KULTTUURIPERINTÖÄ KÄYTTÄJÄ EDELLÄ

Tuloksia Digitaalinen avoin muisti -hankkeesta

Miia Kosonen (toim.)

XAMK
Kaakkois-Suomen ammattikorkeakoulu

Miia Kosonen (toim.)

KULTTUURIPERINTÖÄ KÄYTTÄJÄ EDELLÄ

Tuloksia Digitaalinen avoin muisti -hankkeesta

XAMK KEHITTÄÄ 160

KAAKKOIS-SUOMEN AMMATTIKORKEAKOULU
MIKKELI 2021

Kannen kuva: Mainostoimisto Ilme Oy
Taitto ja paino: Grano Oy
ISBN: 978-952-344-356-3 (nid.)
ISBN: 978-952-344-357-0 (PDF)
ISSN: 2489-2467 (nid.)
ISSN: 2489-3102 (verkkojulkaisu)
julkaisut@xamk.fi

ESIPUHE

Viimeistään Covid-19-pandemian myötä museosektorilla on herätty virtuaalisten näyttelyiden tarpeeseen ja mahdollisuuksiin. Kuluneen vuoden aikana uusia virtuaalimuseoita onkin avattu kiitettävästi ja ne ovat saaneet huomiota myös eri medioissa. Virtuaalisten näyttelyiden juuret juontavat www-sivujen alkuaikaan ja vuoteen 1995, jolloin Oxfordissa sijaitseva tieteen historian museo avasi ensimmäisen verkkonäyttelynsä. Suomessa verkkonäyttelyt yleistyivät 2000-luvun alussa, mutta ne eivät saavuttaneet missään vaiheessa suurta suosiota.

Virtuaalimuseoilla on kuitenkin potentiaalia kehittyä yleistyväksi tavaksi tutustua taiteeseen, esineelliseen kulttuuriin ja kulttuurihistoriaan. 360-kuvausmenetelmien, 3D-mallintamisen ja erilaisten 3D-skannausmenetelmien kehittyessä virtuaalimuseoiden ja -näyttelyiden tuottamisesta tulee edullisempaa, halvempaa ja laadultaan parempaa. Tällä hetkellä monet virtuaalimuseoista ja -näyttelyistä on toteutettu ns. 360-kuvaustekniikalla, jolla on tallennettu fyysisesti olemassa oleva näyttely tai rakennuskokonaisuus digitaaliseen muotoon. Toteutuksen laadullinen taso näissä esityksissä vaihtelee paljon ja asiakaslähtöisyyden näkökulmasta käytettävyys on usein korkeintaan keskinkertaista. Poikkeuksiakin toki löytyy, kuten Zoan Oy:n luomat kohteet Virtuaaliseen Helsinkiin.

Digitaalinen avoin muisti -hankkeessa Digitalia lähti hakemaan ratkaisuja siihen, kuinka virtuaalimuseoita olisi mahdollista toteuttaa kustannustehokkaammin ja laadukkaammin. Tavoitteena oli luoda dynaaminen, kolmiulotteinen museoympäristö Unity-pelimoottorin avulla. Muun muassa tämän kehitystyön tuloksia kuvataan tässä artikkelikokoelmassa. Museosektorilla ratkaisut otetaan innolla vastaan.

