OKM VA-DIGI: osaprojekti 4

Jyry Suvilehto, Aino Ropponen, Joonas Pesonen, Johan Himberg

Tekoälyn ja kehittyneen analytiikan hyödyntäminen valtionavustusprosessissa

SISÄLLYSLUETTELO

1 YHTEENVETO

Valtionavustusten digitalisoinnin esiselvityksen (OKM VA-DIGI) osaprojektissa 4 selvitettiin valtionavustusten digitalisoinnin mahdollisuuksia erityisesti valtionavustusprosesseihin liittyvän tietosisällön analysoinnin näkökulmasta. Tavoitteena oli selvittää tekoälyn ja kehittyneen analytiikan hyödyntämisen mahdollisuuksia, rajoituksia ja vaatimuksia. Osaprojektin tärkeimmät johtopäätökset ovat:

- Tekoälyä ja kehittynyttä analytiikkaa voidaan käyttää päätöksenteon tukena avustuspäätöksiä tehdessä. Analytiikkaa hyödyntämällä voidaan muodostaa kokonaiskuva hakijasta ja hakukohteesta. Tekoälyn avulla voidaan esimerkiksi luokitella hakemuksia hakutekstissä esiintyvien sanojen perusteella sekä tunnistaa samankaltaisia hakemuksia, jolloin ne voidaan ohjata samalle käsittelijälle. Lisäksi analytiikkaa ja tekoälyä hyödyntämällä voidaan tukea esimerkiksi alueellisen tasapuolisuuden toteutumista, väärinkäytösten estämistä sekä projektien valvontaa.
- Tekoälyn avulla hakemuksia voidaan vertailla ja laittaa järjestykseen. Tulevaisuudessa tekoäly voi myös antaa ehdotuksen avustusten jakamiseksi, mutta päätöksenteko jää silti viranomaisen tehtäväksi. Tekoäly ei korvaa viranomaisen harkintaa. Erityisesti ehdotetun toiminnan vaikuttavuuden arviointi etukäteen tulee olemaan jatkossakin ihmisille haastavaa ja tietokoneille käytännössä mahdotonta.
- Järjestelmän tulee kerätä tiedot sekä hakijasta että haun kohteesta mahdollisimman tarkasti, jotta käsittelijät voivat saada kokonaiskuvan kummastakin. Tietosisältö tulee dokumentoida ja tieto tulee tarjota soveliaassa yleisesti tuetussa formaatissa. Kokonaiskuvaa varten järjestelmästä tulee mallintaa pienin yhteinen nimittäjä, eli ne tiedot, jotka ovat saatavilla kaikista hakemuksista ja päätöksistä. Kaikkea monimuotoisuutta pienin yhteinen nimittäjä ei kata, mutta se toimii pohjana kokonaiskuvan muodostamiselle.
- Valtionavustusprosessit eroavat merkittävästi laajuudeltaan sekä hakemuksien määrässä ja euromääräisesti. Samat menetelmät eivät voi kattaa koko ilmiön monimuotoisuutta. Ominaisuuksia suositellaan lisättävän hyöty-kustannusarvioinnin ja kokeilun kautta.

2 TAUSTA

Valtionavustusten digitalisointihanke on osa pääministeri Sipilän "Digitalisoidaan julkiset palvelut" -kärkihanketta. Hankkeen tarkoituksena on digitalisoida avustusten hakeminen ja myöntäminen sekä kehittää tarvittavat tietojärjestelmät hakemiseen, myöntämiseen, valvontaan, seurantaan ja vaikutusten arviointiin.

2.1 Esiselvityksen tavoitteet

Digitalisoinnin tavoitteena on edistää hyvää hallintoa, varmistaa päätösten oikeudenmukaisuutta, vähentää avustusten päällekkäisyyttä, parantaa avustusten tarkoituksenmukaista kohdentumista ja lisätä yhteiskunnallista vaikuttavuutta. Lisäksi tavoitteena on parantaa avustustoiminnassa kertyvän tiedon hyödynnettävyyttä mm. visualisoinnin ja data-analytiikan avulla.

Esiselvitysprojektia johtava opetus- ja kulttuuriministeriö on määritellyt esiselvitykseen kuusi keskeistä aluetta:

1. Valtioneuvoston kanslian VAHVA-hankkeen tuottaman ratkaisun soveltuvuuden arviointi VA-digi hankkeen asianhallinnassa
2. Tulevaisuuden hakuprosessin suunnittelu ja palvelumuotoilu
3. Mittarien määrittely strategisista tavoitteista ja sitä kautta vaikuttavuuden arviointi

4. Sisällön analyysi, miten keinoälyä ja data-analyysiä voidaan hyödyntää avustusprosessin eri vaiheissa
5. Valtionavustustoiminnan laajuuden ja eri käsittelyprosessien kartoitus
6. Tulosten yhteenveto ja jatkosuunnitelma

Tämä raportti selvittää osaprojektin 4 työn ja löydökset. Sen tärkeimmät löydökset tiivistetään koko projektin loppuraporttiin. Työ sivuaa useassa kohtaa erityisesti osaprojekteja 2 ja 3, ja tämä mainitaan soveltuvissa kohdissa.

2.2 Valtionavustustoiminnan kirjo

Osaprojektin 5 loppuraportin mukaan vuodessa tehdään noin 20 000 myönteistä avustuspäätöstä. Lisäksi Elintarviketurvallisuusvirasto Evira ja ja ELY-keskukset tekevät normaalista poikkeavien avustusprosessien kautta noin 40 000 ja 45 000 myöntävää päätöstä vuodessa. Nämä poikkeavat prosessit liittyvät starttirahaan, työntekijän palkkatukeen sekä kasvituhojen ja eläintautien korvauksiin.

Jotkin toimijat tekevät alle 10 myönteistä avustuspäätöstä vuodessa. Keskimäärin päätöksiä tehdään noin 600 vuodessa ja Eviran ja ELY-keskukset poisluettuna eniten päätöksiä tekevät Taiteen edistämiskeskus ja Opetus- ja kulttuuriministeriö, kumpikin noin 3600 vuodessa. Suurin hylkäysprosentti on Suomen Akatemian 75% hakemuksista ja useampi toimija hylkää alle 5% hakemuksista.

Avustustoiminnan hallinnointiin käytettiin arviolta 400 henkilötyövuotta. Reilusti eniten työtä hallinnointiin käytti TEKES, 100 henkilötyövuodella. Kuva 1 näyttää hallinnointiin käytettyjen henkilötyövuosien jakautumisen osa-alueille.

Kuva 1. Henkilötyövuosien jakautuminen valtionavustusten hallinnoinnissa

Projektin korkean tason tavoitteena on pienentää käsittelyn eli mekaanisen työn osuutta, jotta voidaan vapauttaa enemmän työpanosta valvontaan, vaikuttavuuden arviointiin ja valmisteluun. Suurimmat käsittelyä helpottavat hyödyt tulevat hakujärjestelmän toiminnallisuuksista ja käytettävyydestä. Data-analytiikan ja koneoppimisen avulla voidaan tuottaa työkaluja erityisesti valvonnan ja valmistelun tueksi. Vaikuttavuuden arviointia voidaan myös tukea erityisesti varmistamalla tiedon yhteismuotoisuus ja koneellinen käsiteltävyys.

3 TEKOÄLY JA KONEOPPIMINEN

3.1 Käsitteet

Data-analytiikkaan liittyvä terminologia sisältää monta päällekkäistä käsitettä, joita käytetään arkikielessä usein ristiin.

Data-analytiikka on prosessi, jossa olemassa olevaa dataa tutkitaan, siistitään, otetaan käyttöön, muunnetaan ja mallinnetaan siten, että datasta voidaan löytää hyödynnettävää tietoa johtopäätöksien tekemiseksi tai päätöksenteon tueksi.

Business intelligence (BI) on termi, joka käsittää ohjelmistoja, järjestelmiä, työkaluja ja käytänteitä, joilla valjastetaan yrityksen tai yhteisön käytössä oleva informaatio päätöksentekoon ja tehokkuuden optimointiin. Business Intelligence keskittyy tiedon tallennusmuodon yhtenäistämiseen, tiedon varastointiin ja raportointiin. BI käsitteenä rajataan yleensä niin, että raportoinnissa käytetyt laskennalliset menetelmät rajoittuvat käytännössä “peruskoulumatematiikkaan”. Analyysissä tarvittavat laskutoimitukset ovat toistettavissa funktiolaskimella, jos tekijän kärsivällisyys vain riittää. BI:n tulokset ja visualisoinnit on aina tarkoitettu asiaan vihkiytyneen asiantuntijan tulkittaviksi.

