VALTIOVARAINMINISTERIÖ

Kartellitutka-projektin loppuraportti

Kilpailu- ja kuluttajavirasto

VM/2343/02.02.03.09/2018

Versio 1.0

30.9.2019

Sisällys

Dokumentin versiohistoria

Versio	Päiväys	Laatija	Muutoksen kuvaus
0.1	30.7.2019	Marian Jecu	Luonnos
1.0	30.9.2019	Peter Berglund	Raportin viimeistely

1. Yhteenveto

Tämä dokumentti on uuden toimintamallin tai teknologiaratkaisun toiminnan todentamiseen tähtäävän Kartellitutka-projektin loppuraportti.

2. Kokeilun toteutuminen

2.1. Kokeilun tiedot

Proof-of-Concept (POC) -projektin tavoitteena oli tehostaa Kilpailu- ja kuluttajaviraston (jäljempänä KKV) kartellivalvontaa toteuttamalla julkisia lähteitä hyödyntävän teknisen ratkaisun. POC:n avulla on voitu testata käytännössä eri teknologioiden, ohjelmistojen ja menetelmien soveltumista tehostettuun kartellivalvontaan.

Projekti toteutettiin suunnitelman mukaisesti. Valittu toimittaja toteutti projektin käyttäen seuraavia menetelmiä ja työkaluja:

- Python Scrapy-nimistä web scraping -kirjasto
- SAS Visual Investigator
- SAS Visual Text Analytics
- SAS Visual Analytics
- SAS Studio

Lähdedatana käytettiin valitun toimialan tiettyjen toimijoiden verkkosivuilta mm. ajankohtaistiedotteita, lehdistötiedotteita ja blogitekstejä. Lisäksi käytettiin KKV toimittamia PDF-tiedostoja, joissa oli valittujen yritysten vastuuhenkilöiden yrityssidokset ja roolit.

Tiedonhankintaa tehtiin siten, että tietolähteitä kerättiin automaattisesti tutkinnan kohteena olevien yritysten verkkosivuilta sekä uutissivustoilta tiedonhakua tekevillä ohjelmilla, ns. spidereillä. Näitä rakennettiin Python-ohjelmointikielen Scrapy -kirjastojen avulla.

Tiedotteiden ja muun kerätyn tekstimateriaalin määrä vaihteli toimijoittain. Sivustojen rakenne ja toteutus vaikuttivat siihen, miten hyvin datankeruu onnistui.

Projektissa käytettiin tekstianalytiikkaa ja verkostoanalyysia. Tekstianalytiikka kohdistettiin toimijoiden verkkosivuilta kerättyyn materiaaliin ja verkostoanalyysia KKV:n toimittamiin PDF-tiedostoihin, joissa oli valittujen yritysten vastuuhenkilöiden yrityssidokset ja roolit.

Tekstianalytiikassa kerätyn datan luonnollinen kieli pilkottiin automaattisesti yksittäisiksi sanoiksi, joiden eri taivutusmuodot tunnistettiin automaattisesti. Eri sanojen välisiä yhteyksiä tarkasteltiin termikarttojen avulla. Seuraavassa vaiheessa tekstianalytiikka tunnisti teksteistä automaattisesti siellä yleisesti esiintyviä aiheita ja kiinnostavista aiheista luotiin kategorioita. Kategorioiden avulla pystyttiin yksinkertaisten kielellisten sääntöjen avulla tunnistamaan uusista teksteistä kiinnostaviin aiheisiin liittyvät maininnat.

Verkostoanalyysi puolestaan mahdollistaa yritysten ja/tai henkilöiden välisten yhteyksien löytymisen nopeasti. Tämä tapahtui SAS:n automatisoiduilla työkaluilla (Visual Analytics ja Visual Investigator).

Merkittävänä erona hakemukseen nähden voidaan pitää tavoiteltua suppeamman tietomäärän saamista analysoitavaksi. Tavoiteltua suppeampi tietomäärä

johtui tiedon keräämisen tavoiteltua työläämmästä keräämisprosessista. Valitun toimittajan vahvuusalueet olivat datan analysoinnissa.