Mikkelissä 20.5.2021

Olli-Pekka Leskinen, toimitusjohtaja, Sodan ja rauhan keskus Muisti

KIRJOITTAJAT

MARJA-LEENA HYNYNEN, HuK, tietoasiantuntija
Kansalliskirjasto

ANSSI JÄÄSKELÄINEN, TkT, tutkimuspäällikkö
Kaakkois-Suomen ammattikorkeakoulu

MIIA KOSONEN, KTT, tki-asiantuntija
Kaakkois-Suomen ammattikorkeakoulu

LIISA NÄPÄRÄ, FT, suunnittelija
Kansalliskirjasto

TUOMO RÄISÄNEN, FT, ohjelmistosuunnittelija
Kaakkois-Suomen ammattikorkeakoulu

SISÄLTÖ

MUISTIKUVIA ETÄTYÖN VOITTOKULUSTA

Miia Kosonen, KTT, tki-asiantuntija, Xamk

Vuonna 2015 perustettu Digitalia on Kaakkois-Suomen ammattikorkeakoulun, Helsingin yliopiston ja Kansalliskirjaston yhteinen digitaalisen tiedonhallinnan tutkimus- ja kehittämiskeskus. Digitalian kehittämät ratkaisut parantavat mahdollisuuksia hyödyntää digitoituja ja syntysähköisiä aineistoja esimerkiksi tieteelliseen tutkimukseen. Varmistamme, että tieto säilyy käytettävänä pitkälle tulevaisuuteen ja on luotettavaa. Kehitämme ja testaamme pitkäaikaissäilytystä. Parannamme digitaalisen historiallisen aineiston käytettävyyttä ja laatua. Kts. http://digitalia.fi

Digitalian toteuttama Digitaalinen avoin muisti (DAM) -hanke on vuosina 2019–2021 vastannut tarpeeseen kehittää digitaalisten aineistojen käytettävyyttä ja mahdollisuuksia hyödyntää aineistoja. Tuloksista hyötyvät muistiorganisaatiot, yritykset, tutkijat, opettajat ja opiskelijat.

DAM-hanke on edistänyt automaation hyödyntämistä digitaalisten aineistojen käsittelyssä ja käyttöön saattamisessa. Lisäksi digitaalisten aineistojen käyttöön on luotu perinteiset rajat ylittävä käyttäjä- ja kehittäjäyhteisö. Digitaalinen avoin muisti -hankkeen työpaketeissa on keskitytty erityisesti käyttäjälähtöiseen tietoon, tiedon visualisointiin ja sosiaalisen median talteenoton kehittämiseen.

TYÖSKENTELYÄ VERKKOYHTEYKSIN

DAM-hanke käynnistyi syksyllä 2019. Digitalian kannalta historiallista on se, että Covid-19 -pandemian myötä enin osa hankkeesta vietiin läpi puhtaasti etäyhteyksin. Hanketiimillemme digitaalisen maailman toimintatavat ovat luontaisia, joten monen muun toimialan projekteista poiketen pandemian tuomat muutokset eivät vaikuttaneet aikaansaannosten määrään kuin korkeintaan positiivisesti.

Pitkä etätyöjakso ansaitsee tulla nostetuksi tässä esille monestakin syystä. Ensinnäkin Digitalian kaikkia hankkeita läpileikkaava ”sateenvarjotavoite” on edistää digitalisaatiota. Yksi olennainen osa kehitystä on digitaalisen maailman käytäntöjen ja toimintatapojen oppiminen. Kannettu vesi ei kaivossa pysy, ja niinpä juhlapuheista on siirryttävä oikeasti tekemään, kokeilemaan ja sisäistämään uutta. Pandemia-ajalle tyypillinen työskentely puhtaasti verkkoyhteyksillä on ollut tätä oppimista parhaimmillaan – toki pakon edessä, mutta monessa tapauksessa vaikuttavin tuloksin.

Toiseksi, yllättävillä tilanteilla on aina monta puolta. Digitaalinen avoin muisti -hankkeen kannalta kielteisiä vaikutuksia ovat olleet muun muassa sidosryhmäyhteistyön hidasteet, konferenssien peruuntumiset ja hankkeen omien työpajojen poisjäännit. Spontaania kanssakäymistä sekä kollegoiden että yhteistyökumppaneiden välillä on ollut hyvin vähän. Uusille luoville ideoille on siten ollut vähemmän tilaa. Työhyvinvoinnista huolehtiminen on jäänyt jokaisen omalle vastuulle. Lisäksi kaikki etätyöläiset eivät taivu tekniikan hyödyntämiseen yhtä sujuvasti kuin valmiiksi IT-marinoidut, mikä on voinut lisätä epätasa-arvon kokemusta. Panostus sekä teknisesti osaaviin että digitaalisissa kulttuureissa sujuvasti luoviviin ”moderaattoreihin” on työyhteisöissä tarpeen nyt ja tulevaisuudessa. Liian usein työyhteisöissä suhtaudutaan verkkotyöskentelyyn kuin Excelin käyttöön: kaikkihan sen osaavat, vaikka kukaan ei olisi opettanut.

