LATVIJAS UNIVERSITĀTES
DATORIKAS FAKULTĀTE

RASMOTĀJA IZSTRĀDE LATVIJAS ATVĒRTO DATU PORTĀLA ANALĪZEI

BAKALAURA DARBS

Autors: **Justs Ķikuts**
Studenta apliecības Nr.: jk17077
Darba vadītājs: docente, Dr.sc.comp. Anastasija Ņikiforova

RĪGA 2021

ANOTĀCIJA

Rasmošana ir veids kā ātri iegūt nepieciešamo informāciju no tīmekļu vietnēm. Darba mērķis ir izstrādāt rasmotāju Latvijas Atvērto datu portālam.

Darbā tiek apskatīts rasmošanas princips un tās vēsture. Tiek izpētīta rasmotāju darbība un veikts trīs brīvi pieejamu rasmotāju pārskats un analīze. Analīzes rezultāti kalpo par izstrādājamā rasmotāja specifikāciju, kā arī rasmošanas rezultātā iegūtie faili tiek salīdzināti ar izstrādāto rasmotāju. Pirms rasmotāja izstrādes darbā tiek arī apskatīts "atvērto datu" jēdziens un vēsture. Tā kā izstrādātais rasmotājs tiks testēts Latvijas Atvērto datu portālā, tas tiek īsi analizēts.

Darba mērķis ir sasniegts un rasmotājs ir izstrādāts, iegūstot CSV failu ar Latvijas Atvērto datu portālu aprakstošo informāciju. Izstrādātais rasmotājs ir publiski pieejams GitHub repozitorijā.

Atslēgvārdi: atvērtie dati, tīmekļa rasmotājs, CSV fails, Python, datu kopas.

ABSTRACT

WEB SCRAPER DEVELOPMENT FOR LATVIAN OPEN DATA PORTAL ANALYSIS

Web scraping is a way to quickly obtain the necessary information from websites. The goal of the thesis is to develop a web scraper for the Latvian Open Data portal.

The principle of web scraping and its history is looked at in thesis. Operations of web scraper are analysed and an overview of three freely available web scrapers is carried out, the results of which are compared with the developed web scraper. Before the development of web scraper thesis explores the concept and history of the “open data” term. As the developed web scraper will be tested on the Latvian Open Data Portal, it is briefly analysed.

The goal has been reached and the web scraper has been developed - CSV file is obtained with the necessary information. The developed web scraper is publicly available in the GitHub repository.

Keywords: open data, web scraper, CSV file, Python, data sets.

SATURS

APZĪMĒJUMI

API - funkciju un procedūru kopums, kas ļauj izveidot lietojumprogrammas, kuras piekļūst operētājsistēmas, lietojumprogrammas vai cita pakalpojuma funkcijām vai datiem

AS – akciju sabiedrība

CSS - Cascading Style Sheets - stila lapas kaskadēšana ir īpaša stila lapas valoda

CSV - ar komatu atdalītas vērtības, ko parasti izmanto izklājlapām vai vienkāršām datu bāzēm

GET - pieprasa norādīto resursu no servera

ĢIS - ģeogrāfiskās informācijas sistēmas

HTML - hiperteksta iezīmēšanas valoda, ko izmanto tīmekļa lapu attēlošanai pārlūkprogrammās

HTTP - HyperText Transfer Protocol - hiperteksta transporta protokols

JSON - JavaScript Object Notation - JavaScript objektu notācija ir datu apmaiņas formāts

POST - nosūta datus uz serveri

REGEX - regular expression - regulārā izteiksme ir simbolu virkne, kas definē meklējamo izteiksmi

SIA - sabiedrība ar ierobežotu atbildību

SVG - mērogojama vektoru grafika

URL - vienotais resursu vietrādis

UTF-8 – astoņu bitu unikoda pārveidošanas formāts

WebGL – grafiskā bibliotēka JavaScript API 3D grafikas attēlu renderēšanai jebkurā saderīgā tīmekļa pārlūkprogrammā

WMS - uz tīmekļa tehnoloģijām balstīts serviss

XHTML - paplašināmā hiperteksta iezīmju valoda

XLSX - Microsoft Excel Atvērtā XML Izklājlapa

XML - paplašināmā iezīmēšanas valoda

XPATH - vaicājuma valoda mezglu atlasei no XML dokumenta

ZIP - arhīva faila formāts, kas atbalsta bezzuduma datu saspiešanu

7z - arhīva faila formāts, kas atbalsta bezzuduma datu saspiešanu

IEVADS

Atvērtie dati ir daļa no plašas globālas kustības, kas ne tikai veicina zinātni un zinātnisko komunikāciju, bet arī pārveido mūsdienu sabiedrību un to, kā tiek pieņemti lēmumi, jo šobrīd arī veicina ekonomisko un tehnoloģisko attīstību. Mūsdienās strauji pieaug iegūto un apstrādāto datu apjoms. Tas, kas sākās ar "Atvērto zinātni" un tiešsaistes žurnālu skaita pieaugumu, tika paplašināts uz "Atvērtajiem datiem". Tas tiek balstīts uz pieņēmumu, ja ziņojumi vai atskaites par datiem ir atvērti, tad arī iegūtajiem vai pamatojošiem datiem vajadzētu būt atvērtiem. No atvērto datu kopu jēdziena pirmsākumiem, valstu valdības ir izstrādājušas savus atvērto datu portālus, kur jebkurš var piekļūt sev nepieciešamiem vai interesējošiem datiem par brīvu. Arvien vairāk valdības aģentūru, finansējošo organizāciju un izdevēju atbalsta aicinājumu palielināt datu koplietošanu, kur daudziem galvenais mērķis ir atvērtie dati. Atvērtie dati ir viens no vismazāk ierobežojošajiem datu koplietošanas veidiem, atšķirībā no pārvaldītajiem piekļuves datiem, kuriem parasti ir lietošanas noteikumi un dažos gadījumos tiek īstenota pašu datu īpašnieku pārraudzība.

"Pasaules Atvērto Datu Indeksā" Latvija ieņem augsto 14. vietu starp 94 valstīm [1]. Latvijas atvērto datu portāls raiti tiek papildināts ar dažādu veidu datu kopām. Dažu gadu laikā tas ir pieaudzis vairākas reizes. Lai analizētu atvērto datu portālu trešajām ar portālu nesaistītajām pusēm, piemēram, datu entuziastiem, var tikt izmantotas dažādas tehnikas, viena no kurām ir tīmekļa rasmotāja pielietošana.

Tīmekļa rasmošana (web scraping) ir lietotājam nepieciešamo datu atlasīšanas no interneta vietnēm un to turpmākais apstrādes process. Rasmošanas procesā var iegūt datus lietotājam ērtā un vajadzīgā formātā. Darba ietvaros ir izstrādāts rasmotājs - rasmošanas programmatūra un rīks, ar kuru palīdzību tiek iegūti un apkopti dati. Ņemot eksistējošo rasmotāju universālumu, tie dažreiz neļauj nolasīt pietiekoši daudz datu no atvērto datu portāliem, tādējādi neļaujot tos efektīvi izmantot. Ar izstrādātā rasmotāja palīdzību tiks veikta Latvijas Atvērto datu portāla rasmošana. Papildus izstrādātais rasmotājs tiek salīdzināts ar citiem, brīvi pieejamiem rasmotājiem un to iegūtajiem datiem.

Lai gan datu ir daudz un tie ir visiem viegli pieejami, tos ir grūtības apkopot, lai veiktu nepieciešamo analīzi, kuras raksturs tāpat kā atvērto datu atkal izmantošanas nolūks var būt katram individuālam lietotājam cits. Manuāli jeb ar roku apstrādāt pieejamos datus prasa ļoti daudz laika. Problēma ir viegli un ātri atrast un atlasīt nepieciešamos datus no Atvērto datu portāla. Rasmotājs palīdzēs analizēt dažādās datu kopas un pēc nepieciešamības iegūt un apkopot tās ātrāk. Papildus,

ņemot vērā dažādu valsts nacionālo portālu līdzīgo struktūru, izstrādātais rasmotājs būs piemērojams arī citu portālu analīzei. Rasmotāja izstrāde būs arī noderīga autora programmēšanas prasmes pārbaudei un visvairāk pilnveidošanai.

Darba mērķis – izstrādāt rasmotāju Latvijas Atvērto datu portālam.

Darbā izvirzītie uzdevumi:

1. Iepazīties ar atvērto datu jēdzienu un apskatīt atvērto datu vēsturi;
2. Veikt Latvijas Atvērto datu portāla analīzi;
3. Iepazīties ar rasmošanas un rasmotāju jēdzieniem;
4. Aplūkot eksistējošus rasmotājus un veikt Latvijas Atvērto datu portāla rasmošanu;
5. Izstrādāt rasmotāju, kas paredzēts Latvijas Atvērto datu portālam, salīdzināt rezultātus ar brīvi pieejamiem rasmotājiem.

Darbs sastāv no 5 sadaļām, kur 1. sadaļā tiek apskatīts atvērto datu jēdziens, kā arī neliels ieskats atvērto datu vēsturē. 2. sadaļā tiek īsi aprakstīts un analizēts Latvijas Atvērto datu portāls. 3. sadaļā tiek aplūkots rasmošanas un rasmotāja jēdziens, kā arī mazs ieskats to vēsturē. Tiek arī apskatīti pieejamie rasmotāji un tiek mēģināts izgūt datus no Latvijas Atvērto datu portāla. 4. sadaļā tiek veikta un aprakstīta rasmotāja izstrāde. Īsi tiek salīdzināti iegūtie CSV faili un rasmotāju darbība ar pašizstrādāto un trešajā nodaļā apskatītajiem rasmotājiem. 5. sadaļā tiek apkopti darbā iegūtie rezultāti.

1. ATVĒRTIE DATI

Pastāv pieaugoša globāla tendence atvērt datus, lai ikviens varētu brīvi piekļūt tiem, tos atkārtoti izmantot un kopīgot pēc individuālas vēlēšanās. Atvērto datu filozofija jau sen ir izveidojusies, taču pats termins "Atvērtie dati" ir salīdzinoši jauns (1995. gads) [9]. Pieaugot globālajam tīmeklim, atvērtie dati kļūst arvien populārāki. Pēdējo desmit gadu laikā ir veikti vairāki atvērto datu uzlabojumi, kuru pamatā galvenokārt ir valstu valdības, kas ar savu iniciatīvu rosina atvērto datu popularitāti.

1.1. Atvērto datu jēdziens

Atvērto datu definīcija varētu būt apkopota, ka "atvērtie dati" ir dati, kurus var brīvi izmantot, atkārtoti izmantot un izplatīt ikviens bez autortiesību, patentu vai citu kontroles mehānismu ierobežojumiem. Dažos gadījumos tiek prasīts norādīt atsauci uz datu avotiem, ja tas tiek norādīts licencē [2]. Lai gan atvērtie dati ir visiem pieejami, tie tāpat ir aprakstīti ar CC0 licenci, kas ļauj datus brīvi izmantot. CC0 ļauj zinātniekiem, pedagogiem, māksliniekiem un citiem ar autortiesībām vai datu bāzēm aizsargāta satura radītājiem un īpašniekiem atteikties no šīm interesēm savos darbos un tādējādi pēc iespējas pilnīgāk iekļaut tos publiskajā domēnā, lai citi varētu brīvi veidot, uzlabot un atkārtoti izmantot darbus jebkādiem mērķiem bez ierobežojumiem saskaņā ar autortiesību vai datu bāzu tiesību aktiem [3].

The Open Definition (opendefinition.org) precizē jēdziena "atvērts" nozīmi attiecībā uz zināšanām, veicinot stabilu kopumu, kurā var piedalīties ikviens, un ir maksimāla sadarbspēja. "Atvērts" Atvērtajos datos būtiskā nozīmē atbilst Atvērtai programmatūrai (Open Source) un tās definīcijai, kas ir sinonīms ar "bezmaksas".

Atbilstoši [4], atvērtajiem datiem ir jāatbilst šādām izplatīšanas prasībām:

1) atvērtā licence vai statuss: datiem jābūt publiskiem vai tiem jābūt nodrošinātiem ar atvērtu licenci. Visi papildu noteikumi, kas pievienoti datiem, piemēram, lietošanas noteikumi vai licences devēja patenti, nedrīkst būt pretrunā ar darba publiskā domēna statusu vai licences noteikumiem;
2) piekļuve: dati jānodrošina kopumā un tiem ir jābūt bez maksas pieejamiem vai nepārsniedzot saprātīgas vienreizējas reproducēšanas izmaksas. Tiem jābūt

lejupielādējamiem. Visa papildus informācija licences ievērošanai, piemēram, līdzdalībnieku vārdi, kas nepieciešami, lai atsauktos uz darbu, jābūt pievienotai datiem;

3) mašīnlasāmība: datiem ir jābūt formā, kuru viegli un saprotami var apstrādāt dators un kurā var ērti piekļūt atsevišķiem datu elementiem, lai tos varētu viegli un bez problēmām modificēt;
4) atvērts formāts: datiem jābūt atvērtā formātā. Atvērtais formāts paredz, ka tā lietošanai nenosaka nekādus naudas vai jebkādus citus ierobežojumus, un to var pilnībā apstrādāt, izmantojot vismaz vienu brīvu atvērtā pirmkoda programmatūras rīku.

Atbilstoši [4], licencei jābūt saderīgai ar citām atvērtajām licencēm. Licence ir atvērta, ja tās noteikumi atbilst šādiem nosacījumiem:

1) nepieciešamās atļaujas (required permissions): licence neatsaucami atļauj sekojošo:
   1.1) izmantošana (use): licencei jāļauj brīvi izmantot licencētos datus;
   1.2) pārdalīšana (redistribution): licencei jāatļauj licencēto datu pārdali, tostarp pārdošanu, neatkarīgi no tā, vai tas notiek atsevišķi vai kā daļa no dažādu datu avotu kolekcijas;
   1.3) pārveidošana (modification): licencei jāļauj izveidot licencētā darba atvasinājumus un jāļauj izplatīt šādus atvasinātos darbus saskaņā ar tiem pašiem noteikumiem, kas paredzēti oriģinālajam licencētajam darbam;
   1.4) atdalīšana (separation): licencei jāatļauj jebkuras darba daļas brīvu izmantošanu, izplatīšanu vai pārveidošanu atsevišķi no jebkuras citas darba daļas vai no jebkuras darba kolekcijas, kurā tā sākotnēji tika izplatīta. Visām personām, kas saņem jebkādu kādas darba daļas sadali saskaņā ar sākotnējās licences noteikumiem, ir jābūt tādām pašām tiesībām kā tām, kas piešķirtas saistībā ar oriģināldarbiem;
   1.5) kompilācija (compilation): licencei jāļauj izplatīt licencētos datus kopā ar citiem atšķirīgiem datiem, neparedzot ierobežojumus šiem citiem datiem;
   1.6) nediskriminācija (non-discrimination): licence nedrīkst diskriminēt nevienu personu vai grupu;
   1.7) izplatīšana (propagation): saistīto datu tiesībām ir jāattiecas uz visiem, kam tas tiek pārdalīts, bez nepieciešamības piekrist kādiem papildu juridiskiem noteikumiem;

1.8) piemērošana jebkuram mērķim (application to any purpose): licencei jāļauj izmantot, pārdalīt, pārveidot un apkopot datus jebkuram mērķim. Licence nedrīkst liegt nevienam izmantot datus noteiktā darba jomā;

1.9) bez maksas (no charge): licence nedrīkst uzlikt nekādu maksu, honorāru vai citu kompensāciju, vai naudas atlīdzību kā daļu no saviem nosacījumiem;

2) Pieņemamie nosacījumi (acceptable conditions): Licence nedrīkst ierobežot, padarīt neskaidru vai citādi samazināt iepriekš noteiktās atļaujas, izņemot šādus pieļaujamus nosacījumus:

2.1) Attiecinājums (attribution): licencē var pieprasīt, lai darba sadalē tiktu iekļauti ieguldītāji, tiesību īpašnieki, sponsori un autori, ja vien šādas darbības nav apgrūtinošas;

2.2) Integritāte (integrity): licencē var pieprasīt, lai licencēta darba modificētajām versijām būtu cits nosaukums vai versijas numurs atšķirībā no sākotnējā darba, vai citādi norādīt, kādas izmaiņas ir veiktas;

2.3) share-alike: licencē var pieprasīt, lai darba sadale paliktu saskaņā ar to pašu licenci vai līdzīgu licenci;

2.4) paziņojums (notice): licencē var pieprasīt saglabāt autortiesību paziņojumus un licences identificēšanu;

2.5) avots (source): licencē var pieprasīt, lai ikviens, kas izplata darbu, nodrošina adresātiem piekļuvi vēlamajai formātam, lai veiktu izmaiņas;

2.6) tehnisko ierobežojumu aizliegums (tehnical restriction prohibition): licencē var pieprasīt, lai darba sadalē paliktu bez jebkādiem tehniskiem pasākumiem, kas ierobežotu citādi atļauto tiesību izmantošanu;

2.7) neuzbrukšanas princips (non-aggression): licencē var pieprasīt, lai pārveidotāji piešķirtu sabiedrībai papildu atļaujas (piemēram, patentu licences), kas nepieciešamas, lai izmantotu licences atļautās tiesības. Licence var arī noteikt neuzbrukšanas atļaujas, kas ir vērstas pret licencētajām personām attiecībā uz visu atļauto tiesību izmantošanu, piemēram, patentu tiesvedību.

Jebkuriem atvērtiem datiem ir jāievēro iepriekš minētie "Open Definition 2.1" nolikumi. Apkopojot iepriekš aplūkoto, datiem ir jābūt nodrošinātai:

- pieejamībai un piekļuvei: datiem jābūt pieejamiem pilnībā un par brīvu, un viegli lejupielādējamiem internetā. Datiem jābūt pieejamiem arī ērtā un modificējamā formā;

- atkārtotai izmantošanai un izplatīšanai: dati jāsniedz saskaņā ar noteikumiem, kas ļauj tos atkārtoti izmantot, ieskaitot sajaukšanos ar citām datu kopām;
- vispārējai dalībai: ikvienam ir jāspēj izmantot, atkārtoti izmantot un izplatīt - nedrīkst būt diskriminācija attiecībā uz centieniem vai personām, vai grupām. Piemēram, nav pieļaujami "nekomerciāli" ierobežojumi, kas liegtu "komerciālu" lietošanu, vai izmantošanas ierobežojumi noteiktiem mērķiem, piemēram, tikai izglītībai.

