Uudistuva ja osaava Suomi 2021 – 2027 EU:n alue- ja rakennepolitiikan ohjelma

VALINTAESITYS

Etelä-Savon maakunnan yhteistyöryhmän sihteeristö

Kokouspäivämäärä:	23.8.2023
Hakemusnumerot:	402183 Kaakkois-Suomen Ammattikorkeakoulu Oy, 402227 Kansallisarkisto, 402325 Elinkeinoelämän keskusarkisto sr.
Ryhmähanketunnus:	R-00603
Hankkeen julkinen nimi:	Tekoälyllä lisäarvoa digiarkistojen asiakkaille
Hankkeen päätoteuttaja:	Kaakkois-Suomen Ammattikorkeakoulu Oy
Osahankkeen toteuttajat:	Kansallisarkisto, Elinkeinoelämän keskusarkisto sr.
Maantieteellinen kohdealue:	Etelä-Savo /Juva, Kangasniemi, Mikkeli, Mäntyharju, Pertunmaa, Pieksämäki, Puumala, Rantasalmi, Savonlinna, Sulkava, Enonkoski, Hirvensalmi
Toimintalinja:	1 Innovatiivinen Suomi
Erityistavoite:	1.2 Digitalisaation etujen hyödyntäminen kansalaisten, yritysten ja julkishallinnon hyväksi
Hakemuksen saapumispäivämäärä:	16.3.2022
Hankkeen alkamispäivämäärä:	1.9.2023
Hankkeen päättymispäivämäärä:	31.8.2024
Viranomainen:	Etelä-Savon maakuntaliitto
Käsittelijä:	Eveliina Pekkanen

Hankkeen sisältö

Tiivistelmä

Hankkeen tavoitteena on edistää tekoälyn hyödyntämistä aineistojen löydettävyyden ja jatkokäytön kannalta erityisesti loppukäyttäjien näkökulmasta. Tekoäly on vahva muutosvoima ja mahdollisuus, jonka hyödyntäminen vaatii pitkäjänteistä panostusta ja jatkuvaa osaamisen kerryttämistä. Tämä hanke kasvattaa tätä tarvittavaa osaamista ja ymmärrystä tekoälyn hyödyntämispotentiaalista kytkeytyen konkreettiseen menetelmäkehitykseen ja pilotointiin.

Kasvavien aineistomäärien turvaaminen jälkipolville ja niiden hyödynnettävyyden varmistaminen vaatii uudenlaisia teknologisia ratkaisuja tuekseen ollakseen kestävää. Hyödyntäjäorganisaatioiden parissa on todettu olevan todellista tarvetta sekä tekstin että sen sisältämien merkitysten tunnistamisessa riittävän luotettavasti. Hankkeessa kehitetään digitaalisten aineistojen automaattista sisällönanalyysia tehostaen aineistojen kuvailua ja luoden loppukäyttäjille arvokkaita haku- ja yhdistelymahdollisuuksia ontologian ja tekoälyn avulla. Kun arkistoiduille aineistoille saadaan ontologiset tunnisteet, haut on mahdollista kohdentaa paremmin toivottuihin aineistosisältöihin ja hyödyntämispotentiaali kasvaa. Tämä avaa myös uusia mahdollisuuksia kansalaisyhteiskunnassa, liiketoiminnassa ja TKI-toiminnassa.

Hankkeen kolme keskeistä tavoitetta ovat: 1) Edistää tietojen automaattista poimintaa erilaisista aineistoista jatkohyödyntämistä varten 2) Kehittää aineistojen sisällöllisten merkitysten tunnistamista käyttäjätarvetta palvellen 3) Konkretisoida automaattisen sisällönkuvailun lisäarvo. Tavoitteista hyötyvät ensisijaisesti digiarkistojen loppukäyttäjät, KAM-sektorin toimijat sekä jollakin tavalla datanhallintaa tekevät yritykset ja toimijat. Hyödynnettävät ontologiasanastot ovat monikielisiä, mikä edesauttaa myös yritysten pyrkimyksiä tarjota sisältöjä kansainvälisille asiakkaille.

