

TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Siim Markus Marvet

Eesti domeenide statistika ja turbeinfo kogumine

Bakalaureusetöö (9EAP)

Juhendaja: Alo Peets, MSc (arvutitehnika)

Tartu 2021

Eesti domeenide statistika ja turbeinfo kogumine

Lühikokkuvõte:

Aastal 2019 avalikustas Eesti Interneti SA .ee tsoonifaili, mille tulemusena on nüüd avalikult teada nimekiri kõigist .ee lõpuga domeeninimedest, mis on Interneti kaudu kättesaadavad. Käesoleva töö eesmärk on luua monitooringuprogramm ehk ämblik ning kaardistada selle abil Eesti domeenimaastikku. Töös kirjeldatakse ämbliku loomise protsessi, tööpõhimõtted ja toimimise efektiivsust. Kogutud andmete alusel antakse statistiline ülevaade .ee domeenide majutusvalikutest riigi- ja organisatsioonipõhiselt, uuritakse aktiivse veebilehega domeenide arvu ja nende tehnoloogilist arhitektuuri. Töö viimases peatükis kirjeldatakse mõnda valitud murekohta veebilehtede kaitsmisel Eesti suurima majutusettevõtte Zone Media OÜ haldusalas ning keskendutakse selle juures põhiliselt sisuhaldussüsteemile WordPress.

Võtmesõnad:

Internet, domeen, turvalisus, Zone

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Collecting statistics and security data on Estonian domains

Abstract:

The Estonian Internet Foundation released the .ee zonefile in 2019, giving the public a complete list of all .ee domain names, that are accessible through the Internet. The goal of this research is to program a web crawler with the aim to systematically map entire .ee domain landscape. This thesis describes creation of the crawler, its working principles and efficiency. The data gathered with the crawler will be analyzed to give a statistical overview of .ee domains' hosting locations by country and organization, the number of domains with an active website and the software stack of those sites. The last chapter of this research describes some example cases of the current problems with website security, as seen by Estonia's largest hosting provider Zone Media OÜ. This security overview focuses mainly on issues encountered with the popular content management system WordPress.

Keywords:

Internet, domain, security, Zone

CERCS: P170 Computer science, numerical analysis, systems, control

Sisukord

1. Sissejuhatus	4
Mõisted ja terminid	6
2. Domeenide monitooring	7
2.1 Varasemad uuringud	7
2.2 Eesti domeenide nimekiri	9
2.3 Ämblik	10
2.3.1 Ämbliku ülesehitus	11
2.3.2 DNS andmete kogumine	13
2.3.3 Veebipäringute teostamine	15
2.3.4 Tehnoloogilise arhitektuuri tuvastamine	16
2.3.5 Rakenduse testimine	19
3. Statistika kogumine .ee alamdomeenide kohta	22
3.1 Domeenide majutus riigi ja organisatsiooni alusel	22
3.2 Unikaalsete veebilehtede arv	25
4. Turvaülevaade	30
4.1 Aegunud tarkvaraversioonid	30
4.2 Versiooninumbrite tuvastamise täpsus	32
4.3 Näited sagedastest rünnetest Zone haldusalas	34
4.3.1 WordPressi kasutajate loendamine ja jõuründed	34
4.3.2 WordPressi kataloogipuu indekseerimine	35
4.4 Veebilehtede kaitse	37
5. Kokkuvõte	39
6. Viidatud kirjandus	41
Lisad	43
I. Lisa 1: Tartu Ülikooli koduleht	43
II. Lisa 1: Tartu Ülikooli domeeni kohta kogutud info JSON formaadis	44
III. Lisa 3: Domeenide arvult suurimad 100 majutusorganisatsiooni	46
IV. Tuvastatud WordPressi versioonid	49
V. Litsents	50

1. Sissejuhatus

Veidi vähem kui aasta pärast taasiseseisvumist 1992. aasta 3. juunis registreeris Keemilise ja Bioloogilise Füüsika Instituut Eesti tippdomeeni, tänu millele oli võimalik hakata kasutama *.ee* sufiksiga domeeninimesid. Peatselt selle järel registreeriti ka esimesed üheksa *.ee* lõpuga alamdomeeni (*kbfi.ee, goodwin.ee, org.ee, eii.ee, fsoi.ee, ebc.ee, obs.ee, postimees.ee, ioc.ee*), pannes aluse Eesti osalusele ülemaailmses Internetis [1]. Tänapäevaks on üheksast Eesti teise taseme domeenist saanud rohkem kui 139 000 ning iga päevaga suureneb nende arv umbes 50 võrra [2]. Piltlikult tähendab see, et pea iga kümne Eesti elaniku kohta eksisteerib üks domeen: kogus, mis on juba ammu ületanud inimeste võimekuse nende käsitsi jälgimiseks.

Otsides Internetist infot *.ee* domeenimaastiku kohta selgub, et laialdast avalikku uurimust ei ole selle kohta varem koostatud. Eesti Interneti SA kogub statistikat domeenide registreerimise kohta [2] ning haldab väikesemahulist (ligi 850 hosti) turvaseadistuste monitooringut Hardenize tarkvara abil [3], aga see ei ole piisav, et saada ülevaadet Eesti domeenide sisu ja kasutatava tarkvara hetkeseisu kohta.

Lisaks domeenide rohkusele on aastate jooksul kasvanud ka nende turvamise keerukus, kus haldamist lihtsustava tarkvara (nt sisuhalduse süsteemide) kasutamise läbi on osa lehe toimimise põhimõtetest veebiarendajate ja -haldajate eest ära peidetud. See tähendab, et rünnatavaid komponente on ühe veebilehe juures järjest rohkem ning vähese kogemuse või tööressursiga administraatorid ei suuda enam kõigi ründevektoritega järge pidada. Kuna tarkvara on küllaltki standardne ning kasutusel korraga paljudel veebilehtedel, siis ohustavad ka nendes leiduvad identsed turvanõrkused suurt hulka lehti. Tuvastades ühe turvaauku on seda automatiseeritud programmidega võimalik võrdlemisi vähese vaevaga otsida ja ära kasutada kõigis puudutatud veebiserverites.

Töö eesmärk on luua veebi analüüsi tarkvara (edaspidi: monitooringu ämblik), mis suudab kaardistada Eesti domeene nendel asuvate veebilehtede sisu järgi ning koguda piisavalt infot majutuskoha määramiseks. Kuna mõnda sagedasemat turvanõrkust on võimalik tuvastada vaadates kasutajale edastatavat veebilehe koodi, siis on eesmärk ämblikule lisada ka veebirakenduste kaugskaneerimise võimekus. Selle abil oleks võimalik tuvastada nende veebilehtede tehnoloogiline arhitektuur ning viia läbi lihtsamaid ründeid, et testida lehtede kaitsevõimet. Pidev seire ja andmebaasi hoidmine lubaks edaspidi saada lihtsamalt ülevaadet trendidest Eesti veebiarenduses ning toetaks reageerimist uutele turvanõrkustele.

Bakalaureusetöö on jaotatud kolme suuremasse ossa. Esimeses antakse ülevaade ämbliku tööpõhimõtetest ja andmete kogumisest. Teises osas tehakse statistiline ülevaade ämbliku abil kogutud infost Eesti domeenide kohta. Kolmandas osas vaadatakse lähemalt Eesti domeenide turvalisust ja sagedasemaid haavatavusi koos nendega kaasnevate rünnetega, mis veebilehtede pihta toimuvad.

Bakalaureusetöö kirjutamist toetab omapoolsete kogemuste ja riistvaraga Eesti suurim veebimajutusettevõtte Zone Media OÜ.

Mõisted ja terminid

Mõistete tsiteeritud osad (kaldkirjas) pärinevad Andmekaitse ja infoturbe leksikonist AKIT (<https://akit.cyber.ee/>) [4].

Domeen on *halduslikult ühtne kogum võrgustatud arvuteid, mille täisdomeeninimes on ühine sufiks*. Põhjalikum kirjeldus 2. peatüki alguses.

Ämblik (ingl *spider* või *crawler*) on rakendus, mis *sirvib süstemaatiliselt WWW veebilehti*. Töö raames loodud ämblik spetsiifilisemalt on programm, mis imiteerib brauseri abil tehtavaid veebilehe külastusi ning salvestab andmeid külastuste kohta. Tänu automatiseeritusele on selle abil võimalik lihtsasti külastusi imiteerida ja uurida suurel hulgal veebilehtedel (töös valiti selleks sihtrühmaks kõigi .ee domeenide avalehed).

Domeeninimede süsteem (DNS, ingl *domain name system*) - *TCP/IP-võrgu komponentide, teenuste ja ressursside nimede hierarhiline süsteem*. Domeeninimede süsteem on hädavajalik komponent Interneti selgroost, mis tagab teenuste ja nende toimimiseks vajaliku informatsiooni seostamise domeeninimedega. Üks olulisemaid ülesandeid selle juures on domeeninimede lahendamine IP-aadressideks ja vastupidi, tänu millele on võimalik pakettide suunamine Internetis.

Tiipne domeen – Domeenide hierarhia kõrgeima taseme domeen(domeeninime kõige parempoolne osa, näiteks Eesti domeenide puhul *.ee*)

Veebirakenduste kaugskanner on *veebirakenduste nõrkuste ja arhitektuuri puuete kaugotsingu programm*.

Domeenikrabamiseks nimetatakse domeenide registreerimist nende hilisema edasimüügi eesmärgil. Tihti registreeritakse selle käigus hiljuti vabanenud, tuntud kaubamärkide nimega või muid potentsiaalselt ihaldusväärseid domeene ning proovitakse neid algsest kordades kõrgema hinnaga edasi müüa.

Tartu Ülikooli koduleht (<https://www.ut.ee/et>) – Töös on läbivalt kasutatud erinevate protsesside illustreerimiseks ülikooli kodulehte. Kuvatõmmis illustreerimiseks toodud lisas 1.

2. Domeenide monitooring

Kõige laiemas mõistes on domeen virtuaalne haldusüksus. Domeene identifitseeritakse domeeninime abil, mille kasutamiseõigust on võimalik endale osta määratud ajaperioodiks, Tartu Ülikooli puhul on see *ut.ee*. Domeeniga võivad olla seotud veel rakendused, nagu

- Veebileht ja seda teenindav server;
- Nimeserverid, mis aitavad tagada domeeniga seotud rakenduste toimimise Internetis. Näiteks tahtes minna lehele *www.ut.ee*, peab veebilehitseja kõigepealt nimeserverite kaudu välja selgitama, mis on seda domeeni teenindava veebiserveri IP-aadress. Teades IP-aadressi on alles võimalik vastavalt serverilt pärida veebilehe sisu kasutades HTTP-d (Hypertext Transfer Protocol);
- Meiliserverid, mis vastutavad domeeni e-posti toimimise eest: näiteks aadressilt *siim.markus.marvet@ut.ee* meilide saatmine, vastuvõtmine ja hoiustamine või edasisuunamine.

Järgmine peatükk kirjeldab varasemaid uuringuid, domeenide valikut ning annab ülevaate ämbliku tööpõhimõtetest andmete kogumisel.

2.1 Varasemad uuringud

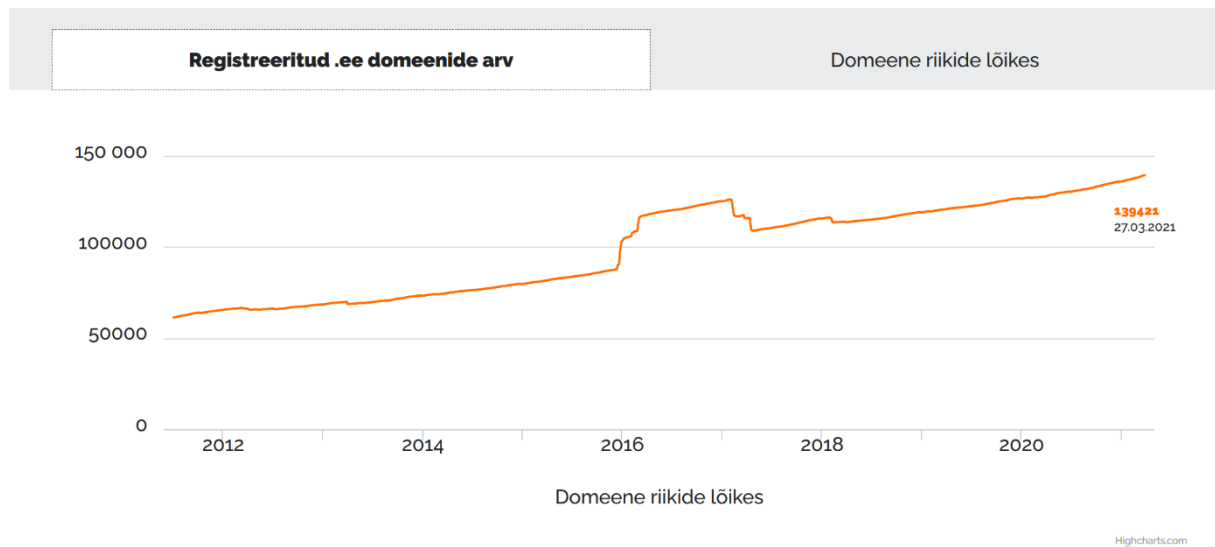
Üks lähedasem leidaolev uuring sellele tööle on Eesti Interneti Sihtasutuse (edaspidi EIS) poolt hallatud turvamonitoring umbes 865-le Eesti hostile Hardenize rakenduse abil [3]. Hardenize rakenduse fookus on domeene majutavate serverite turvastandarditele vastavuse kontrollimine. Selle alla lähevad levinumatest standarditest: veebiserveri suhtluse suunamine HTTPS protokollile, SSL/TSL; meiliserveri SMTP, SPF, DMARC; nimeserveri DNSSEC.

Hardenize .ee töölaualt¹ leiab turvaülevaate kõigi jälgitavate domeenide kohta, domeeninimele vajutates kuvatakse täpsem kirjeldus turvaseadustest ja samuti on võimalus näha agregeeritud infot kogu Eesti peale või sektorite lõikes². Lisaks sellele saab Hardenize kodulehel lasta koostada raport vabalt valitud url'i kohta, mis on hea võimalus enda hallatava domeeni turvalisuse kontrollimiseks.

¹ Hardenize detailne töölaud. <https://www.hardenize.com/dashboards/ee-tld/#/details> (Hardenize veebileht paistab olevat arendusfaasis ja mitte pidevalt kättesaadav. Viimati kontrolliti 05.05.2021, mil leht oli kättesaadav.)