Ir jābūt skaidram, ko nozīmē vārds "atvērts" Atvērtajos datos, jo tas tieši ietekmē sadarbspēju. Sadarbspēja apzīmē dažādu sistēmu un organizāciju spēju strādāt kopā (savstarpējā sadarbība). Šajā gadījumā tā ir spēja savstarpēji izmantot vai savstarpēji sasaistīt dažādas datu kopas. Sadarbspēja ir svarīga, jo tā ļauj sadarboties dažādiem komponentiem. Spēja izveidot sastāvdaļas un "saslēgt kopā" sastāvdaļas ir būtiska lielu un sarežģītu sistēmu veidošanai. Bez sadarbspējas tas kļūst gandrīz neiespējami - par to liecina slavenākais mīts par Bābeles torni, kurā nespēja sazināties (savstarpēji sadarboties) izraisīja torņa celtniecības centienu pilnīgu sabrukumu [5].

Līdzīga situācija ir vērojama attiecībā uz datiem. Datu kodols ir tāds, ka vienu tajā iekļauto "atvērto" materiālu var brīvi sajaukt ar citu "atvērto" materiālu. Aprakstītā sadarbspēja ir absolūti būtiska, lai realizētu galvenos "atvērtības" praktiskos ieguvumus, kas krasi uzlabotu spēju apvienot dažādas datu kopas un tādējādi attīstīt vairāk un labāk dažādus produktus un pakalpojumus. Skaidras atklātības definīcija nodrošina to, ka, iegūstot divas atvērtas datu kopas no diviem dažādiem avotiem, būs iespējams tās apvienot kopā. Tā var izvairīties no problēmas, ka ir daudz datu kopu, bet maz vai vispār nav iespējams apvienot tās lielākās sistēmās, kurās ir īstā vērtība.

1.2. Atvērto datu metadati

Viens no svarīgākajiem atvērto datu aspektiem ir to metadati. Metadati ir strukturēta informācija, kas apraksta, izskaidro, atrod vai kā citādi atvieglo informācijas resursa izguvi, izmantošanu vai pārvaldību [6]. Metadati ir ļoti svarīgi, lai palīdzētu apmeklētājiem efektīvi atrast un izmantot publicētos datus. Labi metadati samazina vajadzību apmeklētājiem meklēt personisko palīdzību, palīdz novērst nepareizu datu interpretāciju un veicina augstāku datu kvalitāti [7]. Metadatus bieži dēvē kā datus par datiem. Bez datu kopas metadatiem publicēto datu katalogs nevarētu pastāvēt. Daudzi atvērto datu portāli ietver nepieciešamos rīkus, lai izveidotu datu kopas metadatus, publicējot jaunus datus. Daži atvērto datu portāli, rediģējot datu kopas, automātiski

atjaunina metadatus. Metadati norāda uz atvērto datu atrodamību, jo tas ir pirmais ko lietotājs redz. Ja apkopotajiem datiem nav labi uzrakstīti metadati, tad lietotāji nevarēs tos tik labi atrast vai padomās, ka tie nav pietiekami labi.

Oficiālajā Eiropas datu portālā (https://data.europa.eu/) var apskatīt katra atvērto datu portāla Eiropā meta-datu kvalitāti. "Metadatu Kvalitātes Nodrošināšana" (The Metadata Quality Assurance) ir paredzēta, lai palīdzētu datu sniedzējiem un datu portāliem pārbaudīt to metadatus salīdzinājumā ar dažādiem rādītājiem. Atvērto datu portāli tiek vērtēti piecās kategorijās: atrodamība (findability), piekļūstamība (accessibility), savietojamība (interoperability), atkārtota izmantojamība (reusability) un kontekstualitāte (contextuality) [8]. 1.1. attēlā ir redzams Latvijas Atvērto datu portāla (https://data.gov.lv/) novērtējums, kas beigās tiek novērtēts kā pietiekams.

Country	Name	Findability	Accessibility	Interoperability	Reusability	Contextuality	Rating
	data.gov.lv (LVA)	33 / 100	86 / 100	59 / 110	30 / 75	5 / 20	Sufficient

1.1. att. **Latvijas Atvērto Datu portāla metadatu kvalitātes novērtējums [8]**

Atbilstoši [7], metadati parasti ir iedalīti divos tipos:

- metadati, kas sniedz datu pārskatu. Šāda veida metadati palīdz lietotājiem atrast datus, veicot meklēšanu internetā, pārvietojoties datu portālos, kuros varētu būt iekļauts konkrēts katalogs;
- metadati, kas sniedz detalizētu informāciju par noteiktām konkrētu datu kopas daļām. Šāda veida metadati ļauj lietotājiem efektīvi izmantot datus, palīdzot izprast dažādos tajā ietvertos elementus un iespējamos ierobežojumus.

Tā kā atvērto datu portālos ir vairāku desmitu pat simtu datu kopas, apmeklētājiem ir noderīgāk un ātrāk, ja viņi var pārlūkot datus pēc kategorijas vai tēmas. Kategorijas ir īsas, ne vairāk kā divi vai trīs vārdi, un tās ļauj grupēt saistītos datus. Kategorijas arī dod iespēju apmeklētājiem iepazīt pieejamos datus, lai smeltos iedvesmu turpmākajiem darbiem, nevis pieprasa izmantot meklēšanas rīku, lai atrastu kaut ko konkrētu.

Kategoriju izveide bieži vien ir būtisks solis, ieviešot atvērtu datu portālu. Kategorijām nav jābūt pastāvīgām. Ir saprātīgi, ka ir trīs līdz četras kategorijas nelielam skaitam datu kopu, un tās katru gadu ir atkārtoti jāizvērtē, kad tiek publicēti papildus dati. Vislielākajos atvērto datu portālos ir vairāk par 12 kategorijām. Ja ir pārāk daudz kategoriju, tas var nozīmēt, ka tās nav pietiekami plašas. Ja to ir pārāk maz, jo īpaši lielos atvērto datu portālos, kuros ir daudz dažādu datu kopu, tas nozīmē, ka kategorijas ir pārāk plašas un mazāk noderīgas apmeklētājiem [9]. Lai gan starp atvērtajiem datu portāliem nav konsekventu kategoriju kopas, šādas kategorijas ir diezgan izplatītas

un var kalpot par sākuma punktu: uzņēmējdarbība, izglītība, vide, finanses, veselības aprūpe, cilvēkresursi (vai sociālie pakalpojumi), īpašums, sabiedriskā drošība, atpūta un transports.

Galvenie metadatu elementi nodrošina vissvarīgāko informāciju, lai palīdzētu apmeklētājiem atrast datus un noteikt, vai tie ir nepieciešami. Daudzi no šiem vienumiem tiks parādīti tieši katalogu navigācijas lapās vai meklēšanas rezultātos [10]:

1) nosaukums: cilvēkiem lasāms nosaukums par datiem. Tam jābūt vienkāršā valodā un pietiekami detalizētam, lai atvieglotu meklēšanu un atklāšanu. Jāizvairās no saīsinājumiem, kas pēc autoru domām ir saistīts ar potenciālo labāko datu kopu atrodamību;
2) apraksts: cilvēkiem lasāms apraksts ar pietiekami detalizētu informāciju datu kopas saprotamībai un potenciālai atkalizmantojamībai. Nosaka datu kopas piemērotību lietotāja vajadzībām vai interesēm;
3) kategorija: datu kopas galvenā tematiskā kategorija, kuru parasti izvēlas no iepriekš definēta saraksta. Daži atvērto datu portāli ierobežo datu kopu līdz vienai kategorijai, citi atļauj vairākus. Palīdz ar datu kopas atrodamību, izmantojot datu portālu filtrus;
4) atslēgvārdi jeb birkas: parasti tie ir atsevišķi vārdi, kas apmeklētājiem palīdz atrast datus. Jābūt iekļautiem terminiem, kurus izmantotu tehniskie un netehniskie lietotāji. Atslēgvārdus var izmantot arī meklētāji, lai palīdzētu apmeklētājiem atrast nepieciešamās datu kopas;
5) modifikācijas datums: pēdējais mainīšanas, atjaunināšanas vai modificēšanas datums;
6) kontaktinformācija: datu kopas izdevēja vārds un uzvārds / nosaukums un e-pasta adrese;
7) licence: bieži datu kopas atvērto datu portālos ir pieejamas publiski, bez atkārtotas izmantošanas ierobežojumiem. To parasti norāda vietnes pakalpojumu sniegšanas noteikumos vai datu politikā, tomēr var būt apstākļi, kad konkrēta datu kopa tiek piedāvāta, izmantojot citu licenci.

Uzlabotie metadatu elementi sniedz noderīgu informāciju, kas ļauj trešās puses programmatūrai izmantot gan datu katalogus, gan datu kopas. Šie vienumi, iespējams, netiek parādīti katalogu navigācijas lapās vai meklēšanas rezultātos, bet ļauj tos koplietot ar citiem atvērtajiem datu portāliem un meklētājprogrammām [10]:

1) biežums: datu kopas atjaunināšanas biežums vienkāršā angļu valodā. Piemēram, “Nekad, “Katru stundu”, “Katru dienu”, “Katru darba dienu”, “Katru nedēļu”, “Katru

pusmēnesi", "Katru mēnesi", "Reizi ceturksnī", "Reizi pusgadā", "Reizi gadā" utt. Tas palīdz apmeklētājiem uzzināt, cik bieži viņiem jāpārbauda jauni dati, un tas ir īpaši noderīgi programmatūras programmētājiem, kuri var iestatīt automātiskās lejupielādes;

2) laika pārklājums: laika diapazons, kas iekļauts šajā datu kopā. Tas var atspoguļot vispārīgu diapazonu visiem ierakstiem vai var atspoguļot agrākos un jaunākos datu ierakstu datumus;
3) telpiskais pārklājums: ģeogrāfiskais apgabals, kuram noteiktā datu kopa ir būtiska. Visbiežāk tiek izmantots vietas nosaukums, īpaši tas, kas saistīts ar skaidrām robežām. Ja datu kopā iekļauta ģeotelpiskā informācija, telpiskais pārklājums var attēlot visu tajā ietilpstošo ģeogrāfiju ierobežojošo taisnstūri vai daudzstūri, lai gan tas ir retums.

1.3. Atvērto datu vēsture

Atvērtība un dalīšanās ar atklājumiem ir bijusi zinātnes pamatā, kopš zinātnisko metodi pirmo reizi aprakstīja Aristotelis [11]. Tomēr vēsturiski ne zinātniskie ziņojumi, ne dati, kas bija šo ziņojumu pamatā, nav bijuši viegli pieejami. Zinātniskie pētījumi tika publicēti žurnālos, kuros piekļuve prasīja apmaksātus abonementus vai bija ieguvums no maksas dalības asociācijā, un datubāzes tika uzskatītas par to personu privāto un intelektuālo īpašumu, kas tās izveidoja. Datu bāzes tika un bieži vien joprojām tiek veidotas un glabātas dažādos veidos, analizētas ar dažādām metodēm un tādējādi var tikt dziļi apslāpētas.

1970. gados Roberts Kings Mertons, kurš tiek uzskatīts par zinātnes socioloģijas dibinātāju, sāka virzīt uz priekšu ideju, ka pētniecībai jābūt brīvi pieejamai visiem. Viņš apgalvoja, ka viena "Mertonijas norma" mūsdienu zinātnes ētikā ir tāda, ka katram pētniekam ir jādod ieguldījums "kopējā podā" un jāatsakās no intelektuālā īpašuma tiesībām, lai zināšanas varētu virzīties uz priekšu [11].

Atvērtās zinātnes kustība 1990. gados veicināja tiešsaistes žurnālu pieaugumu, atspoguļojot zinātnes sākotnējo nolūku atbalstīt pārredzamību un sadarbību pētniecībā un zinātniskajā komunikācijā. Atvērtās zinātnes kustību veicināja novērojums, ka pētniecība bieži tiek apmaksāta ar valsts līdzekļiem, un līdz ar to nodokļu maksātājiem nevajadzētu ierobežot piekļuvi ar atsevišķu samaksu. Tas radīja plašu atbalstu un pieprasījumu pēc atklātas piekļuves zinātniskām publikācijām un pašreizējo tendenci autoriem un žurnāliem pieņemt "Creative Commons" licenci,

kas ļauj cilvēkiem brīvi lasīt un lietot zinātniskas publikācijas ar atbilstošu pielietojumu. Joprojām notiek aprakstītā pāreja ar atvērtās piekļuves žurnāliem un abonēšanas žurnāliem [11].

Termins atvērtie dati pirmo reizi parādījās 1995. gadā Amerikas zinātniskās aģentūras dokumentā. Tajā tika apskatīta ģeofizisko un vides datu atklāšana. Citējot ziņojuma autorus: "Mūsu atmosfēra, okeāni un biosfēra veido integrētu veselumu, kas pārsniedz robežas. Tie veicina pilnīgu un atklātu zinātniskās informācijas apmaiņu starp dažādām valstīm, kas ir priekšnoteikums apskatīto globālo parādību analīzei un izpratnei" [12].

Atvērtās zinātnes kustības atbalstītāji ir spēruši vēl soli tālāk, veicinot plašāku piekļuvi ģenerētiem vai apkopotiem datiem. Atvērto datu pamatā ir doma, ka ne tikai pētījumu rezultātiem un ziņojumiem jābūt atvērtiem, bet arī pamatdatiem, kas tos informē un atbalsta. Nobela prēmijas laureāts Elinors Orstrom konstatējis, ka Atvērtie dati ir jauna veida "publiskie labumi". Domāšana bija tāda, ka atšķirībā no citiem sabiedriskā labuma veidiem, atvērto datu izmantošana nevis noārda kopējo krājumu, bet potenciāli bagātina to [12].

2007. gada decembrī trīsdesmit domātāji un interneta aktīvisti sasauca sanāksmi Sebastopolē, Kalifornijas štatā, ziemeļos no Sanfrancisko. Viņu mērķis bija definēt atvērto publisko datu koncepciju un panākt, lai to pieņem ASV prezidenta kandidāti. Viņu vidū bija divi labi zināmi cilvēki: Tims O'Reilijs un Lorenss Lessigs. Pirmais ir pazīstams ar tehnoloģijām: amerikāņu autors un redaktors ir daudzu avangarda datoru un interneta kustību aizsācējs. Viņš definēja un popularizēja izteicienus, piemēram, atvērto pirmkodu un Web 2.0. Stenfordas universitātes (Kalifornijā) tiesību profesors Lorenss Lessigs ir "Creative Commons" licenču dibinātājs, kas balstīts uz ideju par "copyleft" un bezmaksas zināšanu izplatīšanu. Sebastopoles sanāksmes dalībnieki lielākoties nāk no brīvās programmatūras un kultūras kustībām [12].

Aktīvistu kustības ir daudzu jauninājumu centrā datoru un interneta jomā pēdējos divdesmit gados. Daži no šiem atklājumiem tagad ir pazīstami, piemēram, enciklopēdija Wikipedia. Citi atvērtā pirmkoda veidojumi plašākai sabiedrībai ir mazāk zināmi, lai gan tiem ir būtiska nozīme tiešsaistes pakalpojumos: piemēram, serveru programmatūra Apache tiek izmantota, lai mitinātu lielāko daļu tīmekļa vietņu [12].

Astoņi vienkāršie principi, proti, ka datiem jābūt pilnīgiem, primāriem, savlaicīgiem, pieejamiem, mašīnapstrādājamiem, nediskriminējošiem, nepatentētiem un bez ierobežojošas licences joprojām kalpo par pamatu tam, kas kļuvis par uzkrītošu atklāto datu kustību. 2007. gadā tas izklausījās kā sapnis, bet rezultāts ir pārsniedzis aktīvistu cerības [13].

Septiņu gadu laikā kopš šo principu īstenošanas valstu valdības visā pasaulē ir pieņēmušas atvērto datu iniciatīvas un uzsākušas platformas, kas ļauj pētniekiem, žurnālistiem un uzņēmējiem izmantot noteikto jauno izejvielu un tās iespējas atklāt jaunus atklājumus un iespējas. Atvērtie dati visā pasaulē ir piesaistījuši pilsonisko hakeru entuziastus, veicinot hakatonus, lietotņu konkursus un citus izaicinājumus, kas ir koncentrēti uz tādiem dažādiem jautājumiem, kā veselība, enerģētika, finanses, transports un pašvaldību inovācijas.

Nedaudz vairāk kā pēc gada prezidents Baraks Obama stājās amatā Baltajā namā un parakstīja trīs prezidenta memorandus. Divi no tiem attiecas uz atvērtu valdību, kuras atvērtie dati ir viens no punktiem. Prezidenta piezīmes skaidri nosaka atvērtā pirmkoda kultūru sabiedrības darbības centrā, nodefinējot tās pamatprincipus: pārredzamību, līdzdalību un sadarbību [13].

Kā piemērs iepriekš minēto principu izpildīšanā ir ASV atvērto datu portāls data.gov, kurš tika uzsākts 2009. gadā, un to pārvalda un uztur ASV Vispārējo Pakalpojumu Administrācija, Tehnoloģiju Pārveidošanas dienests (General Services Administration, Technology Transformation Service). Eiropas Savienības atvērto datu portāls tika izveidots 2012. gadā pēc Eiropas Komisijas Lēmuma 2011/833 / ES par Komisijas dokumentu atkārtotu izmantošanu. Latvijas Atvērto datu portāla izstrāde sākta 2014. gadā Eiropas reģionālās attīstības fonda līdzfinansētā projekta Nr. 2.2.1.1/16/I/001 "Publiskās pārvaldes informācijas un komunikācijas tehnoloģiju arhitektūras pārvaldības sistēma" (PIKTAPS) ietvaros [14].