Taidoista kenties olennaisin on kyky pitää yllä vuoropuhelua, innostaa ja houkutella mukaan yhteiseen tekemiseen. Ellei tarvittavaa osaamista löydy, vuorovaikutus typistyy helposti kuivaksi ja tehtäväkeskeiseksi suorittamiseksi ruutua tuijottaen. Etäkokoukset ovat itsessään kuormittavia ja paljon keskittymistä vaativia, eikä niitä pitäisi lainkaan joutua ahtamaan kalenteriin useita peräkkäin. Varmaa tässä vyyhdessä on vain se, että kaikki ovat kuluneen vuoden aikana oppineet jotain uutta – ja se on hieno uutinen tulevaisuutta ajatellen. Perustason netiketistä on vihdoin tullut rutiinia ja etätapaamisten toteuttaminen eri tahojen kanssa ei enää kaadu kroonistuneeseen tekniseen hämmennykseen. Hybridimalli lienee tullut jäädäkseen oppimiseen ja asiantuntijatyöhön.

Kolmanneksi on vielä syytä noteerata ajanjakson 2020–2021 historiallisuus ja laajempi yhteiskunnallinen merkitys. Digitalia osaltaan kehittää ratkaisuja, joilla korona-ajasta kertyvää ainutlaatuista aineistoa voidaan saada talteen ja arkistoitua myöhempien tutkijasukupolvien käyttöön.

TIME MACHINE -YHTEISTYÖ

Etäyhteyksillä toteutettiin myös Digitalian järjestämä parituntinen verkkopaneeli Time Machine -hankkeesta elokuussa 2020. Paneeli korvasi digitaalisen tiedon kesäkoulun, josta oli tehty välivuosipäätös jo ennen koronaa. Asiantuntijoina paneelissa olivat Suomen Time Machine -lähettiläät Tomi Ahoranta ja Juha Henriksson sekä professori Eero Hyvönen Aalto-yliopistosta ja HELDIGistä. Tapahtuma kokosi yli 70 osallistujaa, joista suurin osa oli muisti- ja tutkimusorganisaatioista.

Time Machinen tavoitteena on luoda edellytykset eurooppalaisen kulttuuriperinnön kattavalle digitoinnille, tuottaa ”menneisyyden big dataa”, varmistaa datan pitkäaikaissäilytys ja ennen kaikkea avata mahdollisuuksia tekoälypohjaiselle isojen aineistomassojen käsittelylle. Näin tieto voidaan tuoda saataville uusilla, ainutlaatuisilla tavoilla. Osana hanketta on toteutettu kansalliset selvitykset osallistuvien organisaatioiden toiveista ja toisiaan täyden-

tävästä osaamisesta. Suomen osalta selvitys valmistui keväällä 2021 ja sen tulokset esiteltiin toukokuussa 2021. Digitalian osapuolet Xamk ja Kansalliskirjasto ovat Time Machine -organisaation jäseniä ja jatkavat yhteistyötä tulevina vuosina.

AINEISTOJEN LUOKITTELUN JA VIRTUAALIMUSEON PILOTIT

Kun asiakirja-aineistoa on hyllykilometreittäin, tarkoittaa se haasteita aineistojen luokitteluun. Skannatuista asiakirjoista tulee voida erotella erilaista kirjoitusta (käsin, koneella, molempia rinnakkain) sisältävät sivut tyhjistä sivuista mahdollisimman luotettavasti. Digitalia onkin DAM-hankkeessa kokeillut ja kehittänyt erilaisia sääntö- ja keinoälypohjaisia ratkaisuja tähän ongelmaan. Ratkaisuja on testattu yhteistyössä Kansallisarkiston kanssa heidän toimittamallaan testiaineistolla. Tuomo Räisänen ja Anssi Jääskeläinen esittelevät artikkelissaan eri malleilla saadut tulokset. Pilotti oli onnistunut: siinä päästiin 97–99 % tarkkuuteen tyhjien tunnistuksessa, mikä jo parantaa aineistojen käytettävyyttä merkittävästi.

