Henkilötietoja sisältävien asiakirjojen automaattinen anonymisointi ja sisällönkuvailu -hanke

Kokemustenvaihtotilaisuus

LVM/Lars Sonckin Sali, 12.3.2019

Aki Hietanen, Arttu Oksanen, Saara Packalén, Minna Tamper

OIKEUSMINISTERIÖ JUSTITIEMINISTERIET

Anoppi-hanke

- Tausta ja tavoitteet
  - Miksi hanke on käynnistetty?
  - Keitä kumppaneita tai kenen kanssa hanketta tehdään?
  - Mitä hankkeessa syntyy lopputuloksena?
  - Mitä hyötyä hankkeen lopputuloksista on ja kenelle?
- Tuotteistettavat sovellukset ja teknologia
  - Mitä hankkeessa konkreettisesti tehdään?
  - Mitä teknologiota hankkeessa hyödynnetään?
- Kokemukset, opit ja tuotteiden jatkohyödyntäminen
  - Eteneekö hanke suunnitellusti vai onko eteen tullut karikoita?
  - Mitä hankkeessa on opittu?
  - Miten muut voisivat lopputulosta hyödyntää?

Tausta ja tavoitteet

- Hanke pohjautuu Severi, Linked Data Finland ja Semanttinen Finlex –hankkeisiin ja niissä tehtyihin prototyyppeihin
- Hanke asetettu ajalle 1.10.2018 – 30.9.2020
- Yhteistyöhanke
  - Oikeusministeriö, Helsingin yliopiston HELDIG-keskus, Aalto-yliopisto ja Edita Publishing Oy
- Julkisella sektorilla tuotetaan valtava määrä tietoaineistoa, jonka saaminen avoimesti niin viranomaisten, yritysten kuin kansalaistenkin käyttöön olisi hyödyllistä, mutta tietosuojasäädösten vuoksi se ei ole mahdollista
  - Viranomaiskäytännön ja lainkäytön tutkimus
  - Hallinnon ja oikeuskäytännön läpinäkyvyys
  - Aiempien päätösten/ratkaisujen hyödyntäminen uusien asioiden käsittelyssä
  - Kansalaisten oikeusturvan parantuminen
  - Avoimen datan saatavuus

Tausta ja tavoitteet – tuomioistuinratkaisujen saatavuus

Tuomioistuinten ratkaisemien asioiden vs. verkossa julkaistujen vuosittaiset määrät

- Korkein oikeus
  - n. 2400 ratkaistua asiaa, joista julkaistiin n. 130 kpl
- Korkein hallinto-oikeus
  - n. 6600 ratkaistua asiaa, joista julkaistiin n. 450 kpl
- Hovioikeudet
  - n. 6900 ratkaistua asiaa, joista julkaistiin n. 50 kpl
- Hallinto-oikeudet
  - n. 22 000 ratkaistua asiaa, joista julkaistiin n. 60 kpl

Käytännön tavoitteista

- Hankkeessa otetaan käyttöön automaattinen anonymisointityökalu ANOPPI ja automaattinen annotointisovellus APPI

- Hankkeen keskiössä tuomioistuinten ratkaisut
  - Tietosuojakysymysten vuoksi ratkaisuaineistoja ei voi sellaisenaan julkaista verkossa
  - Tarve anonymisoida/pseudonymisoida/peittää tietoja (esim. ”Henkilö A”) sekä julkaistavissa tuomioistuinten ratkaisuissa että tietopyyntöihin vastattaessa
  - Nykyisin hidasta ja kallista käsityötä, minkä vuoksi vain pieni osa ratkaisuista julkaistaan

- Hankkeessa tuotteistettavien sovellusten avulla
  - Voidaan lisätä julkaistavien tuomioistuinratkaisujen määrää
  - Vähennetään manuaalista työtä henkilötietojen tai salassa pidettävien tietojen peittämisessä asiakirjoista
  - Mahdollistetaan älykkäät haut suurista aineistomassoista ja linkitys dokumenttien välillä
  - Jatkossa arvioidaan käyttömahdollisuudet koko julkishallinnossa

Tuotteistettavat sovellukset ja teknologia (1)

- Anoppi
  - Verkkopalvelu asiakirjatekstien (Word, HTML, XML) automaattiseen anonymisointiin (pseudonymisointiin)
  - Käyttäjän mahdollista muokata automaattisen pseudonymisoinnin tulosta verkkosovelluksessa
- Appi
  - Työkalu automaattiseen annotointiin ja linkitykseen
  - Kontekstuaalinen lukija
  - Asiakirjoja eri teemojen mukaan luokitteleva hakusovellus