Koneoppiminen (engl. machine learning) on tekoälyn osa-alue. Se tarkoittaa laskennallisten mallien sovittamista dataan, tarkoituksena saada malli toimimaan halutulla tavalla samanlaisen, uuden datan kanssa. Data-analytiikka on olennainen osa koneoppimisratkaisujen rakentamista, koska nykyiset koneoppimisratkaisut vaativat vielä hyvin esikäsiteltyä aineistoa toimiakseen.

Tekoäly eli keinoäly (engl. artificial intelligence) on tietokoneohjelma, joka kykenee älykkäisiin toimintoihin. Tyypillisesti älykkäiksi laskettavia toimintoja ovat esimerkiksi oppiminen ja ongelmanratkaisu. Tekoälylle ei ole olemassa tarkkaa määritelmää ja usein termejä tekoäly ja koneoppiminen käytetään ristiin.

Mitä tämä käytännössä tarkoittaa? Data-analytiikasta ja tekoälystä voidaan puhua vapaasti menemättä yksityiskohtiin käytettävistä työkaluista tai menetelmistä.

3.2 Koneoppiminen

Koneoppimisella tarkoitetaan sitä, että tietokonetta ei ohjelmoida perinteisellä sääntöpohjaisella logiikalla niin, että säännöt on annettu, vaan käytetään opetusaineistoa, josta opittua mallia voidaan käyttää myös uuteen, samasta lähteestä muodostettuun samanmuotoiseen dataan. “Oppiminen” on tyypillisesti ymmärrettävä mekaanisena ehdollistumisena, ei korkean tason kognitiivisena, inhimillisenä oppimisena. Nykymuotoinen tekoäly ei kykene yleistämään korkeammalle käsitteelliselle tasolle saamastaan opetusdatasta. Koneoppiminen on rajoitettu vain ja ainoastaan syötedataan (sen tilastolliseen rakenteeseen): mitkään semanttiset merkitykset, joita data-alkioilla asiantuntijalle on, eivät ilman erillistä mallintamistyötä vaikuta lopputulokseen.

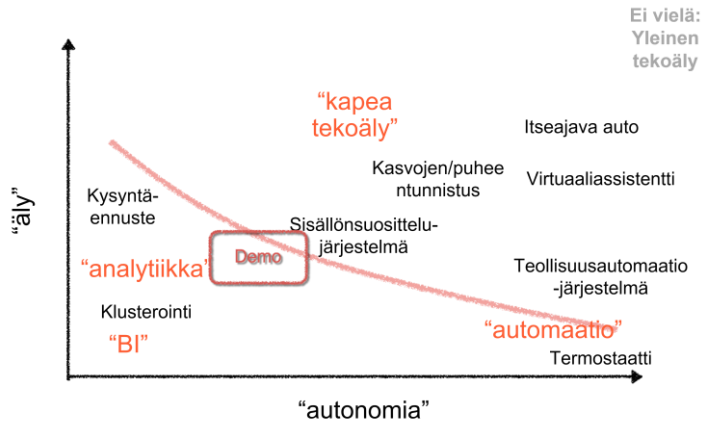
Koneoppimista voidaan käyttää tuottavasti ja tehokkaasti seuraavin reunaehdoin:

- Selkeä tehtävä, joka on teknisesti puettavissa optimointi-, luokittelu- tai muiksi algoritmein ratkaistavissa olevaksi koneoppimistehtäviksi
- Relevantti, tehtävään hyvin valmisteltu data, jonka merkitys tunnetaan
- Systeemityö ja suunnittelu, jolla yksittäiset ratkaisut istutetaan organisaatioon

Tunnetut ja teknisesti haastavat esimerkit, pelit, konenäkö, automaattiajo, puheentunnistus, kääntäminen, ovat luonteeltaan—ja kehitysinvestoinneiltaan—erilaisia kuin yksittäisen organisaation ongelmien ratkaisu

datavarannon perusteella. Sen sijaan monitahoisen reaalimaailman tehtävän autonominen asettaminen, ja sen autonominen ratkaiseminen, on nykyiselle *kapealle tekoälylle* vaikeaa — ellei mahdotonta.

Kuva 2 esittää erilaisten tekoälyratkaisujen suhdatta toisiinsa äly-autonomia-asteikolla. Tässä projektissa toteutettu soveltuvuusselvitys (Luku 5) sijoittuu asteikossa analytiikan ja sisällönsuosittelun väliin.



Kuva 2. Tekoälyratkaisujen sijoittuminen autonomia-äly-asteikolle. Tässä osaprojektissa tehty soveltuvuusselvitys sijoittuu analytiikan ja sisällönsuosittelun väliin ("Demo"). Kuva Reaktor AI:n esityksestä (Himberg, Särelä et al. 2017).

Koneoppimisessa olennaista on, että kaikki samalle järjestelmälle syötetty data on yhtenäismuotoista tai muutettavissa yhtenäiseen muotoon. Mikäli syötedatat eroavat edes vähän, täytyy joko kouluttaa kaksi erillistä järjestelmää tai rajoittaa järjestelmä tarkastelemaan syötedatojen pienintä yhteistä nimittäjää, s.o. dataa, joka kaikista näytteistä on saatavilla. Koneoppimisjärjestelmät koulutetaan opetusdatalla, ja paraskaan järjestelmä ei pysty oppimaan ilman tilastollisesti riittävän edustavaa dataa. Tarvittavan opetusdatan määrä riippuu opittavan kuvauksen vaikeudesta, valitun menetelmän kompleksisuudesta ja asetetuista tarkkuusvaatimuksista. Koneoppimisjärjestelmät ovat luonteeltaan tilastollisia, ne eivät tuota "täydellisiä" tuloksia, vaan tekevät aina jonkin verran virheitä. Tehokkuus syntyy siitä, että koneoppimisratkaisut ovat väsymättömiä, tasalaatuisia ja voivat seuloa laajoja aineistoja.

Koneoppimista käyttävien järjestelmien suunnittelussa on otettava huomioon tämä virhemahdollisuus järjestelmän toiminnallisuudessa. Koneoppimisratkaisut ovat hybridejä, jotka koostuvat asiantuntijoiden laatimista säännöistä ja koneoppimisella sovitetusta osasta.

3.2.1 Koneoppimisen pääluokat

Koneoppiminen voidaan jakaa pääluokkiin eri tavoin. Yleisimmin käsitellään

- **Ohjattu oppiminen** sovittaa mallin koulutusdataan *A* ja siihen liittyviin esimerkkivastauksiin *B*
- **Ohjaamaton oppiminen** etsii A:sta mielenkiintoisia rakenteita ihmisen tulkittavaksi tai alkuperäistä dataa tehokkaammaksi esitystavaksi ohjatulle oppimiselle.
- **Palauteoppiminen**, on ohjatun oppimisen variantti, jossa on koulutusdata *A*, mutta ei suoranaisia esimerkkejä *B.* Tämän sijaan järjestelmä saa aina toimittuaan palautetta toiminnastaan. Palauteoppiminen on tällä hetkellä vähemmän sovellettu, vielä kehittyvä osa-alue, jolla on tärkeä sija robotiikassa.
- **Siirto-oppiminen**: toinen vasta kehittyvä osa-alue jossa koneoppiminen siirtää ongelmasta A opittuja rakenteita uuteen, mutta läheiseen ongelmaan. Esimerkkejä tästä on uudemmissa

konekäännösjärjestelmissä, joissa eri kielien koneoppiminen itse asiassa tehostaa myös toisten mallintamista.

3.2.2 Koneoppiminen ongelmanratkaisussa

Tyypillinen koneoppimisen soveltaminen ongelmanratkaisussa on seuraava.

