Digitaalisen tiedon kesäkoulu 2019

Miia Kosonen
TKI-asiantuntija
Xamk
Digitalia

Järjestyksessään jo neljäs Digitalian toteuttama digitaalisen tiedon kesäkoulu järjestettiin Helsingin yliopistolla 20.–21. elokuuta. Teemapäiviksi valikoituivat Tiedonhallinnan käytännöt ja sääntely sekä Eettinen vastuu. Osallistujia oli ensimmäisessä päivässä 68 ja toisessakin yli 50.

Yhden odotetuimmista ja myös jälkeenpäin kiitetyimmistä esityksistä piti tietohallintoneuvos Tommi Oikarinen Valtiovarainministeriöstä. Selkeässä kokonaisuudessa käytiin läpi tiedonhallinnan yleislainsäädäntö. Etukäteen puheenvuoro oli otsikoitu tarkoituksella löyhästi "ajankohtaiskatsaukseksi", koska ei ollut tiedossa, mihin asti muutosten valmistelussa päästään.

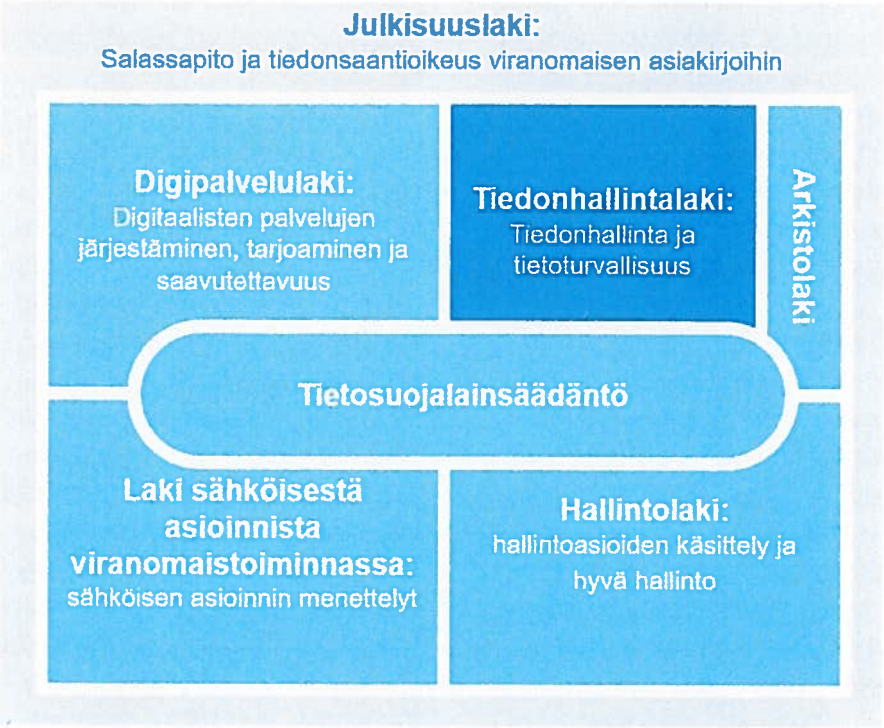
Vaikka puhutaan digitaalisuudesta, paljon kuljetetaan edelleen paperiaineistoa asiakkaan kautta, Oikarinen muistutti. Kuva 1 tiivistää tiedonhallinnan sääntelyn kokonaisuuden. Julkisuuslaki on luonnollisesti primäärilaki, jossa määritellään salassapito ja tiedonsaantioikeus viranomaisen asiakirjoihin.

Eduskunta hyväksyi esityksen tiedonhallintalaista 18.3.2019 ja laki julkisen hallinnon tiedonhallinnasta (906/2019) vahvistettiin 9.8.2019. Laki tulee voimaan vuoden 2020 alusta. Sitä sovelletaan tiedonhallintaan ja tietojärjestelmien käyttöön, kun julkisuuslain piirissä olevat viranomaiset käsittelevät tietoaineistoa.

Uudella tiedonhallintalailla on kolme tavoitetta: yhteentoimivuuden edistäminen, aineistojen laadun varmistaminen (ajantasaisuus, virheettömyys, käyttökelpoisuus, tietoturvallisuus) sekä hyvän hallinnon toteuttaminen tehokkailla tiedonhallintamenettelyillä. Taustalla ovat tietojen saatavuutta ja uudelleenkäyttöä koskeva PSI-direktiivi, palveluperiaate ja hallinnon tehokkuus ja tuloksellisuus.

