

Webcrawling metodo taikymas automatizuotai nelegalių skelbimų paieškai Aplinkos apsaugos departamente

Problema

Aplinkos apsaugos departamente (AAD) atliekama galimai nelegalių skelbimų paieška vykdoma rankiniu būdu, todėl yra lėta, neišnaudojamos informacinių sistemų (IS) efektyvumo galimybės, neefektyviai naudojami žmogiškieji ištekliai.

Tikslas

Pagreitinti ir automatizuoti galimai nelegalių skelbimų paiešką ir pateikti rezultatus peržiūrai. Įrankis turi veikti kaip webservisas, formuojantis duomenis kurie tinka integracijai į kitas sistemas.

Rizikos, priklausomybės

- Automatizuotos sistemos, iš kurios būtų galima gauti informaciją apie asmens/įstaigos turimus/turėtus leidimus skirtingose institucijose, nebuvimas: daugelio AAD prižiūrimų veiklų vykdymui yra reikalingi leidimai iš įvairių institucijų, visos jos informaciją saugo skirtingais formatais ir gali būti sunku integruoti šią informaciją į webcrawler. Be galimybės automatiškai patikrinti skelbime esančią kontaktinę informaciją su leidimų registrais bus gaunamas didesnis skelbimų kiekis, kuriuos teks rankiniu būdu papildomai peržiūrėti AAD darbuotojams;
- Webcrawler įrankio kūrėjai – noras bendradarbiauti, galimybės sukurti AAD tinkamą produktą;
- Skelbimų portalai – noras bendradarbiauti dėl papildomų privalomų laukų skelbimuose sukūrimo;
- Tęstinumo užtikrinimo trūkumas – neaišku kas, kaip prižiūrės įrankį ilgalaikėje perspektyvoje.

Įgyvendinimas

- Surinkti prioretizuotų sričių, kurių internetiniuose skelbimuose galima pritaikyti webcrawler įrankį, sąrašą:
 - Aukštesnį prioritetą turėtų akivaizdžiai nelegalūs skelbimai (pvz.: draudžiama prekybinė veikla, kuriai nėra leidimų, draudžiamų gyvūnų, augalų pardavimas ir pan.).
- Išanalizuoti skelbimų, pasitvirtinusių kaip nelegalių, specifiką (raktažodžiai, nuotraukos, telefono numeriai);
- Sukurti identifikatorių grupes (raktažodžių sąrašus) kiekvienai iš prioritetinių sričių;
- Išsiaiškinti dėl integracijos galimybių į aplinkosaugos leidimų IS (<https://www.alisas.lt/permits>), kad webcrawler galėtų pasiimti leidimų informaciją iš šios IS;
- Išsiaiškinti dėl galimybės pridėti papildomus privalomus laukus skelbimų portaluose tai veiklai, kuriai privaloma turėti aplinkosaugos leidimus;
- Suderinti webcrawler rezultatų formatą su AADIS ir įrankio kūrėjais;
- Suderinti techninius ir organizacinius projekto įgyvendinimo aspektus su webcrawler įrankio kūrėjais;
- Pilotinės webcrawler versijos, kuri apimtų aukštesnio prioriteto sritis, parengimas naudoti ir testavimas;
- Atsižvelgus į naudotojų ir susijusių asmenų atsiliepimus, pagal poreikį, padaryti pakeitimai po pilotinės webcrawler versijos ir paleistas daugiau sričių apimantis webcrawler, kuris turėtų integraciją į ALIS;
- Numatyti webcrawler priežiūros, raktažodžių atnaujinimo procesą.

Įvykdymo kriterijai (acceptance criteria)

- Surinktas prioretizuotų sričių, kurių skelbimuose galima pritaikyti webcrawler įrankį, sąrašas.
- Išanalizuota skelbimų, pasitvirtinusių kaip nelegalių, specifiška ir paruošti jų aprašymai su įvardintais geriausius rezultatus atnešusiais metodais:
 - Išanalizuotos ir aprašytos skirtingų Lietuvos institucijų patirtys naudojant webcrawling metodą;

- Išanalizuotos ir aprašytos specialistų, dirbančių su nelegaliais aplinkosaugos skelbimais, patirtys.
- Sukurtos identifikatorių grupės (raktažodžių sąrašai) kiekvienai iš prioritetinių sričių;
- Skelbimų portaluose pridėti papildomi privalomi laukai toms sritims, kurių veiklai privaloma turėti aplinkosaugos leidimus;
- Užtikrinta galimybė integruoti ALIS leidimų tikrinimą į webcrawler su susijusiomis šalimis ir įrankio kūrėjais;
- Paleistas webcrawler:
 - kuris periodiškai pateiktų galimai nelegalių skelbimų sąrašus;
 - teiktų ataskaitas tinkamu formatu peržiūrai ir importavimui į AADIS bei kitas išorines sistemas.
- Sukurtas dokumentuotas procesas, leidžiantis užtikrinti tęstinumą - techninė priežiūra ir periodiškas raktažodžių atnaujinimas.