2.2. Kokeilun rahoitus, kustannukset ja henkilötyöpäivät

Kokeilun suunnitellut (käyttö- ja kirjausoikeuspäätöksen mukaiset) ja toteutuneet kustannukset euroina ovat eriteltyinä omaan ja ostettuun työhön sekä muihin kustannuksiin seuraavat:

Kustannus	**Suunniteltu €**	**Toteutunut €**
Oma työ (nykyresursseilla tehtävä työ)	5000	11574
Oma työ (kokeiluun erikseen palkattavien resurssien työ)	0	0
Palvelujen ostot	21000	21000
Muut kustannukset	0	0
Kokonaiskustannus	**26000**	**32574**

Kustannukset eriteltynä rahoituslähteittäin euroina ovat seuraavat:

Rahoituslähde	**Suunniteltu €**	**Toteutunut €**
28.70.22 Hallinnon palveluiden digitalisoinnin tuki	21000	21000
51901-KARTELLITUTKA Karetellitutka 166.2019 (oma työ)	5000	11574
Kokonaiskustannus	**26000**	**32574**

Kokeilun kustannukset ovat pysyneet täysin suunnitelman mukaan VM:n rahoituksen osalta. Viraston sisäisen työmäärän toteutuneet kustannukset ovat ylittyneet suunniteltuun nähden. Kokeilun kustannusten ylitys johtuu pääosin seuraavista tekijöistä:

- Projektiin käytettävää omaa työaikaa kyettiin arvioimaan etukäteen ainoastaan suuntaa-antavasti projektin kokeellisesta luonteesta johtuen.
- On lisäksi huomioitava, että projektissa päädyttiin järjestämään kilpailutus. Tarjoajien välisien erojen vaatima oma työaika oli vahvasti riippuvainen kunkin toimittajan tarjoamien palveluiden yksityiskohdista, joten eri toimittajien tarjoamat palvelut edellyttävät erilaista määrää omaa työtä. Oman työn määrä riippui osaltaan kilpailutuksen tuloksesta.

Oman, kokeiluun erikseen palkatun henkilöstön toteutunut kustannus euroina ja henkilötyöpäivinä:

€	**htp**
0	0

2.3. Hankintakäytännöt

Hankinta toteutettiin viraston sisäiset hankintaohjeet huomioiden kilpailuttamalla, vaikka hankintalain kilpailuttamista edellyttävän kynnysarvon (60 000 euroa) ei arvioitukaan ylittyvän.

Kilpailutusta valmistelevassa vaiheessa KKV toteutti markkinakartoituksen, joilla haluttiin selvittää potentiaaliset markkinatoimijat. Kartoitusta koskeva kysely lähetettiin seitsemälle alan toimijalle. Kartoituksen tulosten pohjalta KKV lähetti tarjouspyynnön pienhankinnasta 28.3.2019 viidelle toimijalle.

Määräaikaan mennessä tarjouksia tuli kaksi kappaletta: F-Secure Cyber Security Services Oy:ltä sekä konsortiolta SAS Institute Oy ja Aureolis Oy. Lisäksi KKV sai yhden myöhästyneen tarjouksen KPMG Oy Ab:lta, jota ei otettu tarjousten vertailussa huomioon.

Tarjouskilpailun voitti konsortio SAS Institute Oy ja Aureolis Oy. Tarjousta koskevaa sopimusta allekirjoitettaessa sovittiin yhdessä, että vastuiden selkeyden vuoksi viraston sopimuskumppaniksi määriteltiin konsortion sijaan Aureolis Oy ja SAS Institute Oy määriteltiin Aureolis Oy:n alihankkijaksi. Muutoksella tarjoajien välisiin alihankintasuhteisiin ei ollut vaikutusta tarjouksen sisältöön muilta osin.

Havaintojemme mukaan kokeellisen Proof-of-Concept -projektin toimittajan hankkiminen kilpailuttamalla vaatii merkittävän panostuksen hankinnan kohteen määrittelemiseen. Määrittelemisessä hankaluutta lisää kohteen kokeellisuus, määrittelyä ei välttämättä ole mahdollista tehdä, koska kohteen määritys on osa toimittajan valinnan jälkeisen varsinaisen toteutusprojektin sisältöä.

2.4. Riskienhallinta

Ennen projektin käynnistämistä tiedostettiin erinäisten riskien olemassaolo. Potentiaalisten riskien seurantaa tehtiin jatkuvana osana projektin valvontaa. Riskiluettelo käsiteltiin ja päivitettiin projektiryhmän kokouksissa sekä ohjausryhmissä.