Tāpat kā plašāk izmantojot “atvērtās zinātnes” darbības, spēja ražot un dalīt milzīgus datu apjomus drīz vien tika automatizēta, pateicoties milzīgiem sasniegumiem tehnoloģiju un skaitļošanas jomā. Tagad ir tāds laikmets, kad ik dienas ģenerēto datu apjoms ir pārsteidzošs. Bez šaubām, pieprasījums pēc datu glabāšanas jaudām turpina pieaugt, jo notiek jaunu un sarežģītāku datu ģenerētāju attīstība. Datu masa kļūst arvien pieejamāka, izmantojot digitālās platformas, bezvadu sensorus, virtuālās realitātes lietojumprogrammas un miljardiem mobilo tālruņu. Virzība uz atvērtiem datiem ir globāla parādība, kas atbalsta iespējas un inovācijas tendences datu analīzē, kas ietver “lielos datus”, mākslīgo intelektu un datorizglītību. Arvien biežāk tiek aicināts “atvērt datus pēc noklusējuma”, un valdības arvien biežāk savās tīmekļa vietnēs iekļauj atklātos datus. Vēlme, pieprasījums un gaidas pēc atvērtajiem datiem kļūst par jaunu normu. Latvija nav izņēmums šajā kustība un tāpēc nākamajā nodaļā tiek aplūkots Latvijas atvērto datu portāls.

2. LATVIJAS ATVĒRTO DATU PORTĀLS

Latvijas Atvērto datu portālā pieejamo atvērto datu kopu skaits lēnām tiek papildināts ar jaunām, modificētām vai atjaunotām vecām datu kopām, lai tās kļūtu aktuālākas, un reizēm kvalitatīvākas. Kopumā šobrīd (16.05.2021.) ir pieejamas 481 datu kopas no 86 dažādiem publicētājiem (organizācijām). Šie skaitļi pietiekami strauji pieaug, skatoties uz Latvijas lieluma fona. Latvijas Atvērto datu kopu portāla https://data.gov.lv/ izstrāde sākās 2014. gadā, oficiāli portālu palaižot 2017. gadā. Kā liela daļa citu atvērto datu portālu, tas izmanto atvērtā koda tehnoloģijas kā CKAN atvērto datu kataloga platformu. Šo platformu izmanto arī Eiropas un ASV Atvērto datu portāli. Latvijas Atvērto datu portāls ir projekts, kas ietilpst Eiropas reģionālās attīstības fonda (ERAF) sastāvā.

Projekta Nr. 2.2.1.1/16/I/001 "Publiskās pārvaldes informācijas un komunikācijas tehnoloģiju arhitektūras pārvaldības sistēma" (PIKTAPS) mērķis ir nodrošināt Eiropas reģionālās attīstības fonda (ERAF) 2014.-2020. līdzfinansēto informācijas un komunikācijas tehnoloģiju (IKT) projektu savstarpējo saskaņotību, būtiskāko centralizēto platformu projektēšanu un īstenošanu, kā arī veicināt sabiedrības spējas un ieinteresētību efektīvi izmantot radītos IKT risinājumus. Projekta kopējais plānotais finansējuma apjoms ir 4 500 000 euro, no kā 3 825 000 euro ir ERAF finansējums [14].

2.1. Latvijas Atvērto datu portāla datu kopu sadalījums pa formātiem

Ir būtiski, lai dati būtu viegli pieejami sabiedrībai un jebkurš varētu tiem piekļūt un apskatīt tos sev ērtā un pieejamā formātā. Viens no šādiem formātiem ir CSV (ar komatiem atdalītas vērtības). Analizējot datu kopas, vairākām bija pievienots vairāk kā viens datu kopu formāts, lai lietotājiem būtu lielāka iespēja izmantot vai apvienot datu kopas informācijas uzlabošanai. Pirmajā pielikumā ir redzama 2.1. tabula, kura apkopo 16 visbiežāk izmantotos un pieejamos lejupielādei datu kopu formātus Latvijas atvērto datu portālā [15].

Atbilstoši pirmā pielikuma 2.1. tabulai, visbiežāk izmantotais faila formāts, lai apkopotu datus, ir CSV failu formāts (234 datu kopas). CSV failu formātā dati tiek atdalīti ar komatu vai semikolu, kas ļoti labi der jebkādu ģenerētu vai iegūtu datu apkopošanai. CSV formātu var atvērt, apskatīt un apstrādāt ļoti daudz dažādās programmās. Sākot ar parastu teksta redaktoru, kā Notepad, un beidzot ar Microsoft Excel, kas piedāvā pārveidot CSV failu cilvēkam vieglāk lasāmā

XLSX faila formātā. CSV faila formāts arī ir ļoti labs mašīnlasāmībai, to var viegli ģenerēt un parsēt jeb sintaktiski analizēt tālāk.

Attiecībā uz XLSX faila formātu, tas ir otrs visvairāk izmantotais faila formāts (133 datu kopas). XLSX ir Microsoft izveidots Microsoft Excel atvērtais XML izklājlapas failu formāts, kurš arī ir viens no populārākajiem datu kopu formātiem, lai apskatītu un apstrādātu datus. Trešo vietu arī ieņem Microsoft Excel faila formāts XLS (36 datu kopas), kas ir XLSX formāta paveids.

Daļa datu kopu bija augšupielādētas kā ZIP vai 7z arhivēšanas failu formāti, kas samazina datu kopas izmēru, taču tas palēnina piekļuvi atvērto datu kopai, prasot to iepriekšējo izvilkšanu no arhīva. Datu kopu arhīva failos iekšā ir vairāki CSV, XLSX, JSON vai cita formāta faili. Tāpēc gala skaits kādam noteiktam failu formātam būs lielāks, kā redzams pirmā pielikuma 2.1. tabulā. Piemēram, datu kopā "Maršrutu saraksti Rīgas Satiksme sabiedriskajam transportam" failā "MarsrutuSaraksti05_2021.zip" ir 8 teksta faili TXT, kas ir izveidoti CSV faila formāta veidā (ar komatu tiek atdalītas vērtības). Dažām datu kopām bija nepareizi norādīti faila formātu nosaukumi, tāpēc tie netiek pieskaitīti pie kopējā formāta datu kopu skaita. Piemēram, datu kopa (1) "Latvijas augstskolu 2017.gada absolventi gadu pēc absolvēšanas (2018.gadā)" kā formātus norādīja "csv, xls" un "word", kas nepieskaitās pie neviena standart izveidota formāta nosaukuma, (2) datu kopai "Prokuratūras darbs dažādās atbildības jomās" faila formāts norādīts kā "json,xml", kas arī netika pieskaitīts standart izveidota formāta nosaukumiem.

2.2. Latvijas Atvērto datu portāla datu kopu sadalījums pēc publicētājiem

Latvijas Atvērto datu portālā kopā ir 86 dažādi datu kopu publicētāji. Darbā tiek apskatīti 15 publicētāji ar visvairāk publicētajām datu kopām. To var redzēt otrā pielikuma 2.2. tabulā [16].

Visvairāk datu kopu publicējis ir ĢEOLatvija.lv (42 datu kopas). Vietnē ir pieejamas dažādas datu kopas no Latvijas vienotā ģeoportāla. Tajā ir apvienota dažāda informācija par Latvijas ģeotelpisko informāciju un pakalpojumiem no dažādiem informācijas datu turētājiem. Bieži vien šajās datu kopās ir ievietota saite uz citu lapu, kurā ir arī apkopti dati, vai ievietots WMS fails. Lai asociētu Latvijas Atvērto datu portālā ievietotās datu kopas, ĢEOLatvija.lv izmanto sevis izveidotās birkas, kā "jūra", "karte", "augsne", "ūdens" u.c. Latvijas Atvērto datu portāls atdala birkas no kategorijām un vairākums organizāciju, publicējot datu kopas, izmanto abus. ĢEOLatvija izmanto tikai birkas, kas var potenciāli negatīvi ietekmēt viņu datu kopu atrodamību un kā rezultātā

arī atkalizmantojamību. Iespējams kategorijas ir pārāk vispārīgas, lai norādītu uz ievietotajām datu kopām.

Nākamā seko Centrālā statistikas pārvalde ar 41 datu kopu. Tā ir Latvijas pārvaldes iestāde, kas ir galvenā valsts statistikas darbu veicēja un koordinatore valstī, kuras darbojas Ekonomikas ministrijas pārraudzībā. Centrālā statistikas pārvalde izmanto Latvijas Atvērto datu portālā pieejamās kategorijas, un visvairāk datu kopas ir publicētas šādās kategorijās: "Iedzīvotāji un sabiedrība" (28 datu kopas), "Reģioni un pašvaldības" (28 datu kopas), "Ekonomika un uzņēmējdarbība" (11 datu kopas), "Izglītība un sports" (5 datu kopas). Vienai datu kopai var norādīt vairākas kategorijas, lai lietotājiem būtu vieglāk atrast nepieciešamās datu kopas. Visbiežāk datu kopām tiek izmantoti CSV (35 datu kopas), JSON (8 datu kopas), PDF (8 datu kopas) failu formāti.

Pārējiem publicētājiem ir mazāk par 30 publicētām datu kopām. Pamatā tās ir valsts iestādes, pašvaldības vai uzņēmumi (AS, SIA). Latvijas atvērto datu portālam ir vēl kur augt, bet attīstība virzās uz pareizo pusi. Dati tiek ievietoti portālā, bet to varētu būt vairāk, lai sabiedrībai būtu pieejama informācija jebkādā valdības mēroga jautājumā.

Detalizētāka un padziļinātāka portāla analīze bez papildlīdzekļu izmantošanas ir apgrūtinoša un resursietilpīga, ar ko saskārās arī vairāki valdības atvērto datu portālu pētnieki. Doto problēmu potenciāli ir spējīgs atrisināt rasmotājs, kas ļautu ātri sagatavot valdības atvērto datu portāla satura apkopojušo datu kopu, ko trešās puses lietotāji spētu izmantot saviem nolūkiem, t.i. netērējot laiku uz šo datu izgūšanu, bet gan sākot strādāt ar tiem. Tāpēc nākamajā nodaļā tiek aplūkots rasmotāja jēdziens, tam sekojot izstrādāta rasmotāja apraksts.

3. RASMOŠANA UN RASMOTĀJI

Rasmošana, ko dēvē arī par tīmekļa izguvi, ir paņēmiens, kā iegūt datus no globālā tīmekļa un saglabāt tos failu sistēmā vai datu bāzē vēlākai izgūšanai vai analīzei. Parasti tīmekļa dati tiek iegūti, izmantojot hiperteksta pārsūtīšanas protokolu (HTTP) vai pārlūkprogrammatūru. To veic manuāli lietotājs vai automātiski robots vai tīmekļa pārmeklētājs (web crawler). Rasmošana parasti attiecas uz automatizētiem procesiem datu iegūšanai. Ņemot vērā to, ka globālajā tīmeklī pastāvīgi tiek ģenerēts milzīgs skaits datu, rasmošana ir plaši atzīta kā efektīva un iedarbīga metode lielu datu savākšanai [17].

3.1. Ramošanas princips

Tīmekļa lapas rasmošana ietver tās nolasi un iegūšanu. Nolase ir lapas lejupielāde, ko pārlūks veic lietotājam apskatot lapu. Tāpēc tīmekļa pārmeklēšana ir galvenā rasmošanas sastāvdaļa, lai ielādētu lapas vēlākai apstrādei. Kad vietnes nolase tiek pabeigta, var sākt datu izguvi. Ar lapas saturu interesents var rīkoties pēc ieskatiem (risinot visdažādākā rakstura problēmas). Var to meklēt, pārformatēt, parsēt, datus kopēt izklājlapā utt. Rasmotāji parasti izņem kaut ko no lapas, lai to izmantotu citiem mērķiem kaut kur citur. Piemērs varētu būt vārdu un tālruņu numuru vai restorānu un to adrešu atrašana un kopēšana sarakstā. Lai piemēroties dažādām situācijām, pašreizējās rasmošanas metodes ir pielāgotas. Sākot ar mazākiem uzdevumiem ar cilvēkiem saistītām procedūrām un beidzot ar pilnībā automatizētu sistēmu izmantošanu, kas spēj pārvērst visas tīmekļa vietnes par labi organizētu datu kopu.

Rasmošana tiek izmantota kontaktu atlasei un kā sastāvdaļa lietojumprogrammās, kas tiek izmantotas tīmekļa indeksēšanai, datu ieguvei un izpētei, tiešsaistes cenu izmaiņu uzraudzībai un cenu salīdzināšanai, vietņu izmaiņu noteikšanai, laika apstākļu datu apsekošanai, produktu pārskatu atlasei, lai pētītu konkurenci, tīmekļa miksēšanai un tīmekļa datu integrācijai, tiešsaistes klātbūtnes un reputācijas izsekošanai, nekustamo īpašumu sludinājumu apkopošanai. Tīmekļa rasmošana tiek izmantota dažādos digitālos uzņēmumos, kas paļaujas uz datu apkopošanu. Likumīgas lietošanas gadījumi ietver:

- meklētājprogrammu robotus, kas pārmeklē vietni, analizējot tās saturu un pēc tam tos sarindo;

- tirgus pētījumu uzņēmumus, kas izmanto rasmotājus, lai iegūtu datus no forumiem un sociālajiem medijiem;
- cenu salīdzināšanas vietnes, kurās tiek izmantoti roboti, lai automātiski iegūtu cenas un produktu aprakstus sabiedroto pārdevēju vietnēm [18].

Rasmošana tiek izmantota arī nelikumīgiem mērķiem, tostarp cenu pazemināšanai un ar autortiesībām aizsargāta satura nozagšanai. Tiešsaistes uzņēmums, uz kuru tiek vērsts rasmotājs, var ciest lielus finansiālus zaudējumus, jo īpaši, ja uzņēmums lielā mērā paļaujas uz konkurējošiem cenu noteikšanas modeļiem vai darījumiem satura izplatīšanas jomā [18].

Datu rasmošanas procesu no interneta var sadalīt divos secīgos posmos: iegūt tīmekļa resursus un pēc tam iegūt nepieciešamo informāciju no iegūtajiem datiem. Tīmekļa rasmošanas programma (rasmotājs) tiek sākta, sastādot HTTP pieprasījumu, lai iegūtu resursus no mērķa vietnes. Šo pieprasījumu var formatēt kā URL, kas satur GET vaicājumu, vai kā HTTP ziņojuma daļā, kurā ir POST vaicājums. Kad mērķa tīmekļa vietne ir veiksmīgi saņēmusi un apstrādājusi pieprasījumu, pieprasītais resurss tiks izgūts no tīmekļa vietnes un pēc tam nosūtīts atpakaļ uz rasmotāju. Resurss var būt vairākos formātos, piemēram, tīmekļa lapās, kas veidotas no HTML, datu plūsmās XML vai JSON formātā vai multivides datos, piemēram, attēlos, audio vai video failos. Pēc tam, kad tīmekļa dati ir lejupielādēti, izvilkšanas process turpina parsēt, pārformatēt un strukturēti organizēt datus.

Diemžēl dažreiz rasmošanas mērķi saista ar vieglas komerciālas priekšrocības iegūšanu, kā konkurentu produktu cenas, potenciālo klientu zādzību, mārketinga kampaņu nolaupīšanu, API novirzīšanu vai pat satura un datu zādzību, tomēr mūsdienās, tāpat kā daudzos citos gadījumos, rasmotāji tiek izmantoti arī pozitīvākos nolūkos, piemēram, ļaujot entuziastiem un / vai pētniekiem iegūt datus dažādu zinātnisko vai citu pētījumu veikšanai.

3.2. Rasmotāju princips

Kā iepriekš tika noskaidrots, rasmotājus izmanto, lai iegūtu datus no vietnes satura. Tie piekļūst vietnes datiem un nokopē tos, pārveidojot lietotājam ērtā formātā, piemēram, XLSX vai CSV. Rasmotājs ir specializēts rīks, kas paredzēts, lai precīzi un ātri iegūtu datus no tīmekļa lapas. Rasmotāji ir ļoti atšķirīgi pēc dizaina un sarežģītības, atkarībā no projekta. Svarīga katra rasmotāja sastāvdaļa ir datu lokatori (vai atlasītāji), kas tiek izmantoti, lai atrastu datus, kurus lietotājs vēlas iegūt no HTML faila - parasti tiek lietots XPATH, CSS selektori, REGEX vai to kombinācija.

Ir divi būtiski rasmotāju moduļi — (1) modulis HTTP pieprasījuma sastādīšanai, piemēram, Urllib2 vai Selenium, un (2) modulis informācijas parsēšanai un izgūšanai no neapstrādāta HTML koda, piemēram, The Beautiful Soup vai Pyquery. Urllib2 modulis definē funkciju kopu, kas ļauj strādāt ar HTTP pieprasījumiem, piemēram, autentifikāciju, novirzīšanu, sīkfailiem utt. Selenium ir tīmekļa pārlūkprogrammas apvalks (wrapper), kas paredzēts tīmekļa lietojumprogrammu automatizēšanai testēšanas nolūkos, bet tas noteikti neaprobežojas tikai ar to. Attiecībā uz datu ieguvi Beautiful Soup ir paredzēts HTML un citu XML dokumentu rasmošanai. Python bibliotēka Beautiful Soup tiek arī izmantota rasmotāja izstrādei. Tas nodrošina ērtas Pythonic funkcijas parsēšanas koka navigācijai, meklēšanai un modificēšanai. Tas ir rīku komplekts HTML faila sadalīšanai un nepieciešamās informācijas izgūšanai, izmantojot lxml vai html5lib. Beautiful Soup var automātiski noteikt parsēšanas kodējumu apstrādē un pārvērst to par klienta lasāmu kodējumu [17].

No dažādiem rasmotāju veidiem daži tiek veidoti, lai automātiski atpazītu lapas datu struktūru, piemēram, Nutch vai Scrapy, vai nodrošinātu tīmeklī bāzētu grafisko saskarni, kas novērš manuāli rakstītu ramotāju kodu, piemēram, Import.io. Nutch ir spēcīgs un mērogojams tīmekļa pārmeklētājs, rakstīts Java valodā. Tas dod iespēju veikt smalku konfigurāciju, paralēlu datu savākšanu, robots.txt noteikumu atbalstu un mašīnmācību. Scrapy, kas rakstīta Python, ir vairākkārt lietojams tīmekļa pārmeklēšanas ietvars. Tas paātrina lielo pārmeklētāju projektu veidošanās procesu. Turklāt tas nodrošina arī tīmekļa izveidotu standartu, lai simulētu cilvēka tīmekļa vietnes pārlūkošanas darbību. Lai ne programmētāji varētu iegūt tīmekļa saturu, tīmekļa pārmeklētājs ar grafisko saskarni ir mērķtiecīgi izstrādāts, lai mazinātu tīmekļa rasmotāju sarežģītību. Starp tiem Import.io ir tipisks pārmeklētājs datu izgūšanai no tīmekļa vietnēm, nerakstot nevienu koda rindu. Tas ļauj lietotājiem identificēt un pārvērst nestrukturētas tīmekļa lapas strukturētā formātā. Import.io grafiskā datu identifikācijas saskarne ļauj lietotājam viegli iemācīties iegūt datus. Pēc tam iegūtie dati tiek glabāti īpašā mākoņserverī, un tos var eksportēt CSV, JSON un XML formātā [17].