² Hardenize töölaud. <https://www.hardenize.com/dashboards/ee-tld/> (05.05.2021)

Lisaks turvaseirele koostab EIS statistilist ülevaadet müüdnud .ee domeeninimede kohta ning avaldab seda oma veebilehel [2].



Joonis 1. Registreeritud .ee domeenide arv [2].

Joonisel 1 on toodud graafik registreeritud .ee domeenide arvust ning selle arvu kasvust alatest kuupäevast 03.07.2011, mil oli registreeritud 61 298. Graafikut uuendatakse iga 10 minuti tagant ning töö kirjutamise hetkeks (28.03.2021) oli registreeritud domeenide arv 139 421, mis on ligi 2,3 korda suurem aastakümnetagusega võrreldes. Lisaks registreerimiste üldarvule on samalt lehelt leitav veel registreeringute arv erinevate jaotuste (juriidilisest isikust registreerija tüübi, registreerimise/pikendamise perioodi pikkuse, anonüümsetele registreerijale kuuluvate domeenide arvu, jt) järgi ning eraldi domeenide kustutamiste ja uuendamiste arv.

Otsides google.com otsingumootorist märksõnu „Eesti domeenide statistika“ tuleb vastuseid umbes viie lehekülje jagu ning nende seast paistavad välja ainult EIS-i koostatav statistika (mille kirjeldus on eelmistes lõikudes), sellel põhinevad ajakirjandusartiklid ja üks väiksem Eesti E-kaubanduse Liidu ja Zone koostöös valminud uurimus, mis ühe aspektina kaardistas e-poodide osakaalu .ee domeenide hulgas [5].

Üks põhjus uuringute puudumise taga võib olla see, et kuni 2019. aastani ei olnud avalikkusel võimalik kätte saada terviklikku nimekirja .ee domeenidest ning see oli teada ainult neid haldavale Eesti Interneti Sihtasutusele. Ühes EIS-i blogipostituses mainib organisatsiooni praegune arendusjuht Timo Võhmar, et nende kümneaastase tegutsemisaja jooksul on terve nimekiri antud välja vaid ühel korral ning ka see oli teaduse jaoks ja karmi lepingu põhjal [6].

2.2 Eesti domeenide nimekiri

Tsoonifail on tekstifail, mis aitab tagada domeenide toimimise ning korrektse lahendamise nende taga seisvate rakenduseni. Eesti tsoonifail sisaldab infot domeenide ja neid teenindavate nimeserverite kohta, nimeserverite IP-sid ja DNSEC kirjeid. Joonisel 2 on toodud näitena tsoonifaili read, mis sisaldavad informatsiooni *ut.ee* domeeni kohta.

ut.ee.	21600	IN	NS	ns.ut.ee.
ut.ee.	21600	IN	NS	ns2.ut.ee.
ut.ee.	21600	IN	NS	ns2.eenet.ee.
ut.ee.	3600	IN	NSEC	uta.ee. NS RRSIG NSEC
ut.ee. 8608	3600	IN	RRSIG NSEC 8 2 3600 20210406030422 20210309024140	ee. JkSubCUTtoaiRbiJ0E5enJ5RImYyzLa5cnT+nlQnvLpOdWm0u9cODPTC FNI8VkJZo29iuF0x2jBcvy+GLTOAj/sok6wl91BoqefN5gnlMYRdiEy1s cvVRRRYVg/nN7J9vK/1sLcJ6wdnFb8NAdZczJIYfgBtViIORDFViRaSi rVU=
ns.ut.ee.	21600	IN	A	193.40.5.99
ns2.ut.ee.	21600	IN	A	193.40.5.76

Joonis 2. *ut.ee* domeeniga seotud kirjed tsoonifailis.

Kuni 2019. aastani oli .ee tsoonifail salajane ning selles sisalduvat infot oli võimalik pärida ainult ühe domeeni kirjete kaupa, kusjuures päringuks oli vaja juba teada vastavat domeeninime. Põhjuseid selle salajas hoidmiseks arutab Timo Vöhmar EIS-i blogipostituses 2016. aastal „Miks on tsoonifail “saladus”?“ [7]. Faili avalikustamise vastu tõi ta järgmised argumentid:

- See lihtsustaks sihtmärkide otsimise reklaamteade saajatele ja domeenikrabbajatele;
- lihtsustaks domeeniomanike profileerimist (näiteks oli domeeninime teades võimalik WHOIS päringuga kätte saada domeeni kontaktisikud);
- avalikustaks domeeninimed, mida soovitakse salajas hoida (näiteks ettevõtete poolt eelseisvate kampaaniate jaoks registreeritud domeenid).

Eelmisest artiklist kolm aastat hiljem, 2019. aasta juunis, muutis EIS Eesti tsoonifaili avalikult kättesaadavaks. Avaldamisele järgnenud blogipostituses [6] arutab Timo Vihmar, miks ei pea EIS enam oluliseks faili salajas hoidmist. Lahenduseks eelmises lõigus toodud probleemidele märgib ta, et seoses Euroopa Liidu uue isikuandmete kaitse üldmääruse (GDPR) rakendamisega ei ole enam WHOIS teenusest võimalik kätte saada inimeste isikuandmeid. Domeeninimed varjamise jaoks lõi EIS võimaluse enda domeeninimi tsooni failist välja jätta, kui see registreerida ilma nimeserveriteta, ning domeenikrabbajatega võitlemiseks loodi

oksjonikeskkond, kus vabanenud domeen läheb mitte kiireimale, vaid kõige rohkem pakkuvale huvilisele. Viimaks märgib ta, et tegelikult oli ka varasemas süsteemis võimalik osavamatel inimestel varjatud info kokku korjata, ja viitab Ardi Jürgensi artiklile [8], mis arutab just selliseid meetodeid.

Töös vaadeldud .ee domeenid on võetud sellest Eesti tsoonifailist ning analüüsiks kasutatud fail laaditi alla AFXR protokolliga kasutades 13.03.2021 aadressilt *zone.internet.ee*. Unix käsk .ee tsoonifaili allalaadimiseks ja faili „zone.ee“ salvestamiseks:

```
dig @zone.internet.ee ee. axfr > zone.ee
```

Igalt tsoonifaili realt otsiti domeeninimesid regulaaravaldisega, mis võttis vastu „.ee“-ga lõppevaid sõnesid rea algusest, millele järgnes punkt ja tabulaator (domeeninimele vastab avaldises sulgude sees olev osa):

```
^[^ ]+?.ee).\t
```

Tsoonifail sisaldab 983 kirjet Eesti domeenide kohta, mis jäävad regulaaravaldisega koostatavasse nimekirja, aga on märgitud ka mõne teise domeeni nimeserveriks (need kirjed on samuti vajalikud DNS-i toimimiseks tsoonifaili alusel). Nimeservereid üldiselt ei kasutata veebilehtede serverimiseks, aga kuna ligi pooled (446) neist vastasid siiski HTTP GET päringule mingi veebilehega, siis otsustas autor need nimekirja alles jätta. Ilmselt oli nende puhul tegu cPanel-i või muu sarnase majutuslahendusega, kus kõik teenused (veebi-, meili- ja nimeserver) on koos samas serveris.

2.3 Ämblik

Ämblik on robotprogramm, mille eesmärk on käia läbi Interneti lehekülgi ja kaardistada nende sisu. Termin on tuletus inglisekeelsest sõnast, kus programmi käimine läbi World Wide Web-i on analoogiline ämbliku ronimisega oma võrgus. Töö jaoks oli vaja koostada ämblik, mis suudaks käia läbi .ee lõpuga veebilehti ning kaardistada nende sisu. Ämbliku programmeerimiskeeleks valis autor JavaScripti ja käigukeskkonnaks (ingl *runtime environment*) NodeJS, kuna need pakuvad lihtsat rakendusliidest asünkroonse koodi kirjutamiseks.

NodeJS käitab JavaScriptis kirjutatud koodi sündmuste silmuse (ingl *event loop*) põhimõttel: programmikoodi hakatakse täitma lineaarselt, kuid kui mõnda asünkroonseks märgitud koodiosa (näiteks funktsiooni) ei ole võimalik kohe lõpuni täita, siis jäetakse see pooleli ning jätkatakse ülejäänud koodi täitmisega. Kuni programmi täitmine ei ole lõpuni jõudnud, jälgib

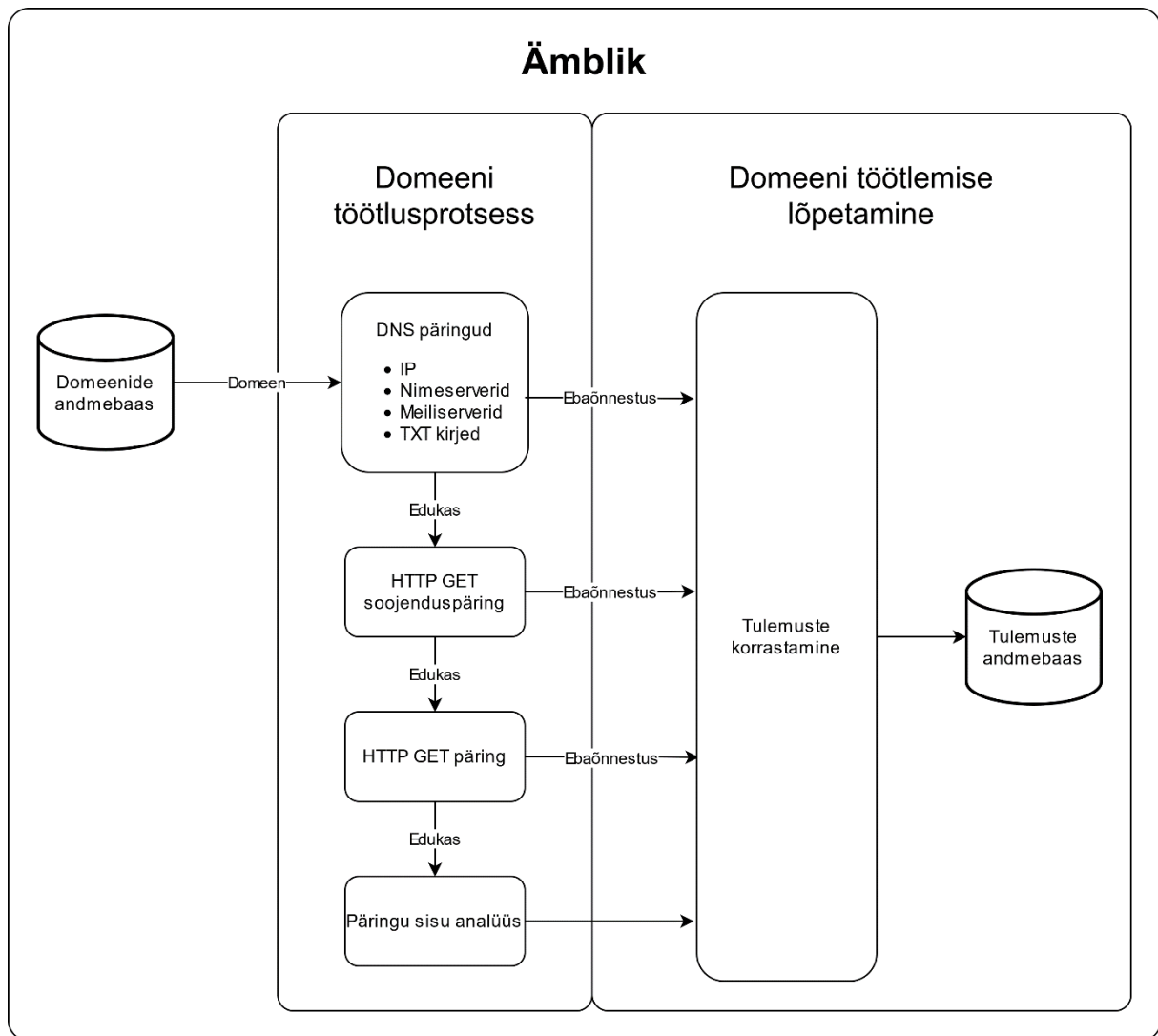
NodeJS käigukeskkond pidevalt sündmuste (nagu näiteks veebipäringu vastuse saabumise) toimumist, mis võimaldaksid pooleli jäetud programmikoodi osi lõpuni täita. Veebipäringute tegemise jaoks annab see suure eelise, kuna veebiserverilt vastuse saamine võtab aega ning ebaõnnestub kohati. JavaScript NodeJS käigukeskkonnas võimaldab teha veebipäringu, kirjeldada ära tööprotsess, mis tuleb teha päringu kättesaamisel, ning jätta vastuse töötlemine hilisemaks. Tänu sellele on võimalik teha mitu veebipäringut korraga, kusjuures programmikood ei jää jõude olekusse ühe päringu vastuse ootamise ajaks, vaid jätkab teiste päringute väljasaatmisega või vastuste töötlemisega.

Ämbliku ehitamisel tulid oluliselt kasuks *Praktika informaatikas* aine jooksul Zones saadud kogemused DNS- ja veebipäringute realiseerimises, sobivate andmebaaside disainimises ning nendega suhtluse korraldamises monitooringurakenduse loomise eesmärgil. Bakalaureusetöö jooksul realiseeriti ka päringutele järgnev analüüs, optimeeriti märgatavalt andmebaasiga suhtlemist ja täiendati veahaldust täpsemate tulemuste saamiseks. Lisaks loodi eraldi moodulina võimekus sagedasemate turvanõrkuste kaardistamiseks.

2.3.1 Ämbliku ülesehitus

Ämblik töötleb domeene korraga 1000 hosti suuruse paki kaupa, mis katsetamise käigus tundus pakkuvat head tasakaalu andmebaasipäringute efektiivsuse ja ajalõpuga (ingl *timeout*) lahenevate veebipäringute hajutamise vahel. Programm on kirjutatud nii, et see võimaldaks suurema läbilaskega võrguühenduse ja andmebaasi korral rakendust lihtsasti üles skaleerida, käivitades paralleelseid isendeid (ingl *instance*), mis töötavad üksteisest sõltumatult. Andmebaasist domeeninimedele valimisel on tagatud, et isend võtab paki, mille töötlemist ei ole veel alustatud või lähiajal lõpetatud.

Joonisel 3 on kujutatud ämbliku tööprotsessi ühe domeeni töötlemisel.



Joonis 3. Ämbliku tööprotsess.

Ämblik teeb iga domeeni jaoks läbi protsessi, kus kõigepealt tehakse päringud DNS andmete kohta ehk selgitatakse välja selle IP-aadressi(d), nimeserverid, meiliserverid ja TXT- kirjed. Kui domeeni IP-aadressi ei olnud võimalik leida, siis eeldab ämblik, et selle jaoks ei ole mõtet ka edasisi veebipäringuid teha: andmed korrastatakse ja pannakse tulemuste andmebaasi.