- Lähtökohtana on liiketoimintaongelma jota ratkaistaan (esim. avustushakemusten käsittelyn tehostaminen). Ongelma jäsennetään tavoitteiksi.
- Etsitään tavoitteiden saavuttamiseksi tehtäviä, jotka voidaan pukea kvantitatiiviseen, mitattavaan muotoon ja josta on koneoppimiseen sopivaa dataa: esimerkki hakemusten lajittelu sopivimmaksi eri käsittelijöille aiempien esimerkkien samankaltaisuuden avulla.
    - Asetetaan kvantitatiivisia mittareita, jotka osoittavat miten hyvin tehtävässä on onnistuttu.
    - Suunnitellaan, miten järjestelmän käyttäjä on parhaiten vuorovaikutuksessa
- Koneoppimistehtävä: Käytännön tehtävä puetaan koneoppimistehtäväksi. Esimerkki: “ennusta luokka (käsittelijä) perustuen faktoroituun tekstiaineistoon”
- Koneoppimisen toteutus: menetelmä (algoritmi), data ja datan esikäsittely
- Implementaatio tekniseen ympäristöön

Koneoppimisessa on hyvä erottaa toisistaan tehtävä (esimerkiksi ennustaminen tai luokittelu) ja itse menetelmät, algoritmit ja ohjelmistot. Sama oppimistehtävä voidaan toteuttaa useilla hyvin erilaisiin lähtökohtiin perustuvilla menetelmillä. Samoin koneoppimismenetelmien ja perinteisten tilastollisten menetelmien rajaveto on häilyvää. Koneoppimismenetelmiä on paljon ja niiden ominaisuudet vaihtelevat. Data-analyytikon ammattitaitoa on valita datan ominaisuuksiin ja kysymykseen parhaiten sopiva menetelmä. Kysymyksen ja vastauksen luonne ja muotoilu eivät muutu.

Koneoppimistehtävät voidaan jakaa karkeasti kolmeen vaikeusluokkaan:

- Tehtävät, jotka ratkeavat hyväksyttävästi useilla yleisesti käytetyillä menetelmillä
- Tehtävät, joiden ratkaisemiseksi on nähtävä paljon vaivaa ja valittava menetelmä huolella
- Tehtävät, joita ei käytettävissä olevan opetusdatan avulla voida ratkaista tyydyttävästi millään menetelmällä

Vaikeusluokka ei välttämättä riipu datan määrästä, vaikka se toki asettaa käytettävälle laskentaympäristölle vaatimuksia. Esimerkiksi kuvantunnistukseen, joka vaikuttaa haastavalta tehtävältä, on nykyisin standardikirjastoja, joita voidaan myös edelleen opettaa tunnistamaan uusia esimerkkejä. Toisaalta vaikeita tehtäviä mille tahansa menetelmälle ovat sellaisen suuren tai pienen aineiston ymmärtäminen, jossa onnistuakseen on ymmärrettävä tietojen merkitys ja tehtävä assosiaatioita reaalimaailmaan, vaikka itse data vaikuttaisi kompleksisuudeltaan ja määrältään vaatimattomalta.

4 TEKOÄLY JA TIEDOLLA JOHTAMINEN VALTIONAVUSTUSTOIMINNASSA

Kehittynyttä analytiikkaa ja tekoälyä voidaan hyödyntää valtionavustustoiminnassa päätöksenteon tukena. Tekoälyn avulla voidaan esimerkiksi luokitella hakemuksia hakutekstissä esiintyvien sanojen perusteella sekä tunnistaa samankaltaisia hakemuksia, jolloin ne voidaan ohjata samalle käsittelijälle. Vastaavasti voidaan havaita, jos samaa kokonaisuutta rahoitetaankin jo jossain muualla tunnistamalla samankaltaisia hakemuksia eri hauista ja aiemmilta hakukerroilta. Tekoäly tarjoaa myös työkaluja tarkastella hakemuksien ja

myönnettyjen avustusten demografista kohdentumista jolloin esimerkiksi alueellisen tasapuolisuuden toteutuminen voidaan ottaa paremmin huomioon.

Yrityspuolella esimerkiksi vakuutushakemusten käsittelyssä käytetään varsinkin pienissä hakemuksissa automaattista päättelyä. Valtionavustuslaki määrittelee valtionavun myöntämisen tapahtuvan viranomaispäätöksellä, joten vastaavan käytännön käyttöönotto vaatisi ilmeisesti muutosta lainsäädäntöön. Tekoälyn avulla hakemuksia voidaan kuitenkin vertailla ja laittaa järjestykseen. Tekoäly voi myös antaa ehdotuksen avustusten jakamiseksi, mutta päätöksenteko jää silti viranomaisen tehtäväksi.

Tehokas analytiikan käyttö edellyttää yhteisen tietoarkkitehtuurin. Datan käyttö perustuu yhtenäismuotoisen tiedon tallentamiseen ja saatavuuteen. Tällaisen tietovarannon päälle voidaan rakentaa hyöty-kustannusanalyysien pohjalta uusia mittareita ja työkaluja sekä tiedolla johtamisen tavanomaisempien välineiden (BI) että tekoälyn avulla.

Analytiikka voidaan jaotella valtionavustustoiminnan kokonaisuuden mukaisesti:

Strategiaprosessin analytiikka auttaa määrittämään strategisia tavoitteita sekä suunnittelemaan avustustoiminnan painopisteitä. Analytiikan avulla voidaan analysoida niiden yhteiskunnallisten ilmiöiden nykytilaa, joihin avustuksilla halutaan vaikuttaa sekä toisaalta arvioida ehdotettujen toimenpiteiden vaikuttavuutta.

Jakoprosessin analytiikka auttaa viranomaisen päätöksentekoa. Lisäksi analytiikka voi tukea tärkeitä jakoprosessin tavoitteita, kuten esimerkiksi alueellisen tasapuolisuuden toteutumista, väärinkäytösten estämistä sekä projektien valvontaa. Tämä analytiikka soveltuu parhaiten automatisoitavaksi korkealle asteelle, jotta jakoprosessissa osallisina olevat tahot voivat vähentää toistuvaa, mekaanista työtä. Automatisoidut päätöksentekoa tukevat indikaattorit hakemuksesta, hakijasta jne. helpottavat valmistelevan virkamiehen työtä.

Arviointiprosessin analytiikka tarjoaa työkaluja arvioida myönnettyjen tukien vaikutusta yhteiskunnallisten ilmiöiden muutokseen.

Arviointi- ja strategiaprosessissa on erityisen tärkeää kehittää mittaristoja, analyysejä ja kokonaiskuvaa yhteiskunnallisista ilmiöistä päätöksentekijöille. Erityisesti ns. viheliäisiä ongelmia käsittelevistä avustusohjelmista ja tavoitteista ei ole aina mahdollista muodostaa yhteismitallisia kvantitatiivisia mittareita, ja siksi viranomaisten tulee käyttää jatkossakin harkintaa. Analytiikkapalvelujen saatavuus tukee tätä harkinnan käyttöä. Analytiikka on parasta toteuttaa jatkuvana prosessina, jossa tärkeimmin priorisoituja haasteita tai kipukohtia käsitellään yksi kerrallaan analytiikalle osoitettujen resurssien puitteissa.

5 SOVELTUVUUSSELVITYS OPETUSHALLITUKSEN DATASTA

Projektin puitteissa tehtiin soveltuvuusselvitys (Proof-of-Concept, PoC) olemassa olevan järjestelmän tietojen hyödyntämisestä koneoppimismenetelmiä käyttäen. Tavoitteena oli tuoda datan käyttö konkreettiselle tasolle.

Soveltuvuusselvitykseen määriteltiin seuraavat osatehtävät.

1. **Hakijoiden tilanteen parempi ymmärtäminen päätösprosessin yhteydessä.** Tässä tavoiteltiin demografisten ja geografisten tekijöiden mukaan tuomista päätösprosessiin päättäjän tueksi. Kun päätökset on tehty, saman tiedon perusteella voidaan raportoida avustusten kohdentuminen alueellisesti ja suhteessa eri hakijatahojen tilanteeseen.

2. **Samankaltaisuuden tunnistaminen.** Tässä tavoiteltiin samankaltaisten hakemuksien tunnistamista päätöksenteon tueksi. Tiedon perusteella samanlaiset hakemukset voidaan ohjata samalle käsittelijälle tai käsittelijän tietoon voidaan tuoda samanlaisia hakemuksia aiemmilta hakukerroilta tai eri hauista päällekkäisyyksien havaitsemiseksi. Samankaltaisuuden mittarina käytettiin paremman puutteessa hakemustekstin samankaltaisuutta.
3. **Asiasanojen tunnistaminen ja hakemusten automaattinen luokittelu** Tässä haluttiin tuoda hakukonetoiminnallisuutta ja etsiä automaattisesti tärkeimpiä asiasanoja hakemusteksteistä eri hakukierroksilta. Tarkoituksena oli tutkia voitaisiinko hakemukset automaattisesti jakaa ennalta määriteltyihin luokkiin tai jopa automaattisesti ehdottaa luokkia, joihin hakemukset voitaisiin jakaa.