Tā kā visiem rasmotājiem ir vienots mērķis - piekļūt vietnes datiem, var būt grūti atšķirt likumīgus un ļaunprātīgus robotus. Likumīgie roboti tiek identificēti ar organizāciju, priekš kuras tiek iegūti dati. Piemēram, Googlebot savā HTTP galvenē identificē sevi kā piederīgu Google. Ļaunprātīgi roboti gluži pretēji atdarina likumīgu datplūsmu, izveidojot viltus HTTP lietotāja aģentu. Likumīgie roboti ievēro vietnes robot.txt failu, kurā ir uzskaitītas tās lapas, kurām robotam ir atļauts piekļūt, un tām, kurām tas nav atļauts. Savukārt ļaunprātīgi rasmotāji pārmeklē vietni

neatkarīgi no tā, ko vietnes operators ir atļāvis. Resursi, kas nepieciešami tīmekļa rasmotāju robotu darbībai, ir ievērojami. Tik daudz, ka likumīgi rasmotāju robotu operatori daudz iegulda serveros, lai apstrādātu milzīgo iegūto datu daudzumu.

3.3. Rasmošanas vēsture

Tīmekļa rasmošana vai datu pārmeklēšana, vai datu ievākšana pastāv jau no tīmekļa pirmsākumiem. Lai gan tas bieži tiek saistīts ar tīmekļa satura iegūšanu, tas ne vienmēr ir kalpojis šim mērķim. Sākotnēji tas tika izstrādāts, lai automatizētu sarežģītus vai "sāpīgus" uzdevumus. Pirmais tīmekļa rasmošanas pielietojums bija ar kopnes ietvaru testēšanu. Izmantojot tādus rīkus kā Selenium, uzņēmumi, piemēram, IP-Label, ir izveidojuši produktus, kas ļauj tīmekļa izstrādātājiem un tīmekļa meistariem katru dienu uzraudzīt vietnes veiktspēju.

Tīmekļa rasmošana ir līdzīga tīmekļa indeksēšanai, kas ir process, ar kuru meklētājprogrammas indeksē tīmekļa saturu. Atšķirība ir robots.txt noteikums, kas nosaka, kurās vietās var doties roboti. Tīmekļa indeksētāji ("labie roboti") ievēro noteikumus; rasmotāji, no otras puses, var tikt izmantoti, lai zagtu jebkuru saturu, ko tie ir ieprogrammēti ienest — cenas, reklāmas, piedāvājumus vai informāciju, kas citādi būtu pieejama tikai maksas abonentiem vai pilnvarotiem biznesa partneriem. Protams, daudzi legālie rasmotāji arī ņem vērā un izpilda robots.txt noteikumu. Tīmekļa pārmeklētājs apmeklē tīmekļa lapas, iegūst datus un atklāj jaunas lapas no "sākuma" lapām. Lai gan lielākā daļa cilvēku uzskata, ka Google, bija pirmais pārmeklētājs, kas pārmeklēja visu tīmekli, tīmekļa pārmeklēšanai kā tehnoloģijai ir diezgan ilga un interesanta vēsture. Lai gan sākotnējie pārmeklētāji varēja pārmeklēt tikai datus, kad mūsdienu tīmekļa pārmeklētāji ir daudz gudrāki, jo tie, izņemot tīmekļa pārmeklēšanu, spēj uzraudzīt tīmekļa lietojumprogrammu ievainojamību un pieejamību.

Sākotnēji internets pat nebija meklējams. Kad nebija nevienas meklētājprogrammas, internets bija tikai FTP (datņu pārsūtīšanas protokol) vietnes savākšanas vieta, kurā lietotāji varēja pārvietoties, lai atrastu konkrētus koplietojamus failus. Tajā laikā cilvēki izveidoja īpašu automatizētu programmu, kas mūsdienās pazīstama kā "Web Crawler" (pārmeklētājs) vai "Web Scraper" (rasmotājs). Tā palīdz atrast un sakārtot internetā pieejamos izplatītos datus. Izstrādātais tīmekļa pārmeklētājs jeb robots ielādē visas lapas, kas ir pieejamas internetā, un pēc tam visu saturu iegūst datu bāzē indeksēšanai. Pirmie pārmeklētāji tika izstrādāti daudz mazākam tīmeklim -

apmēram 1’000’000 tīmekļa lapu, taču mūsdienās dažās populārākajās vietnēs ir līdz pat miljoniem lapu.

Galu galā ar meklētājprogrammas palīdzību tika pievienoti miljoniem tīmekļa lapu, un tā kļuva par miljoniem tīmekļa datu, tostarp audio, video, attēlu un tekstu, mājvietu vairākās formās. Tas pārvērtās par atvērtu datu avotu. Tā kā internets kļuva par datu avotu “jūru”, kurā ir viegli meklēt, cilvēkiem sāka šķist vienkārša publiski pieejamu datu ieguve. Bet problēma radās, kad dažas vietnes atteicās piešķirt lejupielādes iespēju, un manuāla datu kopēšana ir acīmredzami nogurdinoša un neefektīva. Un ar esošo problēmu radās tīmekļa rasmošanas metode. Tīmekļa rasmošanu faktiski nodrošina roboti jeb tīmekļa pārmeklētāji, kas darbojas tāpat kā meklētājprogrammas - ielādē un kopē datus. Tīmekļa rasmošana ir vērsta uz visu konkrētu datu iegūšanu no vietnes, turpretī meklētājprogrammas bieži iegūst lielāko daļu vietņu internetā [19].

3.4. Rasmotāju apskats

Ir izveidoti ļoti daudz bezmaksas un maksas rasmotāju rīku. Tomēr ne visām rasmotāju programmatūrām ir nepieciešamas programmēšanas prasmes. Par samaksu pieejamie ir vairāk domāti kā rasmošanas pakalpojumi, ko piedāvā kāda noteikta mājaslapa vai uzņēmums, vai pat cilvēks. Turpmāk uzskaitīti ir vieni no populārākajiem un par brīvu pieejamiem rasmotājiem, kuriem nav nepieciešamas kodēšanas prasmes. Izvēlētās programmatūras ir ļoti viegli uztvert un tās apmierina lielāko daļu rasmošanas vajadzību ar saprātīgu datu daudzumu. Saraksts iegūts un apkopots Google meklētājā ierakstot “web scraper” un aplūkojot dažus populārākos sarakstus (top lists).

- Webscraper.io – bezmaksas Google Chrome vai Firefox pārlūkprogrammas paplašinājums (extension);
- ParseHub – bezmaksas versija, kurā ļauj izstrādāt līdz pat 5 projektiem;
- Octoparse – bezmaksas versija paredzēta nelielu datu atlasei;
- Data Miner - Google Chrome pārlūkprogrammas paplašinājums, ar kuru bezmaksas var iegūt datus līdz pat 500 lapām mēnesī;
- Scraper - rasmotājs ir ļoti vienkāršs Google Chrome paplašinājums. Lai gan datu ieguve ir ierobežota, tas atvieglo tiešsaistes izpēti, ja dati ir ātri jāiegūst izklājlapas veidā. Tas ir paredzēts kā ērti lietojams rīks pieredzējušiem lietotājiem, kuri ērti strādā ar XPath [20].

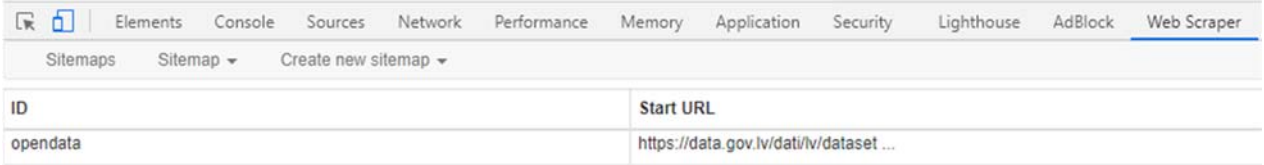
Gandrīz visiem uzskaitītajiem rasmotājiem, izņemot Scraper, kurš ir pilnībā bezmaksas, ir pieejamas bez maksas izmēģinājuma versijas (free trial) vai bez maksas versijas tiek citādāk ierobežotas, kas arī tiek augstāk norādīts. Ir pieejami dažādi apmaksas plāni ar dažādām priekšrocībām. Plašāk tiek apskatīti un izmēģināti pirmie trīs rasmotāji: webscraper.io Google Chrome paplašinājums, ParseHub bez maksas versija un Octoparse bez maksas versija. Visi rasmotāji tiks pārbaudīti, ar tiem izgūstot Latvijas Atvērto datu portāla saturu, to apkopojot CSV failā, tādējādi visiem rīkiem nodefinējot vienu uzdevumu, kas ļaus veikt objektīvākus secinājumus par to darbību un piemērotību uzdevumam. Dati tiek izgūti no saites https://data.gov.lv/dati/lv/dataset, no kuras tiek mēģināts izgūt datu kopu virsrakstu, aprakstu, kategorijas, faila formātu(s), skatījumus un saiti uz datu kopu, kā arī mēģināts pārslēgties uz nākamo lapu, lai izgūtu informāciju no visām publicētajām datu kopām. Iegūtā pieredze un rezultāti arī kalpos par prasību specifikāciju izstrādātajam rasmotājam un vēlāk tiks salīdzināti ar autora pašizstrādāto rasmotāju. Tiks analizēti trūkumi, no kuriem jāizvairās, un priekšrocības, kas būs jārealizē pašizstrādātajā rasmotajā. Šādi tiks arī apskatīts, cik viegli lietojami un cik ļoti piemēroti ir rasmotāji Latvijas Atvērto datu portālam.

3.4.1. Rasmotājs webscraper.io

Rasmotāja webscraper.io bezmaksas versiju var lejupielādēt Chrome internetveikalā kā paplašinājumu pārlūkprogrammai Google Chrome. Atbilstoši [21], bezmaksas versija paredzēta:

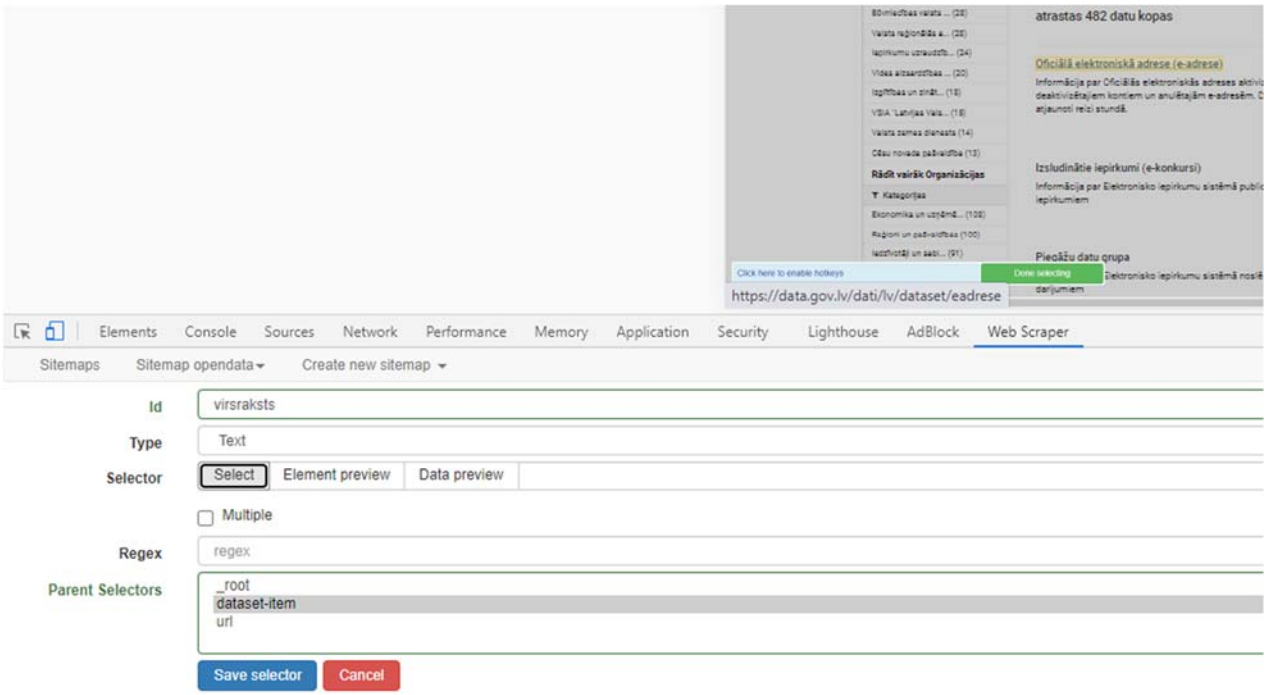
- tikai lokālai izmantošanai;
- dinamiskām tīmekļa vietnēm (Dynamic Websites);
- JavaScript izpildei;
- CSV faila lejupielādei.

Lai izmantotu rasmotāju, jānokļūst Chrome "izstrādātāja rīkos" ejot caur iestatījumiem vai nospiežot taustiņu "F12". Uzspiežot uz cilni "Web Scraper" atveras 3.1. attēlā redzamais logs, kur var arī redzēt jau izveidotu rasmošanas mēģinājumu.



3.1. att. **"Web Scraper" cilne pārlūkprogrammā Google Chrome [autora veidots]**

Ir jāizveido jauna vietnes karte (sitemap), kurā tiek norādīta tīmekļa lapa, kuru nepieciešams rasmot. Atverot izveidoto vietnes karti, var pievienot jaunu selektoru (selector). Ar selektoru var izvēlēties nepieciešamos datus, ko vajag iegūt no tīmekļa lapas. Piemēram, 3.2. attēlā tiek izvēlēts Latvijas Atvērto datu portāla datu kopu virsraksts. Tas notiek, uzspiežot uz virsraksta elementa un atzīmējot "Multiple", lai rasmotājs atrod visus virsrakstus. Tas tiek atkārtots visiem elementiem, kas atbilst lietotāju interesējošai informācijai. Testēšanas gadījumā tika izvēlēta sekojošā informācija: virsraksts, apraksts, kategorija, failu formāts, skatījumi, saite uz datu kopu.



3.2. att. **Selektora pievienošana [autora veidots]**

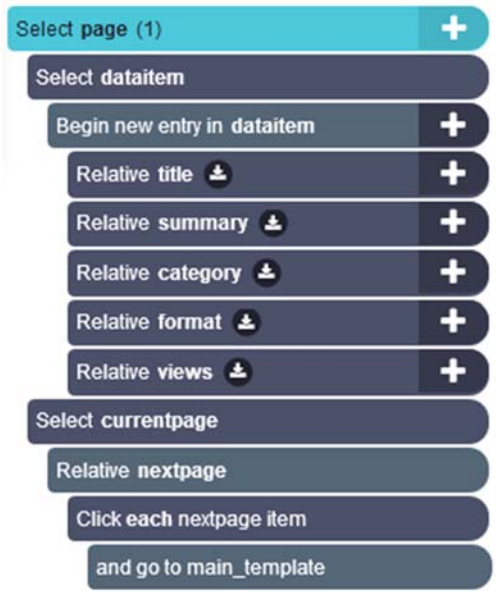
Pēc datu atlases un pārmeklēšanas, websraper.io paplašinājums piedāvā lejupielādēt CSV failu ar iegūtajiem datiem. Piektā pielikuma attēlā var redzēt iegūto CSV failu. Tajā iegūtā informācija ir izmētāta pa faila saturu, ir salikti nevajadzīgi komati un atkāpes un tiek ievietoti nevajadzīgi dati. Microsoft Excel to nevar apstrādāt, lai iegūtu lasāmu izklājlapu. Var secināt, ka iegūtie dati CSV failā nav izmantojami. Tos jebkādi apstrādāt, lai kļūtu lasāmi nav iespējams. Webscraper.io ir viens no sliktākajiem piemēriem.

3.4.2. Rasmotājs ParseHub

Nākamais apskatītais rasmotājs ir ParseHub. To var lejupielādēt ParseHub mājaslapā: https://www.parsehub.com/quickstart. Rasmotājam arī ir pieejama gan bezmaksas, gan maksas versijas. Atbilstoši [22], bezmaksas versija iekļauj:

- līdz pat 200 datu lapu iegūšanu tikai 40 minūtēs;
- līdz pat 200 lappuses iegūšanu vienā izpildes reizē;
- līdz pat 5 projektu publiskošanu, ar kuriem iespējams dalīties;
- ierobežotu atbalstu;
- datu saglabāšanu līdz pat 14 dienām.

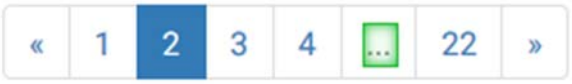
Pēc rasmotāja lejupielādes un atvēršanas ir nepieciešams piereģistrēties vai pieslēgties lietotāja profilam. Viena no rasmotājam priekšrocībām ir lielais skaits gan video, gan rakstiskas, gan interaktīvas pamācības, kurās pa soļiem viss tiek paskaidrots, lai jebkurš lietotājs spētu un mācētu iegūt viņam nepieciešamu atbalstu rasmošanas ceļā. Trešā pielikuma attēlā var redzēt ParseHub galveno lapu, kurā var arī uzsākt jaunu rasmošanas projektu. Nepieciešams norādīt tīmekļa vietni, no kuras tiks iegūti dati. Tālāk līdzīgi iepriekš apskatītajam webscraper.io ir jāizvēlas selektori un jāatzīmē nepieciešamie lauki. 3.3. attēlā var redzēt visus selektorus. Izvēloties laukus (virsraksts, apraksts utt.) nepieciešams izmantot "Relative select", lai izejā dati būtu kopā apvienoti, jeb CSV failā veidotu vienu ierakstu.



