Chair of Network Architectures and Services
School of Computation, Information, and Technology
Technical University of Munich

Thesis B.Sc.

IDP

Leveraging RO-Crates with Internet Scan Data

Motivation

There has been a concerted effort to enhance reproducibility and transparency within the scientific community in the recent years, leading to the introduction and implementation of RO-Crates. RO-Crates are a `JSON-LD`-based format to enrich research data with corresponding meta data about its creation.

https://github.com/ResearchObject/ro-crate/blob/main/docs/assets/img/ro-crate.png

Our team conducts large-scale Internet scans that generate substantial amounts of data periodically. The infrastructure is interconnected, with the output from one scanner or tool serving as the input for another. For instance, the results from our DNS scanner are utilized in our TLS scan. Moreover, the scanners are continuously refined over time, leading to different versions being used. This complexity makes tracing which output corresponds to which scan and input data increasingly challenging.

To address this challenge, this project aims to explore how RO-Crates can be used to enrich our Internet scans with meta data.

Your Task

- Familiarize yourself with RO-Crates, RO-Crate profiles and tooling
- Familiarize yourself with parts of our scanning operation and tools
- Explore the different strategies for implementing RO-Crates into our Internet scans
- Implement the selected strategy for a select subset of our tools

References

- https://www.researchobject.org/ro-crate/
- https://www.researchobject.org/packaging_data_with_ro-crate/
- https://net.in.tum.de/projects/gino/

Contact

Christian Dietze diec@net.in.tum.de
Tim Betzer betzer@net.in.tum.de

https://net.in.tum.de/members/diec/
https://net.in.tum.de/members/betzer/