Hanke mahdollistaa paremman asiakaskokemuksen luomisen loppukäyttäjille ja samaan aikaan se tuo henkilötyöaikaa säästävät tekoälyautomaation hyödyt yritysten ja arkistotoimijoiden saataville. Kehittämistyön tuloksia voidaan luontevasti soveltaa uudenlaisten palveluiden rakentamiseen ja käyttäjäkokemuksen rikastamiseen. Hankkeen jälkeen kaikki tuotetut tekniset komponentit ovat saatavilla avoimena lähdekoodina ja toimiva demoympäristö on käytettävissä Memory Labin infran kautta. Käyttötukea ja opastusta tarjoaa Memory Campuksen toimijaverkosto

Tarve

Hanke kehittää digitaalisten aineistojen automaattista sisällönanalyysia tehostaen aineistojen kuvailua ja luoden loppukäyttäjälle arvokkaita hakumahdollisuuksia. Tekoälyn avulla voidaan tarkastella, tutkia ja hyödyntää tietoa uusin tavoin. Hankkeen myötä erilaisten arkistoitujen aineistojen hyödyntämispotentiaali kasvaa ja uudenlaisia yhdistelymahdollisuuksia avautuu. Hanke tehostaa digitaalista tiedonhallintaa ja mahdollistaa sen soveltamisen suuriinkin aineistomassoihin, joiden manuaalinen läpikäynti tai kuvailu ei olisi rationaalisessa ajassa mahdollista.

Aiemmissa hankkeissa (mm. DALAI) on pilotoitu ja kehitetty automaattisia sisällönanalyysin menetelmiä. Nämä ovat keskittyneet pääsääntöisesti digitointiprosessin laadunvarmistukseen, ei loppukäyttäjien tarpeisiin. Keskeinen pullonkaula, on ollut asiakirjoissa esiintyvän tekstin muuttaminen koneluettavaan muotoon riittävällä laatutasolla. Tekstintunnistus on edellytys käyttäjien toivomien edistyneempien menetelmien käytölle sisällönanalyysissä. Hanke taklaa näitä ongelmakohtia ja jatkaa sisällöntunnistuksen kehittämistä kartoittamalla loppukäyttäjien erilaisia tarpeita ja keskittymällä uusiin analyysimenetelmiin.

Toistaiseksi muistiorganisaatioissa ja niiden asiakaskunnassa on rajoitetusti osaamista tekoälystä ja sen hyödyntämismahdollisuuksista. Tämän hankkeen myötä osaaminen kasvaa kytkeytyen konkreettiseen menetelmäkehitykseen ja kokeiluihin. Hanke edistää myös digitalisaation hyödyntämistä laajemmassa mittakaavassa ja digitaalisten työmenetelmien käyttöönottoa. Esimerkiksi Memory Labin liiketoimintaselvityksen perusteella (Osaango ja Xamk, 2022) datankäsittelyyn ja tekoälyyn liittyvästä osaamisesta on pulaa. Tämä hanke tukee suoraan molempia aihealueita ja näitä hyödyntävien ohjelmistoratkaisuiden syntymistä tuottamalla avoimia ohjelmistokomponentteja ja lisäämällä osaamista alalla.

Tavoite

Hankkeen päätavoitteena on edistää tekoälyn hyödyntämistä aineistojen tarjoamisessa jatkokäyttöön. Hanke kehittää digitaalisten aineistojen automaattista sisällönanalyysia tehostaen aineistojen kuvailua ja luoden loppukäyttäjille arvokkaita haku- ja yhdistelymahdollisuuksia. Hankkeen keskeiset tavoiteet ovat

1) Edistää tietojen automaattista poimintaa erilaisista aineistoista jatkohyödyntämistä varten
2) Kehittää aineistojen sisällöllisten merkitysten tunnistamista käyttäjätarvetta palvellen
3) Konkretisoida automaattisen sisällönkuvailun lisäarvo

Kohderyhmä

1) Digitaalisten arkistojen ja arkistoaineistojen loppukäyttäjät, kuten kansalaiset ja tutkijat, jotka voivat etsiä, löytää ja yhdistellä tietoja aiempaa monipuolisemmin taustalla olevan tekoälyn avustamana.
2) KAM-sektorin toimijat (kirjastot, arkistot, museot, muut Memory Campus -yhteisön organisaatiot) ja näiden toimijoiden asiakasorganisaatiot ml. yritykset, joille hanke mahdollistaa palveluiden kehittämisen paremman asiakaskokemuksen tarjoamiseksi vähemmällä henkilötyöllä.
3) datanhallintaa ja dataperustaista liiketoimintaa kehittävät yritykset ja start upit, jotka saavat hankkeen ratkaisuista uusia mahdollisuuksia datan jatkohyödyntämiseen ja tiedonhallinnan palveluiden kehittämiseen.