3.3. att. **ParseHub izmantotie selektori [autora veidots]**

ParseHub ir iespēja pāriet uz nākamo lapu, lai iegūtu visus datus no pieejamām lapām. Tas sagādāja problēmas, jo Latvijas Atvērto datu portālā poga uz nākamo lapu samaina savu pozīciju. ParseHub nespēja to atrast, jo skaitīja pogu skaitu pirmajā lapā. To var redzēt 3.4. un 3.5. attēlos.

3.4. att. **Navigācijas pogas 1. lapā [autora veidots]**



3.5. att. **Navigācijas pogas 2. lapā [autora veidots]**

Līdzīgai problēmai bija pieejama pamācība, kā norādīt nākamo lapu, bet šajā situācijā tas nepalīdzēja un dati tika iegūti tikai no pirmās lapas. Taču iegūtie un apkopotie dati CSV failā ir ļoti pārlasāmi un labas kvalitātes. Fragmentu no lejupielādētā faila var apskatīt 6. pielikumā. Var pieminēt, ka rasmotājs labi tika galā ar faila formāta nolasīšanu. No autora pieredzes, bieži rasmotājiem ir problēmas ar faila formāta lauka datu izguvi un apstrādi. Iespējams, ja ilgāks laiks tiktu patērēts uz ParseHub rasmotāju, tad arī izdotos pāriet uz nākamo lapu, lai iegūtu datus no visām lapām.

3.4.3. Rasmotājs Octoparse

Pēdējais no apskatītajiem rasmotājiem ir Octoparse. Pieejams Ocotparse mājaslapā: https://www.octoparse.com/download. Rasmotājam ir pieejami dažādi maksājuma plāni, kā arī bez maksas versija. Lai to izmantotu, ir nepieciešams piereģistrēties. Atbilstoši [23], bez maksas versija iekļauj:

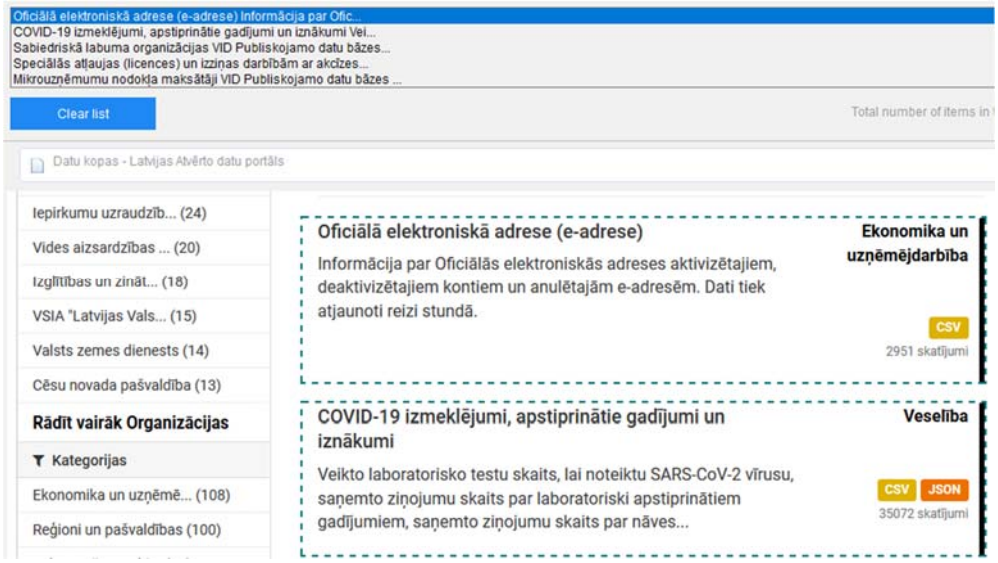
- neierobežotu lapušu skaitu uz vienu pārmeklēšanu;
- neierobežotu skaitu izmantoto datoru;
- līdz pat 10000 ierakstu vienā datu eksportā;
- līdz pat divu vienlaicīgu lokālu datu iegūšanu;
- līdz pat 10 pārmeklētājus;
- ierobežotu sabiedrības atbalstu.

Pēc Octoparse lejupielādes un pieslēgšanās savam profilam, lietotājs nokļūst galvenajā skatā, skatīt ceturto pielikumu. Pirmo reizi ieslēdzot rasmotāju, parādīsies īsa pamācība kā izmantot programmatūru. Ir pieejamas dažādas rasmošanas veidnes, lai iegūtu datus no populārākajām

tīmekļa saitēm, kā “Amazon”, “Tripadvisor” vai “Instagram”. Lai sāktu rasmošanu, nepieciešams sākt jaunu uzdevumu. Tiks izmantots iebūvētais vedņa režīms (wizard mode). Kā iepriekš redzēts, nepieciešams ievadīt tīmekļa saiti, no kuras tiks izgūti dati. Nākamajā solī rasmotājs pieprasa norādīt, kāda veida datu iegūšana tiks veikta:

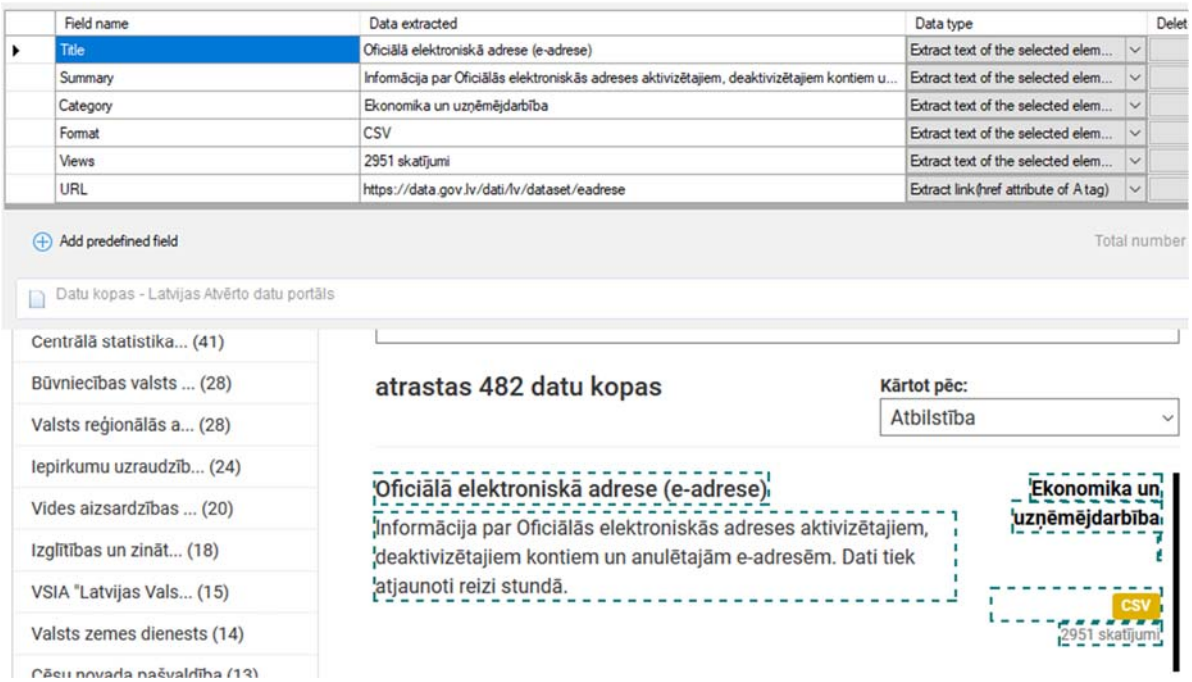
- saraksta/tabulas izgūšana no jebkuras tīmekļa lapas. Atbalsta lapu numerāciju;
- detalizētas tīmekļa lapas datu izgūšana, ja ir saišu saraksts, uz kurām jānoklikšķina. Atbalsta lapu numerāciju;
- datu izgūšana no vienas tīmekļa lapas.

Tiks izvēlēta saraksta izgūšana, lai iegūtu datus no Latvijas Atvērto datu portāla datu kopām. Sākumā ir jāiezīmē HTML elements, kas iekļauj visus nepieciešamos datus. Tas redzams 3.6. attēlā.



3.6. att. **HTML elementa izvēle Octoparse rasmotājā [autora veidots]**

Nākamajā solī tiek izvēlēti visi nepieciešamie dati izvēlētajā elementā. Tas ir redzams 3.7. attēlā. Tiek izvēlēts datu kopas virsraksts, apraksts, kategorija, faila formāts, skatījumi, saite uz datu kopu.



3.7. att. **Datu izvēle Octoparse rasmotājā [autora veidots]**

Kā jau iepriekš tika minēts, rasmotājs atbalsta lapu numerāciju un tas arī tiks izmantots, lai iegūtu datus no visām lapām. Tiek izvēlēts nākamās lapas rādītājs "»" poga. Atšķirībā no ParseHub rasmotāja, Octoparse nākamajās lapās tieši meklē elementu ar simbolu "»" HTML koda tekstā, tādējādi veiksmīgi iegūstot datus no visām lapām. Vienīgo problēmu sagādāja faila formāta lauka datu iegūšana, CSV failā tiek ielikta liela atstarpe, bet viss pārējais ir iegūts precīzi. To var arī aplūkot septītajā pielikumā, kur ir redzams fragments no CSV faila. Redzam, ka ir tajā brīdī iegūtas visas 482 pieejamās datu kopas (neskaita galveni (header)). 3.8. attēlā redzama neveiksmīgā faila formāta lauka datu nolase. Tiek atstātas lielas atstarpes, kas pagrūtina CSV faila lasīšanu un apstrādi.

3.8. att. **Fragments no Octoparse iegūtā CSV faila ar neveiksmīgo faila formāta lauku [autora veidots]**

Starp trīs apskatītajiem rasmotājiem, priekšroka ir jādod Octoparse rasmotājam. Tajā ir ļoti viegli strādāt. Nevajag lielas zināšanas, lai iegūtu sev nepieciešamos datus. Visa datu izvēle notika bez problēmām, kā arī liels pārsteigums bija bez problēmu lapu pārslēgšana, kas abos pārējos sagādāja problēmas vai pat nebija iespējama. Rasmotājs strādāja ātri un bez kļūdām. Iegūtie dati ir viegli pārskatāmi un izmantojami.

4. RASMOTĀJA IZSTRĀDE

Iepriekšējās nodāļās tika izpētīti un aprakstīti dažādi veidi kā iegūt lietotājam nepieciešamos datus. Viens no tiem ir rasmotāju izmantošana. Datus bija nepieciešams iegūti no Latvijas Atvērto datu portāla. Visi trešajā nodaļā apskatītie rasmotāji to izdarīja dažādos kvalitātes līmeņos. Tā kā visus nepieciešamos datus nebija iespējams iegūt ērtā un apskatāmā veidā ar visiem pieejamiem bez maksas rasmotājiem, tika izstrādāts autora rasmotājs. Pašizstrādāts rasmotājs var palīdzēt iegūt datus ne tikai autoram, bet arī trešajām personām, kas ir atvērto datu entuziasti vai datu analītiķi. Rasmotājs paātrinās datu analīzi, jo nebūs nepieciešams iet cauri visām datu kopām manuāli. Visi dati jau tiks apkopoti CSV failā, kuru pēc tam apstrādāt ir daudz vieglāk. Rasmotāja nepieciešamību, kas iegūst visus lietotājam nepieciešamos datus, apliecina arī virkne zinātnisko rakstu, kuros pasaules pētnieki un praktiķi pēta visdažādākos ar valdības atvērto datu portālu saistītus jautājumus [24, 25], kur rasmotāja palietošana ļautu būtiski atvieglot veiktos pētījumus, ļaujot tiem koncentrēties uz datu apstrādi, mazāk laika tērējot uz to izgūšanu un sagatavošanu, kur katrs meklē pieeju interesējošo datu izgūšanai (piem. metadatu analizatori utt., kas ne vienmēr ir spējīgi izgūt visus lietotājam nepieciešamos datus).

Izstrādāto rasmotāju mazliet modificējot varēs izmantot datu ieguvei no citiem valdības atvērto datu portāliem. Daudziem atvērto datu portāliem ir ļoti līdzīga struktūra.

4.1. Izmantotās tehnoloģijas un prasības

Lai izstrādātu rasmotāju, tika izvēlēta Python programmēšanas valoda, jo autoram ir zināšanas un vislielākā pieredze tieši šajā programmēšanas valodā. Python ir augsta līmeņa programmēšanas valoda, kas nozīmē, ka tās pašas funkcionalitātes uzrakstīšanai nepieciešams mazāks rindu skaits un pirmkods ir vieglāk saprotams [26]. Autors izmanto Python3 versiju 3.7. Ir ļoti daudz un dažādas kompilatoru un izstrādes vides darbam ar Python. Var izmantot arī paša Python izstrādātu vidi, taču autors izmantoja alternatīvu vidi PyCharm 2021.1.1 (Professional Edition) bez maksas 30 dienu izmēģinājuma versiju.

Iegūtajā CSV failā jābūt zemāk nosauktajām kolonnām, jo autors uzskata, ka sekojošie dati ir svarīgākie turpmākai datu apstrādei un / vai analīzei, kā arī ar izvēlēto kolonnu apstrādi visbiežāk ir saistīti arī iepriekšminēto pētnieku analīzes darbi:

- “Headline” – datu kopas nosaukums;

- "Summary" – datu kopas pilnais apraksts, dažām datu kopām tas ir ļoti īss, dažos gadījumos identisks ar datu kopas nosaukumu, dažām tas ir ļoti garš (vairāk par 500 vārdiem);
- "Organization" – datu kopas organizācija jeb publicētājs;
- "Category" – visas datu kopas kategorijas, kas ir norādītas citā cilnē (saitē);
- "File format" – visi faila formāti;
- "Views" – datu kopas skatījumi;
- "Tags" – publicētāju izveidotas un datu kopai pievienotas birkas;
- "URL" – saite uz datu kopu.

Izstrādātajam rasmotājam vajadzēs veikt sekojošās darbības:

- iegūt un nolasīt Latvijas Atvērto datu saites HTML kodu;
- atrast vajadzīgās saites no HTML koda, lai iegūtu plašāku informāciju par katru datu kopu (iegūt datus vienu soli "dziļāk");
- atrast nākamās lapas pogu portāla vietnē un iziet cauri visām lapām;
- nolasīt un izgūt datu kopas pilno nosaukumu no datu kopas saites;
- nolasīt un izgūt datu kopas pilno aprakstu no datu kopas saites;
- nolasīt un izgūt datu kopas publicētāja (organizācijas) nosaukumu no datu kopas saites;
- nolasīt un izgūt visas datu kopai saistītās kategorijas no datu kopas saites "Kategorijas" cilnes;
- nolasīt un izgūt faila formātus, kādā ir publicēta datu kopa;
- nolasīt un izgūt datu kopas skatījumus;
- nolasīt un izgūt visas datu kopai ierakstītās birkas no datu kopas saites;
- nolasīt un izgūt saiti uz datu kopu;
- visus iepriekš izgūtos datus apkopot un saglabāt vienā CSV failā, kuru var izmantot datu apstrādei un / vai analīzei.

Valodā Python ir pieejamas vairākas bibliotēkas, ar kuru palīdzību var viegli pārmeklēt tīmekļa lapas, izgūt datus no tām un apkopot lietotājam nepieciešamā formātā. Autors izvēlējās sekojošās bibliotēkas:

- Beautiful Soup - Python bibliotēka datu izvilkšanai no HTML un XML failiem. Tā darbojas kopā ar parsētāju, lai nodrošinātu idiomiskus navigācijas, meklēšanas un

parsēšanas koka modificēšanas veidus. Tas parasti saglabā programmētāju darba stundas vai dienas [27]. Tika izmantota jaunākā beautifulsoup4 4.9.3. versija.

- requests - ļauj ļoti viegli nosūtīt HTTP/1.1 pieprasījumus. Nav vajadzības manuāli pievienot vaicājuma virknes vietrāžiem URL vai norādīt POST datu kodējumu. Keep-alive un HTTP savienojumu apvienošana ir 100% automātiska pateicoties urllib3 [28].
- lxml – ļoti bagāta ar funkcijām un viegli lietojama bibliotēka XML un HTML apstrādei Python valodā [29]. Beautiful Soup dokumentācija iesaka izmantot lxml parsētāju tīmekļa saites pārmeklēšanai [27];
- csv - modulis ievieš klases tabulāru datu lasīšanai un rakstīšanai CSV formātā [30].

4.2. Rasmotāja darbības princips

Lai varētu sākt jebkādu datu ieguvi, sākumā rasmotājam nepieciešams nolasīt pilno saites HTML kodu. Tāpēc tiek izmantota Python requests bibliotēka, lai viegli izpildītu GET pieprasījumu Latvijas Atvērto datu portālam. Kopā ar lxml parsētāju tiek iegūts izvēlētās saites HTML kods no portāla. HTML kods atbilst tam, ko redz lietotājs atverot saiti. HTML koda iegūšanas funkciju var redzēt 4.1. attēlā. Pirmais pieprasījums tiek veikts "Datu katalogs" saitei https://data.gov.lv/dati/lv/dataset. 4.1. attēlā redzamajā koda fragmentā "url" mainīgā vietā var arī ievietot citu portāla saiti. Var izmantot portālā pieejamos filtrus un atlasīt datu kopas tikai ar to filtru, piemēram, izfiltrēt datu kopas, kurām ir norādīta kategorija – "Veselība". Tiek iegūta saite https://data.gov.lv/dati/lv/dataset?groups=veseliba, kas tiek ievietota mainīgā "url" vietā un tā iegūstot CSV failu ar datu kopām, kurām norādīta kategoriju "Veselība".

```
url = 'https://data.gov.lv/dati/lv/dataset'

def getdata(url):
    source = requests.get(url).text
    soup = BeautifulSoup(source, 'lxml')
    return soup
```

4.1. att. HTML **koda iegūšanas funkcija no Latvijas Atvērto datu portāla [autora veidots]**

Uzreiz vajadzēja atrisināt problēmu ar pāriešanu uz nākamo lapu, lai izgūtu datus no visām lapām, ko nav izdevies panākt ar apskatītajiem webscraper.io un ParseHub rīkiem. Problēma ir sekojoša, HTML kodā netiek norādīta klase vai kāds cits identifikators nākošās lapas saitei. Vienīgais, kas atšķir nākamās lapas saiti no pārējām lapu saitēm, ir uzrakstītais simbols "»". Augstāk minētā problēma tika novērota vairākās vietās HTML kodā, lai iegūtu nepieciešamos elementus. 4.2. attēlā var redzēt izgriezumu no HTML koda un lapu izvēles pogām.