Kui DNS andmete pärimine oli edukas, siis jätkab ämblik domeeni uurimist, tehes järgmisena HTTP GET soojenduspäringu domeeni pihta. Soojenduspäring on vajalik, kuna ämblik salvestab ka veebiserveri vastamiskiirust ning vähese külastatavusega lehed võivad olla ooteolekus, võttes esimeseks vastamiseks ebatavaliselt pikalt aega (nn cold start). Kui GET päring ebaõnnestub, proovitakse seda kaks korda uuesti saata, ning kui need samuti ebaõnnestuvad, siis loetakse veebileht kättesaamatuks ja sellekohane info sisestatakse andmebaasi.

Kui soojenduspäring õnnestus, läheb ämblik edasi põhilise päringu juurde. Sisuliselt on see identne soojenduspäringuga, kuid selle puhul mõõdetakse ka aega esimese baidi saabumiseni (ingl *Time To First Byte*) ehk kui pikalt läks veebiserveril aega, et hakata külastajale infot edastama. Põhilise päringu õnnestumisel suunatakse see edasi analüüsi, mille jooksul eraldatakse veebipäringu vastusest HTML, JavaScript-ide url-id ning proovitakse Wappalyzer tarkvaraga tuvastada serveri tarkvaraarhitektuur.

2.3.2 DNS andmete kogumine

Domeeninimesüsteem on vajalik domeeni toimimise seisukohalt, et aidata seadmetel Internetis leida oma sihtmärk. Andmed, mille järgi seda tehakse, annavad meile aga natukene ka aimu domeeni üldisest arhitektuurist ja selle majutusasukohast. Joonisel 4 on toodud DNS kirjed, mille ämblik on kogunud *ut.ee* domeeni kohta.

```
{
  "A": "193.40.5.73",
  "NS": "ns.ut.ee, ns2.EENet.ee, ns2.ut.ee",
  "MX": "berta.it.da.ut.ee, frida.it.da.ut.ee",
  "TXT": [ "ZOOM_verify_IPj5TTFVSyKF5kjMAbobow",
    "google-site-verification=psomsSN1Sc7FH8x6mzfoA_OV9YwzbLGyQ1cRhu28_xc",
    "google-site-verification= qT73sEfWHrViVKos6YCtoiR2v3rTJD078NMGwr78fz0",
    "kKOeal7QtSBh9KqU122jcRVhA0O6z8DTS2zukE/acWVPnQp5TvVUQSOFZnTBEdw/c
    Wo8nvh21DSZs2HqfniNLQ==",
    "v=spf1 mx a:sntp1.it.da.ut.ee a:sntp2.it.da.ut.ee a:bounces.ut.ee
    include:spf.protection.outlook.com include:_spf.ut.ee ~all" ]
}
```

Joonis 4. *ut.ee* domeeni kohta kogutud DNS kirjed.

Vaatame kõigepealt domeeni A-kirjeid ehk seda esindavat IP-aadressi. Selle IP-aadressi kaudu on näiteks HTTP GET päringuga võimalik kätte saada ülikooli veebileht. Lisaks on selle järgi võimalik määrata hosti asukoht riigi järgi ning teatavates piirides on võimalik määrata ka

organisatsioon, mille haldusalasse vastav IP-aadress kuulub. Organisatsiooni määramiseks kasutab ämblik MaxMind-i GeoLite2 ASN andmebaasi³.

Kasutusesolevad nimeserverid ja meiliserverid annavad aimu sellest, kui palju domeeniomanik ise rakenduste haldamisega tegeleb, näiteks on Tartu Ülikooli andmete puhul näha, et ülikoolil on nii enda hallatavad nimeserverid kui ka meiliserverid.

TXT-kirjed (või tekstikirjed) on mõeldud ükskõik, mis tekstikujul informatsiooni hoiustamiseks domeeni kohta ning on seega kõige umbmäärasemad, kuid neid on võimalik siiski kasutada domeeni tarkvaraarhitektuuri määramisel. Klassikaliselt kasutatakse tekstikirjeid näiteks tõestamiseks erinevatele rakendustele, et sina oled päriselt domeeni omanik.

Kuna tekstikirjed näevad välja väga segased, siis vaatame Tartu Ülikooli näitel, millist infot oleks võimalik nendest kätte saada. Alustame kõige populaarsematest kirjetest ehk *google-site-verification*.

```
"google-site-verification=psomsSN1Sc7FH8x6mzfoA_OV9YwzbLGyQ1cRhu28_xc",  
"google-site-verification= qT73sEfWHrViVKos6YCtoiR2v3rTJD078NMGwr78fz0",
```

Google pakub teenust Google Search Console⁴, mille kaudu on võimalik saada infot oma domeeni nähtavuse kohta Google otsingutes ning mida inimesed on Google-st otsinud, et jõuda sinu leheküljele. Selle jaoks on vaja kõigepealt luua Google kasutaja ning seejärel tõestada, et mingi domeen käesolevale kasutajale. Domeeni omaniku staatuse tõestamiseks genereerib Google vastavale kasutajale spetsiifilise räsi ning palub selle panna enda domeeniga seotud tekstikirjetesse. Kuna domeeni DNS kirjeid peaks saama muuta ainult domeeni omanik (või selle haldaja), siis loetakse seda piisavaks, et tõestada omanik olemist.

Järgmisena vaatame kirjet ZOOM_verify.

```
"ZOOM_verify_IPj5TTFVSYKF5kjMAbobow"
```

Guugeldamise tulemusel paistab, et tegu on taaskord eelmisele sarnase eesmärgiga kirjega ehk seekord on tõenäoliselt ostetud ülikooli jaoks Zoomi litsents ning selle sidumiseks ülikooli domeeniga on lisatud litsentsile vastavalt genereeritud tekstikirje.

³ MaxMind ASN andmebaas. <https://dev.maxmind.com/geoip/geoip2/geolite2-asn-csv-database/> (10.04.2021)

⁴ Google Search Console kirjeldus. <https://search.google.com/search-console/about> (10.04.2021)

Nüüd vaatame kirjet SPF (Sender Policy Framework) kohta.

```
"v=spf1 mx a:smtp1.it.da.ut.ee a:smtp2.it.da.ut.ee a:bounces.ut.ee  
include:spf.protection.outlook.com include:_spf.ut.ee ~all"
```

SPF kirje aitab takistada meiliteesklast (ingl *email spoofing*), andes nimekirja (alam)domeenidest, millelt on lubatud ülikooli domeenilt e-kirju saata. Väikese infokilluna näeme siin kirjes ka reeglit *include:spf.protection.outlook.com*, mis Microsoft-i SPF-i seadistamise dokumentatsiooni⁵ järgi on vajalik Exchange Online teenuse kasutamiseks.

Kõige krüptilisem näeb välja viimane käsitlemata kirje, millel ei paista olevat selget seost ühegi kindla rakendusega.

```
"kKOeal7QtSBh9KqU122jcRVhA0O6z8DTS2zukE/acWVPnQp5TvVUQSOFZnTBEdw/cWo8  
nvh21DSZs2HqfniNLQ=="
```

Guugeldamise tulemusena võib leida viited sellele, et 512-biti pikkust Base64 tekstikirjet kasutatakse domeeni sidumiseks Microsoft Exchange Federation-i rakendusega⁶. See ei ole muidugi garantii tarkvara kasutatamisele, kuna selline räsi tekib ka näiteks populaarse sha512 algoritmi tulemusel, kuid ründaja jaoks võib too info anda juba potentsiaalseid lähenemissuundasid. Praegusel juhul on ülikooli arvutiabi lehelt⁷ võimalik leida, et mingis variandis on Microsoft Exchange rakendus tõesti ülikoolil kasutusel. Selliste leidude olulisust võib iseloomustada näiteks asjaoluga, et töö kirjutamise ajal leiti Microsoft Exchange tarkvarast kriitiline turvanõrkus⁸ ning taolisi mustreid võidakse otsida sihtmärkide tuvastamise käigus. Sellega oleme vaadanud ühte komponenti domeeniga seotud tarkvaraarhitektuuri tuvastamisest.

2.3.3 Veebipäringute teostamine

HTTP ehk hüperteksti edastuse protokoll (ingl *HyperText Transfer Protocol*) on standard veebiserveriga suhtluse korraldamiseks. GET on üks HTTP päringu liikidest, mida kasutatakse

⁵ Microsofti juhend SPF kirjete seadistamiseks. <https://docs.microsoft.com/en-us/microsoft-365/security/office-365-security/set-up-spf-in-office-365-to-help-prevent-spoofing> (10.04.2021)

⁶ Expta Consulting juhend Microsoft Exchange Federation seadistamiseks. <https://blog.expta.com/2011/07/how-to-configure-exchange-2010-sp1.html> (10.04.2021)

⁷ Tartu Ülikooli arvutiabi leht. <https://wiki.ut.ee/display/AA/Microsoft+Office+365> (10.04.2021)

⁸ Hafnium grupi poolt kasutatud Microsoft Exchange turvanõrkuse kirjeldus.

<https://www.microsoft.com/security/blog/2021/03/02/hafnium-targeting-exchange-servers/> (10.04.2021)

mingi veebiressursi (näiteks veebilehe HTML-faili) küsimiseks. Tehes sellise päringu veebilehe URL-i pihta tuleb vastuseks selle HTML sisu. Tavaline veebilehitseja teeb aga lisaks sellele veel päringuid, et laadida alla kõik lehe kuvamiseks vajalikud failid. HTTP-le lisaks eksisteerib ka selle turvalisem krüpteeritud variant HTTPS ehk turvaline hüperteksti edastuse protokoll (ingl *Hypertext Transfer Protocol Secure*).

Veebileht koosneb ülesehituselt suurest hulgast failidest, mille keskmes on HTML dokument. Selles on kirjas veebilehe tekstiline sisu, üldine struktuur ning lisaks lingid teistele failidele, mis on vaja veebilehe korrektseks kuvamiseks. Nende lingitud failide hulgas on põhilised CSS failid, mis kirjeldavad detailsemalt veebilehe kujundust, JavaScript-failid, mida kasutatakse veebilehe dünaamiliseks muutmiseks, ning pildid ja meediafailid. Töö jaoks on olulised lehe keskmes asuv HTML-fail ning hiljem turvalisust analüüsides JavaScript-failid. Kuna aga HTML-is sisalduvad lingid kõigile teistele failidele, siis on ämblikul mõtekam alla laadida ainult HTML-fail ning hiljem eraldi JavaScript-failid. Seeläbi saab vältida ebavajaliku liiklust ning võrguühendust mitte üle koormata. Näiteks on Tartu Ülikooli kodulehe laadimiseks arvuti veebilehitsejas vaja edastada 4.6MB informatsiooni, millest 3,8MB on pildid ja meediumifailid (~81,7%), 623KB JavaScript-failid (~13,2%), 40KB (~0,85%) lehe HTML sisu ning ülejäänud 203KB on muud lehe välimust kujundavad failid (~4,3%).

Ämbliku üks tööülesanne on käia läbi kõik nimekirjas olevad Eesti domeenid, teha HTTP GET päring nende pihta ning töödelda sellele vastuseks saadud infot. Iga domeeni puhul vaadatakse ainult selle pealehte ehk päring tehakse URL-i `http://{domeeninimi}` pihta ning andmeid kogutakse ainult vastuseks tuleva lehe pealt ja mitte kõigilt alamlehtedelt või alamdomeenidelt. Küll läheb ämblik kaasa kõigi ümbersuunamistega algse URL-i pealt, nagu näiteks suunamine `https` protokollile või ka täiesti teisele alamlehele/domeenile. Nii saab imiteerida päris veebilehe külastust ning uurida, kas domeenid järgivad head tava, suunata liiklus üle turvalisele protokollile `http` -> `https` ja ega ei toimu külastajate ümbersuunamist mõnele pahatahtlikule domeenile.

2.3.4 Tehnoloogilise arhitektuuri tuvastamine

Kui kõik veebipäringud siiani on õnnestunud, siis läheb ämblik edasi analüüsi juurde. Analüüsiks on kasutatud vabavaralist rakendust Wappalyzer [9], et tuvastada veebilehe tehnoloogiline arhitektuur. Wappalyzer on vabavaraline ja avatud lähtekoodiga rakendus, mida täiendatakse jooksvalt vabatahtlike abil. Selle programmiga on võimalik teostada rakendustele omaste tekstimustrite otsingut, kasutades rakendusega kaasatulevat suurt andmebaasi

regulaaravaldistest. Tekstisisendid, mida selle jaoks kasutatakse, on põhiliselt veebilehe HTML sisu (millest on võimalik ka välja lugeda JavaScript failinimed, kui need ei ole sogastatud), GET päringu vastuse päised ja DNS tekstikirjed (mille analüüsimist me oleme juba varasemast peatükis vaadanud). Võimalik oleks tuvastamist veel täiendada, lisades analüüsi JavaScript-failide sisu, kuid praegusel hetkel jääb selle lisamine töö skoobist välja.

Üks Wappalyzer-i poolt tuvastatud rakendustest ülikooli kodulehel on „Drupal 7“. Järgnevalt on kirjeldatud selle tulemuseni jõudmist. Drupal on PHP-s kirjutatud sisuhalduse süsteem, mis lihtsustab lehe administreerimist ja sellele sisu loomist. Nagu varasemalt mainitud, kasutab Wappalyzer suurt hulka reegleid, mille abil on võimalik tuvastada erinevaid tarkvaralahendusi. Järgnevalt on toodud reegel Drupal tarkvara kohta ning pärast reeglit ka kirjeldus selle interpreteerimiseks.

```

"Drupal": {
  "cats": [
    1
  ],
  "cpe": "cpe:/a:drupal:drupal",
  "description": "Drupal is a free and open-source web content management framework.",
  "headers": {
    "Expires": "19 Nov 1978",
    "X-Drupal-Cache": "",
    "X-Generator": "Drupal(?:\\s([\\d.]*)?)\\;version:\\1"
  },
  "html": "<(?:link|style)[^>]+\\\"/sites/(?:default|all)/(?:themes|modules)/\"",
  "icon": "Drupal.svg",
  "implies": "PHP",
  "js": {
    "Drupal": ""
  },
  "meta": {
    "generator": "^Drupal(?:\\s([\\d.]*)?)\\;version:\\1"
  },
  "oss": true,
  "scripts": "drupal\\.js",
  "website": "https://drupal.org"
},

```

Joonis 5. Wappalyzer-i reegel Drupal tarkvara tuvastamiseks.