5.1 Tausta

Soveltuvuusselvitystä varten pyydettiin dataa operatiivisista järjestelmistä käytettäväksi. Dataa saatiin Opetus- ja kulttuuriministeriön SALAMA-järjestelmästä sekä Opetushallituksen (OPH) valtionavustusjärjestelmästä. Näistä soveltuvuusselvitykseen valittiin OPH:n data.

OPH:n järjestelmässä oli 9/2015-6/2017 päättyneistä 54 hakukierroksesta noin 4000 soveltuvuusselvitykseen kelpaavaa päättötilassa (hyväksytty/hylätty) olevaa hakemusta. Näiden hakemuksien yhteenlaskettu budjetti oli yhteensä noin 290 milj. € ja jaettiin noin 95 milj. €. Hakukierrosten koko vaihteli kahdesta hakemuksesta 325 hakemukseen. Haut olivat paitsi aiheiltaan myös kooltaan, jakomäärältään ja hyväksymisprosentiltaan poikkeavia. Yhdessä haussa jaettiin 4.5 milj. € yhdelle kahdesta hakijasta ja toisessa 2 milj. € jaettiin 288 hakijalle. Hyväksymisprosentti (avustushakemusten kappalemäärästä) vaihteli välillä 18...100 %. On selvää, että miljoonia euroja myönnettäessä esittelijän tulee selvittää perusteellisesti hakijan tilanne, hakemuksen toteutettavuus ja muut asian ratkaisemiseen vaadittavat seikat, sekä käyttää aikaa varojen käytön seurantaan. Vastaavasti pienten (myönnettiin n. 100 alle 1500 € summaa) kokoluokan hakemuksien ei kannata käyttää liikaa työaikaa, etteivät hakemuksen käsittelemisen kustannukset nouse korkeammaksi kuin haettu summa.

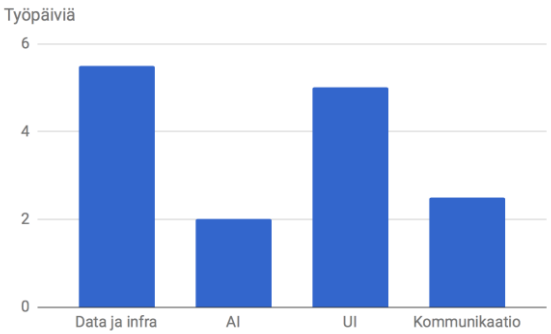
Osassa avustushakuja on numeerinen arviointiasteikko, osassa ei, ja asteikko vaihtelee. Päätösten perustelujen ja kommentit ovat vapaata tekstiä. Hakukaavakkeet ovat eri hakukierroksilla erilaisia ja kenttien asiasisältö vaihtelee.

5.2 Toteutus

Soveltuvuusselvityksen toteutus hankittiin Reaktor Innovations Oy:lta. Soveltuvuusselvitys toteutettiin analytiikka- ja koneoppimiskomponenteista rakennettuna WWW-sovelluksena. Sen työjärjestys oli seuraava:

1. Tehtävänasettelu ja **käsitemallin** ymmärtäminen
2. **Datan migraatio** ja muokkaaminen analytiikkasovellukselle
3. **Analytiikka**
    1. Hakuprosessin läpimenoajat ja työmäärä (prosessimttarointi, **process intelligence)**
    2. Hakukierrosten **kohdentuminen**
4. **Tekoäly**: **automaattinen aihepiirien muodostaminen** tekstistä ja hyödyntäminen eri tavoin
    1. Projektikuvausten vapaatekstikenttien hyödyntäminen (**NLP**)
    2. Koneoppimismenetelmä tyypillinen **sisällönsuosittelujärjestelmissä** (topic model)
5. **Käyttöliittymien** suunnittelu ja toteutus

Kuva 3 esittää soveltuvuusselvityksen tekemiseen käytetyn työajan jakautumisen. Kokonaisuudessaan työhön meni noin 15 henkilötyöpäivää.



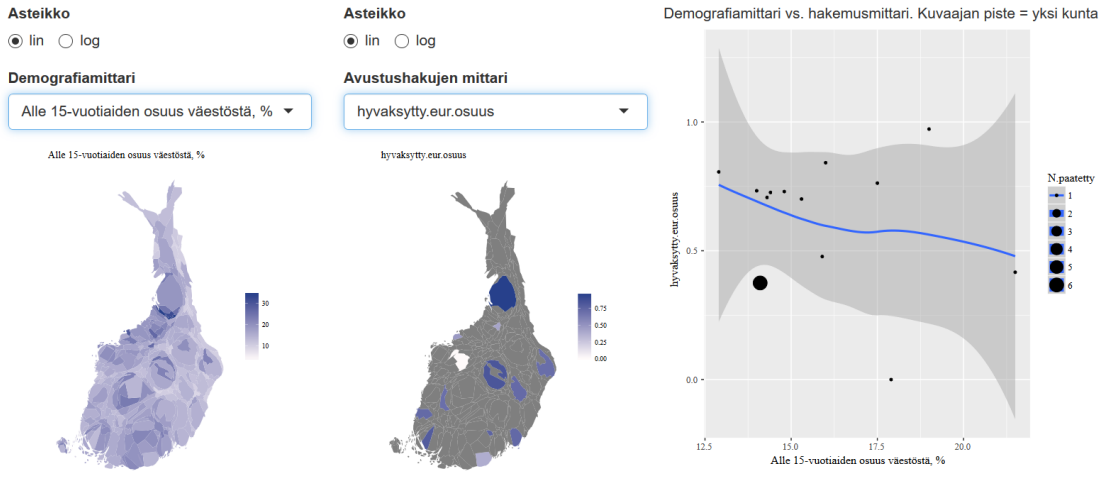
Kuva 3. Soveltuvuusselvityksen tekemiseen käytetyn työajan jakautuminen

Työn toteutus ja osatehtävien ratkaiseminen on kuvattu tarkemmin seuraavissa aliluvuissa.

5.2.1 Osatehtävä 1: kohdentumismittarien demonstraatio

Tilastokeskuksen kunta-aineiston ja muutamien hakemuksista saatavien euro- ja kappalemääriin perustuvien tietojen vertailu demonstroi hakemusten kohdentumisen kvantitatiivista arviointia - ja sen haasteita. Hakemuksen vaikutusarviointia varten tarvittaisiin relevantteja mittareita, joita seurata. Tässä demonstraatiossa voi tarkastella, miten kuntien geodemografiset tiedot vertautuvat eri hakukierroksilla jaettuun tai anottuun rahaan. Valitsemalla useita kierroksia, tiedot aggregoidaan kartta ja riippuvuustarkasteluun.

Riippuvuustarkastelu on vain idean demonstraatio: 1) ensinnäkin geodemografisiksi muuttujiksi on otettu helposti tarjolla olleita kuntatietoja, spesifejä mittareita näistä tuskin löytyy 2) datan määrä yksittäisellä hakukierroksella on niin vähäinen, ettei pitkälle meneviä johtopäätöksiä voi vetää, 3) tilastollinen mallinnus vaatisi tosiasiassa paneutumista mallinnukseen huomattavasti enemmän, 4) alueellinen kohdentuminen on teknisestikin ongelmallista; datassa ei ole helposti koneluettavassa muodossa tarjolla paikkaa (esim. koulua, jossa 'raha käytetään'). Kohdennuksen subsitituutiksi on otettu Y-tunnuksella yhdistetty kotipaikka PRH:n yritystiedoista. Kuva 4 näyttää esimerkin riippuvuustarkastelusta.



Kuva 4. Kuvakaappaus soveltuvuusselvityksen www-toteutuksesta: esimerkki riippuvuustarkastelusta demografiamittarin (alle 15-vuotiaiden osuus väestöstä, %) ja avustushakujenmittarin (hyväksytty.eur.osuus) välillä.

5.2.2 Osatehtävät 2 ja 3: aihemallinnus

Ohjaamattomassa oppimisessa aineistosta muodostuu “mielenkiintoisia piirteitä”. Klassinen esimerkki on faktorianalyysi, jossa esimerkiksi persoonallisuustestivastauksista lasketaan tilastollisesti “faktoreita” jotka kuvaavat pääasiallista vaihtelua vastauksissa, ja siten toivottavasti kiinnostavia persoonallisuuden piirteitä.