4.2. att. HTML **koda izgriezums lapu izvēles saitēm [autora veidots]**

Rasmotājā vajadzēja izveidot funkciju, kas atradīs simbolu "»" no lapu saraksta un no tā iegūs saiti uz nākamo lapu. Izveidotā funkcija redzama 4.3. attēlā.

```
def getnextpage(soup):
    page = soup.find('ul', class_='pagination')
    try:
        if '»' in page.text:
            links = page.find_all('a')
            for link in links:
                if '»' in link.text:
                    url = 'https://data.gov.lv' + str(link.get('href'))
            return url
        else:
            return
    except:
        if page not in soup:
            pass
```

4.3. att. **Nākamās lapas saites iegūšanas funkcija [autora veidots]**

Funkcija sākumā HTML kodā (1) atrod sarakstu ar saitēm uz lapām, tad (2) pārbauda vai tajā sarakstā ir simbols "»", kas liecina par saiti uz nākamo lapu, (3) iegūst nākamās lapas saiti, kas tālāk ārpusē tiek padots HTML koda iegūšanas funkcijai. Ir jāveic viens izņēmums, ja tiek iegūti dati no filtrētas portāla saites, kurā ir tikai viena lapa. Ja tas tiek konstatēts, funkcija tiek izlaista un rasmotājs var turpināt savu darbību.

Pēc funkciju deklarēšanas tiek izveidots CSV fails ar Python csv bibliotēkas palīdzību. Tiek palaists "writer", kas CSV failā ieraksta pirmo rindu - galveni. Galvene satur kolonu nosaukumus: "Headline", "Summary", "Organization", "Category", "File format", "Views", "Tags", "URL".

Rasmotājs izsauc HTML koda izgūšanas un nākamās lapas saites iegūšanas funkcijas. Kamēr tiek iegūts mainīgais "url", rasmotājs turpina strādāt. Tā rasmotājs neiestrēgs mūžīgā datu izgūšanas ciklā. Tāpat arī tiek izveidots cikls, kas apskata katru datu kopu. Secīgā kārtībā ciklā tiek izpildītas sekojošās darbības katrai datu kopai (cikla pilnais kods redzams 8. pielikumā):

1) tiek iegūta saite uz datu kopu;
2) tiek izveidots GET pieprasījums datu kopas saitei un tiek parsēts saites HTML kods;
3) tiek izgūts pilnais nosaukums no datu kopas saites, alternatīvi tiek iegūts nosaukums no "Datu kataloga", ja nav iespējams to dabūt no datu kopas saites;
4) tiek izgūts pilnais nosaukums no datu kopas saites, alternatīvi tiek iegūts nosaukums no "Datu kataloga", ja nav iespējams to dabūt no datu kopas saites;
5) tiek izgūts pilnais apraksts no datu kopas saites, alternatīvi tiek iegūts apraksts no "Datu kataloga", ja nav iespējams to dabūt no datu kopas saites;
6) tiek izgūts datu kopas publicētājs jeb organizācija no datu kopas saites, alternatīvi tiek atstāts tukšs laukums;
7) tiek iegūta saite uz "Kategorijas" cilni, kurā ir norādītas visas kategorijas tekošajai datu kopai. No iegūtās saites tiek izgūtas visas kategorijas un saglabātas vienā mainīgajā, atdalot katru ar komata zīmi. Alternatīvi kategorija tiek izgūta no "Datu kataloga";
8) tiek izgūts/i datu kopas faila/u formāts/i, tiek saglabāts vienā mainīgajā, atdalot katru ar komata zīmi. Alternatīvi tiek atstāts tukšs laukums;
9) tiek izgūta/as datu kopas birka/as no datu kopas saites, alternatīvi tiek atstāts tukšs laukums;
10) tiek izgūts skatījumu skaits;
11) cikla beigās visa iegūtā informācija par vienu datu kopu tiek ierakstīta CSV faila rindā.

Kad vairs netiek atrasts nākamais "url" mainīgais un visi dati ir izgūti un ierakstīti CSV failā, CSV fails tiek aizvērts un rasmotājs pārtrauc savu darbību. Iegūto CSV failu var atvērt ar Microsoft Excel palīdzību, izmantojot "From text/CSV" rīku. Nepieciešams norādīt, ka failam ir UTF-8 kodējums un atdalītājs ir semikols, kā arī pirmā rinda ir galvene. CSV faila fragmentu var redzēt 4.4. attēlā. Iegūtā informācija ir pilnīga un kvalitatīva. Datu apstrādei un / vai analīzei salīdzinoši viegli lietojama.

Headline	Summary	Organization	Category	File form	Views	Tags	URL
Oficiālā elektroniskā adrese (e-adrese)	Informācija par Oficiālās elektroniskās adreses aktivizētajiem, deaktivizētajiem kontiem un anulētajām e-adresēm. Dati tiek atjaunoti reizi stundā.	Valsts reģionālās attīstības aģentūra	Ekonomika un uzņēmējdarbība, Iedzīvotāji un sabiedrība, Reģioni un pašvaldības, Valsts pārvalde,	CSV,	2961 skatījumi	PIKTAPS, e-adrese,	https://data.gov.lv/dati/lv/datase
COVID19 vakcinācijas	Datu kopa satur informāciju par vakcināciju pret COVID19. Tajā ietverta informācija par vakcināciju veikušo ārstniecības iestādi, vakcinācijas datumu, vakcīnas preparātu, vakcīnas sērijas numuru, vakcinācijas posmu, vakcīnas kārtas numuru, preparāta daudzumu ml, vakcīnas ievadīšanas veids un indikācijām vakcinācijai. Dati tiek izgūti no E-Veselības vakcinācijas datu modeļa. Dati tiek apkopoti katru darba dienu, par iepriekšējo periodu. Ja vakcinācijas dati E-Veselības vakcinācijas datu modelī, tiek pievadīti vēlāk vai laboti, tad atjaunojot datus mainās iepriekšējo datumu rādītāji.	Nacionālais veselības dienests	Veselība,	XLSX,	22549 skatījumi	covid-19,	https://data.gov.lv/dati/lv/datase vakcinacijas
Vakances	Uzņēmumu un organizāciju izsludinātās aktuālās vakances Latvijas valsts sektorā.	Nodarbinātības Valsts Aģentūra	Ekonomika un uzņēmējdarbība, Valsts pārvalde,	CSV,	2642 skatījumi	te ir darbs, vakance, vakances,	https://data.gov.lv/dati/lv/datase
Patiesie labuma guvēji	Ziņas par tiesību subjektu aktuālajiem patiesajiem labuma guvējiem (PLG) - fiziskajām personām. Tiesiskais pamats fizisko personu datu publiskošanai – likuma "Par Latvijas Republikas Uzņēmumu reģistru" 4.10 panta divpadsmitā daļa (spēkā ar 01.11.2020). Kolonnu skaidrojumi: id - Vienotais resursa identifikators legal_entity_registration_number - Tiesību subjekta reģistrācijas numurs forename - Vārds surname - Uzvārds latvian_identity_number_masked - Daļa no personas koda birth_date - Dzimšanas datums (ja nav personas koda) formātā YYYY-MM nationality - Valstspiederība (ISO 3166 Alpha-2 kods) residence - Dzīvesvietas valsts (ISO 3166 Alpha-2 kods) registered_on - Reģistrācijas datums last_modified_at - Ieraksta pēdējās labošanas laiks	LR Uzņēmumu reģistrs	Ekonomika un uzņēmējdarbība,	CSV,	723 skatījumi	patiesais labuma gu..., plg,	https://data.gov.lv/dati/lv/datase labuma-guveji
Maksātnespējas procesi	Uzņēmumu reģistra Maksātnespējas reģistrā ierakstīto maksātnespējas un tiesiskās aizsardzības procesu informācija. Maksātnespējas procesu atvērto datu lauku skaidrojumi – spied šeit.	LR Uzņēmumu reģistrs	Ekonomika un uzņēmējdarbība,	CSV, XLSX,	311 skatījumi	maksātnespējas process,	https://data.gov.lv/dati/lv/datase ejas-procesi

4.4. att. **Pašizstrādātā rasmotāja iegūtais CSV faila fragments [autora veidots]**

Salīdzinot ar pārējiem iegūtajiem CSV failiem no trešajā nodaļā apskatītajiem rasmotājiem, pašizstrādātais rasmotājs ieguva vissprecizāko un vispilnāko informāciju. No trešajā nodaļā apskatītajiem rasmotājiem, vislabākais bija Octoparse rasmotājs, taču ar Octoparse nebija iegūti dati, ejot "dziļāk" datu kopas saitēs, lai izgūtu pilnāku informāciju, kā pilno datu kopas aprakstu, vai papildus informāciju, kā organizācijas, birkas un visas datu kopas kategorijas, kuru iegūšana varētu sagādāt lielas problēmas vai pat būt neiespējama. Tomēr liela priekšrocība, strādājot ar Octoparse vai citu rasmotāju, ir tā grafiskā saskarne. Lietotājs var viegli pārvietoties caur rasmotāja programmu, lai izgūtu sev nepieciešamo informāciju. Lietotājam arī nav jāpārzina programmēšana, pietiek ar parastām datora izmantošanas prasmēm. Tāpēc autors plāno paturpināt rasmotāja izstrādi un pievienot grafisko vidi, lai būtu vieglāk izmantot pašizstrādāto rasmotāju.

Apskatīto rasmotāju problēma ir to unviersālums. Protams, ja nepieciešams izgūt datus no dažādām tīmekļa vietnēm, to nevar nosaukt par problēmu. Tomēr, ja ir jāiegūst dati no Atvērto datu portāliem, tostarp Latvijas Atvērto datu portāla, pašizstrādātais rasmotājs to izdara labāk. Ja ir nepieciešama kāda papildus informācija, to var viegli ierakstīt, papildinot kodu. Ja kāda informācija konkrētajā gadījumā nav nepieciešama, to var vienkārši izdzēst no koda. Lai gan rasmotājs pamatā ir paredzēts datu izgūšanai no Latvijas Atvērto datu portāla, rediģējot dažas koda rindas tas kļūst piemērots arī citiem valdību atvērto datu portāliem ar citādāku struktūru. Ja struktūra ir tāda pati, kas ir lielākai daļai portālu, būs nepieciešams pielabot tikai niecīgas nianses kā parametru vērtības. Tāpēc var secināt, ka izstrādātais rasmotājs ir puslīdz universāls Atvērto datu portāliem.

4.3. Problēmas un trūkumi

Autors saskārās ar neskaitāmām problēmām rasmotāja izstrādes procesā. Galvenā problēma bija elementu klases vai citu identifikatoru neizmantošana HTML kodā, piemēram, iepriekšējā nodaļā minēta problēma, kas saistīta ar saites uz nākošo lapu korektu apstrādi. Vienīgais veids, kā noteikt vajadzīgo saiti, ir tekstā ierakstītais teksts "»". Ja pēc tekstā ierakstītā nevarēja noteikt nepieciešamo elementu, bija jāmeklē elementi t.i. elementu grupas, piemēram, "div" vai "ul", kurās iekšā atrodas nepieciešamais izgūstamais elements. Tas tika darīts ar jau nosaukto nākamās lapas saiti, organizāciju nosaukumu, kurai HTML kodā ir vienāda klase ar datu kopas nosaukumu. Lai izgūtu datu kopas skatījumu skaitu, vajadzēja norādīt, ka tas ir otrais pēc kārtas "span" elements noteiktā "div" grupā. Nebija norādīts neviens cits identifikators skatījumu elementam, tāpēc vajadzēja skaitīt elementus pēc kārtas.

Ilgi tika risināta arī visu datu kopai saistīto kategoriju izgūšana. Kategoriju saraksts atrodas citā cilnē, kurš arī nav identificēts HTML kodā. Tika izmantots līdzīgs paņēmiens kā nākamās lapas saites iegūšanai. Tika atrasta saite, kurā ir teksts "Kategorijas" vai "Groups" saites angļu valodas versijai. Tiek ieiets saitē un iegūts pilnais HTML kods, kurā ir arī visas datu kopai saistītās kategorijas. Ja kategorija nav norādīta, saite rādīsies tukša. Tad rasmotājs izgūst nosaukumu "Cits" no "Datu katalogs" saites.

Problēmas radīja datu kopa "My First Dataset". Aizejot uz datu kopas saiti datu kopa netiek atrasta. Parādās kļūda "404 Not Found". Tā kā nekas netiek atrasts, rasmotājs arī nevar neko nolasīt. Pie datu kopas "My First Dataset" rasmotājs izdeva kļūdu un pārtrauca savu darbību, tāpēc autoram vajadzēja veikt labojumus kodā, lai tiktu ņemts vēra izņēmums. Astotajā pielikumā redzamajā cikla kodā var redzēt, ka katrai datu kopai tiek veikti izņēmumi. Datu kopas nosaukumam un aprakstam tiek izveidoti jauni mainīgie, lai izgūtu datus no "Datu katalogs" saites. Citiem datiem tiek ielikts tukšs lauks. Pēc portālā iebūvētajiem filtriem var redzēt, ka datu kopu ir publicējusi "Valsts reģionālās attīstības aģentūra". Šo kļūdu būtu nepieciešams izlabot datu kopas publicētajam vai Latvijas Atvērto datu portāla administratoriem.

Dažos gadījumos portālā redzamais datu kopas apraksta formatējums var atšķirties no CSV failā izgūtā, jo portālā tiek ignorētas tekstā ievietotās atkāpes, savukārt HTML kodā var redzēt atkāpes. Tāpēc teksta atkāpes arī parādās CSV failā.

Ir noteikti pieejami vairāki veidi, kā uzlabot rasmotāja darbību, padarot to ātrāku un uzticamāku. Astotajā pielikumā redzamajā ciklā var izņemt visus izdrukāšanas "print" izsaucienus. Tie bija nepieciešami autoram izstrādes procesā, lai pārliecinātos par rasmotāja darbību.

Kopumā rasmotājs atbilst autora zināšanām un pieredzei Python programmēšanas valodai. Tā kā autora darbā tiek apskatīts atvērto datu jēdziens, izstrādātais rasmotājs ir visiem publiski pieejams. Tas ir publicēts GitHub repozitorijā - https://github.com/Latviano123/rasmotajs. Ja ir vēlēšanās lejupielādēt tikai iegūto CSV failu (26.05.2021.), to var izdarīt repozitorijā.

REZULTĀTI

Autora darbs tika iedalīts divās lielās daļās - teorētiskajā un praktiskajā daļās. Teorētiskajā daļā autors apskatīja atvērto datu jēdzienu un ar to saistītus jautājumus. Tika veikts īss apskats atvērto datu vēsturē. Autoram vajadzēja iepazīties ar kādiem datiem būs jāizstrādā rasmotājs, tāpēc tika veikta īsa Latvijas Atvērto datu portāla analīze. Analīzes rezultātā tika noteikti Latvijas Atvērto datu portāla datu kopu populārākie formāti, kas ir CSV, XLSX un XLS failu formāti, kā arī apkopoti top publicētāji, kas ir ĢEOLatvija.lv un Centrālā statistikas pārvalde. Autors arī iepazinās ar portālu un tā darbību.

Turpinot teorētisko daļu tika apskatīts rasmošanas un rasmotāju princips, lai autoram būtu saprotama izstrādājamā programmas darbība. Tika arī veikts īss ieskats rasmošanas vēsturē. Iepazīstoties ar pieejamo informāciju, autoram bija saprotams, kā būs nepieciešams izstrādāt savu rasmotāju.

Praktiskās daļas sākuma daļā tika veikts īss apskats brīvi pieejamiem rasmotājiem, lai vēlāk varētu salīdzināt ar pašizstrādātā rasmotāja iegūtajiem rezultātiem. Tika apskatīti trīs rasmotāji – webscraper.io, ParseHub un Octoparse. Visi strādāja pēc viena principa, lietotājam ar selektoru palīdzību izvēloties sev nepieciešamos datus izgūšanai. Pēc rasmotāju apskates un datu izgūšanas mēģinājuma no Latvijas Atvērto datu portāla rezultāti bija ļoti dažādi. Webscarper.io rasmotāja iegūtais CSV fails nebija lietojams. ParseHub ļoti labi un kvalitatīvi izguva datus no pirmās lapas, bet bija problēmas ar nokļūšanu uz nākamo, tāpēc CSV fails nav pilnīgs. Octoparse ļoti labi pārsteidza autoru ar labi izstrādāto grafisko saskarni un viegli iegūstamajiem datiem. CSV failā izgūtie dati bija kvalitatīvi un pilnīgi.

Pēc visas izpētes un izpildītā darba autors noprojektēja un izstrādāja savu rasmotāju, kas ir paredzēts Latvijas Atvērto datu portālam. Rasmotājs ir uzprogrammēts Python valodā, izmantojot Python 3 3.7 versiju. Papildus rasmotāja izstrādei tika izmantotas četras bibliotēkas – beautifulsoup4, requests, lxml un csv. Autoram vajadzēja tikt galā ar dažādām kļūdām un nestandarta situācijām. Rasmotājs tika veiksmīgi izstrādāts un pēc tā darbības beigām tiek iegūts CSV fails, kurā tiek apkopoti sekojošie dati no Latvijas Atvērto datu portāla: pilnais datu kopas nosaukums, pilnais datu kopas apraksts, datu kopas publicētājs jeb organizācija, kategorijas, failu formāti, skatījumi, birkas un saite uz datu kopu.

SECINĀJUMI

Iepazīstoties un apkopojot informāciju par atvērtajiem datiem un to vēsturi, kas iegūta no dažādiem literatūras avotiem, tika veikti vairāki secinājumi. Atvērtajiem datiem ir jābūt bez maksas, ērti un brīvi visiem pieejamiem. Atvērto datu kustība jau ilgi pastāvēja pirms termins "atvērtie dati" tika oficiāli izmantots 1995. gadā. Datu brīva piekļuve ir ļoti svarīga mūsdienās, jo daudzu uzņēmumu darbība un dažādu pētījumu veikšana tieši balstās uz atvērtajiem datiem.