Toimenpiteet

Hankkeen keskeiset toimenpiteet ovat seuraavat:

Tavoitteeseen 1 liittyen (9/2023–5/2024):

- Parannetaan olemassa olevien automaattisten analyysimenetelmien toimivuutta
- Selvitetään mahdollisuuksia olemassa olevien automaattisten tekstintunnistuksen (OCR) mallien jatkokouluttamiseen tekstintunnistuksen laadun parantamiseksi arkistoaineistosta. Verrataan eri malleilla saatuja tuloksia ja toteutetaan paras malli. Tekstintunnistuksen ratkaisu julkaistaan avoimen lähdekoodin toteutuksena.
- Kehitetään segmentointia eli asiakirjatyyppien rakenteistamista tietojen poimintaa varten. Rakenteistettu malli kuvaa mistä kohdasta dokumenttia löytyy mitäkin käyttäjälle hyödyllistä tietoa ja tukee siten tietojen automaattista poimintaa. Segmentointiratkaisu julkaistaan avoimen lähdekoodin toteutuksena.
- Yhdistetään tekstintunnistus- ja segmentointiratkaisu työnkuluksi, jolla voidaan toteuttaa mahdollisimman automaattinen tietojen poiminta
- Pilotoidaan tietojen poimintaa soveltuvalla aineistotyypillä (historiallisilla aineistoilla kuten perunkirjoilla tai sotilaskantakorteilla, ja yritysaineistoilla) ja tutkitaan mahdollisuuksia laajentaa muihin yleisemmin käytettyihin asiakirjatyyppeihin
- Käydään jo avoimena olevat kantakortit läpi uudella AI-menetelmällä ja testataan löytyneitä tietoja tutkijoilla rinnakkain.

Päävastuu tavoitteesta yksi on Kansallisarkistolla. Xamk osallistuu segmentoinnin kehittämiseen. Elka tuottaa opetusaineistoja ja osallistuu pilotointiin.

Tavoitteeseen 2 liittyen (12/2023–8/2024):

- Selvitetään arkistojen asiakkaiden tarpeita automaattiseen sisällönkuvailuun ja aineistojen yhdistämiseen kokemusnäkökulmasta laadullisena fokusryhmähaastatteluna
- Kehitetään malli, joilla erilaisia arkistoaineistojen sisällönkuvailuja voidaan rikastaa semanttisesti ja yhdistää toisiinsa palvelemaan asiakkaiden aineistotarpeita (liiketoiminta, kehittämistyö jne.). Sisältöanalyysien tulokset linkitetään olemassa oleviin ontologioihin ottamalla käyttöön Memory Labissa ontologiapalvelin. Toimenpiteessä hyödynnetään Annifia, joka on Kansalliskirjastossa kehitetty avoimen lähdekoodin työkalu. Ratkaisu tukee automaattisesti esim. monikielisyyttä eli löytää tuloksia myös eri kielellä kuin alkuperäinen haku (kts. metsänhoito esimerkki).
- Pilotoidaan semanttisten analyysien, asiasanoituksien rikastamisen ja haun mahdollisuuksia arkistoissa perustuen olemassa olevien ontologiasanastojen lisäksi avoimiin tekoälymenetelmiin ja pyritään mahdollistamaan aineistojen koneellinen linkkaus tunnistettujen entiteettien ja ontologiatietojen avulla.