Aina palveluperiaate ei toteudu. Oikarinen kertoi tapauksesta, jossa oli kokonaan kieltäydytty luovuttamasta asiakkaalle häntä koskevia asiakirjoja siksi, että hän ei tiennyt niiden diaarinumeroa. Pyynnöt on luonnollisesti voitava yksilöidä, mutta lain mukaan "tiedon pyytäjää on diaarin ja muiden hakemistojen avulla *avustettava* yksilöimään asiakirja". (Laki viranomaisten toiminnan julkisuudesta, 13 §, kursivointi lisätty.)

Lain käsitteistö on tarkoitus yhdenmukaistaa myöhemmin erityislainsäädäntöön. Tiedonhallintalaki kohdistuu aineistoihin, jotka koostuvat asiakirjoista, tai tiedoista, joista voidaan muodostaa asiakirjoja. Kiinnostava kysymys onkin tällöin tiedon ja asiakirjan suhde.



Kuva 1. Tiedonhallinnan sääntely. Kuva: Tommi Oikarinen, VM.

Näitä ei määritellä laissa. Tietona pidetään "asiakirjaan rinnastettavia tietoja", joista voidaan muodostaa hakuperusteilla asiakirjoja. **Tietoaineisto** on viranomaisten asiakirjoista muodostuva tietokokonaisuus. **Tietovaranto** on tietoaineistoista viranomaisten tehtävien tuloksena muodostuva looginen kokonaisuus.

Tiedonhallintayksikkö puolestaan on viranomainen, jonka tehtävänä on järjestää tiedonhallinta uuden lain vaatimusten mukaisesti. Tämä pyrkii selkeyttämään nykytilannetta, jossa tiedonhallinta on yleensä järjestetty organisaatiotasolla, vaikka organisaatio koostuisi useammasta viranomaisesta. Tiedonhallintayksiköitä ovat esimerkiksi valtion virastot ja laitokset, valtion liikelaitokset, kunnat ja kuntayhtymät.

Tiedonhallintayksikön johdon on huolehdittava selkeästä vastuutuksesta, asianmukaisista työvälineistä, valvonnasta, ohjeistuksesta ja koulutuksesta. Jokainen viranomainen toteuttaa samoja velvoitteita ja kokonaisuutta hallitaan th-yksikön kautta. Sen johto vastaa myös **tiedonhallintamallista**, joka on lain mukanaan tuoma uutuus. Tiedonhallintamallin osia ovat prosessit, tietovarannot ja tiedot, tietojen arkistointi, tietojärjestelmät sekä tietoturvallisuustoimenpiteet.

Perustuslakivaliokunta oli Oikarisen mukaan käynyt mielenkiintoisen keskustelun siitä, millainen uusi tiedonhallintayksikkö pohjimmiltaan on "olentona". Toteutustapoihin ei pitäisi mennä liian syvälle, koska käytännössä se lisää joustamattomuutta. Myös kesäkoulussa kuultujen kommenttien perusteella olennaista on oikeus saada tietoa, ei organisointitapa. Jokaisen tiedonhallintayksikön tulee kuitenkin täyttää vähintään samat perustason vaatimukset. Samalla edistetään yhteiskunnan kokonaisturvallisuutta ja lisätään viranomaisten keskinäistä luottamusta.

Yleisöstä tiedusteltiin myös arkistolain uudistuksesta. Tarkempaa aikataulutietoa ei vielä valitettavasti ollut, koska hallitusohjelman toteutussuunnitelma puuttuu.

Osaango Oy:tä edustava Marjukka Niinioja jatkoi edellisten Liikearkistopäivien osallistujille jo osittain tutulla esityksellä dokumenttien hilloamisesta ja rajapintojen (APIen) hyödyllisyydestä. Palautteen mukaan osalle kuulijoista jäi edelleen epäselväksi, mikä se API olikaan selkokielellä, joten paikattakoon tämä puute tässä: ohjelmointirajapinnan (Application Programming Interface, API) avulla erilaiset sovellukset ja ohjelmat voivat tehdä pyyntöjä ja vaihtaa tietoja – eli keskustella keskenään. Kun luodaan digitaalisten palvelujen ekosysteemejä ja palvelupolkuja, tarvitaan siis rajapintoja. Samaa rajapintaa voi käyttää moniin eri sovelluksiin.

Niinioja kehotti kiinnittämään enemmän huomiota dokumentin itsensä sijaan sen sisältöön. Digikehittäjälle tuottaa päänvaivaa se, että tietyn dokumentin löytääkseen on tiedettävä missä arkistossa se on, miten sen rajapintaa käytetään ja mitkä ovat dokumentin metatiedot. Haut kestävät ikuisuuden ja löytyneet dokumentit on vaikea yhdistää oikeaan organisaatioon tai henkilöön ja esimerkiksi tuoteinformaatio oikeaan tuotteeseen tai komponenttiin. Alustojen ja APIen avulla nämä ongelmat ovat ratkaistavissa. Esimerkiksi Samlinkin casessa saatiin asiakirjapohjiin tarvittavat perustiedot automaattisesti ja tuloksena olivat valmiit PDF-A:t, sähköinen allekirjoitus ja vienti arkistoon sovelluksen toimesta. Kuten Digitaliassakin on usein todettu, ihmiset eivät inhoa mitään niin paljon kuin käsin tehtävää metatiedottamista. Ja "jos jotain on toistettava useammin kuin kerran, se täytyy automatisoida" (Jääskeläinen, 2018).

