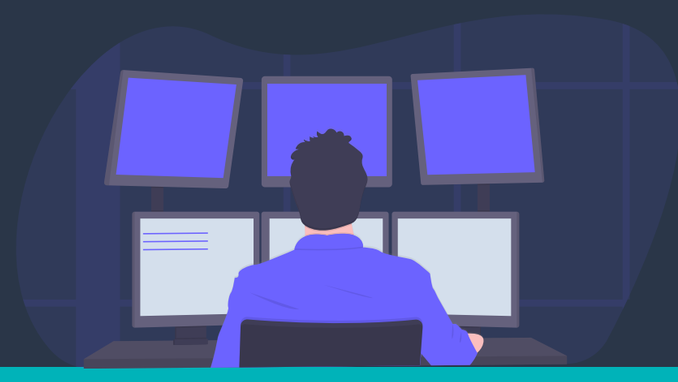
EESTI INTERNETI KAARDISTAMINE

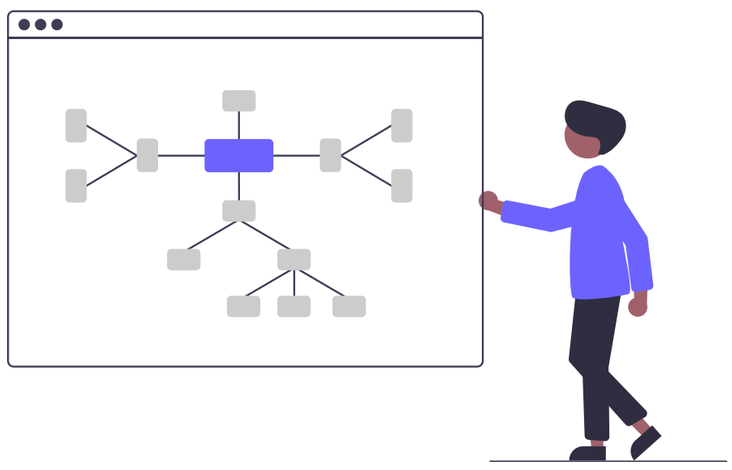
Autor: Siim Markus Marvet
https://github.com/siimmarkus/Eesti-domeenide-statistika



SISSEJUHATUS

Kui palju on .ee lõpulisi veebilehti? Kus neid majutatakse? Mis tarkvara kasutavad? Mida saab turvalisuse kohta öelda?

Eesti osalus ülemaailmses Internetis sai alguse 1992. aasta suvel, mil registreeriti esimesed üheksa .ee lõpuga domeeni. Tänaseks on see arv ületanud juba 136 000, aga mingit täpset ja avalikku ülevaadet ei ole selle massi kohta võimalik leida.
Selle projekti käigus loodi web crawler (või eesti keeles ämblik), mis suudab imiteerida tavalist brauseriga tehtavat külastust veebilehele, koguda üldisemat meta-infot ja uurida lehekülje sisu.



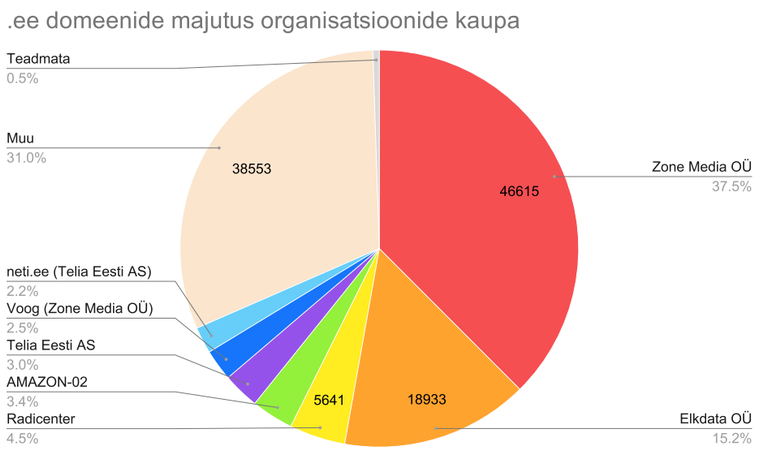
KUI PALJU JA KUS NEED VEEBILEHED ASUVAD?