Asiasanoitetaan luokiteltuja ja poimittuja dokumentteja, haetaan merkitys sanastoista ja tallennetaan URI metatietoihin.
- Demotaan semanttisten tunnisteiden linkitystä esimerkiksi jollain ÄES-kärjen aineistolla. Demon toteutukseen haetaan yrityskumppania tai yhteisökumppania esim. Liikearkistoyhdistyksen kautta

Päävastuu tavoitteesta kaksi on Xamkilla. Kansallisarkisto ja Elka tuottavat opetusaineistoja, joita tarvitaan kehittämisen pohjana ja osallistuvat pilotointeihin ja kehitettyjen toimintojen testaukseen

Tavoitteeseen 3 liittyen (9-12/2023, 3-8/2024):

- Arvioidaan tavoitteissa 1&2 kehitettyjen automaattisten menetelmien suorituskykyä ja lopputulosta käyttäjätestauksen avulla. Toteutetaan käyttäjälähtöinen vertailu, jossa arvioidaan testiarkistossa aiempien ja nyt kehitettyjen menetelmien lisäarvoa loppukäyttäjille. Vertailuun testaajiksi rekrytoidaan edustajia paitsi henkilöloppukäyttäjistä (kansalainen, tutkija) myös yhdistys/yritystoimijoista ja pyritään samalla luomaan testaajapoolia. Tehtyjen testien ja demojen pohjalta voidaan arvioida määrällisesti ja laadullisesti, kuinka hyvin tietojen automaattinen poiminta ja merkitysten linkitys toimii
- Tehdään määrittelytyö aineistojen vastaanottolaiturin (Memoriaali) ja sisällönanalyysityökalujen (DALAI ja tämä hanke) yhdistämiseksi helposti siirrettäväksi ja käyttöönotettavaksi järjestelmäksi.
- Dokumentoidaan tehdyn menetelmäkehityksen ja käyttäjäautomaattisen sisällönkuvailun rooli ja arvioidaan sen potentiaalia digiaineistojen hallinnassa edellisten tavoitteiden ja käyttäjätestauksen pohjalta. Laaditaan tästä avoin raportti suosituksineen.

Päävastuu tavoitteesta kolme on Xamkilla. Kansallisarkisto ja Elka osallistuvat käyttäjätarveselvitykseen sekä testaajarekrytointiin omista asiakasverkostoistaan. Kansallisarkisto ja Elka tuottavat testiarkistoon aineistot ja analyyseja, joita tarvitaan vertailussa. Elka osallistuu määrittelytyöhön aineistojen vastaanoton ja sisällönanalyysin yhdistämiseksi.
Kaikki osallistuvat raportointiin ja suositusten laadintaan.

Rahoitusmallin flat rate –osuutta on ajateltu käyttää mm. käyttäjätestauksen kuluihin, menetelmäkehitykseen liittyviin kuluihin esim. lisensoinnista sekä hankehenkilöstön matkakuluihin liittyen sidosryhmäyhteistyöhön.

Tulokset

Hankkeen tavoitteet tukevat elinkeinojen ja TKI-toiminnan uudistumista edistämällä digitalisaation hyödyntämistä ja laajentamalla datan käyttömahdollisuuksia. Hankkeen TKI-työssä hyödynnetään uusimpia tekoälymenetelmiä aineistojen sisällönanalyysiin ja yhdistämiseen. Hanke tuo Etelä-Savon alueelle uutta tekoälyosaamista ja vahvistaa alueen painoarvoa edistyneiden digitaalisten menetelmien kehittäjänä, osaajana ja soveltajana.

Hankkeen konkreettisina tuloksina syntyy:
Tavoitteeseen yksi liittyen

- Tietojen poimintaan tarkoitetut tekstintunnistus- ja segmentointiohjelmistokomponentit on yhdistetty työnkuluksi joka voidaan ottaa käyttöön missä tahansa API-ratkaisuja hyödyntävässä järjestelmässä. Komponentit ovat testattavissa Memory Labin ympäristössä ja niiden lähdekoodit ovat vapaasti saatavilla. Ohjelmistokoodi julkaistaan avoimena koodina.
  Latausmääriä voidaan seurata.
- Uusi OCR tunnistuksen malli valmiina hyödynnettäväksi
- Kantakorttien tuloksien vertailu vanhalla ja uudella menetelmällä saatavilla raporttimuodossa