Esimerkki ohjaamattomasta oppimisesta on OPH:n aineistoon perustuvassa soveltuvuusselvityksessä käytetty aihemallinnus (topic model) ja siihen perustuva dokumenttien järjestäminen. Soveltuvuusselvityksessä on otettu hakemusten projektikuvauskenttien tekstisisältö ja laskettu näistä aihemalli. Ennen aihemallinnusta sanat on palautettu sanakirjamuotoon eli lemmoiksi. Aihemallinnuksen pohjaksi lasketaan jokaisen dokumentin lemmojen frekvenssit, joista ohjaamattoman oppimisen menetelmä (Latent Dirichlet Allocation) muodostaa aiheita (sanajoukkoja) ja vastaavasti aiheiden painotukset eri hakemuksissa.

Ohjaamattoman oppimisen menetelmien kirjo on erittäin laaja. Asiantuntijat voivat hyödyntää tuloksia aineiston ja ilmiöiden ymmärtämisessä, mutta sitä usein käytetään myös teknisenä välivaiheena, jota ei ole tarkoitus hyödyntää suoraan. Erilaisten faktorointien, klusterointien, automaattisesti generoitujen sääntöjen tms. haasteena on se, että niitä täytyy tulkita. Relevantti tulkitseminen vaatii paneutumista alan asiantuntijalta, data-analyytikko ei siihen yksin yleensä voi kyetä.

Tämän osaprojektin soveltuvuusselvityksessä aiheita, “teemoja” on kuvattu sanoilla, jotka ovat erottelevia muihin teemoihin nähden. Yleisimmät sanat ovat hakemuksissa laajalti samoja. Tämän jälkeen teemat voidaan nimetä kuvaavasti - teemoja voidaan käyttää dokumenttimassaa kuvaavina piirteinä, hauissa ja dokumenttien järjestämisessä. Niillä on usein myös teknisesti alkuperäistä aineistoa helpompaa rakentaa ohjatun oppimisen malleja: ne muuttavat rakenteetonta tekstidataa kvantitatiiviseen muotoon, joka on otollista monille yleisesti käytetyille menetelmille.

Demonstraatiossa on esitetty, miten aiheita, “teemoja” voidaan käyttää kuvaamaan ja järjestämään hakemuksia. On huomattava, että tematisointi ei käytä suoraan informaatiota hakukierroksista tai hakijoista vaan vain projektia kuvaavista, valikoiduista tekstikentistä. Samalla metodiikalla ja saman sukuisilla algoritmeilla toimivat useat kaupallisessa käytössä olevat ns. suosittelujärjestelmät, joissa esimerkiksi mediasisältöä haetaan ja luokitellaan ilman käsin määriteltyjä hakusanoja.

Taulukkoon 1 on kerätty demonstraatiossa löydetyt kahdeksan teemaa. ”Nimi” on asiantuntijan antama nimi teemalle. ”Erotteleva” on joukko sanoja jotka esiintyvät sadan painavimman termin joukossa vain tässä aiheessa. ”Tyypillinen” on joukko sanoja, jotka esiintyvät sadan painavimman termin joukossa tässä ja 1–2 muussa aiheessa.

Aihemallinnuksen jälkeen hakemusten aiheprofiili eli kunkin hakemuksen liittyminen teemoihin voidaan määritellä. Kuvassa 5 esitetään muutamien OPH:n avustushakemusten aiheprofiilit. OPH:n aineiston hakemuksista osa kuului selkeästi yhteen teemaan, mutta monet kuuluivat useampaankin teemaan. Hakemusten samankaltaisuutta voidaan arvioida tarkastelemalla niiden aiheprofiilien samankaltaisuutta. Kuvissa 6 ja 7 esitetään otos aiheprofiilien keskiarvoista avustuskierros- ja päätoimialakohtaisesti.

Taulukko 1. Soveltuvuusselvityksessä tunnistetut kahdeksan teemaa ja niihin liittyvät erottelevat ja tyypilliset sanat

Aihe	Nimi	Erotteleva	Tyypillinen
X1	'kerhoja'	kerhotoiminta, kerho, koulupäivä, harrastus, hyvinvointi, harrastaa, harrastustoiminta, liikunta, syrjäytyminen, koti, ohjaaja, harrastusmahdollisuus, sektori, toive, mahdollinen, oppilaskunta, lisääminen, kolmas, vapaaaika, laadukas, huomio, mielekäs, vähän, yläkoulu, kerhotarjonta, monipuola, huomioida, lisää, liikkua, kysely, löytää, kasvu, jatkaa, perhe	oppilas, lapsi, monipuolinen, mukaan, huoltaja, itse, osallisuus, ohjata, jokainen, palaute, toimija, vanhempi, suunnitella, erityisesti, tutustua, antaa, sisältö, vahvistaa
X2	'ammattiopintoja'	järjestäjä, työpaikka, opinnotpolku, prosessi, tapahtua, reformi, erityinen, ala, tunnistaminen, henkilökohtaistaa, työssäoppiminen, joustaa, aikuinen, kuvata, edustaja, aineisto	opiskelija, työelämä, tutkinto, ammatillinen, ohjaus, opinnot, tarvita, henkilöstö, yksilöllinen, digitaalinen, yritys, organisaatio, laatu, toimintatapa, palaute, mukana, käytänne, toimija, levittää, mukaisesti, ohjata, muutos, strategia
X3	'lukioita'	lukio, lukiolainen, turku, urheilija, musiikkiopisto, suorittaa, syksy, jatkoopinnot, helsinki, lukiokoulutus, valita, musiikki, aste, kevät, sähköisä, lukuvuosi, matematiikka, yhteiskoulu, oulu, normaalikoulu, oppimateriaali, korkeakoulu	opiskelija, kurssi, opiskelu, sähköinen, opinnot, opiskella, yliopisto, vuosi, mukana, oppiaine, projekti, työelämä, peruste, monipuolinen, toimintakulttuuri, valtakunnallinen, antaa, verkko, tutustua, yritys, erityisesti, tapa
X4	'opettajantukemista'	osallistuja, seurata, liite, eritellä, kerätä, toteutuminen, kouluttaja, verkkosivu, markkinointi, työyhteisö, koulutusosio, jälkeen, tuotos, kohderyhmä, koulutushanke, markkinoida, opetushallitus, rekrytointi, toteutus, julkaista, uutinenkirje, oph, rehtori, täydennyskoulutus, syntyä, koulutuskokonaisuus, tehtävä, vaikuttavuus, jatkua, rekrytoida, johtaminen, facebook, kertoa	palaute, media, levittää, yliopisto, ammatillinen, sisältö, antaa, tarvita, sähköinen, käytänne, jokainen, ohjaus, tapahtuma, organisaatio, valtakunnallinen, verkko
X5	'varhaiskasvatusta'	varhaiskasvatus, tutoropettaja, päiväkoti, yksikkö, esiopetus, kouluttaa, tutor, osallistaa, kehittyminen, leikki, toteuttaminen, varhaiskasvatussuunnitelma, esi, vuorovaikutus, suunnitelma, jakaminen, valmius	lapsi, henkilöstö, toimintakulttuuri, osallisuus, pedagogiikka, arki, vahvistaa, muutos, mukainen, henkilökunta, edistää, toimintatapa, digitaalinen, käytänne, väline, huoltaja, jokainen, ohjata, mukaisesti, peruste, vanhempi, ryhmä
X6	'kansainvälisyyttä'	opisto, kansalaisopisto, taide, maahanmuuttaja, kansainvälinen, asiantuntija, yhteistyökumppani, ulkomaa, tuntiopettaja, kiina, määrä, laatia, asiakas, liikkuvuus, vaihto, sivistystyö, koulutusvienti, vapaa, yksi, viestintä, kumppani	opiskelija, kansavälinen, kurssi, vuosi, henkilöstö, suomalainen, ammatillinen, strategia, kehitys, toimija, henkilökunta, kansainvälisyys, mukaan, tutkinto, venäjä, mukana, media, edistää, yritys, laatu, tapahtuma, suunnitella
X7	'kieliä'	kieli, kulttuuri, kestää, englanti, vieras, kielitaito, ruotsi, kaksikielinen, saksa, teema, vierailu, maa, maailma, kielirikasteinen, yhteys, välinen, vaikuttaa, merkitys, luonto, ohjelma, tärkeä, peruskoulu, kulttuurinen, matka, ymmärtää, välillä, aloittaa, ihminen, kasvaa, globaalikasvatus, asia, kiinnostus	oppilas, tutustua, projekti, kehitys, kansavälinen, opiskelu, suomalainen, luokka, vahvistaa, opiskella, ryhmä, media, vuosi, mukaan, venäjä, tulevaisuus, ympäristö, arki, kansainvälisyys, tuoda, tapahtuma
X8	'oppimisinnovaatiota'	teknologia, tila, oppija, rakentaa, oppiminenkokonaisuus, kokeilla, laajaalainen, monialainen, laite, kokeilu, tutkia, ratkaisu, hyödyntäminen, uudenlainen, avoin, työskentely, innovatiivinen, kokonaisuus, yhteisöllinen	oppilas, digitaalinen, toimintakulttuuri, ympäristö, oppiaine, luokka, tapa, väline, suunnitella, itse, projekti, levittää, sisältö, sähköinen, mukainen, yksilöllinen, pedagogiikka, monipuolinen, ryhmä, edistää, tulevaisuus, tarvita, tuoda