Aineistojen visualisoinnin osalta Digitalian tavoitteena oli tuoda teksti-, kuva-, 3D- ja videomuotoista materiaalia uusinta virtuaalitodellisuustekniikkaa hyödyntävään 3D-ympäristöön. Pilotissa tehtiin yhteistyötä Suomen Elinkeinoelämän Keskusarkiston yhteydessä toimivan Designarkiston kanssa.

Virtuaalisten museoelämysten rakentamisen esteenä on usein teknisen tietotaidon puute, resurssipula ja digitaalisessa muodossa olevan materiaalin vähyys. DAM-hankkeen virtuaalimuseopilotti ratkaisee kaksi ensimmäistä ongelmaa täysin, sillä ratkaisun soveltaminen ei vaadi kallista infraa tai edistynyttä teknistä osaamista. Myös kolmanteen ongelmaan on saatavilla apua, sillä hankkeessa testattiin työnkulkua esineestä 3D-mallin kautta museoon. Anssi Jääskeläinen avaa artikkelissaan tarkemmin virtuaalimuseon toteutusta ja tuloksia.

KÄYTTÄJÄLÄHTÖISET DATAPALVELUT JA HISTORIALLISTEN AINEISTOJEN OPETUSKÄYTTÖ

Käyttäjälähtöisyyden alkupiste on digitaalisten aineistojen kohdalla siinä, että ymmärretään tutkijoiden tarpeita näihin aineistoihin ja dataan liittyen. Näin päästään kehittämään aineistojen käytettävyyttä. Myös käyttäjä- ja kehittäjäyhteisön ylläpitämiseen tarvitaan uutta osaamista. Liisa Näpärä Kansalliskirjastosta kuvaa artikkelissaan DAM-hankkeessa tehtyjä toimenpiteitä näiden tiedontarpeiden täyttämiseksi. Tietoa kerättiin tutkijoilta itseltään, mutta lisäksi haettiin benchmarkingin avulla vertaisoppia muiden maiden kansalliskirjastoilta. Selvityksen pohjalta luotiin Kansalliskirjaston library lab -palvelukonsepti eli käyttäjälähtöiset datapalvelut.

Hankkeen yhtenä tavoitteena oli parantaa mahdollisuuksia hyödyntää laajoja digitaalisia aineistoja. Tätä silmälläpitäen Kansalliskirjastossa lähdettiin tuottamaan digitaalisia opetuspaketteja, joilla voidaan tutustuttaa yläkoulujen ja lukioiden opettajia ja oppilaita Digin lehtiaineistoihin. Opetussuunnitelmaan sidotut aineistopaketit julkaistiin Finna Luokkahuone -sivustolla toukokuussa 2021.

Marja-Leena Hynynen ja Liisa Näpärä Kansalliskirjastosta kuvaavat artikkelissaan tarkemmin opetuspakettien tarvetta, toteutusta, sisältöä ja kouluista saatua palautetta. Yhteistyö opettajien kanssa osoittautui hedelmälliseksi, sillä kirjaston vahvan aineistotuntemuksen rinnalle saatiin koulumaailman pedagoginen näkökulma. Opetuspaketit tutustuttavat digitaalisten aineistojen käyttöön sekä opetustilanteissa että niiden ulkopuolella, tarjoavat opetussuunnitelman mukaista sisältöä ja kannustavat oppilaita samalla lähdekriittisyyteen.

VIRANOMAISTEN TWITTER TALTEEN

Hitaus tai vastahakoisuus linjata viranomaisten sosiaalisen median aineistojen käsittelyperiaatteita ei saa olla esteenä toimeen tarttumiselle – muutoin takamatka muodostuu entistä pidemmäksi ja arvokasta aineistoa saattaa ehtiä kadota kokonaan. Tuomo Räisänen ja Miia Kosonen kuvaavat artikkelissaan Digitalian toteuttamaa teknistä ratkaisua organisaatioiden virallisten Twitter-tilien sisällön talteenottoon.