Projektin koosta johtuen riskien hallintamalli oli melko yksinkertainen. Riskitasoa arvioitiin tunnistettujen riskien vaikuttavuuden ja todennäköisyyden suuruuden osalta sekä hallintaa ROAM-mallilla (ROAM = Resolved, Owned, Accepted, Mitigated). Riskejä tunnistettiin seuraavasti: Sisällön laajuus, työmäärän pitävyys, Henkilöriskit (Tilaaja), Henkilöriskit (Toimittaja), Kehitysympäristön toimivuus sekä Ongelmat tiedonkulussa.

Kokeilun riskien tilanne kokeilun päättyessä:

Riski	Lopullinen tila	Toimenpiteet	Toimenpiteiden vaikutus
Sisällön laajuus: Julkisista lähteistä kerätty data ei vastaa KKV:n tarpeita tai dataa ei löydy.	Löytynyt data on relevantti KKV:n tutkimukselle. Dataa saatiin riittävästi johtopäätösten tekemiseen.	Määriteltiin tarkasti projektissa käytettävät lähteet ja hakusanat.	Tarkan määrittelyn perusteella kerätty data on ollut melko hyödyllistä ja sitä on löytynyt valituista julkisista lähteistä.
Kehitysympäristön toimivuus: KKV on riippuvainen valitun toimittajan ohjelmistosta/ympäristöstä	Käytettävät datankeruutyökalut ovat open-source -pohjaisia ja projektia varten räätälöityjä. Analyysityökalut ovat sen sijaan SAS:n omia työkaluja ja näiden ominaisuuksien hyödyntäminen edellyttää lisenssien hankkimista	Projektille varattiin riittävät resurssit toimittajan tiedonkeruu- ja analyysiympäristöistä.	Datankeruu-työkalut ovat käytettävissä ilmaiseksi jatkossakin, mutta niitä täytyy räätälöidä tapauskohtaisesti. Analyysityökalut edellyttävät investointeja tai kilpailevien tuotteiden hankkimista.
Työmäärän pitävyys:	Toimittajan työmäärä pysyi	Projektin laajuut-	Toimittajan työmäärä

Ilman huolellista työn rajaamista työmäärä ei pysy määriteltyjen resurssien sisällä.	määritellyn työmäärän sisällä. Viraston oma työ ylittyi suunnitellusta.	ta rajoitettiin haettavia tietolähteitä rajoittamalla.	pysyi suunnitelman mukaisena.
Henkilöriskit (Tilaaja): Projekti ei saa tilaajalta riittäviä henkilöresursseja.	Tilaaja pystyi antamaan projektin tarvitseman työpanoksen.	Projektiryhmä käytti työhön suunniteltua enemmän työaikaa.	Tilaajan työmäärä ylittyi suunnitellusta, mutta toimittaja sai samalla tarvitsemansa tuen projektissa.
Henkilöriskit (Toimittaja): Projekti ei saa tarvitsemiaan resursseja toimittajalta.	Toimittaja antoi riittävät henkilöresurssit projektille.	Projektin laajuutta rajoitettiin.	Toimittajan henkilöresurssit saatiin riittämään projektin tarpeisiin.
Ongelmat tiedonkulussa: Projektin tilaajan ja toimittajan välinen tiedonkulku on puutteellista.	Tiedonkulku toimi hyvin koko projektin ajan.	Perustettiin erillinen projektityötila ja projektiryhmä kokoontui vähintään viikoittain.	Sekä tilaajalla että toimittajalla oli koko ajan selkeä käsitys projektin tavoitteista ja tilasta.

2.5. Kokeilun tavoitellut hyödyt ja niiden toteutuminen

Projektin tavoitteena oli tehostaa kartellivalvontaa ja nostaa avoimien tietolähteiden käyttöä uudelle tasolle, jolloin kartellien kiinnijäämisriski kasvaisi. Projekti toteutettiin ns. Proof-of-Concept -mallilla eli tarkoituksena oli testata toimintatapaa esimerkkitapauksella.

Projektin tavoitteena oli myös mahdollistaa työajan säästö automatisoimalla aineiston keräämistä.