Kõigepealt on Jooniselt 5 näha, et Drupal-ile on iseloomulikud mõned veebipäringu päised (hallil taustal): aegumise päises „Expires“ on vaikeväärtusena pandud kaasa „19 Nov 1978“, lisaks sellele on veel kaks mittestandardset päist „X-Drupal-Cache“ ning „X-Generator“ jaoks. „X-Generator“-i puhul on veel märkimisväärne, et sellega paneb Drupal kaasa enda *major* versiooninumbri. Topelt kaldkriips ja semikoolon reegli „\\;“ regulaaravaldises tähistab Wappalyzer-i jaoks, et eelnev *capture group* regulaaravaldises vastab mingile otsitavale väärtusele, praegusel juhul versiooninumbri. Rohkem infot reeglite kohta saab lugeda Wappalyzer-i dokumentatsioonist [9].

Vaatame ülikooli kodulehele tehtud GET päringu vastuse päiseid.

The screenshot shows the 'Headers' tab of a browser's developer tools. The request is a GET to <https://www.ut.ee/et>. The status is 200 OK. The response headers are listed below, with several highlighted by red boxes:

- Cache-Control: no-cache, must-revalidate
- Connection: Keep-Alive
- Content-Encoding: gzip
- Content-Language: et
- Content-Length: 40692
- Content-Type: text/html; charset=utf-8
- Date: Wed, 14 Apr 2021 10:17:58 GMT
- Drupal-Pagecache-Memcache: MISS
- Expires: Sun, 19 Nov 1978 05:00:00 GMT**
- Keep-Alive: timeout=5, max=150
- Link: <<https://www.ut.ee/et>>; rel="canonical",<<https://www.ut.ee/et>>; rel="shortlink"
- Server: Apache
- Strict-Transport-Security: max-age=63072000
- Vary: Accept-Encoding
- X-Content-Type-Options: nosniff
- X-Drupal-Cache: MISS**
- X-Frame-Options: ALLOW-FROM <https://www2.cv.ee/>
- X-Generator: Drupal 7 (<http://drupal.org>)

Joonis 6. Ülikooli kodulehe GET vastuse päised.

Joonisel 6 näeme, et Drupal-ile omased päised on siin olemas ning nagu oodatud, sisaldab „X-Generator“ kirje ka *major* versiooninumbrit 7. Lisas 2 on toodud JSON formaadis kogu informatsioon, mis oli ülikooli domeeni kohta võimalik tuvastada siiani kirjeldatud meetodite abil.

2.3.5 Rakenduse testimine

Rakendust testiti arvutil, millel oli 4-tuumaline 8-lõimega protsessor (Intel i7-6700k), 32GB operatiivmälu, SSD püsिमälu (Kingston SUV400S37480G) ja 50Mbit/s internetiühendus. Testimise jaoks käivitati PM2⁹ rakenduse abil paralleelselt kaheksa isendit monitooringu ämblikust, kus ühe korraga töödeldava paki suurus oli 1000 domeeni. Kokku võttis kõigi 136 285 domeeni skannimine aega kaks ja pool tundi, mis tähendas, et sekundis töödeldi umbes 38 hosti või vastupidiselt vaadates, kulus ühe hosti töötamiseks keskmiselt 60 millisekundit. Seejuures tuleb aga arvestada, et need numbrid käivad agregeeritud kogumi kohta (paralleelselt skaneeriti: 8 isendit * 1000 hosti = 8000 hosti) ning omaette võttis ühe hosti töötlemine oluliselt

⁹ PM2 koduleht: <https://pm2.keymetrics.io/> (21.04.2021)

pikemalt. Protsessori töökoormuse kasv oli vaevumärgatav, kuna valdav osa ajast kulust päringute vastuste ootamisele ning nende hilisem töötlemine oli küllalt hajutatud. Keskmiselt võttis üks programmi isend 1000-domeenilist pakki töödeldes 200-300MB mälu, mis kaheksa isendi peale kokku oli 1,6-2,4GB. Võrguliiklus jäi üldiselt vahemikku 10-20Mbit/s, kuid tõusis maksimaalse 50 Mbit/s lähedusse hetkedel, kui mõni isenditest saatis oma päringutepakki välja või hakkas esimesi vastuseid kätte saama. Andmebaasiga suhtlemisel jäi pooljuhtketta lugemiskiirus samuti oluliselt alla poole võimekusest. Seega paistab, et skanneerimise kiirust oleks võimalik suurendada rohkemate isendite paralleelse käivitamisega, kuid juba praeguse kiiruse juures hakkasid ilmnema hostarvuti või -serveri jõudlusest sõltumatud probleemid.

Ühena neist selgus, et kuigi üldine võrgu suutvus (ingl *network capacity*) ei olnud valdava osa ajast kaugeltki mitte ammendatud, siis võis päringute maht käia üle jõu või minna üle lubatud piiride DNS serverile (mis testides Telia võrgus vaikesätetega oli ilmselt üks nende serveritest). A-kirjete päringutest (mis teostati eraldi veebilehe IP-aadressi lahendamisest) olid edukad 94 174 ja 33 523 (~24%) lõppesid veakoodiga *ETIMEOUT*. Vaatamata sellele õnnestusid veebilehe päringud 122 221 domeeni jaoks, mille jooksul oli samuti vaja lahendada domeeninimi IP-aadressiks. Seda probleemi ei esinenud ämblikku Zone sisevõrgus virtuaalserveril katsetades. Selle põhjal on autori hüpotees, et koduvõrgus kasutatavatel DNS serveritel võis olla implementeeritud *Response Rate Limit* (RRL) või sellesarnane lahendus, mis takistas liiga tihedalt tulevatele päringutele vastamist. RRL-i rakendatakse põhiliselt DNS kaudu tehtavate ümmistusrünnete ärahoidmiseks. Üht sellist juhtu käsitleb Riigi Infosüsteemi Amet 2020. aasta maikuu ülevaates *DDoS NXNSAttack* ründe kohta, kus soovitatakse ettevõtete DNS teenuses rakendada RRL võimekus [10]. Kuna veebilehe päringutel oli korduskatsete vahel vähemalt 15 sekundit, siis ilmselt hajutas see päringute saatmist DNS serverile piisavalt, et ajalõppu ei juhtuks.

Teise murekohana ilmnis, et Zone serverist skaneeringut teostades hakkas pärast umbes 10000 domeeni skannimist järjest kasvav hulk päringutest lõppema *ETIMEDOUT* veaga ning töö lõpuks aegus pea iga teine päring. See arv - pärast mida probleem ilmnis - ei paistnud ka sõltuvat sellest, kui mitu isendit ämblikust korraga töötasid ehk kui pika aja jooksul need 10000 päringud välja saadeti (mõistlikkuse piirides). Zones uurides ei paistnud, et neid päringuid hakataks piirama sealse võrgu siseselt (näiteks mõnes tulemüüris). Kuna autori koduvõrgust skaneerides seda probleemi ei esinenud, siis viidi töö lõpupooles kõik monitooringud läbi sealt.

Kolmanda probleemina toimub ämblikus kohati mäluleke või mälu kurnamine, mille allikat ei õnnestunud kindlaks teha. Kui keskmiselt võtab 1000-domeenist pakki töötlev ämblik 200-

300MB mälu, siis kohati tõuseb see ühe protsessi maksimaalse lubatud 2GB-ni, mille juhtudes tehakse rakendusele taaskäivitus. Kuna probleem esines vähem kui korra iga 2-3 täis-skanneeringu tagant ehk selle tõttu kadumaminev 1000-ne pakk moodustab ainult 0,3% monitooringu kogumahust ning selle vea reprodutseerimise katsed olid edutud, jäi selle edasine uurimine töö skoobist välja. Järgneva statistika peatüki aluseks kasutatud täismonitooringus mäluleket ei toimunud.

3. Statistika kogumine .ee alamdomeenide kohta

Järgnev peatükk annab statistilise ülevaate Eesti domeenidest ning nendel kasutusesolevast tarkvaraarhitektuurist olulisemate kategooriate kaupa. Viimane skaneering andmete uuendamiseks on tehtud 18.04.2021, HTML ja JavaScript-failid on laetud alla sellele järgneval päeval.

Skaneeringutega koguti infot 136 285 .ee domeeni kohta, nendest

- 122 221 (~89,7% kõigist) puhul oli veebipäring edukas,
- 10 292 (~7,5%) jaoks ebaõnnestus *nslookup* ja seda põhiliselt *ENOTFOUND* veakoodiga (domeeni kohta ei leitud nimeserverist infot),
- 2659 (~2,0%) ei vastanud päringule ja lõppesid veakoodiga *ETIMEDOUT* (ooteaeg 30 sekundit),
- 431 (~0,3%) keeldusid veebipäringust veakoodiga *ECONNREFUSED*,
- 682 (~0,5%) korral ei õnnestunud päring mõne muu harvem esineva vea tõttu.

Viimase skaneeringu tulemused on leitavad OneDrive kaustast failis *crawl_results.json*¹⁰, ainukese muudatusena on selle „wappalyzer“ tuvastuste osas asendatud täpsed versiooninumbrid tärnidega.

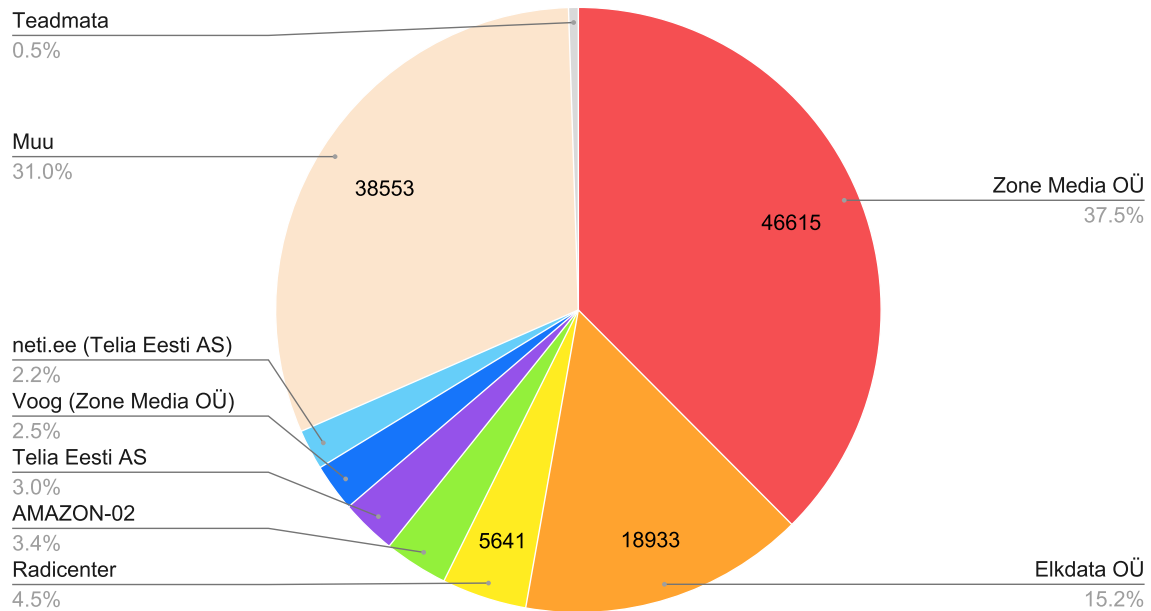
3.1 Domeenide majutus riigi ja organisatsiooni alusel

Domeene majutavaid organisatsioone määrati esimese nimeserveri A-kirjes oleva IP-aadressi alusel MaxMind-i GeoLite2 ASN andmebaasist (vaata pt 2.3.2). A-kirjed koguti ülejäänud skaneerimisest eraldi, kuna seal esinesid probleemid, mida kirjeldati peatükis „Rakenduse testimine“. Vasted leiti 124 176 domeeni kohta ning kuuluvus mingi organisatsiooni haldusalasse oli võimalik tuvastada 123 527 (~99,5% kõigist A-kirjetest) domeeni puhul. Kokku tuvastati kuuluvus 1470 organisatsiooni haldusalasse ning tuvastamiste esirinnas on veebimajutus-ettevõtted. Järgnevalt on toodud tabel domeeni arvu poolest seitsme suurima organisatsiooni kohta. Kokku on grupeeritud domeenid, mille puhul ei olnud võimalik organisatsiooni määrata (< Teadmata >) ja organisatsioonid, mille domeenide osakaal kõigist jäi alla 2% (< Muu >). Lisaks on mõnel juhul toodud organisatsiooni järel sulgudes teine nimi, kui nende infrastruktuur paikneb teise organisatsiooni juures. Kuna määramist tehti nimeserveris oleva A-kirje järgi, siis ei pruugi leitud majutuskoht alati kokku minna tegelikuga

¹⁰ Bakalaureusetöö OneDrive kaust: https://1drv.ms/f/s!AtwLP0bMpZK76AE-PE5_IC3lF8ZP

(näiteks kui kasutatakse proksit mõne teise organisatsiooni haldusalas). A-kirje puudus või ei olnud võimalik kätte saada 12 109 domeenil ning need on majutuse arvestusest välja jäetud.

.ee domeenide majutus organisatsioonide kaupa



Joonis 7. .ee domeenide majutus organisatsioonide kaupa.

Joonisel 7 on visualiseeritud Eesti seitsme suurima domeenide arvuga organisatsiooni osakaal kogu domeenide arvust. Sama info on täpsemalt toodud järgnevalt tabelis 1 ning kirjeldatud pärast tabelit.

Tabelis 1 on toodud kaheksa suurimat klassifikatsioonigruppi autonoomsüsteemi organisatsiooni määramisel: 46 615 kuuluvad Zone Media OÜ haldusalasse, 18 933 Elkdata OÜ haldusalasse (laiemalt tuntud kui veebimajutus.ee), 5641 Radicenter-i haldusalasse, 4245 Amazoni haldusalasse, 3700 Telia Eesti AS haldusalasse, 3163 Voog haldusalasse, 2677 neti.ee haldusalasse ning 38553 domeeni kuuluvad mõne väiksema osakaaluga organisatsiooni haldusalasse. Sellele lisaks oli 649 domeeni, A-kirjele ei leitud andmebaasist majutust.

Tabel 1. Domeenide arv seitsme suurima organisatsiooni haldusalas.