Tavoitteeseen kaksi liittyen

- Fokusryhmähaastattelujen tulokset avaava avoimesti saatavilla oleva raportti. Raportti luo tilannekuvan arkistojen käyttäjien tarpeista sisällönanalyysin suhteen ja tukee kehittämistä tässä hankkeessa ja laajemmin.
- Sisältöanalyysejä ja OCR-tietoja ontologialla laajentava ohjelmistokomponentti, joka voidaan ottaa käyttöön missä tahansa API-ratkaisuja hyödyntävässä järjestelmässä. Komponentti on testattavissa Digitalian / Memory Labin ympäristössä ja vapaasti saatavilla. Ohjelmistokoodi julkaistaan avoimena koodina. Latausmääriä voidaan seurata.
- Demo, joka konkretisoi semanttisen analyysin tulokset jollakin ÄES alalla. Julkaistaan mahdollisimman avoimena datana, riippuen aineiston käyttöoikeuksista. Memory Labissa toimiva ontologia palvelin on vapaasti kehittäjien käytettävissä

Tavoitteeseen kolme liittyen

- Raportti käyttäjätestauksesta, jossa on selvitetty automaattisen tietojen poiminnan sekä semanttisen analyysin tehokkuus ja hyödyllisyys loppukäyttäjien näkökulmasta. Määrälliset ja laadulliset tulokset kuvaavat kehitettyjen teknisten ratkaisujen toimivuutta ja kypsyyttä. Raportin latausmääriä ja viittauksia voidaan seurata.
- Suositukset automaattisen sisällönanalyysin, tietojen poiminnan ja semanttisen yhdistämisen hyödyntämisestä erityyppisille aineistoille sekä jatkokehitysaihioiden tunnistaminen.
- Määrittely aineistojen vastaanottotyökalun, automaattisten analyysityökalujen, manuaaliseen sisällönkuvailun ja joukkoistuksen yhdistämisestä yhdeksi konttiratkaisuna toimivaksi ja helposti siirrettäväksi ohjelmistoratkaisuksi.

Vaikutukset

Hanke auttaa luomaan Etelä-Savoon uutta tekoäly- ja digiosaamista ja uusia liiketoimintamahdollisuuksia. Aineistojen automaattisen analyysin myötä voidaan toteuttaa täysin uudenlaisia julkisia ja yksityisiä palveluita sekä dataperusteista liiketoimintaa. Vaikutuksia todennetaan loppukäyttäjätutkimuksella sekä aineistoon liitettyjen metatietojen ja ontologiatietojen määrän muutoksella. Hankkeella edistetään aineistojen digitalisoitumista. Näin yhä suurempi osa niin julkishallinnon aineistoista kuin yrityssektorinkin aineistoista on hyödynnettävissä digitaalisessa muodossa.
Hanke kiihdyttää myös TKI-toimintaa avaamalla uusia mahdollisuuksia aineistojen saatavuuteen ja yhdistelyyn.

Toiminnan jatkuvuus

Hankkeeseen osallistuvat organisaatiot sekä sidosryhmäkumppanit saavat hankkeen kautta uusinta tietoa teknologisista mahdollisuuksista ja pystyvät arvioimaan niiden vaikutusta ja potentiaalia omassa toiminnassaan. Tämä tietämys leviää laajemmin hyödynnettäväksi käyttäjätestauksen sekä Memory Campus –ekosysteemin toiminnan kautta. Näin kaikissa potentiaalisissa hyödyntäjä organisaatioissa kasvaa laajemminkin ymmärrys siitä millaisia palveluita menetelmien päälle voidaan kehittää ja millaisia liiketoimintamahdollisuuksia ne voivat avata.

Hankkeen ratkaisut ovat pk-yritysten avoimesti hyödynnettävissä. Menetelmiä voi soveltaa moniin erilaisiin aineistoihin. Yritykset voivat ratkaisujen avulla analysoida omia aineistojaan tai tarjota palveluitaan muille. Ratkaisujen käyttöönottoon on tarjolla tukea Memory Campuksen ja Memory Labin kautta myös hankkeen päätyttyä. Elka tukee omien asiakasyritystensä toimintaprosessien digitalisoitumista hankkeen tulosten pohjalta. Kansallisarkisto kehittää hankkeen tulosten pohjalta omia digitaalisia aineisto- ja tietopalveluitaan edelleen.