Publicēto datu kopu skaits katru dienu mainās Latvijas Atvērto datu portālā ar tendenci uz datu kopu skaita palielināšanos. Datu kopas tiek pievienotas, izdzēstas un rediģētas ik dienu.

Tīmekļa rasmošana ir ļoti labs, ērts un ātrs veids kā iegūt lietotājam nepieciešamo informāciju lasāmā formātā, piemēram CSV failā. Lai rasmošana būtu veiksmīga, ir nepieciešama speciāla programma – rasmotājs. Rasmotājs nolasa tīmekļa vietni un izgūst no tās nepieciešamos datus. Rasmotāji var ļoti palīdzēt datu entuziastiem un pētniekiem ar ātro datu ieguvi, lai viņi varētu vairāk laika veltīt datu analīzei. Tika dziļāk apskatīti trīs bez maksas pieejamie rasmotāji. Pēc visu rasmotāju testēšanas var secināt, ka vislabāk ar doto uzdevumu tika galā Octoparse rasmotājs. Tomēr iegūtais CSV fails autoru ne līdz galam apmierināja, jo nebija iespējas iegūta visu vēlamo informācija.

Testēto rasmotāju problēma bija to universālums, tāpēc autors izstrādāja savu rasmotāju, kas paredzēts tieši Latvijas Atvērto datu portāla rasmošanai. Kopumā rasmotāja izstrāde notika veiksmīgi un visas autoram norādītās prasības tika izpildītas. Protams, tika sastaptas dažas problēmas un šķēršļi, bet tie tika pārvarēti. Ja rasmotājs tiek mazliet modificēts, pielāgojoties citai tīmekļa saitei un citam HTML kodam, to var izmantot, lai iegūtu nepieciešamos datus no kādas citas valdības atvērto datu portāla. Autors uzskata, ka izstrādātais rasmotājs var būt palīgs datu entuziastiem un pētniekiem, kas analizē Atvērto datu portālus. Atbalstot atvērtības principus, rasmotāja pirmkods ir publicēts GitHub repozitorijā: https://github.com/Latviano123/rasmotajs brīvai izmantošanai, tā turpinot atvērto datu kustību. Autors plāno nākotnē uzlabot rasmotāju, kā arī izstrādāt grafisko saskarni vieglākai izmantošanai.

Visi darba sākumā izvirzītie uzdevumi tika izpildīti un mērķis tika sasniegts.

IZMANTOTĀ LITERATŪRA UN AVOTI

1. Place overview [tiešsaiste] – [atsauce 28.04.2021.] – pieejams: https://index.okfn.org/place/

2. The Open Definition [tiešsaiste] – [atsauce 29.04.2021.] – pieejams: https://opendefinition.org/

3. "No Rights Reserved" [tiešsaiste] – [atsauce 29.04.2021.] – pieejams: https://creativecommons.org/share-your-work/public-domain/cc0/

4. Open Definition 2.1 [tiešsaiste] – [atsauce 30.04.2021.] – pieejams: https://opendefinition.org/od/2.1/en/

5. What is Open Data? [tiešsaiste] – [atsauce 30.04.2021.] – pieejams: https://opendatahandbook.org/guide/en/what-is-open-data/

6. Open Data & Metadata Quality. [tiešsaiste] – [atsauce 06.05.2021.] – pieejams: https://www.europeandataportal.eu/sites/default/files/d2.1.2_training_module_2.2_open_data_quality_en_edp.pdf

7. Open Data Metadata Guide [tiešsaiste] – [atsauce 06.05.2021.] – pieejams: https://centerforgov.gitbooks.io/open-data-metadata-guide/content/

8. Overview [tiešsaiste] – [atsauce 07.05.2021.] – pieejams: https://data.europa.eu/mqa?locale=en

9. Categories [tiešsaiste] – [atsauce 07.05.2021.] – pieejams: https://centerforgov.gitbooks.io/open-data-metadata-guide/content/categories.html

10. Dataset Metadata [tiešsaiste] – [atsauce 08.05.2021.] – pieejams: https://centerforgov.gitbooks.io/open-data-metadata-guide/content/dataset-metadata.html

11. A brief history [tiešsaiste] – [atsauce 11.05.2021.] – pieejams: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6781855/

12. A brief history of Open Data. [tiešsaiste] – [atsauce 11.05.2021.] – pieejams: http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data#:~:text=The%20term%20open%20data%20appeared,from%20an%20American%20scientific%20agency.&text=The%20theory%20that%20bears%20his,be%20freely%20accessible%20to%20all

13. A brief history of open data. [tiešsaiste] – [atsauce 12.05.2021.] – pieejams: https://fcw.com/articles/2014/06/09/exec-tech-brief-history-of-open-data.aspx

14. Projekts "Publiskās pārvaldes informācijas un komunikācijas tehnoloģiju arhitektūras pārvaldības sistēma" (PIKTAPS) [tiešsaiste] – [atsauce 14.05.2021.] – pieejams: https://www.varam.gov.lv/lv/projekts/projekts-publiskas-parvaldes-informacijas-un-komunikacijas-tehnologiju-arhitekturas-parvaldibas-sistema-piktaps

15. Datu katalogs. [tiešsaiste] – [atsauce 14.05.2021.] – pieejams: https://data.gov.lv/dati/lv/dataset?_res_format_limit=0

16. Organizācijas. [tiešsaiste] – [atsauce 14.05.2021.] – pieejams: https://data.gov.lv/dati/lv/organization

17. Zhao, Bo. "Web scraping." *Encyclopedia of big data* (2017) [tiešsaiste] – [atsauce 15.05.2021.] – pieejams: https://www.researchgate.net/profile/Bo-Zhao-3/publication/317177787_Web_Scraping/links/5c293f85a6fdccfc7073192f/Web-Scraping.pdf

18. Web scraping. [tiešsaiste] – [atsauce 15.05.2021.] – pieejams: https://www.imperva.com/learn/application-security/web-scraping-attack/#:~:text=Web%20scraping%20is%20the%20process,replicate%20entire%20website%20content%20elsewhere

19. The History of web scraping. [tiešsaiste] – [atsauce 16.05.2021.] – pieejams: https://www.xbyte.io/the-history-of-web-scraping.php

20. Scraper [tiešsaiste] – [atsauce 17.05.2021.] - pieejams: https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd?hl=en

21. Automate data extraction with Web Scraper Cloud [tiešsaiste] – [atsauce 18.05.2021.] - pieejams: https://webscraper.io/pricing

22. A plan for all of your web scraping needs [tiešsaiste] – [atsauce 18.05.2021.] - pieejams: https://www.parsehub.com/pricing

23. Octoparse Premium Pricing & Packaging [tiešsaiste] – [atsauce 19.05.2021.] - pieejams: https://www.octoparse.com/pricing

24. Zuiderwijk, Anneke, Rhythima Shinde, and Marijn Janssen. "Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results?" *International Review of Administrative Sciences* 85.4 (2019) [tiešsaiste] – [atsauce 24.05.2021.] – pieejams: https://journals.sagepub.com/doi/pdf/10.1177/0020852317739115

25. Quarati, Alfonso, and Monica De Martino. "Open government data usage: A brief overview." *Proceedings of the 23rd International Database Applications & Engineering*

Symposium. 2019. [tiešsaiste] – [atsauce 24.05.2021.] – pieejams: https://dl.acm.org/doi/abs/10.1145/3331076.3331115

26. Programmēšanas pamati ar valodu Python, Jānis Zuters, Latvijas Universitāte, 2019-2020 [tiešsaiste] – [atsauce 24.05.2021.] – pieejams: http://home.lu.lv/~janiszu/courses/python/python3.pdf

27. Beautiful Soup Documentation [tiešsaiste] – [atsauce 24.05.2021.] – pieejams: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

28. Requests: HTTP for Humans™ [tiešsaiste] – [atsauce 24.05.2021.] – pieejams: https://docs.python-requests.org/en/master/

29. lxml - XML and HTML with Python [tiešsaiste] – [atsauce 24.05.2021.] – pieejams: https://lxml.de/

30. csv — CSV File Reading and Writing [tiešsaiste] – [atsauce 26.05.2021.] – pieejams: https://docs.python.org/3/library/csv.html#module-csv

PIELIKUMI

1. Pielikums

Latvijas Atvērto datu portāla pieejamo formātu tabula

2.1. tabula

Latvijas Atvērto datu portāla pieejamie datu kopu formātu top 16 (16.05.2021.) [15]

Nr.p.k.	Formāts	Formāta skaidrojums	Datu kopu skaits
1	CSV	Ar komatu vai semikolu atdalītas vērtības, ko parasti izmanto izklājlapām vai vienkāršām datu bāzēm.	234
2	XLSX	Microsoft Excel atvērtā XML izklājlapa	133
3	XLS	Microsoft Excel izveidots izklājlapu formāts	36
4	WMS	Plaši izmantots karšu un ĢIS datu formāts	31
5	JSON	JavaScript objektu notācija ir datu apmaiņas formāts	27
6	OData	Definē protokolu datu vaicājumiem un atjaunināšanai	19
7	DOCX	Microsoft Word atvērtā XML dokuments	17
	SHP	Ģeotelpisko vektoru datu formāts (ĢIS) programmatūrai	17
9	ZIP	Arhīva faila formāts	15
10	XML	Paplašināmā iezīmēšanas valoda	12
11	PDF	Faila formāts, kurā visi izdrukātā dokumenta elementi ir tverti kā elektronisks attēls	10
12	HTML	Hiperteksta iezīmēšanas valoda	7
13	ODS	Calc programmas izklājlapa	3
	REST	Arhitektūras stils	3
	SAV	Faila formāts SPSS datu failiem	3
	TSV	Ar tabulatoru atdalītas vērtības, līdzīgs CSV	3

2. Pielikums

Latvijas Atvērto datu portāla organizāciju tabula

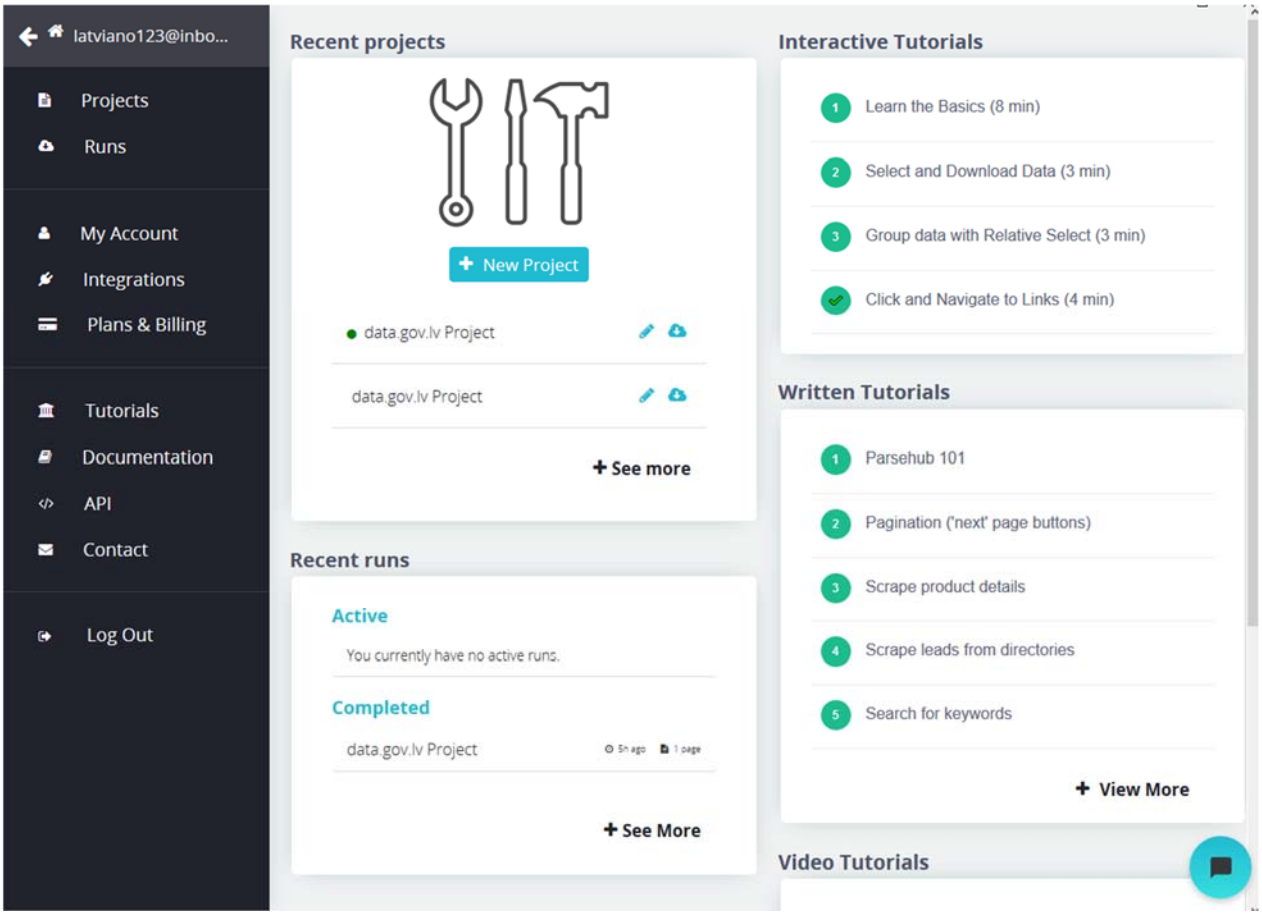
2.2. tabula

Latvijas Atvērto datu portāla top 15 organizācijas pēc publicēto datu kopu skaita (16.05.2021.) [16]

Nr.p.k.	Organizācija	Datu kopu skaits
1	ĢEOLatvija.lv	42
2	Centrālā statistikas pārvalde	41
3	Būvniecības valsts kontroles birojs	28
	Valsts reģionālās attīstības aģentūra	28
5	Iepirkumu uzraudzības birojs	24
6	Vides aizsardzības un reģionālās attīstības ministrija	20
7	Izglītības un zinātnes ministrija	18
8	VSIA "Latvijas Valsts ceļi"	15
9	Cēsu novada pašvaldība	13
	Rīgas dome	13
	Valsts zemes dienests	13
12	Labklājības ministrija	12
	Latvijas Institūts	12
	Pilsonības un migrācijas lietu pārvalde	12
15	LR Uzņēmumu reģistrs	10
	Valsts ieņēmumu dienests	10