Projektin aikana KKV oppi käyttämään Scrapy Python-kirjastoa, joka soveltuu datankeruun verkkosivuilta rakennettavien ohjelmanpätkien (spidereiden) avulla. Näitä on melko helppoa päivittää eri tutkinnallisiin tarkoituksiin. Vaikka spidereiden tekeminen tai päivittäminen ei yleensä edellytä syvällistä ohjelmointiosaamista silloin kun haettavien tietojen rakenne ei merkittävästi muutu, tarvitaan tapauskohtaiseen räätälöintiin työaikaa. Projektissa kutakin yksittäistä tietolähdettä koskevan spiderin räätälöintiin meni arviolta 2 tuntia. Räätälöintiin kuluva aika väheni jonkin verran työn edetessä, mutta ei merkittävästi.

Spidereiden rakentaminen ei ole tehokkain tapa tiedon hakemiseen. Vaihtoehtoisia tapoja tiedon hakemiseen ovat mm. web crawler -työkalut, joilla verkkosivustojen sisältöä voidaan kerätä automaattisesti ilman ohjelmointitaitoja, vaikka myös ne edellyttävät tietolähdekohtaista räätälöintiä. Web crawler -työkalujen tehokas käyttö edellyttää kuitenkin työaikaa perehtymiseen ja testaamiseen. Eräs vaihtoehto on yhdistää web crawlereiden ja spidereiden käyttöä tarpeen mukaan.

Tiedon hakemiseen voisi oman tuotannon lisäksi soveltua myös ulkoistus sellaiselle toimittajalle, jolla on tässä projektissa valittuja toimittajia paremmat edellytykset tiedon hakemiseen.

SAS:n tarjoamat kerätyn datan analyysityökalut ovat projektin aikana osoittautuneet mielenkiintoisiksi, joskin lisenssien hankinta saattaa osoittautua kalliiksi.

Jotta datankeruuseen käytettävä aika olisi mahdollisimmat lyhyt ja tehokas, edellyttää se toiminnan automatisoinnin mahdollisimman pitkälle. Kartellitutkaprojektin kokemusten perusteella voidaan todeta, että Scrapy-kirjastoja hyödyntävät spiderit eivät ole tehokkain tapa automatisoida datankeruu julkisista

lähteistä, sillä näiden rakentaminen ja ylläpito vaatii perehtymistä asiaan ja jonkin verran ohjelmointitaitoa.

Selkein puute suunnitelmaan nähden on se, että tarjouskilpailussa valitun toimittajan kanssa oli välttämätöntä voimakkaasti rajata tietolähteitä. Odotuksiamme nostattivat toisen toimittajan kanssa käydyt keskustelut ”koko suomalaisen Internetin” lukemisesta viidessä minuutissa.

Datankeruuvaiheeseen meni myös suurin osa käytettävissä olevasta ajasta. Analyysivaiheessa olisi ollut tarpeen saada suurempi datasetti käsittelyyn ja käyttää pääosa ajasta siihen. Nyt mahdollisista analyysin hyödyistä voi vasta vetää varovaisen positiivisia johtopäätöksiä.

Kuvaa alla olevaan taulukkoon kehitettävän prosessin vaikuttavuus- ja asiakashyötypotentiaali hakemuksen mukaan ja arvioi sen toteutumista kokeilun jälkeen:

Arvio kehitettävän prosessin vaikuttavuus- ja asiakashyötypotentiaalista		
Tavoiteltava yhteiskunnallinen vaikuttavuus	Hyötyjen realisoituminen hakemuksen mukaan	Arvio hyötyjen realisoitumisen toteutumisesta, jos kokeilussa rakennettu muutos otetaan tuotantoon
Kartellien kiinnijäämisriski kasvaa	Tavoitteena oli proof-of-concept -toteutuksella todentaa koneoppimisen soveltuminen avoimista tietolähteistä saatavan tiedon soveltuminen kartellimaisen toiminnan jättämien jälkien havaitsemiseen ja yhdistelemiseen. Tarkoituksena oli löytää tehokkaita työkaluja, jotka voitaisiin myöhemmin ottaa kilpailuvalvonnassa käyttöön.	POC-projektissa ei löydetty sellaista skaalautuvaa ratkaisua, jota voisi suoraan hyödyntää. Keskeinen este hyötyjen saamisessa johtuu haettavien tietolähteiden tietojen hakutyökalujen räätälöintiin kuluvasta ajasta. Analyysivaiheessa saatiin varovaisen positiivisia tuloksia, mutta käytettävissä olleen datan määrän vuoksi tarvittaisiin lisää kokemuksia kattavien johtopäätösten tekemiseksi.