Memory Campuksen ekosysteemi hyödyntää tuloksia toiminnassaan. Tavoitteena on, että alueen muistiorganisaatiot ovat eturintamassa hyödyntämässä uudenlaista teknologiaa. Hankkeen tekniset tuotokset jäävät vapaasti hyödynnettäviksi ja niitä ylläpidetään Memory Labissa, jonka ympäristössä niitä voi myös koekäyttää ja testata vapaasti. Tämä tuottaa mahdollisuuksia myös yrityksille, jotka jo hyödyntävät tai voisivat hyödyntää arkistoaineistoja liiketoiminnassaan.

Kustannukset ja rahoitussuunnitelma

Ryhmähankkeen kokonaiskustannukset: 382 636 €
Haettu EU- ja valtion rahoitus: 286 977 €
Omarahoitus: 95 659 €

Kaakkois-Suomen ammattikorkeakoulu Oy

Hankkeen kokonaiskustannukset: 166 978 €
Palkkakustannukset: 119 270 €
Flat rate (40 %): 47 708 €
Haettu EU- ja valtionrahoitus: 125 234 € (75 %)
Omarahoitus: 41 744 € (25 %)

Kansallisarkisto

Hankkeen kokonaiskustannukset: 161 987 €
Palkkakustannukset: 115 705 €
Flat rate (40 %): 46 282 €
Haettu EU- ja valtionrahoitus: 121 490 (75 %)
Omarahoitus: 40 497 € (25 %)

Elinkeinoelämän keskusarkisto sr.

Hankkeen kokonaiskustannukset: 53 671 €
Palkkakustannukset: 38 336 €
Flat rate (40 %): 15 335 €
Haettu EU- ja valtionrahoitus: 40 253 € (75 %)
Omarahoitus: 13 418 (25 %)

Hakemuksen käsittely ja lisätietoja hakemuksesta

Ryhmähankehakemus on saapunut Etelä-Savon maakuntaliiton kevään 2023 Uudistuva ja osaava Suomi 2021–2027 EAKR-hakuun toimintalinjalle 1 Innovatiivinen Suomi ja erityistavoitteeseen 1.2 Digitalisaation etujen hyödyntäminen kansalaisten, yritysten ja julkishallinnon hyväksi. Ryhmähankkeen hakemukset ovat saapuneet rahoittajalle haun määräaikaan 17.3.2023 mennessä.

Rahoittajan arvio hankkeesta

Hanke on EU:n alue- ja rakennepolitiikan ohjelman mukainen ja se toteuttaa toimintalinjan 1 Innovatiivinen Suomi erityistavoitetta 1.2 Digitalisaation etujen hyödyntäminen kansalaisten, yritysten ja julkishallinnon hyväksi. Hankkeen tavoitteena on edistää tekoälyn hyödyntämistä aineistojen löydettävyyden ja jatkokäytön kannalta erityisesti

loppukäyttäjien näkökulmasta. Hanke auttaa luomaan Etelä-Savoon uutta tekoäly- ja digiosaamista ja uusia liiketoimintamahdollisuuksia. Aineistojen automaattisen analyysin myötä voidaan toteuttaa täysin uudenlaisia julkisia ja yksityisiä palveluita sekä dataperusteista liiketoimintaa. Hankkeella edistetään aineistojen digitalisoitumista. Näin yhä suurempi osa niin julkishallinnon aineistoista kuin yrityssektorinkin aineistoista on hyödynnettävissä digitaalisessa muodossa. Hanke kiihdyttää myös TKI-toimintaa avaamalla uusia mahdollisuuksia aineistojen saatavuuteen ja yhdistelyyn.

Hanke liittyy Etelä-Savon maakuntastrategian 2030 tavoitteeseen Uudistuvat elinkeinot ja älykäs erikoistuminen sekä Etelä-Savon älykkään erikoistumisen strategian läpileikkaavaan digitalisaation teemaan.

Rahoittajan esitys

Hanketta esitetään hyväksyttäväksi.

Ratkaisun mahdolliset perustelut ja jatkotoimenpiteet
Etelä-Savon maakuntaliiton pisteytys 27.4.2023
Etelä-Savon maakuntaliiton hankeryhmän kokous 31.5.2023Etelä-Savon maakuntaliiton yhteistyöryhmän sihteeristö 23.8.2023