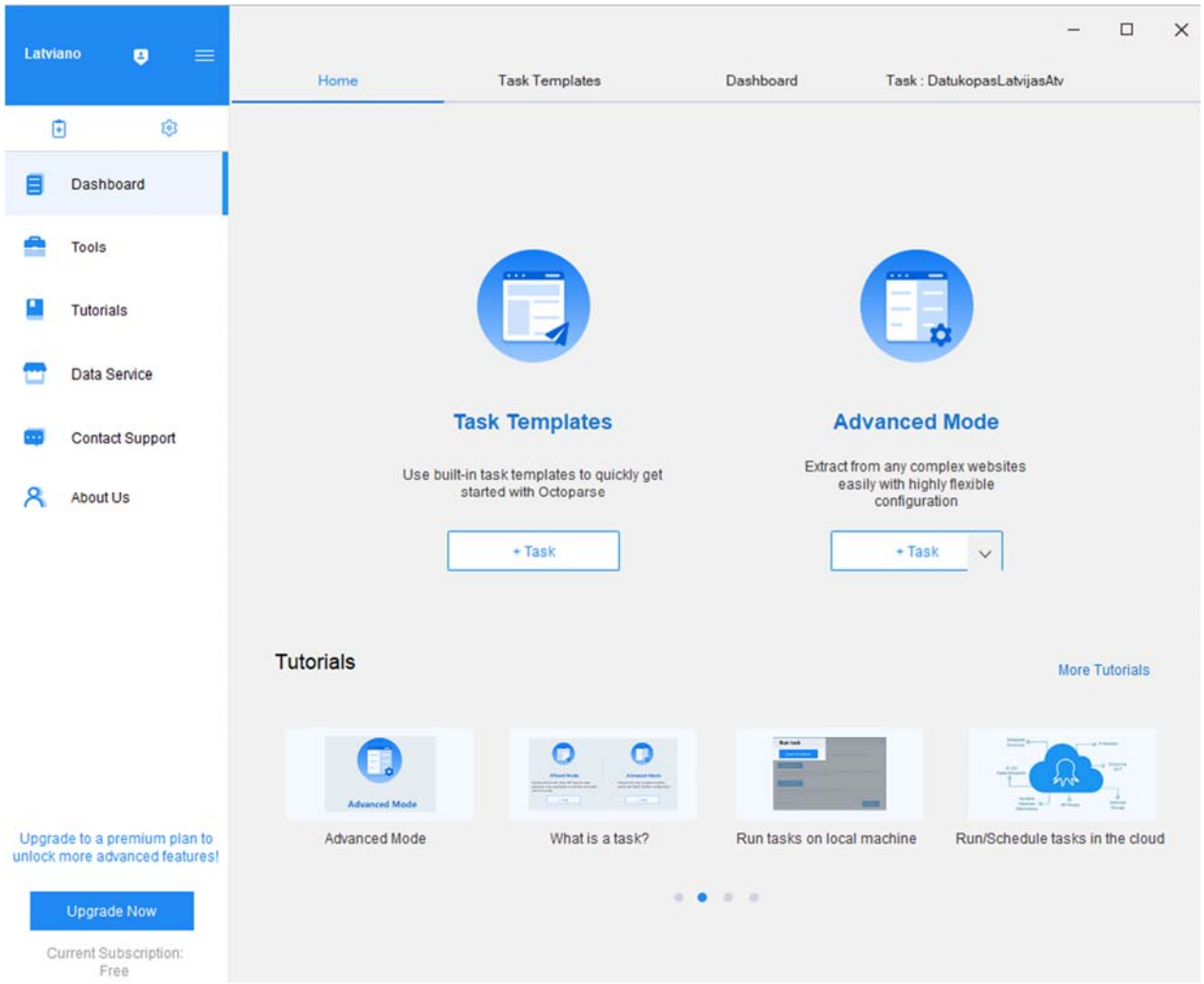
3. Pielikums

ParseHub rasmotāja galvenais skats



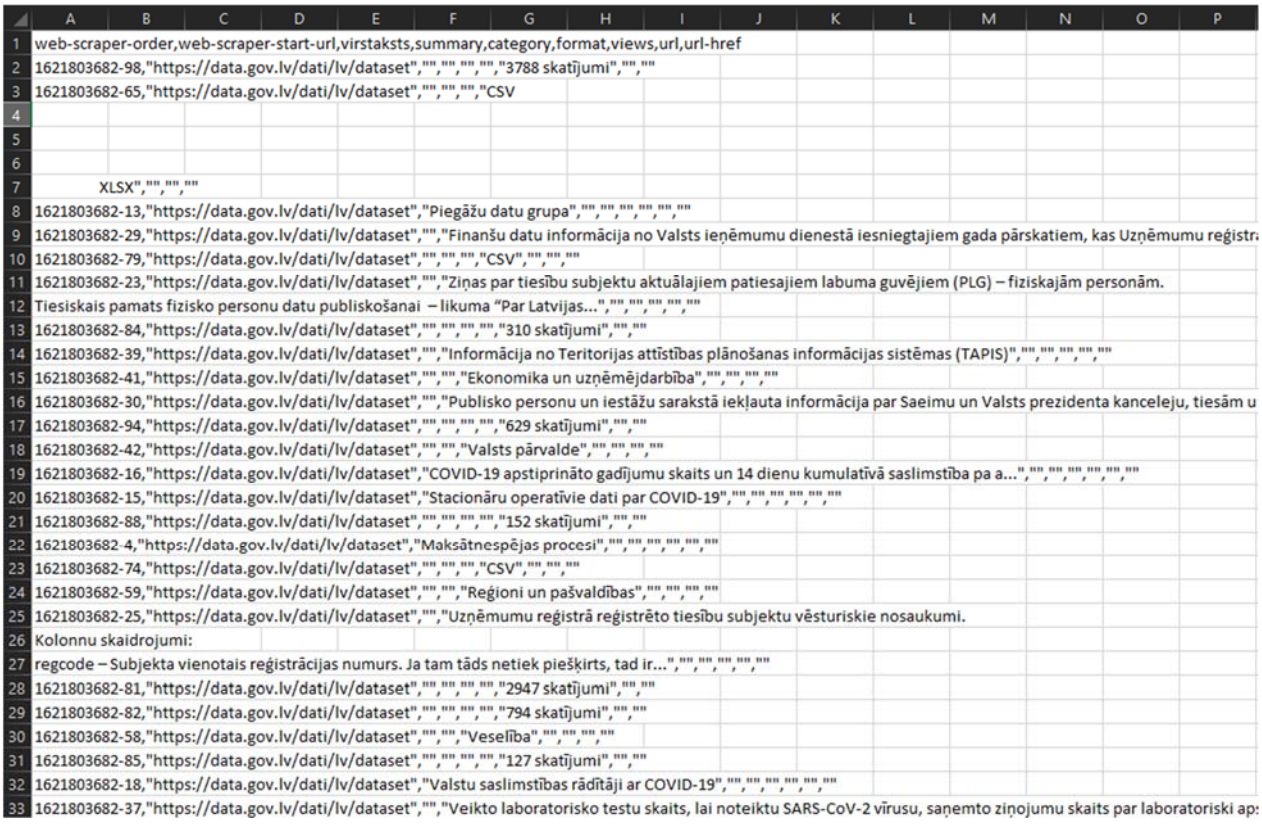
4. Pielikums

Octoparse rasmotāja galvenais skats



5. Pielikums

Fragments no webscraper.io bezmaksas rasmotāja iegūtā CSV faila



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	web-scraper-order,web-scraper-start-url,virstaksts,summary,category,format,views,url,url-href															
2	1621803682-98,"https://data.gov.lv/dati/lv/dataset","","","","","3788 skatījumi","",""															
3	1621803682-65,"https://data.gov.lv/dati/lv/dataset","","","","CSV															
4																
5																
6																
7	XLSX","","",""															
8	1621803682-13,"https://data.gov.lv/dati/lv/dataset","Piegāžu datu grupa","","","","","",""															
9	1621803682-29,"https://data.gov.lv/dati/lv/dataset","","Finanšu datu informācija no Valsts ieņēmumu dienestā iesniegtajiem gada pārskatiem, kas Uzņēmumu reģistr															
10	1621803682-79,"https://data.gov.lv/dati/lv/dataset","","","","CSV","","",""															
11	1621803682-23,"https://data.gov.lv/dati/lv/dataset","","Ziņas par tiesību subjektu aktuālajiem patiesajiem labuma guvējiem (PLG) – fiziskajām personām.															
12	Tiesiskais pamats fizisko personu datu publiskošanai – likuma "Par Latvijas...","","","","","",""															
13	1621803682-84,"https://data.gov.lv/dati/lv/dataset","","","","","310 skatījumi","",""															
14	1621803682-39,"https://data.gov.lv/dati/lv/dataset","","Informācija no Teritorijas attīstības plānošanas informācijas sistēmas (TAPIS)","","","","","",""															
15	1621803682-41,"https://data.gov.lv/dati/lv/dataset","","","Ekonomika un uzņēmējdarbība","","","",""															
16	1621803682-30,"https://data.gov.lv/dati/lv/dataset","","Publisko personu un iestāžu sarakstā iekļauta informācija par Saeimu un Valsts prezidenta kanceleju, tiesām u															
17	1621803682-94,"https://data.gov.lv/dati/lv/dataset","","","","","629 skatījumi","",""															
18	1621803682-42,"https://data.gov.lv/dati/lv/dataset","","","Valsts pārvalde","","","",""															
19	1621803682-16,"https://data.gov.lv/dati/lv/dataset","COVID-19 apstiprināto gadījumu skaits un 14 dienu kumulatīvā saslimstība pa a...","","","","","",""															
20	1621803682-15,"https://data.gov.lv/dati/lv/dataset","Stacionāru operatīvie dati par COVID-19","","","","","",""															
21	1621803682-88,"https://data.gov.lv/dati/lv/dataset","","","","","152 skatījumi","",""															
22	1621803682-4,"https://data.gov.lv/dati/lv/dataset","Maksātnespējas procesi","","","","","",""															
23	1621803682-74,"https://data.gov.lv/dati/lv/dataset","","","","CSV","","",""															
24	1621803682-59,"https://data.gov.lv/dati/lv/dataset","","Reģioni un pašvaldības","","","",""															
25	1621803682-25,"https://data.gov.lv/dati/lv/dataset","","Uzņēmumu reģistrā reģistrēto tiesību subjektu vēsturiskie nosaukumi.															
26	Kolonnu skaidrojumi:															
27	regcode – Subjekta vienotais reģistrācijas numurs. Ja tam tāds netiek piešķirts, tad ir...","","","","","",""															
28	1621803682-81,"https://data.gov.lv/dati/lv/dataset","","","","","2947 skatījumi","",""															
29	1621803682-82,"https://data.gov.lv/dati/lv/dataset","","","","","794 skatījumi","",""															
30	1621803682-58,"https://data.gov.lv/dati/lv/dataset","","","Veselība","","","",""															
31	1621803682-85,"https://data.gov.lv/dati/lv/dataset","","","","","127 skatījumi","",""															
32	1621803682-18,"https://data.gov.lv/dati/lv/dataset","Valstu saslimstības rādītāji ar COVID-19","","","","","",""															
33	1621803682-37,"https://data.gov.lv/dati/lv/dataset","","Veikto laboratorisko testu skaits, lai noteiktu SARS-CoV-2 vīrusu, saņemto ziņojumu skaits par laboratoriski ap															

6. Pielikums

Fragments no ParseHub rasmotāja iegūtā CSV faila

	A	B	C	D	E	F
1	dataitem_title	dataitem_title_url	dataitem_summary	dataitem_category	dataitem_forma	dataitem_views
2	Oficiālā elektroniskā adrese (e-adrese)	https://data.gov.lv/dati/lv/dataset/eadrese	Informācija par Oficiālās elektroniskās adreses aktivizētajiem, deaktivizētajiem kontiem un anulētajām e-adresēm. Dati tiek atjaunoti reizi stundā.	Ekonomika un uzņēmējdarbība	CSV	2947 skatījumi
3	Vakances	https://data.gov.lv/dati/lv/dataset/vakances	Uzņēmumu un organizāciju izsludinātās aktuālās vakances Latvijas valsts sektorā.	Ekonomika un uzņēmējdarbība	CSV	2561 skatījumi
4	Iepirkumu rezultāti (e-konkursi)	https://data.gov.lv/dati/lv/dataset/iepirkumu-rezultatu-datu-grupa	Informācija par Elektronisko iepirkumu sistēmā publicētājiem līgumiem	Valsts pārvalde	CSV	2226 skatījumi
5	Iepirkumu grozījumi (e-konkursi)	https://data.gov.lv/dati/lv/dataset/iepirkumu-grozijumu-datu-grupa	Informācija par Elektronisko iepirkumu sistēmā publicēto iepirkumu grozījumiem	Valsts pārvalde	CSV	794 skatījumi
6	Patiesie labuma guvēji	https://data.gov.lv/dati/lv/dataset/patiesie-labuma-guveji	Ziņas par tiesību subjektu aktuālajiem patiesajiem labuma guvējiem (PLG) – fiziskajām personām. Tiesiskais pamats fizisko personu datu publiskošanai – likuma "Par Latvijas...	Ekonomika un uzņēmējdarbība	CSV	709 skatījumi
7	Maksātnespējas procesi	https://data.gov.lv/dati/lv/dataset/maksatnespejas-procesi	Uzņēmumu reģistra Maksātnespējas reģistrā ierakstīto maksātnespējas un tiesiskās aizsardzības procesu informācija. Maksātnespējas procesu atvērto datu lauku skaidrojumi – spied...	Ekonomika un uzņēmējdarbība	CSV XLSX	310 skatījumi
8	Tiesību subjektu vēsturiskie nosaukumi	https://data.gov.lv/dati/lv/dataset/tiesibu-subjektu-vesturiskie-nosaukumi	Uzņēmumu reģistrā reģistrēto tiesību subjektu vēsturiskie nosaukumi. Kolonnu skaidrojumi: regcode – Subjekta vienotais reģistrācijas numurs. Ja tam tāds netiek piešķirts, tad ir...	Ekonomika un uzņēmējdarbība	CSV XLSX	127 skatījumi
9	Biedrību un nodibinājumu darbības jomas	https://data.gov.lv/dati/lv/dataset/biedribu-un-nodibinajumu-darbibas-jomas	Uzņēmumu reģistrā reģistrēto biedrību un nodibinājumu darbības jomas atbilstoši Ministru Kabineta noteiktajai klasifikācijai. Kolonnu skaidrojumi: regcode – Subjekta vienotais...	Ekonomika un uzņēmējdarbība	CSV XLSX	254 skatījumi
10	Dati par sabiedrību ar ierobežotu atbildību dalībniekiem	https://data.gov.lv/dati/lv/dataset/members	Dati par sabiedrību ar ierobežotu atbildību dalībniekiem.	Ekonomika un uzņēmējdarbība	CSV	290 skatījumi
11	Tiesību subjekta valdes locekļu, pārstāvēttiesīgo biedru vai citu pārstāvētti...	https://data.gov.lv/dati/lv/dataset/officers	Tiesību subjekta valdes locekļu, pārstāvēttiesīgo biedru vai citu pārstāvēttiesīgo amatpersonu dati.	Ekonomika un uzņēmējdarbība	CSV	152 skatījumi

7. Pielikums

Fragments no Octoparse rasmotāja iegūtā CSV faila

	Title	Summary	Category	Format	Views	URL
476	Dati par jaunlaulātajiem un laulībām Rīgā	Apkopoti dati par Rīgā noslēgto laulību skaitu, laulību reģistrācijas iestādi, laulāto vecumu, kā arī to, kura laulība pēc skaita tā ir laulātajam.Apkopotie datu lauki:id –...	Iedzīvotāji un sabiedrība	OData	445 skatījumi	https://data.gov.lv/dati/lv/dataset/dati-par-ja
477	Rīgā deklarēto personu skaits	Datu kopa, kurā reizi diennaktī tiek attēlots Rīgas pašvaldībā deklarēto personu skaits, to izmaiņas pa dienām. Datus iespējams izmantot, lai analizētu deklarēšanās tendences...	Iedzīvotāji un sabiedrība	OData	446 skatījumi	https://data.gov.lv/dati/lv/dataset/riga-deklar
478	Statistika par saziņu ar Rīgas pašvaldību	Datu kopa par Rīgas pašvaldības klientu pieteikumiem. Datus iespējams izmantot, lai analizētu tendences un sagatavotu pārskatus par kurām tēmām, kurās iestādēs vēršas klienti,...	Iedzīvotāji un sabiedrība	OData	579 skatījumi	https://data.gov.lv/dati/lv/dataset/statistika-p
479	Uzaicināto un uzņemto bērnu skaits Rīgas pašvaldības pirmsskolas izglītības i...	Datu kopa, kurā reizi diennaktī tiek attēloti pašvaldības pirmsskolas izglītības iestādēs uzaicinātie un uzņemtie bērni pa kalendārajiem gadiem. Tekošajā kalendārajā gadā tiek...	Iedzīvotāji un sabiedrība	OData	687 skatījumi	https://data.gov.lv/dati/lv/dataset/uzaicinato-
480	Valsts finansējuma un Rīgas pašvaldības piemaksas par uztura korekcijas izlie...	Datu kopa, kurā reizi mēnesī tiek attēlots valsts finansējuma un Rīgas pašvaldības piemaksas par uztura korekcijas izlietojums skolēnu ēdināšanai Rīgas pašvaldības vispārējās...	Iedzīvotāji un sabiedrība	OData	184 skatījumi	https://data.gov.lv/dati/lv/dataset/valsts-finar
481	Rīgas pašvaldības piešķirtais ēdināšanas pabalsts maznodrošināto un trūcīgo ģ...	Datu kopa, kurā reizi mēnesī tiek attēlots Rīgas pašvaldības piešķirtais ēdināšanas pabalsts maznodrošināto un trūcīgo ģimeņu skolēniem Rīgas pašvaldības vispārējās izglītības...	Iedzīvotāji un sabiedrība	OData	281 skatījumi	https://data.gov.lv/dati/lv/dataset/pasvaldibas
482	Dati par Rīgas privāto pirmsskolas izglītības iestāžu tāmēm	Datu kopa, kurā reizi diennaktī tiek attēlotas Rīgas privāto pirmsskolas izglītības iestāžu tāmes.Apkopotie datu lauki:id – ieraksta identifikatorsinstitution_id – iestādes...	Izglītība un sports	OData	257 skatījumi	https://data.gov.lv/dati/lv/dataset/dati-par-pr
483	Rīgas pašvaldības finansējuma izlietojums bērnu uzraudzības pakalpojumam	Datu kopa, kurā reizi mēnesī tiek attēlots Rīgas pašvaldības finansējuma izlietojums bērnu uzraudzības pakalpojumam. Finansējums tiek piešķirts bērniem, kam netika nodrošināta...	Iedzīvotāji un sabiedrība	OData	157 skatījumi	https://data.gov.lv/dati/lv/dataset/pasvaldibas

Rasmotāja datu izgūšanas cikla kods

```
for article in soup.find_all('li',class_='dataset-item'):

    urls = 'https://data.gov.lv' + str(article.h3.a.get('href'))
    print(urls)

    r_urls = requests.get(urls).text
    zupa = BeautifulSoup(r_urls, 'lxml')

    try:
        headline = zupa.find('h1', class_='heading').text
        print(headline)
    except:
        headline = None
        headline1 = article.h3.a.text
        print(headline1)

    try:
        summary = zupa.find('div', class_='notes').text.strip()
        print(summary)
    except:
        summary = None
        summary1 = article.find('div',class_='dataset-content').div.text
        print(summary1)

    try:
        organization = zupa.find('section', class_='module-
content').h1.text.strip()
        print(organization)
    except:
        organization = None

    col = article.find('div', class_='col-sm-3')
    span_tag = col.select('span')

    try:
        categ = zupa.find(lambda tag: tag.name == 'a' and ('Groups' in
tag.text or 'Kategorijas' in tag.text))
        cat = 'https://data.gov.lv' + str(categ.get('href'))
        r_cat = requests.get(cat).text
        zupa_cat = BeautifulSoup(r_cat, 'lxml')

        j = 0
        category = ''
        categ = zupa_cat.find('ul', class_='media-grid')
        for cat_h3 in categ.find_all('h3'):
            category += cat_h3.text + ', '
            j += 1
            if j == len(categ.find_all('h3')): print(category, end='\n')
    except:
        category = span_tag[0].text.strip()
        print(category)

    try:
        i = 0
        forma = ''
```

```
        col_ul = col.find('ul', class_='dataset-resources')
        for col_li in col_ul.find_all('li'):
            forma += col_li.a.text + ', '
            i += 1
            if i == len(col_ul.find_all('li')): print(forma, end='\n')
    except:
        form = None

    try:
        k = 0
        tag = ''
        sec_tag = zupa.find('ul', class_='tag-list')
        for tag_li in sec_tag.find_all('li'):
            tag += tag_li.a.text + ', '
            k += 1
            if k == len(sec_tag.find_all('li')): print(tag, end='\n')
    except:
        tag = None

    views = span_tag[1].text.strip()
    print(views)

    print()

    csv_writer.writerow([headline or headline1, summary or summary1,
organization, category, forma, views, tag, urls])
```

Bakalaura darbs „Rasmotāja izstrāde Latvijas Atvērto datu portāla analīzei” izstrādāts LU Datorikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti.

Autors: Justs Ķikuts 31.05.2021.

Rekomendēju darbu aizstāvēšanai

Vadītāja: docente, Dr.sc.comp. Anastasija Ņikiforova 31.05.2021.

Recenzents: docents, Dr.sc.comp. Leo Trukšāns

Darbs iesniegts Datorikas fakultātē 31.05.2021.

Dekāna pilnvarotā persona: vecākā metodiķe Ārija Sproģe

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē

__.06.2021. prot. Nr. ___.

Komisijas sekretāre: docents Dr.sc.comp. Aivars Niedrītis