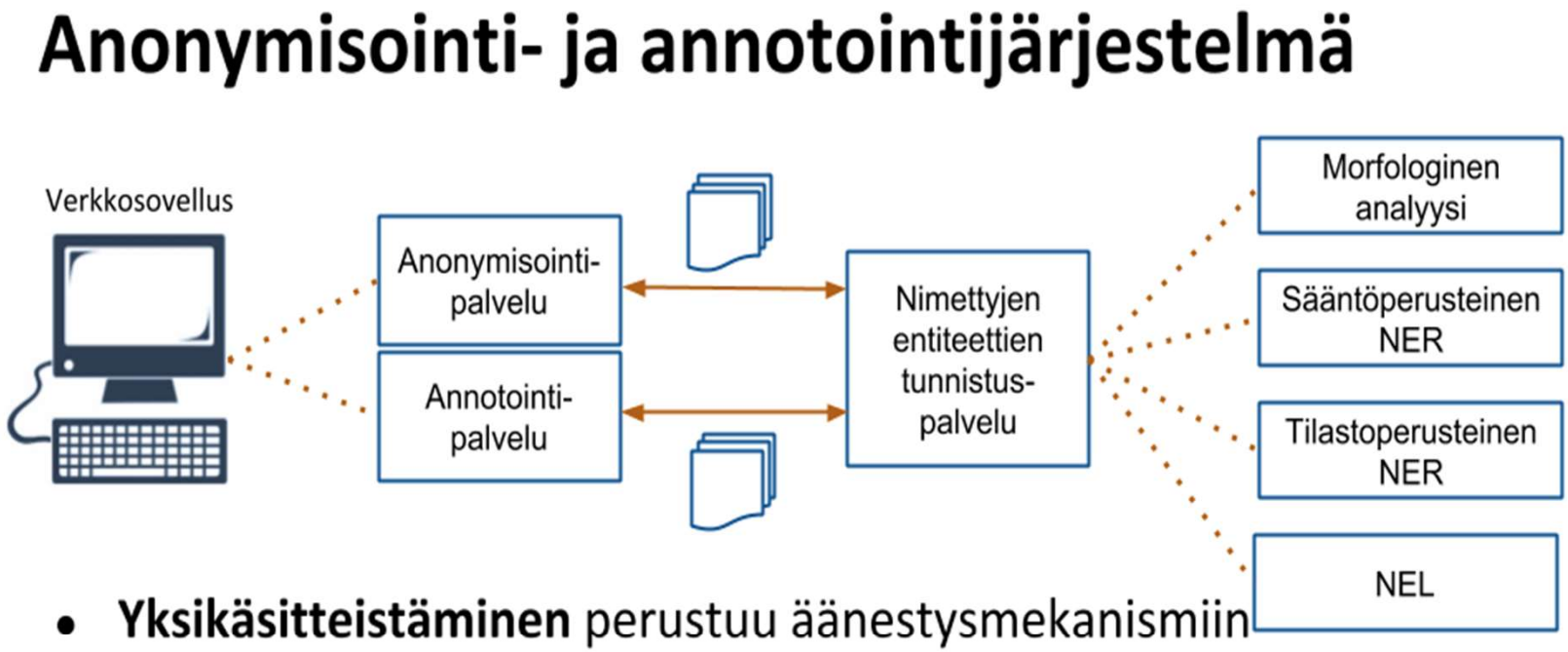
Tuotteistettavat sovellukset ja teknologia (2)

- HTML5/JavaScript, Scala/Java, Python, SPARQL, RDF

- Luonnollisen kielen käsittely
  - FiNER, sääntöperusteinen nimettyjen entiteettien tunnistin
  - ARPA, linkittää tekstin ontologioihin
  - TurkuNLP Finnish Dependency Parser, tekstin jäsennin
  - Säännöllisiä lausekkeita hyödyntävä tunnistin päivämäärille, numeraalisille tunnisteille ym.
  - Erillinen henkilönnimien tunnistin

- Koneoppiminen (kertyneeseen dataan perustuva oppiminen)
  - Gensim, TensorFlow (täsmentyy myöhemmin)

Arkkitehtuuri



- **Yksikäsitteistäminen** perustuu äänestysmekanismiin
  - Taustalla useita eri luonnollisen kielen käsittelyyn ja nimettyjen entiteettien tunnistukseen (NER) sekä linkittämiseen (NEL) kehitettyjä työkaluja
  - Semanttisia annotointeja tuotetaan käyttäen NEL:iä

Kokemukset, opit ja tuotteiden jatkohyödyntäminen

- Hanke saanut positiivisen vastaanoton ja herättänyt kiinnostusta
  - EU:n tietosuoja-asetuksen (GDPR) mukanaan tuoma tarve ja kiinnostus anonymisointivälineiden kehittämiseen muuallakin EU:ssa
  - Yhteistyö EU-maiden kanssa käynnistymässä
- Asiakirjojen pseudonymisointiin ei ole yhtä ainoaa toimintatapaa
  - Eri organisaatioissa tarpeet ja toimintaprosessit erilaisia
  - Tuomioistuimissa lähtökohtaisesti asianosaisten henkilönimet peitetään, kuitenkin esim. tuomareiden ja asiantuntijoiden nimet jätetään näkyviin
  - Tilannekohtainen harkinta välttämätöntä, esim. paikannimen ja/tai ammattinimikkeen peittämisen tarpeellisuus tapauskohtaista
  - Sovellusten integroiminen toimintaprosesseihin
  - Erityispiirteiden huomioon ottaminen Anoppi-sovelluksessa
- Sovellukset tulevat perustumaan avoimeen lähdekoodiin, mikä mahdollistaa niiden monipuolisen jatkokäytön

OIKEUSMINISTERIÖ

www.oikeusministerio.fi

Facebook: facebook.com/oikeusministerio
Twitter: twitter.com/oikeusmin @oikeusmin

Käyntiosoite: Eteläesplanadi 10, Helsinki
Postiosoite: PL 25, 00023 Valtioneuvosto

Vaihde: 029 516 001

Sähköposti
Anoppi-hanke@om.fi

Hanketiedot
https://oikeusministerio.fi/hanke?tunnus=OM042:00/2018
SeCo-tutkimusryhmän verkkosivut
https://seco.cs.aalto.fi/projects/anoppi/