Erityisesti ns. seitsikkokaupungit ovat olleet kiinnostuneita sosiaalisen median talteenotosta. Jo Digitalian kesäkoulussa 2019 kuultiin paljon kiittävää palautetta saanut esitys aiheesta Tampereen kaupungin toimesta. Pilottina DAM-hankkeessa olivatkin kuntien Twitter-aineistot. Aineistoa on koottu jo yli sadan kunnan tileiltä, kaikkiaan noin 250 000 twiittiä. Varsinaista arkistointia ei kuitenkaan tehdä, ennen kuin arkistoinnin juridiset reunaehdot ovat selvillä, ja Kuntaliitto onkin luvannut selvittää asiaa.

Digitalian osapuolilla on sosiaalisen median talteenotossa selkeä työnjako: Kansalliskirjasto tallentaa suomalaista kulttuuriperintöaineistoa ja toteuttaa verkossa muun muassa teemakeräyksiä, jotka liittyvät ajankohtaisiin, merkittäviin ilmiöihin ja tapahtumiin. Xamk voi tuottaa yksittäisille organisaatioille helppoja avoimeen lähdekoodiin perustuvia työkaluja koota virallisten tilien sisältö talteen. Kuten edellä virtuaalimuseon tapauksessa, pienemmätkin toimijat voivat käyttää tällaista ratkaisua ilman syvällistä teknistä osaamista tai isoa hintalappua. Aineistoa voidaan hyödyntää organisaation oman toiminnan kehittämiseen, siihen voidaan kohdistaa hakuja ja tehdä yksinkertaisia analyyseja sisällöstä. Jos arkistointilupa saadaan, tulevaisuudessa aineisto palvelee paremmin myös tutkijakäyttöä.

KATSE TULEVAAN

Digitalian oli tarkoitus kokeilla DAM-hankkeessa laajaa ohjausryhmäämme osallistavaa toimintamallia, missä perinteisten ja valitettavasti usein hyödyttömien kokousten sijaan testataan kehitettyjä ratkaisuja ja saadaan samalla palautetta jatkokehitykseen. Tämä käyttäjälähtöisyyttä edustava käytäntö jää odottamaan tulevia hankkeita. Niissä laajennamme tiedon visualisointiin ja sosiaalisen median talteenottoon kehitettyjä ratkaisuja, kuten myös historiallisten aineistokokoelmien hyödyntämistä eri kohderyhmien tarpeisiin.

Covid-19 -vuosista on kertynyt ainutlaatuista aineistoa nykyisille ja tuleville tutkijasukupolville. Myös tämän mitä ajankohtaisimman aineiston talteenottoon, säilyttämiseen ja jatkokäyttöön Digitalian osapuolet Xamk ja Kansalliskirjasto voivat tarjota arvokasta apuaan.

Tämän koontijulkaisun artikkelit avaavat näitä mahdollisuuksia tarkemmin, kukin omasta näkökulmastaan mutta toisiaan täydentäen. Antoisia lukuhetkiä artikkelikokoelmamme parissa!

KUINKA TUNNISTAA TYHJÄT SIVUT: KEINOÄLY VS. SÄÄNNÖT

Tuomo Räisänen, FT, ohjelmistosuunnittelija, Xamk
Anssi Jääskeläinen, TkT, tutkimuspäällikkö, Xamk

Kun asiakirja-aineistoa on hyllykilometreittäin, sen luokittelu ihmistyönä on mahdotonta. Digitalia onkin DAM-hankkeessa kokeillut ja kehittänyt erilaisia sääntö- ja keinoälypohjaisia ratkaisuja aineiston luokitteluongelmaan. Ratkaisuja on testattu yhteistyössä Kansallisarkiston kanssa heidän toimittamallaan testiaineistolla. Tässä artikkelissa kuvaamme toteutuksen ja tulokset selkokielellä.