avustus_name	yritysmuoto	paatoimiala	'kerhoja'	'ammattiopintoja'	'lukioa'	'opettajantukemista'	'varhaiskasvatusta'	'kansainvälisyyttä'	'kieliä'	'oppimisinnovatiota'
All	All	All		All		All	All	All		All
Vapaan sivistystyön laatu- ja kehittämisavustukset	Kunta	Julkinen yleishallinto	0	0	0.1	0	0.27	0.63	0	0
Vapaan sivistystyön laatu- ja kehittämisavustukset	Kunta	Julkinen yleishallinto	0	0.06	0.12	0	0.26	0.55	0	0
Muiden kuin opetustuntikohtaista valtionosuutta saavien taiteen perusopetuksen järjestäjien harkinnanvarainen valtionavustus vuonna 2016	Aatteellinen yhdistys	Muut koulutusta antavat yksiköt	0.01	0.01	0.01	0.01	0.28	0.65	0.01	0.01
Vapaan sivistystyön laatu- ja kehittämisavustukset	Kunta	Julkinen yleishallinto	0	0	0.13	0.02	0.31	0.53	0	0.01
Vapaan sivistystyön laatu- ja kehittämisavustukset	Kunta	Julkinen yleishallinto	0.02	0	0	0	0.29	0.69	0	0
Vapaan sivistystyön laatu- ja kehittämisavustukset	Kunta	Julkinen yleishallinto	0	0	0	0	0.33	0.67	0	0
Vapaan sivistystyön laatu- ja kehittämisavustukset	Kunta	Julkinen yleishallinto	0	0	0	0	0.25	0.75	0	0

Kuva 5. Kuvakaappaus soveltuvuusselvityksen www-toteutuksesta: avustushakemusten aiheprofiilit. Kuvasta on rajattu pois hakijan tietoja sekä projektin nimi.

value	'kerhoja'	'ammattiopintoja'	'lukioa'	'opettajantukemista'	'varhaiskasvatusta'	'kansainvälisyyttä'	'kieliä'	'oppimisinnovatiota'
A) Ammatillisen koulutuksen 2017 kansainväliSYYSpäivien järjestäminen hyvien käytäntöjen levittämiseksi B) Opetushallituksen kansainvälisyysverkostohankkeiden seminaarin järjestäminen	0	0.081	0	0.02	0	0.796	0.053	0.049
Alkuperäiskarjan geenipankkitoiminta	0	0	0	0	0	0.969	0	0.03
Alkuperäiskarjan hoidon tuki	0	0	0	0	0	0.999	0	0
Ammatillisen erityisopetuksen kehittämis- ja palvelutoiminnan sekä työelämälähtöisyyden vahvistaminen	0	0.964	0	0.007	0.01	0	0	0.018
Ammatillisen koulutuksen 2016 kansainväliSYYSpäivien järjestäminen hyvien käytäntöjen levittämiseksi	0	0.133	0	0.013	0	0.789	0.064	0
Ammatillisen koulutuksen kansainvälistyminen	0	0.317	0.003	0.009	0	0.634	0.032	0.004
Ammatillisen koulutuksen koulutusvientikokeilujen koordinointi	0	0.385	0	0	0	0.614	0	0
Ammatillisen koulutuksen koulutusvientikokeilujen markkinointimateriaali	0	0	0	0	0	1	0	0
Ammatillisen koulutuksen koulutusvientikokeilut	0	0.189	0.007	0.003	0.004	0.797	0	0
Ammatillisen koulutuksen laadun kehittäminen	0	0.924	0.003	0.024	0.01	0.032	0.004	0.003
Esi- ja perusopetuksen tietoverkkojen hankinnat	0.022	0.006	0.018	0.002	0.276	0.002	0.055	0.62
Innovatiivisten oppimisympäristöjen edistäminen esi- ja perusopetuksessa sekä lukiokoulutuksessa	0.018	0.033	0.225	0.01	0.089	0.017	0.062	0.547
Innovatiivisten oppimisympäristöjen edistäminen esi- ja perusopetuksessa sekä lukiokoulutuksessa 2017	0.029	0.018	0.021	0.003	0.059	0.005	0.067	0.797
Innovatiivisten oppimisympäristöjen edistäminen varhaiskasvatuksessa 2017	0.028	0.001	0	0	0.709	0	0.021	0.241
Kärkihanke 1 - hakuryhmä 1: Tutoropettajien koulutus ja osaamisen kehittäminen	0.002	0.164	0.002	0.412	0.413	0.002	0.002	0.002
Kärkihanke 1 - hakuryhmä 3: Kokeilu-, kehittämis- ja innovaatiotoiminta	0.015	0.014	0.043	0.019	0.18	0.016	0.034	0.679
Kärkihanke 6 - Joustavan perusopetuksen (JOPO) toiminnan kehittäminen	0.114	0.176	0.282	0.024	0.282	0.008	0.026	0.088
Koulun kerhotoiminnan kehittäminen	0.919	0.001	0.003	0.002	0.014	0.003	0.039	0.018

Kuva 6. Kuvakaappaus soveltuvuusselvityksen www-toteutuksesta: avustuskierrokselle (value) tulleiden hakemusten aiheprofiilien keskiarvot

value	'kerhoja'	'ammattiopintoja'	'lukioa'	'opettajantukemista'	'varhaiskasvatusta'	'kansainvälisyyttä'	'kieliä'	'oppimisinnovatiota'
Alemman perusasteen koulutus	0.448	0.057	0.073	0.041	0.04	0.011	0.183	0.147
Ammattialajärjestöjen toiminta	0	0.049	0.072	0.746	0.023	0.063	0	0.045
Ammattiyhdistysten toiminta	0	0	0.125	0.808	0	0	0	0.066
Esittävät taiteet	0.011	0.003	0	0.778	0.044	0	0.035	0.129
Ikääntyneiden hoitolaitokset	0	0.968	0	0	0.031	0	0	0
Ikääntyneiden palveluasuminen	0	0.14	0	0.859	0	0	0	0
Julkinen yleishallinto	0.339	0.053	0.046	0.083	0.202	0.044	0.047	0.187
Kansanopistot, kansalaisopistot, työväenopistot yms.	0.04	0.406	0.046	0.193	0.022	0.143	0.121	0.03
Kaukolämmön ja -kylmän erillistuotanto ja jakelu	0.962	0	0.01	0.028	0	0	0	0
Keskiasteen ammatillinen koulutus	0.004	0.558	0.029	0.185	0.01	0.146	0.014	0.055
Kirjojen kustantaminen	0.001	0.744	0.001	0.001	0.001	0.001	0.252	0.001
Korkea-asteen koulutus yliopistoissa ja ammattikorkeakouluissa	0.021	0.054	0.024	0.775	0.033	0.01	0.012	0.07
Koulutuskeskukset	0.01	0.152	0.006	0.688	0.031	0.015	0.017	0.081
Koulutusta palveleva toiminta	0.012	0.239	0.022	0.477	0.098	0.021	0.002	0.128
Lääketieteellinen tutkimus ja kehittäminen	0.071	0.002	0	0.863	0.063	0	0	0
Lasten ja nuorten laitokset ja ammatillinen perhehoito	0	0.513	0.014	0.017	0.012	0.29	0.007	0.147
Lasten päiväkodit	0.153	0.037	0.156	0.109	0.113	0.332	0.101	0
Lukiokoulutus	0.224	0.02	0.395	0.065	0	0.018	0.136	0.141
Muualla luokittelemattomat muut järjestöt	0.084	0.087	0.047	0.628	0.032	0.042	0.033	0.046
Muu ammatillinen, tieteellinen ja tekninen toiminta	0	0	0	0.556	0.376	0	0	0.067
Muu liikkeenjohdon konsultointi	0	0	0	0.736	0.116	0	0.005	0.142
Muu luonnontieteellinen tutkimus ja kehittäminen	0.08	0	0	0.667	0.02	0.141	0	0.092
Muut koulutusta antavat yksiköt	0.169	0.577	0.023	0.075	0.014	0.009	0.068	0.066
Muut muualla luokittelemattomat sosiaalihuollon avopalvelut	0.096	0.016	0	0.686	0.132	0.07	0	0

Kuva 7. Kuvakaappaus soveltuvuusselvityksen www-toteutuksesta: aiheprofiilien keskiarvot päätoimialoittain.