Pienryhmäkeskusteluissa käsiteltiin organisaatioiden tiedonhallinnan kehittämistä tähän automaattisempaan ja keskustelevampaan suuntaan, ja toisaalta sen haasteita. Erään ryhmän mukaan suurinta osaa arvokkaasta informaatiosta ei voi jakaa järkevästi, koska lait ja asetukset ovat vastassa. Toisessa ryhmässä puolestaan korostettiin tarpeiden ymmärtämistä: milloin prosessista on järkevää huolehtia koneen voimin ja milloin taas ihmisten.

Anssi Jääskeläinen jatkoi esittelemällä Digitalian hankkeissa kehitettyjä ratkaisuja isojen aineistomassojen käsittelyyn. Paljon ylistetyt APItkaan eivät itsessään ratkaise mitään, vaan takana täytyy olla fiksu toteutus. Digitalian oma API julkaistaneen syksyn aikana.

Peruslähtökohta paperisten asiakirjojen digitoinnissa on Jääskeläisen mukaan se, ettei tunneta tai osata tarvittavaa tekniikkaa kovin hyvin. Näin ollen pyydetään tarjoukset ja lopuksi valitaan halvin toimittaja. Mikä on lopputulos? Tuhansia jpg-tiedostoja tai vastaavanlainen kaaos. Digitoinnin laatukriteereitä käy läpi esimerkiksi Teemu Hännisen kokoama Digitointiopas (2019), josta on tarkempi esittely toisaalla tässä lehdessä. Esityksessä listattiin joitakin vaatimuksia aineistoille, jotta

Tutkimusjohtaja Noora Talsi Xamkilta avaamassa kesäkoulua. Kuva: Miia Kosonen.

automaattinen tunnistus OCR-työkaluilla toimisi mahdollisimman hyvin: kuvien koon tulisi olla 150-300 dpi, laatu mielellään pakkaamaton, ja mahdollisimman suuri kontrasti taustan ja sisällön välillä. Käsin tehdyt lisämerkinnät olivat vaikeimpia tunnistaa.

Jääskeläinen esitteli Digitalian ratkaisun asiakirjojen anonymisointiin. Se on toteutettu NER- (Named Entity Recognition) ja sääntöpohjaisesti. Mallissa luetaan teksti, tunnistetaan anonymisoitavat kohteet NER-tunnistuksen ja sanalistojen avulla, puretaan pdf-rakenne ja etsitään ne fyysiset kohdat, joissa edellä tunnistetut kohteet ovat. Näiden päälle piirretään laatikot, tallennetaan koko sivu kuvaksi ja tehdään uudelleen OCR-luku. Lisäksi on kokeiltu pdf-redactoriin perustuvaa anonymisointia suoraan dokumentista. Digitaliassa odotetaan mielenkiinnolla tuloksia OM:n syksyllä 2020 päättyvän Anoppi -hankkeen (Henkilötietoja sisältävien asiakirjojen automaattinen anonymisointi ja sisällönkuvailu) kieliteknologisesta tekoälystä.

Esillä olivat myös automaattinen sisällönanalyysi ja automaattinen metatiedottaminen. Ratkaisua voi testata Digitalian tietopankissa https://digitalia.xamk.fi/content-analyser.

Lisäksi on syytä mainita digitaaliset aineistot käyttöön -hankkeemme tulosten koonti, joka ilmestyi alkukesällä 2019 Xamk kehittää -sarjassa (80). Ratkaisuja digitaalisten aineistojen käytettävyyden parantamiseen -artikkelikokoelma on ladattavissa Theseuksesta: https://www.theseus.fi/handle/10024/226884.

Ensimmäisen kesäkoulupäivän päätti ansiokkaasti Tampereen kaupungin digitointiprojektin projektipäällikkö ja asiakirjahallinnon suunnittelija Juha Laine. Aiheena oli kuntien itse tuottaman sosiaalisen median aineiston arkistointi. Digitaalinen arkistointi on kaupungin organisaatiossa pitkällä, sillä Donna-järjestelmä

on ollut käytössä vuodesta 2008 ja arkistossa on yli 1,7 M asiakirjaa. Uusimpana aluevaltauksena Tampereella lähdettiin selvittämään, mitä, miksi ja miten sosiaalisesta mediasta voidaan arkistoida.