Digitoitavan tai jo digitaalisessa muodossa olevan materiaalin määrä lasketaan hyllykilometreissä jo Suomenkin tasolla (Massadigitoinnin suunnitteluprojektin loppuraportti). Jos sivuja käsitellään 5 000 kappaletta päivässä ja yhden sivun avaaminen, päätös ja siirto oikeaan paikkaan veisi ainoastaan 10 sekuntia, yhdeltä henkilöltä menisi ilman taukoja noin 14 h käsitellä koko aineistomäärä. Todellisuudessa ihminen voi taukojen, virheiden ja tympääntymisen saattelemana käsitellä maksimissaan 500–1000 sivua päivässä ja tällaisenkin työmäärän teettäminen ihmisellä on ajan haaskausta. Automatiikan apua tarvitaan siis jälleen.

Digitalian DAM-hankkeen ja Kansallisarkiston yhteistyönä toteuttamassa pilotoinnissa keinoälyn (AI, artificial intelligence) ja sääntöihin perustuvan päättelyn avulla pyrittiin lähtökohtaisesti erottelemaan skannatuista asiakirjoista (bittikarttakuvista) erilaista kirjoitusta sisältävät sivut mahdollisimman luotettavasti tyhjistä sivuista. Pilottimme tulokset ovat lupaavia. Ne parantavat merkittävästi aineistojen käytettävyyttä ja vähentävät huomattavasti seulontaan käytettävän ihmistyövoiman tarvetta.

On todettava, että yhdelläkään automaattisella menetelmällä ei tulla Digitalian näkemyksen mukaan koskaan saavuttamaan täyttä 100 % tarkkuutta. Kuitenkin sekä Digitalian että Kansallisarkiston tulokset osoittavat, että varsin hyvään 97–99 % tarkkuuteen tyhjien tunnistuksessa voidaan päästä monellakin eri menetelmällä tai niiden yhdistelmällä. Lisätutkimuksilla ja kehittämisellä 99.9 % voisi olla saavutettavissa, mutta tällöinkin on kysyttävä, olisiko järkevämpää antaa se viimeinen epäselvä prosentti ihmistarkastukseen sen sijaan että pyrkisimme automatisoimaan myös sen.

ONGELMAKENTTÄ

Kansallisarkistoon arkistoidun materiaalin digitoinnissa syntyy pääpiirteittäin neljänlaista materiaalia. Sivuja, joissa on:

1. konekirjoitettua tekstiä
2. käsinkirjoitettua tekstiä
3. molempia edellisistä
4. tyhjiä sivuja.

Oletuksena siis oli, että tässä pilotoitavassa Kansallisarkiston aineistossa ei olisi mukana kuvia. Pilotin loppuvaiheessa, suuremmalla aineistomassalla testattaessa, muutamia kuvia kuitenkin löytyi. Sääntöjen uudelleen kirjoittaminen tai keinoälyn uudelleen opettaminen ei tässä vaiheessa ollut enää mahdollista, joten kuvatiedostot tunnistuvat lähes aina tyhjiksi etenkin sääntöpohjaista päättelyä käytettäessä.

Jo ensimmäisissä kokeiluissa ennakko-oletuksemme siitä, että tyhjien tunnistaminen olisi helpointa, vahvistui. Niinpä päädyimme muokkaamaan tehtävää siten, että ensi vaiheessa erotellaan tyhjät ja muu aineisto toisistaan. Kansallisarkiston tapauksessa lisämääreenä oli, että yhtäkään tietoa sisältävää sivua ei saisi tunnistua tyhjäksi (kriittinen virhe).

Tehtävän muokkaamisesta huolimatta kyseessä on siis klassinen aineiston luokittelutehtävä. Vaikka ihmiselle tämä luokitteluongelmamme on sangen helppo ja silmäilemällä ratkottavissa, Kansallisarkiston tapauksessa materiaalin määrä lasketaan kuitenkin hyllykilometreissä ja päivässä syntyvän materiaalin määrä on kymmeniä tuhansia kappaleita. Automatiikan apua tarvitaan.

Lisäksi on huomioitava, että jokaisen sivun tallentaminen tallekappaleena sekä käyttökappaleina ja metatietoina vie tilaa. Vuodessa kumuloituvan aineiston kohdalla voidaan puhua kymmenien, jopa satojen teratavujen määristä ja mahdollinen tyhjien sivujen poistaminen pienentäisi täten myös säilytyskustannuksia.