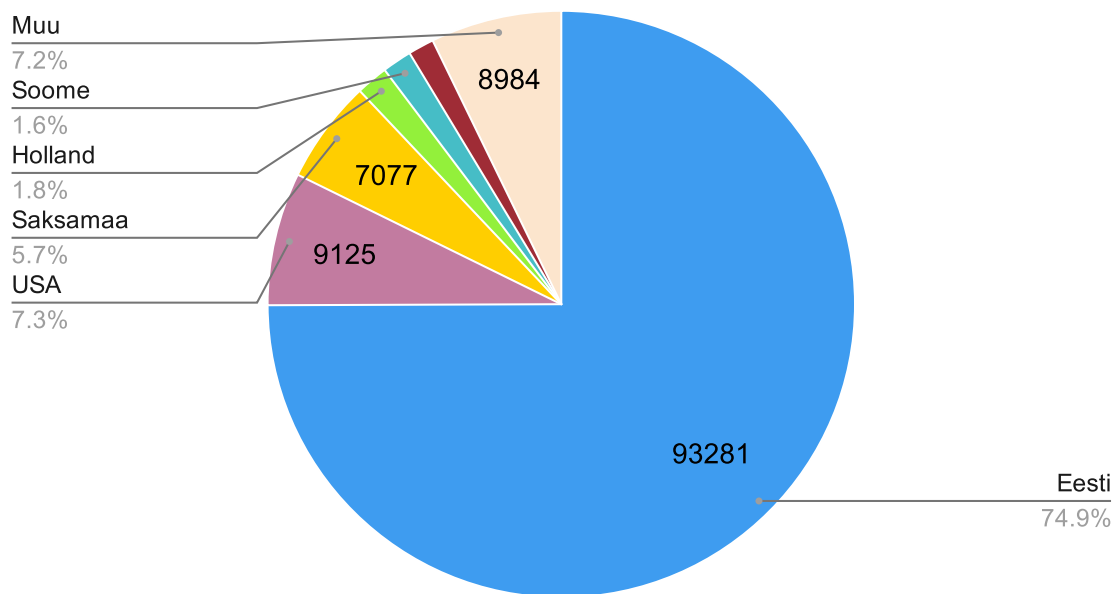
organisatsioon	domeenide arv	osakaal kõigist
Zone Media OÜ	46615	37.5%
Elkdata OÜ	18933	15.2%
Radicer	5641	4.5%
AMAZON-02	4245	3.4%
Telia Eesti AS	3700	3%
Voog (Zone Media OÜ)	3163	2.5%
neti.ee (Telia Eesti AS)	2677	2.2%
< Muu >	38553	31%
< Teadmata >	649	0.5%
Kokku:	124176	

Tabel 100 suurima organisatsiooniga on toodud töö lisa 3 ning terve nimekiri koos autonoomsüsteemi-numbritega on leitav OneDrive failis *hosting_organizations.csv*¹¹.

A-kirjetes olevate IP-aadresside järgi on samuti võimalik määrata riik, mille haldusalasse see kuulub. Siin kehtivad täpsuse osas samad piirangud, nagu olid ka organisatsioonide määramisel ehk kui A-kirjes olev IP-aadress vastab mõnele vahendajale, siis loetakse vastav domeen majutatuks selle vahendaja riigis. Üks suuremaid näiteid sellest on Cloudflare, mis pakub erinevaid proksiteenuseid 2286-le .ee domeenile (autonoomsüsteemi info järgi on CLOUDFLARENET haldusalas) ning nende puhul oleks oluliselt keerulisem või isegi võimatu tegelikku majutusriiki määrata.

¹¹ Bakalaureusetöö OneDrive kaust: https://1drv.ms/f/s!AtwLP0bMpZK76AE-PE5_IC31F8ZP

.ee domeenide majutus riigiti



Joonis 8. Domeenide asukoht riigiti.

Ligi kolmveerand ehk 93 281 uuritud domeenidest paiknevad IP-aadressi järgi Eestis, sellele järgneb Ameerika Ühendriigid 9125 domeeniga, Saksamaa 7077 domeeniga, Holland 2193 domeeniga, Soome 2032 domeeniga, Läti 1775 domeeniga ning muudes riikides paikneb kokku 8984 domeeni. Saksamaa puhul on see number mõnevõrra petlik, kuna seal asuvas Amazoni andmekeskuses on majutatud enamik pargitud .ee domeene (hinnanguliselt vähemalt 2200), mis on müügis turuplatvormi *dan.com* vahendusel.

Terve tabel jaotustest on OneDrive failis *hosting_countries.csv*¹².

3.2 Unikaalsete veebilehtede arv

Domeeni registreerimine ei tähenda, et sellele seatakse üles ka mingi veebileht: domeeni võib registreerida näiteks mõne tulevikuprojektide jaoks, endale reserveerimiseks. Tihti suunatakse sarnase kirjapildiga või organisatsiooniga seostatud märksõnu sisaldavad domeenid kokku ühele domeenile. Näitena võib tuua Tartu Ülikooli, mille veebileht asub domeenil *ut.ee*, kuid sellele domeenile suunavad ümber ka *tu.ee*, *ty.ee* ja *tartuülikool.ee*. Seega võib tsoonifailist leitud domeenide arv 136 285 olla mõnevõrra petlik ning oleks huvitav vaadata, kui palju on tegelikult sisult erinevaid .ee lõpuga veebilehti.

¹² Bakalaureusetöö OneDrive kaust: https://1drv.ms/f/s!AtwLP0bMpZK76AE-PE5_IC3IF8ZP

Sellele lisaks toimub domeenide parkimine ehk ihaldusväärsema nimega domeenide ostmine nende edasimüümise eesmärgil. Sellised domeenid on üldiselt: lühikesed (2-3 sümboli pikkused) tähekombinatsioonid, mis võivad moodustada mõne olulisema akronüümi; sõnastikusõnad, mis võiksid olla olulised mõnele ettevõttele (mõned näited *dan.com* müügiplatvormilt: *gaasiseadmed.ee*, *tervisesalong.ee*, *mahetoit.ee*); varasemalt registreeritud domeenid, mille omanik on lasknud aeguda ja võiks hiljem tahta tagasi osta. Üks suuremaid platvorme domeenide ostmiseks ja müümiseks on *dan.com*, kus erinevate .ee domeenide rendihinnad jäävad üldiselt vahemikku 100€-400€ ühe kuu eest ning väljaostuhinnad vahemikku 1000€-5000€. Konteksti jaoks maksab suuremate teenusepakkujate juures praegu vaba .ee domeeni ostmine aastaks (ilma käibemaksuta) Zones 8,37€, veebimajutus.ee-s 8,25€, Radicenteris 8,35€. Sellistele domeenidele on üldiselt paigutatud vaike-veebilehed, mis viitavad ostmistingimustele, kuid mida ei ole üldiselt mõtet lugeda „aktiivseks veebileheks“.

Vaadates statistika peatüki algusest veebilehtede arvu, mis ei vastanud GET päringule, saaksime me võrdlemisi kitsa tulemuse üldiselt unikaalsete veebilehtede arvu kohta, kuna nagu arutatud, vastavad ka paljud „mitte-aktiivsed“ domeenid veebilehega. Kokku laaditi alla HTML-failid 120 584 domeenilt, mida oli kokku 6,5GB ehk umbes 54KB faili kohta. Nimekirjas oli 29 678 (~24%) faili (k.a duplikaadid), millel leidis vähemalt üks identne HTML-fail mõnelt teiselt domeenilt ning unikaalseid faile oli nende vahel kokku 3815. Siin mängisid lisaks domeenide ümbersuunamisele ka suurt rolli tühjad lehed ja erinevate rakenduste/teenuspakkujate vaikelehed.

Viimasest arvust jääb välja märgatav kogus veebilehti, mis on peaaegu identsed ja erinevad vaid väga vähesel määral üksteisest näiteks sisaldades lehe HTML-is vastava domeeni nime (sellised on paljud vaikelehed, mis ütlevad näiteks „Domeen { } on müügis“, „Domeenil { } puudub serveriteenus“). Siit tuli järgmine katsumus, milleks oli veebilehtede sarnasuse hindamine. Klassikaliselt saab tekstide omavahelist sarnasust hinnata algoritmidega nagu Levenshtein-i kaugus, kuid kuna selle arvutuslik keerukus kasvab sisendtekstide pikkuste põhjal eksponentsiaalselt, siis ei ole neid realistlikult võimalik kasutada siin failide võrdlemiseks. Kuna faile oli kokku 120 584, siis nende omavahelisi võimalikke kombinatsioone paari kaupa, mida tuleb läbi vaadata, on kokku $C_{120\ 584}^2 \approx 7 * 10^9$ ehk 7 miljardit.

Selle jaoks on võimalik kasutada hägusräsimit (ingl *fuzzy hashing*), kus räsifunktsioon on loodud nii, et sarnased sisendid annaksid väljundina sarnased räsud ning nende tulemuste omavahel võrdlemine annaks aimu algsete failide sarnasusest. Üks arvutikriminalistikas ja

pahavara tuvastamises sagedalt kasutatud leidev räsimis-programm on ssdeep¹³, mis kasutab *context triggered piecewise hashing*-ut. Selle algoritmi töö käigus jagatakse sisend väiksemateks plokkideks, arvutatakse räsid nendele plokkidele ning ühendatakse plokkide tulemusi *rolling hash* põhimõttel kokku üheks suureks räsiks. Seda protsessi kirjeldab täpsemalt ssdeep algoritmi autor J. Kornblum oma töös „*Identifying almost identical files using context triggered piecewise hashing*“ [11]. Hiljem on võimalik kahte räsi ssdeep programmi abil omavahel võrrelda: programm vaatab, kui suur osa plokkidest on muutunud ning anda selle põhjal hinnangu 0-100, kui suure tõenäosusega olid tegu natuke muudetud variantidega samast failist. Oluline on siin märkida, et väljundiks saadud hinnang on heuristiline tõenäosus, mitte failide sisu kattuvus protsendina.

Programmi ssdeep abil arvutati kõigi failide omavaheline sarnasus, mis võttis neljatuumalisel protsessoril räsede arvutamiseks ja 7 miljardi võrdluse tegemiseks aega umbes kaks tundi. Tulemusena leidis programm nullist suurema tõenäosusega sarnasuse ~132 miljoni failipaari vahel, nendest ~124 miljonit või 94% olid kindlusega > 90. Järgmise sammuna oli vaja sarnased failid jagada klastritesse, et neid oleks lihtsam kategoriseerida. Selleks otsiti faile, millele leiti kõige rohkem sarnaseid vasteid, võeti need klatri esindajaks ning pandi ülejäänud vasted samasse klastrisse. Seda sammu korrati järjest failide peal, mida ei olnud veel kuhugi ära paigutatud.

Samuti oli vaja leida heuristilise kindluse miinimumpiir, kust maalt edasi enam faile kokku ei määrata (liiga madala kindlusega hakkavad vaevu sarnanevad isendid kokku paigutama). Kuna andmebaasis olevate domeenide hulk oli liiga suur, et neid kõiki läbi kontrollida, tuli siin alampiiri valimisel teatud määral lähtuda subjektiivsest hinnangust. Katsetamise käigus paistis kindlus ≥ 70 (ehk tõenäosus, et mõni klatri fail on natuke muudetud variant esindajast) olevat hea tasakaal, kus sattusid kokku sarnased failid. Lastes lubatud kindluse alla 50, hakkasid moodustuma klastrid erinevate lehe-ehitamise (ingl *site-builder*) rakendustega loodud domeenidest, kus lehed võivad peale vaadates erineda välja näha (pildid, tekst, menüü), kuid kuna nad on loodud ühe malli järgi, siis on ülejäänud HTML-failide sisu suurel määral kattuvad. Kindlusega ≥ 70 leidis koopia 62 620 domeenil (k.a koopiad) ning nendest moodustus 7234 klatri, millest suuremad on toodud tabelis 2.

¹³ GitHubi repositoorium projektile ssdeep. <https://ssdeep-project.github.io/ssdeep/index.html> (29.04.2021)

Tabel 2. Suuremad sarnaste veebilehtedega domeenide klastrid.

klaster	esindaja	domeenide arv	ettevõtte - selgitus
1	<i>0a.ee</i>	12949	Zone - veebiserver puudub
2	<i>0365.ee</i>	7823	Tühi leht
3	<i>onn.ee</i>	3001	Tühi leht
4	<i>100fm.ee</i>	2247	Zone – veebiserver puudub (varem on olnud)
5	<i>10eurot.ee</i>	1810	veebimajutus.ee - veebiserver puudub
6	<i>teave.ee</i>	1761	dan.com - domeen müügis
7	<i>kasskass.ee</i>	1128	Zone - statusCode 403
8	<i>100beers.ee</i>	833	Radicenter - veebiserver puudub
9	<i>123media.ee</i>	508	Interneto vizja - veebiserver puudub
10	<i>ehitusprojekt24.ee</i>	440	Voog - tühi või vähese sisuga leht
Kokku:		32500	

Töö jooksul vaadati läbi kõikide klastrite esindajad, millesse kuulus vähemalt 100 domeeni (kokku 34 klastrit) ning lisati neile selgitus selle kohta, kuidas domeenid sarnanevad. Tabelis 2 on toodud 10 suurimat klastrit, kuid kogu jaotus koos tabelist väljajäävate selgitustega on leitav OneDrive kaustast failis *clusters/clusters_match_70.csv*¹⁴. Juba suurimad 10 moodustavad üle poole kõigist tuvastatud domeenidest ning nendest joonistuvad välja selged põhjused sarnasusteks: vaikelehed domeenidel, millel pole veel oma veebilehte (kategoriseeritud kui „veebiserver puudub“); veakoodid ja erinevad tühjad või minimaalse sisuga lehed. Domeenid on jagunenud kahte erinevasse klastrisse „tühi leht“, kuna vaatamata tõsisema sisu puudumisele `<body></body>` tähiste vahel, on ülejäänud failid natukene erinevad.