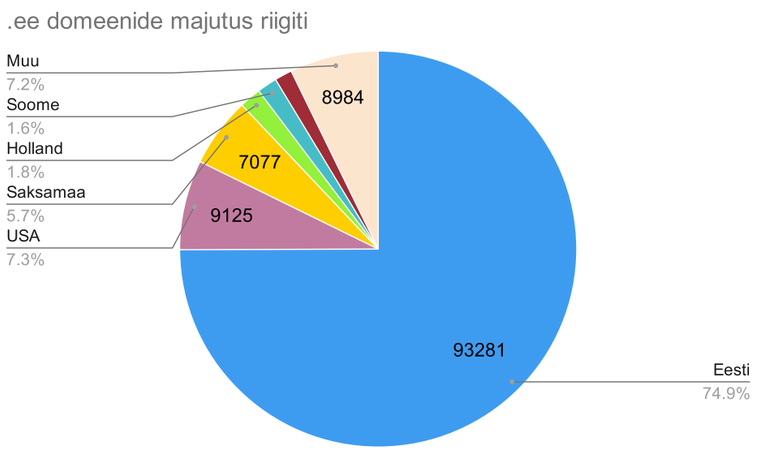
Tänu 2019. aastal Eesti Interneti SA poolt avalikustatud .EE tsoonifailile on nüüd kättesaadav peaaegu terviklik nimekiri kõigist registreeritud .ee lõpuga domeenidest. Arvuliselt oli neid uuringu koostamise ajal (18.04.2021) natuke üle 136 000 ning iga päevaga lisandub sellele umbes 50 uut domeeninime.

Peaaegu 124 000 (91%) jaoks oli domeeninime süsteemist leitav A-kirje (veebiserveri IP aadress), mida teades sai omakorda pöörd-aadressiteisenduse (reverse resolving) ja avalike andmebaaside abil määrata nende majutusorganisatsioon ja -riik.

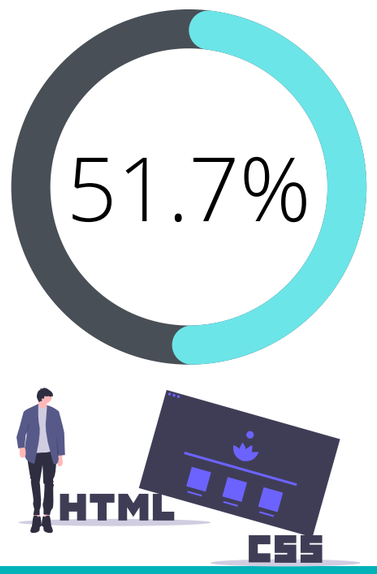
Organisatsioonide esirinnas on ootuspäraselt veebimajutusettevõtted, kusjuures pool kõigist Eesti domeenidest on kahe firma - Zone Media ja Elkdata (veebimajutus.ee) - haldusalas. Organisatsioonide jaotus on toodud vasakpoolsel diagrammil, kus kategooriasse "Muu" on kokku grupeeritud kõik alla 2% turuosaga organisatsioonid.

Riikide poolest on huvitav näha, et Eesti ligi 75% järel tuleb USA, mille põhjuseks paistsid olevat proksiteenus Cloudflare, Amazon AWS ning väiksemad lehe-ehitamis teenused. Kolmandal kohal riikide populaarsuselt on Saksamaa, mille jaoks mängis suurt rolli Amazon AWS-i ühe Euroopa suurema andmekeskuse asukoht.