Kuvaa alla olevaan taulukkoon kehitettävän prosessin vaikuttavuus- ja asiakashyötypotentiaali hakemuksen mukaan ja arvioi sen toteutumista kokeilun jälkeen:

Arvio kehitettävän prosessin tuottavuuspotentiaalista		
Taloudelliset hyödyt	Hyötyjen realisoituminen	Arvio hyötyjen realisoitumisen toteutumisesta, jos kokeilussa rakennettu muutos otetaan tuotantoon
Viraston arvion mukaan tehokkaalla tiedonkeruulla ja analysoinnilla voidaan saavuttaa laadukkaampien tuloksien lisäksi myös arviolta yhden henkilötyövuoden verran manuaalista selvitystyötä vuosittain eli 60 000 euroa/vuosi säästettynä työaikana.	POC-projektissa arvioitiin, että voitaisiin säästää yhden henkilötyövuoden verran työaikaa manuaalisista verkkohauista kilpailuvalvonnan asiantuntijoilta.	POC-projektissa toteutetut toimenpiteet eivät tuota realisoitavia hyötyjä, koska datankeruuvaihe oli ennakoitua työläämpi toteuttaa.

3. Kokeilun päättäminen

3.1. Kokeilun opit

POC-projektissa ei löydetty sellaista skaalautuvaa ratkaisua, jota voisi suoraan hyödyntää. Keskeinen este hyötyjen saamisessa johtuu haettavien tietolähteiden tietojen räätälöintiin kuluvasta ajasta. Analyysivaiheessa saatiin varovaisen positiivisia tuloksia, mutta käytettävissä olleen datan määrän vuoksi tarvittaisiin lisää kokemuksia kattavien johtopäätösten tekemiseksi.

3.2. Kokeilun kokemusten jakaminen

POC-projektissa saatuja kokemuksia on ensisijaisesti tarkoitus hyödyntää viraston sisäisessä kehitystyössä. Projektissa käytettiin kilpailuvalvonnassa vireillä olevaa tapausta, joten projektin loppuraportissa on salassa pidettäviä tietoja, eikä sitä voi sellaisenaan jakaa.

Kilpailuvalvonnan kartellien vastaista työtä tehdään tiiviinä verkostona EU:n jäsenvaltioiden kilpailuviranomaisten kesken. Verkoston digitaalista tutkintaa ja tekoälyä koskeva vuosikokous pidetään Helsingissä 3.-4.10. (European Competition Network, Digital Investigation and Artificial Working Group). Projektin ohjausryhmässä toiminut Peter Berglund esittelee POC-projektin kokemukset kokouksessa. Esittelyn yhtenä tarkoituksena on myös saada ryhmältä suosituksia kehitystyön jatkamiseen.

Kansallisten viranomaisten kanssa on mielekästä jakaa kokemuksia sellaisten toimijoiden kanssa, joilla on vastaavia tarpeita. Virasto on kokoustanut projektin yhteydessä TUKESin Nettidogi-projektin henkilöiden kanssa (yhteyshenkilönä Tuiri Kerttula). Tarkoitus on vielä järjestää projektin päättymisen jälkeen kokous, jossa vaihdettaisiin kokemuksia molempien projektien tuloksista.

3.3. Kokeilun hyödyntäminen

POC-projektissa ei löydetty sellaista kartellitoiminnan jälkiä automaattisesti haravoivaa, automaattista ja skaalautuvaa ratkaisua, jota voisi suoraan hyödyntää. Keskeinen este POC-projektissa tuotettujen työkalujen hyödyntämiselle johtuu haettavien tietolähteiden tietojen hakutyökalujen räätälöintiin kuluvasta ajasta.

Analyysivaiheessa saatiin analyysityökalujen hyödyistä varovaisen positiivisia tuloksia, mutta analysointia varten käyttöön saadun datan pienen määrän vuoksi tarvittaisiin lisää kokemuksia kattavien johtopäätösten tekemiseksi.

Projektin asiakirjat ja projektin aikana kehitetyt työkalut sekä kerätty data on siirretty kilpailuvalvontaan myöhempää kehittämistä varten.