Selle põhjal võib anda üldistava hinnangu, et kõigist ligi 136 000 .ee domeenist 15 700 (11%) ei olnud üldse võimalik HTTP GET päringuga kätte saada, 37 900 (28%) domeenidel olevad veebilehed kuvasid mingit laadi vaike- või vealehte ning ülejäänute hulgas oli 24 700 (18%) veebilehel heuristilise tõenäosusega ≥ 70 vähemalt üks identne leht (näiteks ümbersuunamiseks mõeldud alias-domeen). Seega paigutub unikaalsete .ee veebilehtede arv umbes 64 900

¹⁴ Bakalaureusetöö OneDrive kaust: https://1drv.ms/f/s!AtwLP0bMpZK76AE-PE5_IC3IF8ZP

lähedusse ehk 47,7% domeenide koguarvust (k.a. mittevastavad) või 51,7% aktiivsete veebilehtede arvust.

$$\begin{aligned} \textit{unikaalseid} &= \text{kõik} - \text{ei vastanud} - \text{omab duplikaati} + \text{unikaalseid duplikaatide hulgas} \\ &= 136000 - 15700 - 62620 + 7234 \approx 64900 \end{aligned}$$

4. Turvaülevaade

Ämblike kasutusvaldkondade hulgas on oma koht ka turvamonitoringutel, seda nii Interneti ohutumaks muutmisel kui ka rünneteks sihtmärkide otsimisel. Töö jooksul kogutud andmete maht on liiga suur, et seda kõike oleks võimalik siin paari peatüki jooksul süvitsi läbi vaadata ning seetõttu tuuakse hoopis põhjalikumad näited mõnede laialt kasutatavate tarkvarade põhjal, et illustreerida, kuidas on võimalik monitoringut kasutada turvanõrkuste avastamiseks ja Interneti ohutumaks muutmiseks. Näideteks toodud tarkvarade valikul on lähtunud Zone kogemusest oma haldusala turvamisel ning kirjeldatud nende silmis mõnda suuremat murekohta praegu Eesti võrgu kaitsmisel.

4.1 Aegunud tarkvaraversioonid

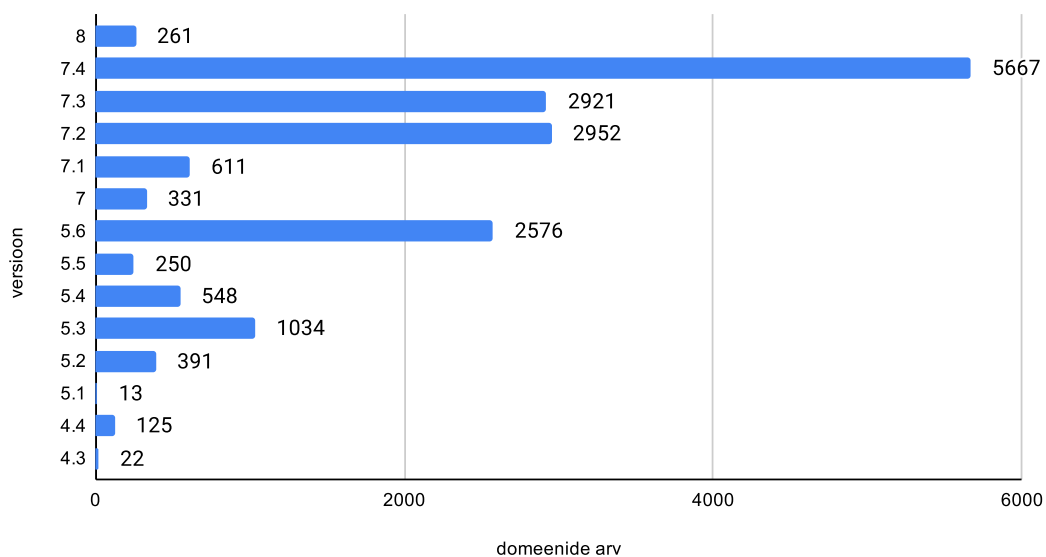
Töö varasemates peatükkides on kirjeldatud tehnikaid, kuidas on võimalik tuvastada veebilehe tarkvaraarhitektuuris olevaid rakendusi nii lihtsalt rakenduse nime tasandil kui ka kohati koos - rohkem või vähem - spetsiifilise versiooninumbri. Versiooninumbri teadmine teeb aga oluliselt lihtsamaks teadaolevate turvanõrkuste ärakasutamise aegunud ja uuendamata rakendustes.

Töö jooksul tuvastati veebilehtedelt kokku 467 erinevat rakendust ning nende peale kokku 4721 erinevat versiooninumbrit. Terve nimekiri tarkvaradest koos versiooninumbrite ja tuvastuste arvuga domeenidelt on leitav OneDrive failis *software_version.json*¹⁵.

Selle infoga on juba võimalik edasi kaardistada veebilehtede turvalisust ning otsida üldisi murekohti. Leidude illustreerimiseks võime võtta PHP programmeerimiskeele, mida veebitarkvara uuriva organisatsioon W³Techs andmetel kasutab 79,2% veebilehtedest nende monitooritava 10 miljoni lehe hulgas, kus serveripoolne programmeerimiskeel oli võimalik kindlaks teha [12]. Kuna PHP on kasutusel nii suurel enamikul kõigist veeblilehtedest, siis annab selle uurimine võrdlemisi laia ülevaate tarkvara ajakohasena hoidmisest kõigil domeenidel. Eesti domeenidel tuvastati kokku koos täpse versiooninumbri 17 709 PHP paigaldust. Erinevad versioone (*major.minor.build*) omaette oli 237, mille jaotused grupeerituna *major* ja *minor* numbril tasemel on toodud järgneval joonisel ja tabelis.

¹⁵ Bakalaureusetöö OneDrive kaust: https://1drv.ms/f/s!AtwLP0bMpZK76AE-PE5_IC3IF8ZP

PHP major + minor versioonid



Joonis 9. PHP versioonid .ee lehtedel.

Tabel 3. PHP versioonide avaldamine ja toetamise lõpp.

PHP Versioon	Avaldamiskuupäev	Toetamise lõpp	Tuvastuste arv	Tuvastuste osakaal
4.3	27. Dec 2002	31. Mar 2005	22	0.1%
4.4	11. Jul 2005	7. Aug 2008	125	0.7%
5	13. Jul 2004	5. Sep 2005	0	0%
5.1	24. Nov 2005	24. Aug 2006	13	0.1%
5.2	2. Nov 2006	6. Jan 2011	391	2.2%
5.3	30. Jun 2009	14. Aug 2014	1034	5.8%
5.4	1. Mar 2012	3. Sep 2015	548	3.1%
5.5	20. Jun 2013	21. Jul 2016	250	1.4%
5.6	28. Aug 2014	31. Dec 2018	2576	14.6%
6.0	Ei avaldatud			
7	3. Dec 2015	10. Jan 2019	331	1.9%
7.1	1. Dec 2016	1. Dec 2019	611	3.5%
7.2	30. Nov 2017	30. Nov 2020	2952	16.7%
7.3	6. Dec 2018	6. Dec 2021	2921	16.5%
7.4	28. Nov 2019	28. Nov 2022	5667	32%
8	26. Nov 2020	26. Nov 2023	261	1.5%

Tabelis 3 toodu avaldamiskuupäevad on leitavad PHP veebilehel avaldamiste arhiivis¹⁶, toetuse lõpetamiste kuupäevad lehtedelt *Unsupported Branches*¹⁷ ja *Supported Versions*¹⁸.

Nagu on näha jooniselt 9 ja tabelist 3, varieeruvad praegu kasutusesolevad versioonid 2002. aastal avaldatud 4.3-st, värskeima 2020. avaldatud 8.0-ni, sealjuures moodustavad lõppenud toetusajaga (≤ 7.3) versioonid 50% kõigist PHP paigaldustest, mida oli võimalik tuvastada. Lõppenud toetusperiood tähendab ka, et neid PHP versioone kasutavad domeenid ei ole saanud turvauuendusi juba mitu aastat ning otsides NIST-i *National Vulnerability Database*-st¹⁹ turvanõrkusi PHP keeles, tuleb 05.05.2021 seisuga vasteid 659 (otsing tehti *Common Platform Enumeration* identifikaatori *cpe:/:php:php:* alusel).

The Open Web Application Security Project (OWASP)²⁰ on rahvusvaheline spetsialistidest koosnev mittetulundusühing, mille eesmärk on avatud lähtekoodi projektidena välja töötada veebirakenduste turvastandardeid. Nende loodud *Application Security Verification Standard* (ASVS) hiliseima, 4.0 versiooni peatükis „*V14 Configuration*“ [13] on välja toodud põhimõtted veebirakenduse konfiguratsiooni turvalisuse kontrollimiseks. Selle standardi punktis 14.3.3 on nõutud, et „HTTP päised või ükski muu osa HTTP vastusest ei tohi paljastada süsteemikomponentide detailset versiooni informatsiooni“²¹ (tõlge). Selle põhjendusena on toodud välja just asjaolu, et peidetud versiooninumber raskendab märgatavalt rakenduste teadaolevate turvanõrkuste ärakasutamist.

4.2 Versiooninumbrite tuvastamise täpsus

Versiooninumbrite tuvastamine on kõige lihtsam regulaaravaldiste abil, mis otsivad rakenduste poolt kindlates kohtades vastuses kuvatavat infot. Nägime juba *ut.ee* puhul päist „X-Generator“, milles oli lehe Drupali versioon, kuid see ei ole kaugeltki mitte omane ainult Drupalile. Sarnast päist „X-Powered-By“ kasutab ka viimases peatükis vaadeldud programmeerimiskeel PHP, mis isegi kriitilisemalt, paneb kaasa versiooni *major minor* ja *build täpsusega*.

¹⁶ PHP avaldatud versioonide arhiiv. <https://www.php.net/releases> (04.05.2021)

¹⁷ PHP lõpetatud toetusega versioonide arhiiv. <https://www.php.net/eol.php> (04.05.2021)

¹⁸ PHP toetusega versioonide arhiiv. <https://www.php.net/supported-versions> (04.05.2021)

¹⁹ NIST *National Vulnerability Database*. <https://nvd.nist.gov/vuln/search> (05.05.2021)

²⁰ OWASP koduleht. <https://owasp.org/> (05.05.2021)

²¹ ASVS 4.0 nõue 14.3.3. <https://github.com/OWASP/ASVS/blob/v4.0.2/4.0/en/0x22-V14-Config.md#v143-unintended-security-disclosure-requirements> (05.05.2021)

▶ GET https://www.██████████/

Status **200 OK** ⓘ
Version HTTP/1.1
Transferred 49.14 KB (48.80 KB size)

▼ Response Headers (350 B) Raw

- ⓘ Connection: keep-alive
- ⓘ Content-Type: text/html; charset=UTF-8
- ⓘ Date: Wed, 05 May 2021 21:30:38 GMT
- Link: <https://www.██████████/wp-json/>; rel="https://api.w.org/"
- Link: <https://www.██████████/>; rel=shortlink
- ⓘ Server: nginx
- ⓘ Transfer-Encoding: chunked
- X-Pingback: https://www.██████████/xmlrpc.php
- X-Powered-By: PHP/5.3.3

Joonis 10. PHP X-Powered-By päis.

Joonisel 10 on toodud näide PHP-d kasutavast veebilehest, kus ei ole päises keelatud täpse versiooni kuvamine. Googeldades ei ole võimalik leida ühtegi tehnilist põhjust selliste päiste kasutamiseks ning konsensus foorumitest paistab olevat, et nende põhilised kasutusvaldkonnad on rikkeotsigu (ingl *troubleshooting*) abistamine ning ettevõtetele enda reklaamimine ja erinevate tarkvaraversioonide turuosa hindamiseks²². Päises „server“ olevat infot serveritarkvara kohta suutsid brauserid kohati kasutada teadaolevate ühilduvusbugide vältimiseks, kuid tänaseks on sellest saanud pigem pärandlahendus (ingl *legacy feature*).

Siiski ei õnnestu regulaaravaldisega versiooninumbrite otsimine alati, kuna nende kasutamine failinimedes ja sisus ei pea alluma kindlatele reeglitele. Failis on näha, et valdaval enamikul kordadest tuvastati jQuery teegi versioon levinud *sequence-based identifier* (nt 1.12.5) formaadis, kuid kohati segunes nende hulka ka teisi, mis nii palju infot ei anna. See juhtus näiteks failinimedest versiooninumbrite otsimisel, kus üldiselt kasutati url-i lõpus valikulist päringut `?ver=` soovitud faili küsimiseks versiooninumbri järgi, kuid mõned üksikud implementatsioonid kasutasid seda faili küsimiseks kas üleslaadimise (unix) ajatempli või avaldamiskuupäeva järgi.

²² Stackoverflow küsimus „X-Powered-By“ päise olulisuse kohta. <https://stackoverflow.com/a/1288385> (06.05.2021)

<code>/wp-includes/js/jquery/jquery.min.js?ver=3.5.1</code>	(versiooninumber)
<code>/wp-content/cache/.../jquery.waypoints.js?ver=1619097972</code>	(unix ajatempel)
<code>/wp-content/themes/.../js/jquery.superslides.js?ver=20120206</code>	(avaldamiskuupäev)

Joonis 11. Versioonitunnused jQuery failide url-ides.

Joonisel 11 on toodud näited kolmest implementatsioonist, kus esimene, klassikaline, sisaldas versiooninumbrit, mida regulaaravaldistega otsiti ning teised olid vähem traditsioonilised implementatsioonid, mis põhjustasid vähesel määral müra versioonituvastustes.

4.3 Näited sagedastest rünnetest Zone haldusalas

Skannerite toel tehtavate rünnete tugevus seisneb võimes kasutada sama turvanõrkust ära suurel hulgal sihtmärkidel, mis kasutavad üksteisega identset tarkvara. Üks populaarsematest sellistest tarkvaradest on sisuhalduse süsteem (edaspidi CMS ehk ingl *Content Management System*) WordPress, mis lihtsustab veebilehe omanikele selle ehitamist ning hilisemat haldamist. Veebitarkvara uuriv organisatsioon W³Techs toob välja, et nende monitooritavate 10 miljoni veebilehe hulgas kasutas 4.mai 2021.a. seisuga WordPressi 41,3% veebilehtedest [14]. Samas uurimuses on leitud, et 36,3% lehekülgedest ei kasuta üldse CMS-i, mis tähendab, et kõigi teiste platvormide turuosa on kokku 22,4% ehk umbes pool WordPressi omast ning nende hulgas suurimad olid Shopify, mis tuvastati 3,5% lehtedest ning Joomla 2,1% lehtedest. Nii suurest kasutusest tulenevalt on ilmne, miks WordPress on ka ahvatlev sihtmärk kurjategijatele ja meile nende kaitsmisel. Põhimõtte jaoks on lisas 4 toodud ka WordPressi tuvastatud versioonide jaotus .ee domeenidel.

4.3.1 WordPressi kasutajate loendamine ja jõuründed

WordPressi on sisse ehitatud erivahend *permalinks*²³, mille eesmärk on loodud alamlehtedele anda unikaalne ja püsiv identifikaator, mille põhjal oleks leht leitav isegi juhul, kui selle url on muutunud. Selle eesmärk on hea: jagades viidet enda sisule võib autor soovida, et sisu oleks leitav ka pikema aja möödudes, kus WordPressis kasutatav kaustastruktuur võib olla muuunud. Näitena tuuakse *permalinks* kasutamise juhendis võimaluse jagada postitust selle ID järgi *http://example.com/?p=N*, kus N asemel on postituse järjekorranumber. Teise võimalusena

²³ Wordpress. Using Permalinks. <https://wordpress.org/support/article/using-permalinks/> (06.05.2021)

lubab WordPress kuvada kõik ühe kasutaja poolt tehtud postitused, pärides alamlehte */author/autori_kasutajatunnus* näiteks *http://example.com/author/siim*.