5.3 Havainnot

Soveltuvuusselvityksen tärkeimmät havainnot:

Havainto 1: Datan pitää olla koneluettavaa ja hyvin dokumentoitua

Data-analytiikkaa ja koneoppimissovelluksia varten datan on oltava koneluettavaa ja hyvin dokumentoitua. Jos data ei ole valmiiksi koneluettavaa, data-analyytikko joutuu muokkaamaan dataa käsin, mikä on usein työlästä ja aikaa vievää. Tämän osaprojektin soveltuvuusselvityksen toteutus hankittiin Reaktor Innovations Oy:lta. Reaktor on toteuttanut myös OPH:n hakujärjestelmän, joten Reaktorin data-analytiikkakonsultti pystyi tiedustelemaan järjestelmän suunnittelussa käytettyjä käsitteitä ja tietokannan rakennetta talon sisältä. Koska järjestelmän suunnittelussa ei ole otettu huomioon data-analytiikan tarpeita, konsultilta kesti 5 työpäivää saattaa data muotoon, jossa sitä on helppo analysoida ja jatkokäsitellä. Mikäli konsultti ei olisi voinut kysyä järjestelmää kehittävältä tiimiltä ratkaisuista ja toimintalogiikasta, aikaa olisi voinut mennä moninkertaisesti enemmän.

Hakujärjestelmän suunnittelussa on tärkeää ottaa huomioon data-analytiikan tarpeet: datan pitää olla koneluettavaa, oikeassa formaatissa olevaa ja hyvin dokumentoitua. Kun analytiikan tarpeet otetaan huomioon jo suunnitteluvaiheessa, analytiikan tekeminen on helpompaa ja nopeampaa. Yhteismitallisen, oikeassa formaatissa olevan ja dokumentoidun datan tuottaminen jatkokäyttöä varten tulee olla järjestelmän toiminnallinen vaatimus. Datan tekninen formaatti voidaan valita toteutusvaiheessa, mutta sen rakenne tulisi olla ns. *tidy data*-[1]formaatissa, jota voidaan pitää data-analytiikan alan epävirallisena standardina.

[1] Wickham, Hadley. "Tidy data." Journal of Statistical Software 59.10 (2014): 1-23.

Havainto 2: Yhtenäiset tietueet

Yksi OPH:n järjestelmän toiminnallinen vaatimus on, että jokaiselle haulle on voitava määrittää oma hakulomake. Tämä on perusteltua, koska haut ovat erilaisia, eikä kaikissa lomakkeissa voida kysyä samoja asioita. Hakulomakkeiden muokkausmahdollisuus on kuitenkin johtanut siihen, että samat tiedot ovat eri hakulomakkeissa eri nimisissä kentissä. Tämä vaikeuttaa analytiikkaa, koska tarvittavat tiedot joudutaan etsimään käsin eikä tietojen yhdistäminen koneellisesti onnistu.

Tähän liittyvä havainto on selkeä: hakulomakkeiden kentistä tulee kehittää yhteinen, hyvin dokumentoitu tietomalli. Lomakkeita suunniteltaessa on järjestelmään kirjattava mitä tietomallin kohtaa mikäkin lomakkeen kohta vastaa. Tämä mahdollistaa koneluettavan, yhteismitallisen datan tuottamisen analytiikan ja koneoppimisen tarpeisiin. Malli tulee rakentaa joko olio-ohjelmoinnin peruskäsitteitä käyttäen perintämallina siten, että esimerkiksi kuntakoodi kysytään ja merkitään aina samalla tavalla. On tärkeää huomata, että hakija ja haettavan avustuksen kohde eivät ole sama asia. Yleishyödyllisten yhdistysten ja yhteisöjen kotipaikka on tyypillisesti Helsingissä postinumeroalueella 00100 riippumatta siitä, mihin kukin näiden haku kohdistuu. Tämän takia on tarpeellista kirjata erikseen hakijan ja hakukohteen kuntatieto ja vastaavasti esimerkiksi kouluihin liittyvissä hauissa kohdekoulun koulukoodi.

Halulomakkeiden täyttäminen kannattaa mahdollisuuksien mukaan automatisoida. Esimerkiksi Y-tunnuksella voidaan hakea hakijan tietoja suoraan autoritäärisestä lähteestä Patentti- ja rekisterihallitukselta (PRH). Tämä parantaa tunnistetietojen luotettavuutta ja vähentää hakijan työtä.

Kaikille hakemustyypeille pitää määrittää pienin yhteinen nimittäjä, joka tiedetään kaikista hakemuksista. Oletettavasti pienimpään yhteiseen nimittäjään sisältyvät ainakin

- Hakija, luettava nimi
- Hakija, uniikki tunniste
- Hakukohde (tyyppi)
- Hakukohde (uniikki tunniste/vast)
- Hakukierros
- Hakemuksen otsikko
- Tiivistelmäteksti hakemuksesta
- Haettu summa
- Myönnetty summa
- Hakuajankohta
- Päätösajankohta
- Myönnetty/ei myönnetty
- Asiasanat

Asiasanojen lisääminen on aluksi joko hakijan tai hakemuksen käsittelijän tehtävä. Kun dataa kertyy tarpeeksi, voidaan harkita siirtymistä asiasanojen ehdottamiseen.

Havainto 3: Hakukierrokset ovat erilaisia (segmentoituneita)

Näytteitä eli yksittäisiä hakemuksia oli 4000, mikä on suhteessa datan kompleksisuuteen melko vähän. Tästä seuraa useita asioita:

Opetusdataa on vähän. Tämä rajoittaa suoraan mahdollisten koneoppimisjärjestelmien tehtävänasettelua, koska monisyiset ennuste- tai diagnostiikkatavoitteet tarvitsevat enemmän opetusdataa.

Datan tuontia uuteen järjestelmään vanhoista tulee harkita. Varsinaiset koneoppimisjärjestelmät tarvitsevat opetusdataa. Mikäli monimutkaisia koneoppimisjärjestelmiä halutaan tehdä, tulee niille olla laadukasta opetusdataa historiasta. Myös ihmisten tekemät analyysit hyötyvät siitä, että vanhoista järjestelmistä tuodaan tietoa.

Prosesseja on monta, koneoppimistehtäviä on useita.

Jos data ei ole semanttisesti yhtenäistä, kuten OPH:n hakukierrosesimerkissä, on lisäksi huomattava, että tehtävänasetteluja on potentiaalisia jo pelkästään siksi, että hakukierrokset ovat luonteeltaan erilaisia. Tämä on huomioitava paitsi tehtävänasettelussa, myös koneoppimisen integroinnissa ja eri tehtäville käytettävissä olevan koulutusdatan määrän arvioinnissa. Mikäli hakukierrosta vastaavia hakuja ei ole ollut, ei koulutusdataa koneoppimisjärjestelmille käytännössä ole.

6 YHTEENLIITTYMÄT OSAPROJEKTIN 2 KANSSA

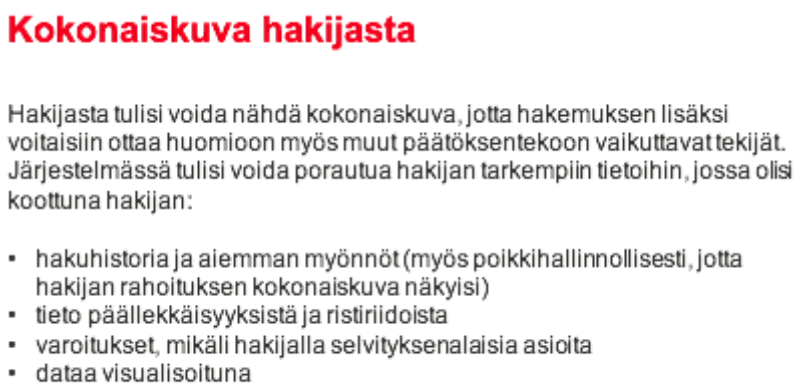
Tulevaisuuden hakuprosessia ja sen käytettävyyttä selvittäneen osaprojektin 2 nousi esille teemoja ja konkreettisia ehdotuksia, jotka liittyvät datan käyttämiseen tai käytettävyyteen. Osaprojektin 2 raportti suositellaan luettavaksi kokonaisuudessaan. Tähän on poimittu erityisiä kohtia raportista. Lainauksista ei saa kattavaa kuvaa osaprojektin tuotoksesta ja siksi se suositellaan luettavaksi kokonaan.