UNIKAALSED VEEBILEHED



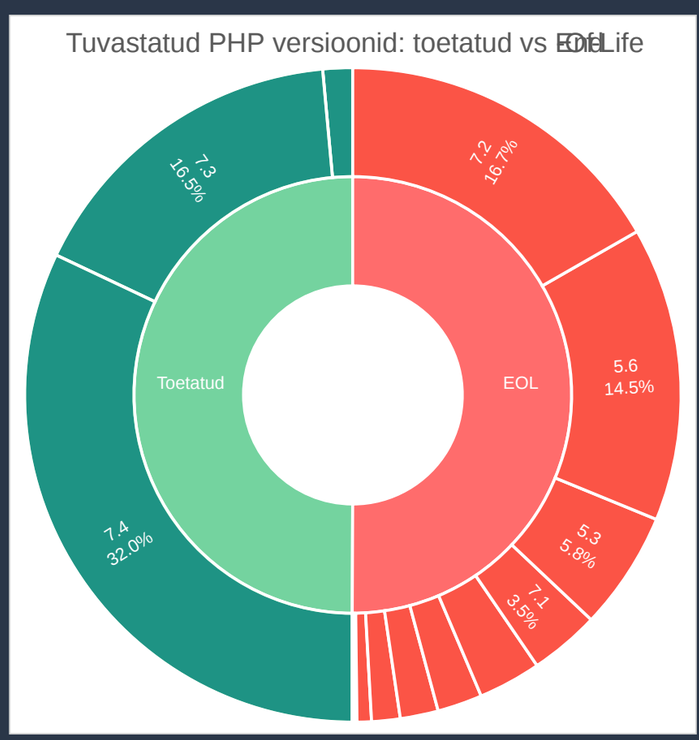
Eraldi huvitav küsimus tekib sellest, et kui domeene on registreeritud 136 000, siis kui paljudel nendest on ka tegelikult unikaalne veebileht ehk ei suuna ümber mujale või ole koopia mõnest teisest (nt vealeht). Kuna mahukate HTML-tekstifailide sarnasuse hindamine klassikaliste algoritmidega nagu Levenshtein-i kaugus ei tuleks ajalise keerukuse mõttes kõne allagi, siis tuli siin minna sammu võrra keerulisemaks ja kasutada hägusräsimist (fuzzy hashing).

Selle eesmärk on, et kahe sarnase faili räsid tuleksid samuti sarnased ning lisaks omavahel võrreldavad, et saaks hinnata muudatuste mahtu. Käesolevas uurimuses kasutati ssdeep algoritmi, mida kasutatakse tihti sarnasuse määramiseks e-posti spämmifiltrites ja pahavara tuvastamisel. Võrreldes kõigi domeenide avalehti omavahel selgus, et nende seas oli umbes 64 900 unikaalset ehk 47,7% kõigist 136 000 domeenist või 51,7% nendest 120 000-st, millel oli aktiivne veebileht.

TURVALISUS

Kogutava metainfo ja veebilehtede sisu järgi oli eesmärk proovida ka määrata nende kasutatav tarkvaraarhitektuur. Selleni jõudmiseks kasutas ämblik Wappalyzer rakendust, mis otsib eelnimetatud infost sagedasi mustreid, mis viitavad spetsiifilise tarkvara kasutamisele - kohati isegi koos versiooninumbriga. Selle abil õnnestus tuvastada igalt .ee veebilehelt keskmiselt 4-5 rakendust (kollektiivselt 467 erinevat rakendust ja nende 4721 konkreetset versiooni).

Huvitava näitena võib vaadata populaarse programmeerimiskeele PHP tuvastusi, mida oli kokku ~17 700 ning mille kõigil (v.a. seitsmel) paigaldustel oli võimalik tuvastada ka versiooninumber. Selle valimi alusel saab arvestatava ülevaate rakenduste uuendamise harjumustele Eesti: peaaegu täpselt 50% juhtudest oli tegu End-Of-Life (EOL) versiooniga, mis ei saa enam turbepaiku (security patch) kriitiliste turvanõrkuste parandamiseks ja avavad sellega ukse erinevatele rünnetele.



Lisaks ülejäänule otsiti skaneerimise käigus veebihaldus-tarkvara WordPress valesti konfigureeritud paigaldusi, mis lubasid Wordpressi kataloogipuu läbimist autoriseerimata kasutajatele ning on olnud viimase aasta paari suurema e-poodide kliendiandmete lekke põhjuseks. Uuringu käigus tuvastati selline paigaldus 209 domeenilt ning nendest teavitati CERT-EE, kes omakorda asus sama päeva jooksul vastavaid domeeniomanikke sellest teavitama.

TARTU ÜLIKOOL 1632 · ZONE · IT Akadeemia toetab Skype™