Kombineerides viimased kaks võimalust, on võimalik aga otsida ühe autori postitusi tema *permalink*-i abil ehk näiteks *http://example.com/?author=1*. Kuna kõik WordPressi kasutajad saavad endale *permalink* numברי, ning neid antakse järjest alustades arvust 1, siis on võrdlemisi lihtne viia läbi kasutajate enumereerimise rünne, kus saadetakse sellised päringud esimese *n* autori kohta ning loetakse ümbersuunamise urlist välja kasutajatunnused (populaarse arvuna *n* jääb Zone's olevate veebilehtede logifailidest silma esimese kümne kasutajanime otsimine). Üldine skeem nendeks rünneteks on järgnev:

- Enumereeritakse autorid, kas esimese *n*-indeksi järgi või kuni jõutakse *not found* vasteni (error 404);
- seejärel sooritatakse masspostitused wp-login.php pihta proovides levinud paroole;
- täiendavalt kasutatakse XML-RPC liidest, mis võimaldab ühe päringuga testida mitmeid kasutaja-parooli kombinatsioone.

Töö raames kirjutati eraldi skanner, et näha, kui palju kasutajanimed on sellisel viisil võimalik tuvastada. WordPress oli paigaldatud 36 702 domeenile ning nende skaneerimise tulemusel tuvastati esimese kümne autori enumereerimisel vähemalt üks kasutajatunnus 9106-lt domeenilt. See number on tegelikkuses aga tõenäoliselt märgatavalt suurem, kuna skanneri kalibreerimise lõpuks (tulemuste salvestamise skaneeringu ajaks) ei vastanud manuaalsel kontrollimisel enam ligi pooled lehtedest ka brauseri kaudu tehtud külastusele. Teise IP-aadressi pealt kontrollides olid veebilehed jätkuvalt üleval ja vastasid */?author=nr* päringutele, millest võib järeldada, et skanneri kasutatud proksi oli jõudnud mingil hetkel IP-aadresside musta nimekirja. Sellele vaatamata demonstreerib see vajadust turvalise salasõna järele ning lisakihina on soovitatav kaksikautentimise kasutamine, kuna kasutajatunnused on WordPressi abil ehitatud lehtedel lihtsasti tuvastatavad ning kasutajatunnuseid teades on kurjategijate jõuründed juba poole efektiivsemad.

4.3.2 WordPressi kataloogipuu indekseerimine

Teine WordPressi pool-peidetud lahendus on lehele üles laaditud failide hoidmine */wp-content/uploads* kaustas, mis toimib ühtlasi külastatava alamlehena paigaldustes, kus nende kataloogide nägemine ei ole autoriseerimata kasutajatele keelatud. Kui tavalise blogilehe korral ei ole sellest suurt ohtu, siis e-poodide puhul võib selle tulemuseks olla konfidentsiaalsete

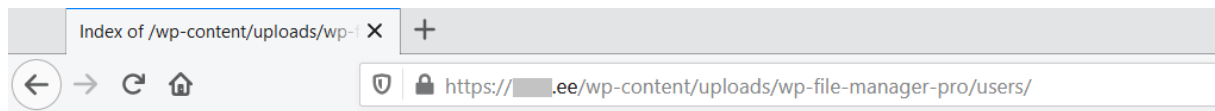
andmete lekkimine. Veebruaris 2021, tuli avalikust kahe veebipoe andmelekked^{24,25}, mille tõttu olid Internetis leitavad ligi 5000 kliendi nimed, meiliaadressid, telefoninumbrid ja kontaktaadressid. Postimehe artiklis on toodud tsitaat andmekaitse inspeksiooni avalike suhete nõunik Signe Heiberg-ilt, kes soovitas nende juhtumite valguses: „Kuna mõlemad e-poeed kasutavad WordPressi platvormi, soovitab inspeksioon olla WordPressi kasutajatel tähelepanelik ning veenduda, et selle kataloogide seadistused ei võimaldaks ligipääsu kõrvalistele isikutele“ [15]. Selle põhjal võib eeldada, et nendel juhtudel lekkisid kliendiandmed kas samast *uploads* või mõnest pistikmooduli spetsiifilisest kaustast, mille ligipääs ei olnud piiratud.

Mõlemas artiklis viitavad poodide omanikud sellele, et failide leidmiseks pidi kellelgi olema suur tahtmine nende lehti rünnata või sealt vigu otsida, kuid tegelikkuses nii päris ei ole ja selle töö teevad ära võimsad indekseerimisrobotid. Otsingumootorid nagu Google toimivad tänu töös loodud ämblikule sarnastele (aga oluliselt keerulisema ülesehitusega) indekseerimisrobotitele, mis käivad süstemaatiliselt läbi kogu Internetti, külastavad iga veebilehte, selle ligipääsetavaid alamlehti ning kõiki faile ja loovad selle sisu alusel andmebaase, mille abil on inimeste otsigutele võimalik (näiteks guugeldamisele) vasteid pakkuda. Sellest on välja kujunenud omaette tehnika, tuntud kui ohmuotsing (ingl *Google dorking*²⁶ või *Google hacking*), mille käigus kasutatakse keerulisemaid otsinguparameetreid ja märksõnu Google indekseeritud sisust sinna kogemata sattunud informatsiooni otsinguks.





²⁴ Postimees. Mineral Garden veebipoe andmeleke. <https://tehnika.postimees.ee/7176653/veebipoe-mineral-garden-omanik-andmelekkest-uurime-kas-tegu-oli-kuberrunnakuga> (06.05.2021)

²⁵ Postimees. Tootemaailm veebipoe andmeleke. <https://tehnika.postimees.ee/7178785/veel-uhe-eesti-veebipoe-klientide-andmed-rippusid-internetis-uleval> (06.05.2021)

²⁶ Wikipedia. Google hacking. https://en.wikipedia.org/wiki/Google_hacking (06.05.2021)



Index of /wp-content/uploads/wp-file-manager-pro/users/

<u>Name</u>	<u>Last modified</u>
 Parent Directory	30-Jun-2020 15:19
 A [redacted]	05-Apr-2019 16:02
 E [redacted]	30-Jun-2020 15:19
 E [redacted]	18-Nov-2019 10:41
 O [redacted]	05-Jan-2020 19:01
 eesnimi.perenimi@domeen	08-Apr-2019 06:56

Proudly Served by LiteSpeed Web Server at [redacted].ee Port 443

Joonis 12. Domeen WordPressi installatsiooniga, kus on failipuu indekseerimine lubatud.

Joonisel 12 on toodud näitena ühe.ee domeeni WordPressi paigaldus, kus on lubatud failide indekseerimine ja alustades kaustast `/wp-content/uploads` on võimalik leida näiteks `wp-file-manager` pistikmooduli kaudu loodud kaustad kasutajate jaoks (kuna kaustad on nimetatud inimeste kasutajatunnuste järgi, siis on need ja täpne e-posti aadress peidetud).

Täiendatud versiooniga autorite loendamiseks loodud skannerist, tuvastati WordPressi kasutatavate .ee domeenide hulgast 209, mille `/wp-content/uploads` kaust ehk lehele üles laaditud failid olid avalikult kättesaadavad. Nimekiri tuvastustest edastati CERT-EE-le, kes on praeguseks juba alustanud selle põhjal domeeniomanike teavitamist.

4.4 Veebilehtede kaitse

Oleme vaadanud mõnda, põhilist viisi kuidas praegu kasutatakse monitooringu programme nii ründe- kui ka kaitserollis. Teades, mida nendega otsitakse, on lihtsamini võimalik ka enda veebilehte kaitsta.

Esimese ja überpääsmatu soovitusena tuleb oma veebirakendus ja server hoida ajakohasena kõigi tarkvarauuendustega, et vältida teadaolevate turvanõrkuste ärakasutamist nende vanemates versioonides. Veebilehe üldiseid turvaseadistusi ja nende puudujääke saab (tasuta) kontrollida Hardenize veebilehel²⁷, lastes seal koostada raporti oma domeeniga seotud üldiste turvaseadistuste kohta. Lisaks kitsaskohtadest teadaandmisele kuvatakse raportis ka lühike

²⁷ Hardenize. <https://www.hardenize.com/> (06.05.2021)

selgitus iga analüüsitud standardi kohta, mille abil on võimalik langetada edasine otsus nende olulisuse või mitteolulisuse kohta enda domeenil.

Töös tarkvaraarhitektuuri ja versioonide tuvastamiseks kasutatud rakendus Wappalyzer on leitav GitHubi repositooriumist²⁸. Rakendust ennast või selle brauseri pistikmoodulit kasutades on võimalik hinnata, kui palju veebilehe arhitektuurist on lihtsasti tuvastatav ning takistada seejärel tehnilise toimimise seisukohalt ebavajaliku info saatmisel HTTP vastuste sisus. Nagu viimastes peatükkides kirjeldatud, on igasugune versiooninumbri kuvamine tugevalt mittesoositud ning paljastab tänaseks ilma erilise põhjuseta rünnatavaid komponente veebilehe arhitektuuris.

Turvalisuse hetkeoludega kursis olemiseks on võimalik endale tellida CERT-EE kübervaldkonna uudiskirja, milles tehakse igal hommikul kokkuvõtte avalikes allikates ilmunud küber- ja IT-uudistest nii Eestis kui ülejäänud maailmas.

²⁸ GitHubi repositoorium. Wappalyzer. <https://github.com/aliasio/wappalyzer> (06.05.2021)

5. Kokkuvõte

Käesoleva bakalaureusetöö raames loodi veebiskaneerimise võimekusega programm ehk ämblik, mille abil on võimalik imiteerida veebilehtede külastusi, uurida nende käitumist külastamisel, määrata domeenide majutus ja tuvastada veebilehtede võimalikult täpne tarkvaraarhitektuur. Töös kirjeldati ämbliku ülesehitust, tehnoloogiavalikuid ja avalikult kättesaadavaid lahendusi nende eesmärkide saavutamiseks. Lisaks demonstreeriti Tartu Ülikooli kodulehe näitel, kuidas on võimalik veebilehe tarkvaraarhitektuur versiooninumbrite täpsuseni määrata. Ämbliku abil kaardistati kogu .ee domeenimaastikku (ligi 136 000 hosti), kirjeldati selle efektiivsus ning raskusi suuremahulise skaneeringu läbiviimisel.

Kogutud andmete põhjal anti ülevaade Eesti domeenide majutusest riigi ja majutusorganisatsiooni põhiselt, hinnati unikaalset sisu omavate .ee veebilehtede arvu 64 900 lähedusse (47,7% domeenide koguarvust) ja uuriti väga sarnaste lehtede suuremaid klastreid. Veebilehtede sisu võrdlemiseks tutvustati räsimestehnikat *ssdeep*, mille abil on võimalik mõistliku täpsuse ja ajakuluga hinnata mahukate tekstifailide omavahelist sarnasust.

Viimaks kirjeldati monitooringuprogrammide olemasolust tulenevaid turvaohтусid ning toodi selle juures näteks, et kõigist PHP paigaldustest, mille täpne versioon oli ämblikuga võimalik määrata, oli 50% hooldusiga juba lõppenud (*end of life*). Toodi näiteid veebimajutusettevõtte Zone Media OÜ haldusalas toimuvatest rünnetest ning testiti WordPress rakendusega .ee domeenide vastuvõtlikust kahele passiivsele ründele. Autorite loendamise (*author enumeration*) tehnikaga prooviti tuvastada WordPressi paigalduste administraatorite kasutajanimed, et simuleerida valmistumist edasiseks jõuründeks (*bruteforce attack*). Selle käigus leiti kasutajanimed 9106 domeeni jaoks, mis on 24% kõigist paigaldustest. Teise ründena otsiti WordPressi paigaldusi, millel on autoriseerimata kasutajatel lubatud kataloogipuu läbimine. Leiti 209 sellist domeeni ning nende nimekiri edastati CERT-EE-le, kes alustasid juba domeeniomanike teavitamist vastavast konfiguratsiooniveast.

Praegu tegeleme Zones töö jooksul kogutud andmete analüüsiga, et leida kaaperdatud ja kuritahtlikku sisu serverivaid veebilehti ning edastada nende nimekiri CERT-EE-le. Edasine plaan on toimiv ämblik kolida üle Zones eraldi virtuaalserverile ning seadistada pidevalt monitoorima kõiki Eesti alamdomeene ja teisi Zone haldusalasse kuuluvaid domeene.

Ämbliku lähtekood jääb ettevõtlus- ja turvakaalutlustel kinniseks, kuid kõik kogutud andmed (millest on peidetud turvakriitiline informatsioon) avaldatakse koos tööga OneDrive kaustas²⁹.