Tutkijat ovat Laineen mukaan pohtineet sosiaalisen median aineistojen käytön eettisyyttä, datan saatavuutta, käyttöä erilaisissa tutkimuksissa, sisällön merkitystä jne., mutta eivät säilyttämistä. Tästä otamme Digitaliassa onkeen ja tulevien hankkeiden valmistelussa on huomioitu myös ratkaisut sosiaalisen median arkistointiin.

Asiakirjahallinnon suunnittelijoiden tapaamisessa kävi ilmi, ettei mikään suurista kaupungeista ollut käytännössä toteuttanut someaineistojen talteen ottamista ja säilyttämistä. Sosiaalinen media on kuitenkin kunnille tärkeä viestintäkanava. On virallisia tilejä, yksiköiden tilejä, poliittisen johdon ja virkamiesten tilejä. Kuitenkaan kunnilla ei ole ollut ohjetta aineistojen säilyttämisestä: mitä on säilytettävä, kuinka pitkään ja miten. "Ajattelimme, että tämä on ihan selvää, mutta ei sitten ollutkaan." Selvitystä jatkettiin suurten kaupunkien yhteistyönä. Siitä muodostui kokonaan uudentyyppisen aineiston säilytysarvon määrittelytyö.

Perusperiaate sosiaalisessa mediassa on, että omalta käyttäjätililtä jaetut sisällöt saa ladata itselleen. Lisäksi kuntien on rajattava, mitkä tilit tai aineistot säilytetään ja miten. Palvelujen erilaiset (ja nopeasti muuttuvat) käyttöehdot on hyvä huomioida: esimerkiksi anonymisointi ei ole relevanttia tviiteille, koska uudelleen käytettäessä ne on aina esitettävä alkuperäisessä asussaan.

Seulontahakemus Kansallisarkistolle laaditaan Kuntaliiton nimissä ja yhteistyössä Kansallisarkiston ja Tietosuojavaltuutetun toimiston kanssa. Tällä hetkellä asia on Kuntaliiton Tuula Sepon pöydällä ja suurten kaupunkien asiakirjahallinto on mukana. Signaali on ollut positiivinen. Eniten tällä hetkellä auki on se, mitä tietoa otetaan mukaan, ei niinkään GDPR tai vastaavat kysymykset. Määrittelyssä haetaan keskeisten tilien pysyväissäilyttämistä siten, että mukana ovat kaikki tiedot (ilman anonymisointia) eli myös kommentit ja tykkäykset. Jos valokuvat, videot, tiedotteet yms. säilytetään myös muualla, pyritään välttämään päällekkäisyyttä. Johtavien politiikkojen ja viranhaltijoiden tilit voidaan arkistoida näiden suostumuksella.

Tekniset ratkaisut otetaan pohdintaan, kunhan päätös saadaan. Aineistojen käsittelyyn on jo olemassa välineitä. Tavoitteena on löytää kunnille yksi yhteinen ratkaisu ja aineistot säilytetään organisaatioiden omissa digitaalisissa arkistoissa. Määrittely- ja ylläpitokysymykset ovat niin ikään auki. Laineen mukaan "vielä ei kannata pidättää hengitystä, että koska tämä tulee", mutta jäämme Digitaliassa mielenkiinnolla odottamaan – ja tarvittaessa myös kehittämään sopivia avoimen lähdekoodin työkaluja yhdessä.

Toinen kesäkoulupäivä keskittyi eettisen vastuun kysymyksiin tekoälykehityksestä yksityisyyteen ja yritysvastuuseen. Allekirjoittanut tiedusteli, miten Google, Amazon, Facebook, Apple, Microsoft ja yritysvastuun käsite soveltuvat yhteen. Luennoitsijan mukaan kyseessä on kiinnostava ilmiö tutkittavaksi.

Nykypäivän data on sosiaalista, joten emme luovu vain omasta yksityisyydestämme. Henrik Rydenfelt muistutti, että käytännössä annamme nettijäteille jatkuvasti luvan rikkoa yksityisyyttä, koska näillä foorumeilla asiat vain tapahtuvat – valtion tai kunnan sijaan kokonaisuutta pyörittävät yksityiset firmat. Samaan aikaan on ymmärrettävä yksityisyyden uusi sävy: nämä yritykset eivät ole varsinaisesti kiinnostuneita juuri minusta, vaan meistä kaikista.

Tätä kaikkeutta ei ole syytä käydä tässä artikkelissa sen tarkemmin läpi, sillä on epäeettistä viedä viimeinenkin palstatila Failista. Palaute ja toiveet tulevista koulutuksista/opasmateriaaleista ovat tervetulleita. Kiitos Digitalian puolesta kaikille osallistujille ja puhujille, ja tapaamisiin tulevissa kesäkouluissa!