new datasets; 2) collect partial or total existing datasets; or 3) combine datasets. Thus, when it comes to DNS data, researchers face many challenges including but not limited to: privacy, confidentiality, coverage, frequency, complexity, and availability [3]. Consequently, data sharing and/or combination is limited. Moreover, with the increase adoption of the principle of *minimum disclosure* [2], it become more challenging to have a full view of the resolver-authoritative exchanges. Thus, restricting the characterization of real-world and global DNS behaviour.

2 DNS DATASETS

To understand the critical role of the DNS; numerous studies have been conducted. 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[1]. DNS datasets are available on a variety of 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 and RIPE Atlas as well as the reduction of PlanetLab usage. However, DNSDB and ICANN's datasets are the oldest (since 2012) used long-term DNS datasets. Researchers also relied on one-of-a-kind datasets: DITL, DNS-Coffee and Census.

While some of the aforementioned data sources frequent papers, the majority of authors choose to rely on self-instrumented, one-off measurements. Therefore, in some works, researchers combine different datasets to synergize coverage. Nevertheless, the adoption of centralisation and minimisation is increasing, stressing challenges addressed in [3].

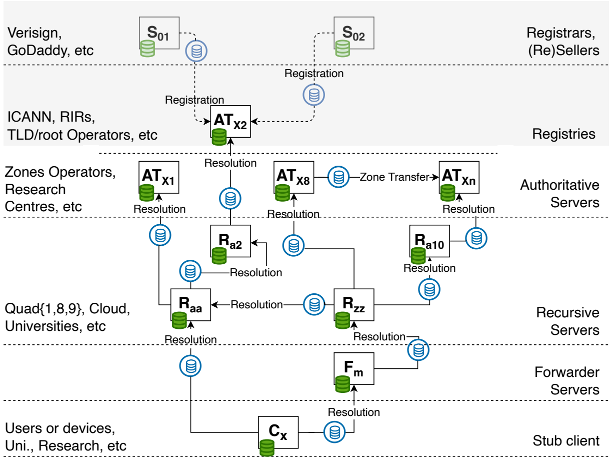

Figure 2: DNS ecosystem overview. Business relationships are on the top (grey section) and allow DNS metadata sharing. Other DNS data can be collected at-rest, on-the-fly or send onwards from the server (either on the wire or shared with a third party) [1].

REFERENCES

[1] 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

[2] Burton S. Kaliski Jr. 2022. Minimized DNS Resolution: Into the Penumbra. *Internet Protocol Journal* 25, 3 (2022).

[3] 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.

[1] 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