²⁹ Bakalaureusetöö OneDrive kaust: https://1drv.ms/f/s!AtwLP0bMpZK76AE-PE5_IC3IF8ZP

6. Viidatud kirjandus

- [1] Akadeemilise andmeside areng Eestis. EENet.
https://www.eenet.ee/EENet/akadeemilise_andmeside_areng_Eestis (25.03.2021).
- [2] .ee domeenide statistika. Eesti Interneti SA. <https://www.internet.ee/abi-ja-info/statistika> (25.03.2021).
- [3] .EE Dashboard. Hardenize. <https://www.hardenize.com/dashboards/ee-tld/> (25.03.2021).
- [4] Andmekaitse ja infoturbe leksikon. Cybernetica. <https://akit.cyber.ee/> (05.05.2021).
- [5] Eesti e-kaubanduse statistika. Eesti E-kaubanduse Liit. <https://e-kaubanduseliit.ee/eesti-e-kaubanduse-statistika> (28.03.2021).
- [6] Võhmar T. Eesti Interneti SA blogipostitus: „Vabastage tsoonifail!“. 04.07.2019.
<https://www.internet.ee/eis/uudised/vabastage-tsoonifail> (28.03.2021).
- [7] Võhmar T. Eesti Interneti SA blogipostitus: „Miks on tsoonifail ’saladus’?“. 09.06.2016.
<https://www.internet.ee/eis/uudised/miks-tsooni-fail-on-saladus> (02.04.2021).
- [8] Jürgens A. Zone.ee blogipostitus: „EE tsoonifail lõpuks avalik, aga mis hinnaga?“. 25.06.2019.
<https://blog.zone.ee/2019/06/25/ee-tsoonifail-lopuks-avalik-aga-mis-hinnaga/> (03.04.2021).
- [9] GitHub repository for project Wappalyzer. AliasIO.
<https://github.com/AliasIO/wappalyzer> (14.04.2021).
- [10] DNS teenuses avastatud haavatavus / DDoS NXNSAttack. Riigi Infosüsteemi Amet. 2020.
https://www.ria.ee/sites/default/files/2020-05_dns_haavatavusest_nxnsattack_0.pdf

- [11] Kornblum J. Identifying almost identical files using context triggered piecewise hashing. Digital Investigation, Vol 3, Supplement, 2006, p 91-97, ISSN 1742-2876. <https://doi.org/10.1016/j.diin.2006.06.015>
- [12] Usage statistics of server-side programming languages for websites. W3Techs. https://w3techs.com/technologies/overview/programming_language (07.05.2021).
- [13] OWASP V14: Configuration Verification Requirements. <https://github.com/OWASP/ASVS/blob/v4.0.2/4.0/en/0x22-V14-Config.md> (05.05.2021)
- [14] Usage statistics of content management systems. W3Techs. https://w3techs.com/technologies/overview/content_management (04.05.2021)
- [15] Randlo T. Artikkel: „Veel ühe Eesti veebipoe klientide andmed rippusid internetis üleval.“. 14.02.2021. Postimees. <https://tehnika.postimees.ee/7178785/veel-uhe-eesti-veebipoe-klientide-andmed-rippusid-internetis-uleval> (04.05.2021)

Lisad

I. Lisa 1: Tartu Ülikooli koduleht

(<https://www.ut.ee/et>)

The screenshot shows the homepage of Tartu University (Tartu Ülikool) in Estonian. The browser address bar shows [ut.ee/et](https://www.ut.ee/et). The website features a blue navigation bar with the following menu items: Sisseastumine, Õppimine, Täiendusõpe, Teadus, Ettevõtlus, Vilistlaselu, and Ülikoolist. The main content area is divided into several sections:

- Top Left:** A large banner for a photo contest titled "Fotovõistlus TARTU ÜLIKOOL PILDIS" running from March 29 to September 30. It includes logos for sponsors like "Toetajad", "TASUKU", and "PhotoPoint".
- Top Right:** A "UUDISED" (News) section with a sub-header "Tartu Ülikool ootab uude paindlikusse magistriõppesse juhtimiskogemusega spetsialiste" dated 12.04.2021.
- Middle Left:** A "SISSEASTUMINE" (Admission) section with a sub-header "Jätakuvalt saab kandideerida Tartu Ülikooli kahe valdkonna võõrkeelsesesse".
- Middle Right:** A news item titled "Lõputööd kirjutavad üliõpilased pääsevad vajaduse korral esmaspäevast raamatukokku" dated 10.04.2021.

The website also includes a search bar, social media icons, and a "TÜ üksuste kontaktandmed" button.

II. Lisa 1: Tartu Ülikooli domeeni kohta kogutud info JSON formaadis

Domeeni ut.ee kohta ämblikuga kogutud info.

```
{
  "host": "ut.ee",
  "uri": "http://ut.ee",
  "ip": "193.40.5.73",
  "hosting": "Hariduse Infotehnoloogia Sihtasutus AS3221",
  "ns": "ns.ut.ee,ns2.EENet.ee,ns2.ut.ee",
  "mx": "berta.it.da.ut.ee,frida.it.da.ut.ee",
  "txt": "ZOOM_verify_IPj5TTFVSYKF5kjMAbobow,
    google-site-verification=psomsSN1Sc7FH8x6mzfoA_OV9YwzbLGyQ1cRhu28_xc,
    google-site-verification=qT73sEfWHrViVKos6YCtoiR2v3rTJD078NMGwr78fz0,
    kKOeal7QtSBh9KqU122jcRVhAOO6z8DTS2zukE/acWVPnQp5TvVUQSOFZnTBEdw/cWo8nhv
    21DSZs2HqfniNLQ==,
    v=spf1 mx a:smtpl.it.da.ut.ee a:smtpl2.it.da.ut.ee a:bounces.ut.ee include:spf.protection.outlook.com
    include:_spf.ut.ee ~all",
  "statusCode": 200,
  "curlError": "",
  "redirect": "https://www.ut.ee/et",
  "cms": "Drupal",
  "wappalyzer": [
    {"app": "AddThis", "version": "", "id": "addthis", "categories": ["widgets"]},
    {"app": "Apache", "version": "", "id": "apache", "categories": ["web-servers"]},
    {"app": "Drupal", "version": "7", "id": "drupal", "categories": ["cms"]},
    {"app": "Google Tag Manager", "version": "", "id": "google-tag-manager", "categories": ["tag-managers"]},
    {"app": "Microsoft Word", "version": "", "id": "microsoft-word", "categories": ["editors"]},
    {"app": "Google Search Console", "version": "", "id": "google-search-console", "categories": ["analytics", "seo", "search-engines"]}
  ],
  "scripts": [
    "https://www.ut.ee/sites/default/files/js/js_SiZsS87V9qdUK_IQE9fVrbDcFH9JsTQZwiuUfWz6K8.js",
    "https://www.ut.ee/sites/default/files/js/js_zS-CmNFGyegtLYJmqFRpxQvvQrfPIFrOMq_3T3C8sZE.js",
    "https://www.ut.ee/sites/default/files/js/js_i_DnzZWC_61KwXjuSvXoB_Hh4jooAL1TSKTQrf1NV10.js",
    "https://www.ut.ee/sites/default/files/js/js_8_HHt596mYI2fsL_7rwEuH1hZg5r-K9MmQNshJ2sY_Y.js",
    "https://www.ut.ee/sites/default/files/js/js_h_9i15rXV8qZi6VEwzmntjeu-FLRiFEW_FLsiD3o5mA.js",
    "https://www.gstatic.com/charts/loader.js",
  ]
}
```

```
"https://www.ut.ee/sites/default/files/js/js_Dl5wcWkxwOWeTJqTnLAHqs5l2gpA3R9kQqZSiNfzaeU.js",
```

```
"https://www.ut.ee/sites/default/files/js/js_uk2KpWolI9m8gcqN9iZFOF11n_sp_gnx_aP9g0loBZo.js",
```

```
"https://www.ut.ee/sites/default/files/js/js_0ywzyM9DHmFK_7BG8hgPTQlwy-SY8IF4Ig_BJ8DWTv4.js",
```

```
"https://www.ut.ee/sites/default/files/js/js_roUHtfMTPRM93ImZCBTfXQUr_BGPmPOXJzhAZT8bkFw.js",
```

```
"//s7.addthis.com/js/300/addthis_widget.js#pubid=xa-528b62e9687dab44",
```

```
"https://www.ut.ee/sites/default/files/js/js_5iYHp2_iuUR3CLcwBvSrQNcjCRUUQb-fVNtsP9adNSs.js"
```

```
]
```

```
}
```

III. Lisa 3: Domeenide arvult suurimad 100 majutusorganisatsiooni

Tabelis 4 on toodud suurimad majutusorganisatsioonid domeenide arvu järgi, koos nende AS (autonoomsüsteemi) numbritega.

Tabel 4. Domeenide arvult suurimad 100 majutusorganisatsiooni.

Nr.	Organisatsioon	Domeenide arv	Osakaal
1	Zone (Zone Media OU AS49604)	46615	37.5%
2	Elkdata (Elkdata OU AS61189)	18933	15.2%
3	Radicenter (CITIC Telecom CPC Netherlands B.V. AS3327)	5641	4.5%
4	AMAZON-02 AS16509	4245	3.4%
5	Telia Eesti AS AS3249	3700	3%
6	Voog (Zone Media OU AS49604)	3163	2.5%
7	neti.ee (Telia Eesti AS AS3249)	2677	2.2%
8	CLOUDFLARENET AS13335	2286	1.8%
9	Virtuaal.com OU AS205930	2274	1.8%
10	Wavecom (Aktsiaselts WaveCom AS34702)	2241	1.8%
11	Wix.com Ltd. AS58182	1424	1.1%
12	Hetzner Online GmbH AS24940	1409	1.1%
13	UAB Rakrejus AS62282	1406	1.1%
14	DIGITALOCEAN-ASN AS14061	989	0.8%
15	GOOGLE AS15169	938	0.8%
16	UpCloud Ltd AS202053	932	0.8%
17	Compic OU AS39823	874	0.7%
18	OVH SAS AS16276	794	0.6%
19	CSC AS19574	726	0.6%
20	unknown	649	0.5%
21	Elisa Eesti AS AS2586	608	0.5%
22	Shoproller (Zone Media OU AS49604)	551	0.4%
23	Avenir Telematique SAS AS24935	519	0.4%
24	Kernel AS AS39038	486	0.4%
25	Information System Authority AS8240	442	0.4%
26	Intermedia (Transip B.V. AS20857)	441	0.4%
27	CITIC Telecom CPC Netherlands B.V. AS3327	410	0.3%
28	Contabo GmbH AS51167	394	0.3%
29	Beehosting (CITIC Telecom CPC Netherlands B.V. AS3327)	389	0.3%
30	AS INFONET AS8728	372	0.3%
31	Hariduse Infotehnoloogia Sihtasutus AS3221	359	0.3%

32	SQUARESPACE AS53831	358	0.3%
33	Team Internet AG AS61969	354	0.3%
34	UNIFIEDLAYER-AS-1 AS46606	350	0.3%
35	PROLEXIC-TECHNOLOGIES-DDOS-MITIGATION-NETWORK AS32787	350	0.3%
36	Elitec (AS INFONET AS8728)	330	0.3%
37	MICROSOFT-CORP-MSN-AS-BLOCK AS8075	329	0.3%
38	WEEBLY AS27647	321	0.3%
39	INCAPSULA AS19551	300	0.2%
40	Hostinger International Limited AS47583	291	0.2%
41	GANDI SAS AS29169	260	0.2%
42	TELE2 AS1257	258	0.2%
43	DBweb (Telia Eesti AS AS3249)	258	0.2%
44	Linode, LLC AS63949	244	0.2%
45	FASTLY AS54113	238	0.2%
46	OU Web Hosting Solutions AS202759	238	0.2%
47	AMAZON-AES AS14618	231	0.2%
48	NSS (CITIC Telecom CPC Netherlands B.V. AS3327)	230	0.2%
49	P.a.g.m. Ou AS198068	227	0.2%
50	Datacenter Luxembourg S.A. AS24611	222	0.2%
51	Web2 (Zone Media OU AS49604)	222	0.2%
52	DREAMHOST-AS AS26347	216	0.2%
53	LeaseWeb Netherlands B.V. AS60781	187	0.2%
54	AUTOMATTIC AS2635	183	0.1%
55	Server Farm LLC AS202635	167	0.1%
56	NAMECHEAP-NET AS22612	155	0.1%
57	QSC AG AS15598	149	0.1%
58	Variti LLC AS64432	147	0.1%
59	Alibaba (US) Technology Co., Ltd. AS45102	142	0.1%
60	Elkdata (Telia Eesti AS AS3249)	136	0.1%
61	InterNetX GmbH AS15456	130	0.1%
62	Sia Nano IT AS43513	129	0.1%
63	Webzilla B.V. AS35415	128	0.1%
64	Infonet Dc Oy AS205950	121	0.1%
65	Telia Lietuva, AB AS43811	114	0.1%
66	Ekspress Grupp AS AS199328	110	0.1%
67	NETSEC AS45753	109	0.1%
68	AS-CHOOPA AS20473	108	0.1%
69	SIA Digitalas Ekonomikas Attistibas Centrs AS12993	103	0.1%
70	KEI.PL Sp. z o.o. AS29522	101	0.1%

71	Host Europe GmbH AS8972	100	0.1%
72	AS STV AS61307	98	0.1%
73	Host Europe GmbH AS21501	95	0.1%
74	1&1 Ionos Se AS8560	90	0.1%
75	GleSYS AB AS43948	87	0.1%
76	OU Interframe AS196743	86	0.1%
77	AS Postimees Grupp AS207254	86	0.1%
78	Data Invest sp. z o.o. S.K.A AS50599	86	0.1%
79	Nameshield SAS AS20756	86	0.1%
80	Aktsiaselts WaveCom AS34702	86	0.1%
81	Edicy (Zone Media OU AS49604)	85	0.1%
82	Elisa Teleteenused AS AS13272	85	0.1%
83	TimeWeb Ltd. AS9123	84	0.1%
84	Beget LLC AS198610	84	0.1%
85	A2HOSTING AS55293	81	0.1%
86	Ddos-guard Ltd AS57724	77	0.1%
87	Levira AS AS50794	77	0.1%
88	Online S.a.s. AS12876	76	0.1%
89	First Colo GmbH AS44066	76	0.1%
90	POWER LINE DATACENTER AS132839	73	0.1%
91	Uus Programm AS204595	72	0.1%
92	SONICTEST Ltd. AS57873	71	0.1%
93	HVC-AS AS29802	70	0.1%
94	Register S.p.A. AS39729	69	0.1%
95	Astrec Data OU AS201601	68	0.1%
96	Telia Inmics-Nebula Oy AS29422	67	0.1%
97	OU cloud.ee AS61952	65	0.1%
98	Geenet OY AS203081	64	0.1%
99	UAB Cherry Servers AS16125	64	0.1%
100	Domain names registrar REG.RU, Ltd AS197695	63	0.1%

IV. Tuvastatud WordPressi versioonid

Tabelis 5 on toodud teadaolevate WordPressi versioonide jaotus .ee domeenide hulgas. Rohelisel taustal praegune versioon, kollasel taustal versioonid, mis saavad veel uuendusi ja punasel taustal lõpetatud hooldustsükliga versioonid.

Tabel 5. Tuvastatud WordPress versioonid.

versioon	domeenide arv
5.7	11131
5.6	3608
5.5	2622
5.4	2085
5.3	1575
5.2	1270
5.1	612
5.0	447
4.9	1817
4.8	519
4.7	632
4.6	238
4.5	275
4.4	261
4.3	186
4.2	180
4.1	139
4.0	84
4	2
3.9	100
3.8	88
3.7	24
3.6	42
3.5	101
3.4	47
3.3	44
3.2	32
3.1	50
3.0	34
2.9	17
2.8	19
2.6	2
2.2	2
2.1	2

V. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Siim Markus Marvet**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
Eesti domeenide statistika ja turbeinfo kogumine,
mille juhendaja on **Alo Peets**,
reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Siim Markus Marvet

07.05.2021