Hakijan tulee tunnistautua. Mikäli hakija hakee edustamansa yhteisön puolesta, on helpointa tarkistaa nimenkirjoitusoikeus automaattisesti PRH:n tiedoista. Samalla saadaan perustiedot hakevasta yhteisöstä, eikä ole perusteltua vaatia käyttäjää täyttämään niitä käsin ellei tiedoissa ole puutteita. Jos puutteita on, ne pitäisi ensisijaisesti korjata alkuperäiseen lähteeseen.

Hakijan tiedoista tulee erottaa haettavan kohteen tiedot. Yhteisöllä saattaa olla toimintaa monella alueella ja hakemus kohdistuu näistä vain yhteen tai muutamaan. Tämä tieto tulee kysyä, mieluiten siten, että sen oikeellisuus tarkistetaan koneellisesti. Helpointa on hakea alueen tiedot postinumeron perusteella tai koulun tiedot koulutunnisteen mukaan ulkoisesta tietokannasta ja pyytää lomaketta täyttävää henkilöä varmistamaan, että tiedot ovat oikein.

Erityisesti hakemusten käsittelyprosessi tulee mallintaa siten, että kun virkailija lukee hakemusta ja pyytää tarkennuksia, tästä jää selkeä jälki. Ajan myötä järjestelmään voidaan liittää ohjaavia toiminnallisuuksia, jotka käyttävät aiemmista hakukierroksista kertynyttä dataa. Järjestelmä voi esimerkiksi hakijoita täyttämään tietyt kentät erityisen huolellisesti, jos niistä on syntynyt paljon lisäselvityspyyntöjä käsittelyvaiheessa.

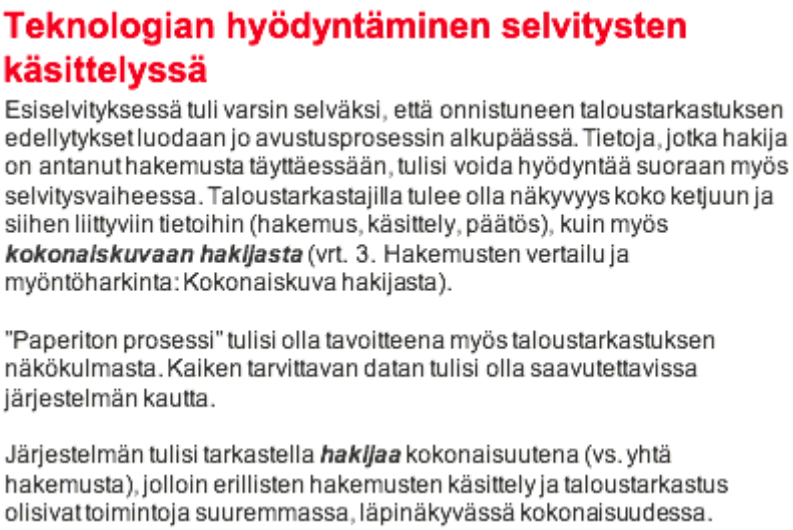
Yksinkertaiset hakulomakkeiden täyttöä ja käsittelyä tehostavat toimenpiteet ovat lupaavin tehostuskohde.



Kokonaiskuva hakijasta on helppo muodostaa, kun hakija on yksikäsitteisesti tunnistettu kaikissa hakemuksissa. Lisäksi kun haun kohde on tunnistettu, on mahdollista muodostaa myös **kokonaiskuva haun kohteesta**. Tällä nähdään kohteelle aiemmin myönnetyt avustukset riippumatta niiden hakijasta ja voidaan varmistua, että kaksi tahoa ei hae samaan kohteeseen valtionavustusta eri hauissa.

Koneoppimisella voidaan suodattaa hakemukset, jotka eivät täytä kriteerejä. Kuitenkin kriteerien määritelmä voi tässä yhteydessä olla käsitteellisesti vaikea koneoppimiselle. Kriteereiden luonnetta ei erityisesti avattu tässä yhteydessä. Mikäli kriteerit on helppo määritellä kerätyn datan perusteella, ei koneoppimista välttämättä tarvita.

Myöntöjaon esittäminen tuskin on mahdollista esimerkkidatan vähyyden takia. Myöntöpäätösten perusteet vaihtelevat joka hakukerralla painotuksista johtuen ja siksi kaksi hakukierrosta tulevat tuskin koskaan olemaan täysin samanlaisia. Toisaalta jos myönnettävä summa on pieni eikä esittelijän työaikaa kannata haaskata, järjestelmä voi esimerkiksi ehdottaa kaikkien muodollisesti pätevien hakemusten myöntämistä, jos hakija on hakenut tukea yhdelle kohteelle ja kohteelle on haettu tukea vain kyseisellä hakemuksella.



Mitä suurempi osa tiedoista haetaan automaattisesti, sitä varmemmin ne ovat niin oikeita kuin mahdollista. Jos haun kohde on yksikäsitteisesti määriteltävissä, tulee myös kohdetta tarkastella osana valmistelua.

7 SUOSITUKSET

1) Ensimmäisessä vaiheessa tekoälyä kannattaa hyödyntää erityisesti kokonaiskuvan muodostamisen apuna. Aiheprofiilien avulla hakemuksia voidaan luokitella ja löytää keskenään samankaltaisia hakemuksia. Riippuvuustarkastelun avulla voidaan tarkastella hakemusten ja myönnettyjen avustusten demografista kohdentumista.
2) Yksinkertaisia sääntöpohjaisia ja kevyesti oppivia koneoppimismalleja voidaan hyödyntää lomakkeiden täytön automaattisessa tukemisessa. Nämä kannattaa todennäköisesti antaa käyttäjälle tiedoksi neuvojen ja automaattisten korjaussuositusten muodossa.
3) Tekoälyn ja kehittyneen analytiikan hyödyntäminen vaatii sen, että opetusdataa uudesta järjestelmästä on kertynyt tarpeeksi. Tiedon tuomista vanhoista järjestelmistä tulee harkita, jotta data-analyysiin saadaan historiatietoa ja koneoppimiseen opetusdataa. Mitä pienempi määrä hakemuksia, sitä suurempi hyöty tekoälystä pitäisi saada, että sen käyttöönotosta voi olla hyötyä. Data-analytiikka kokonaisuudessaan tarjoaa työkaluja myös pienien avustusprosessien käyttöön ja valtionavustusten kokonaisuuden hallintaan. Koneoppimisen käyttö ja automatisoidut järjestelmät pienten hakujen yhteydessä tulee harkita tapauskohtaisesti. Kaikissa ei ehkä saada järjestelmän kehitykseen käytetylle rahalle vastinetta.
4) Tiedot haun kohteesta ja hakijasta pitää pyytää koneluettavasti ja mahdollisimman yksilöivästi, jotta kummastakin voidaan hakea tietoa muista järjestelmistä monimutkaisempia analyyseja sekä kokonaiskuvan muodostamista varten.
5) Pienin yhteinen nimittäjä hakijoiden, hakemusten ja päätösten suhteen tulee määritellä. Tätä osaa datasta voidaan käyttää koko prosessin kattavan analytiikan tekoon. Datamallin tulee tukea hakuprosessikohtaista dataa, sekä tämän datan tarjoamista tidy data -muodossa analyysia ja jatkoselvittelyjä varten.

   Pienimmän yhteisen nimittäjän löytämiseksi olisi syytä iteroida datamallia samalla kun järjestelmäkehitystä kilpailutetaan. Lähtökohtina voidaan pitää olemassa olevia datamalleja, mutta niiden saaminen voi olla haasteellista. Helpointa on lähteä liikkeelle julkisesti saatavilla olevista päätöstiedoista, joita on listattu osaprojektin 5 loppuraportin liitteessä.