Optimaalisessa tilanteessa tyhjä sivu sisältää pelkkiä samanvärisiä pikseleitä. Entäpä valkoisen eri sävyt, kohinaa sisältävät sivut, sivut joissa taustapuoli näkyy läpi, tyhjän taulukon sisältävä sivu ja niin edelleen? Ihmissilmälle sivut ovat luonnollisesti tyhjiä, koska niissä ei ole asiasisältöä, mutta koneelle tilanne ei ole helppo. Jos kuvia parannetaan, taustapuoli ja viivat tehostuvat, jolloin niistä löytyy ”tekstiä”. Jos kuvia ei tehosteta, haaleimmista oikeasti tietoa sisältävistä sivuista ei tunnisteta mitään.

LÄHTÖKOHDAT

Kansallisarkistolta saimme pilottia varten kaikkiaan noin 15 000 sivua sangen monimuotoista materiaalia: pöytäkirjoja suomen ja ruotsin kielellä, erilaisia lomakkeita sekä allekirjoituksilla tai merkinnöillä varustettuja sivuja kokonaan tai osittain käsin kirjoitettuna. Mukana oli myös kopioita kokonaisista aukeamista ja tietenkin myös tyhjiä sivuja. Osa aineistosta oli ruutupaperille kirjoitettu ja osassa oli repeytymiä.

Taulukko 1 kertoo tarkemmin saamiemme tiedostojen määrän. Jotain keinoälyn toiminnasta tietävät havaitsevat suoraan, että aineisto ei ole tasapainossa. Epätasapainoa voidaan luonnollisesti korjata over sampling (ylinäytteistys) -menetelmillä, mutta se ei ole sama asia kuin aidosti balanssissa oleva opetusaineisto.

Taulukko 1. AI:n opetus- ja testiaineistojen määrät ja luokat

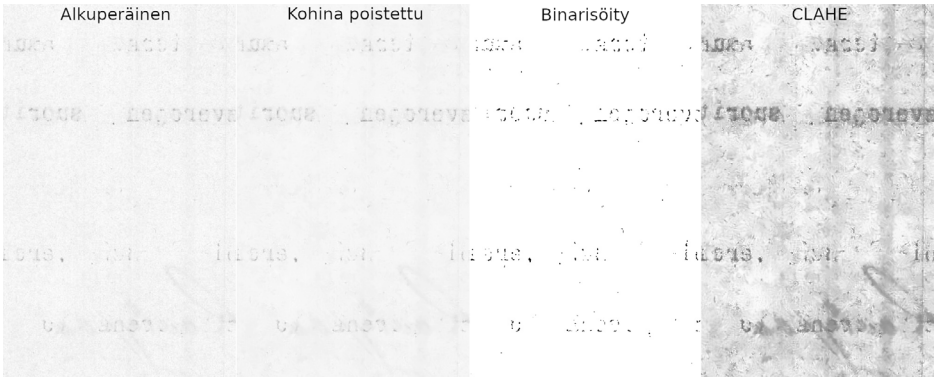
Luokka / Aineisto	Opetusaineisto	Testiaineisto
Käsin kirjoitetut	148	201
Sekä että	4200	3931
Konekirjoitetut	1477	1899
Tyhjät	1617	1760
Yhteensä	**7442**	**7791**

Ratkaisumme oli tarkoitettu tehtäväksi sekä sääntöpohjaisella että keinoälypohjaisella ratkaisulla rinnakkain, jolloin tuloksia voisi mahdollisesti yhdistellä vieläkin paremman lopputuloksen saavuttamiseksi. Molemmat ratkaisumme perustuivat kuitenkin Tesseractin tuottamaan sivukohtaiseen numeeriseen dataan, joten sekä sääntöpohjainen että keinoälypohjainen ratkaisumme teki jotakuinkin samoja virheitä. Virheet liittyivät pääasiassa taulukkojen sahalaitaisiin mustan ja valkean reunuksiin tai sivuihin, joilla oli vain vähän käsinkirjoitettua tekstiä. Peräkkäin ajojakin kokeiltiin, mutta edellä mainitun syyn takia tilastollisesti merkittäviä poikkeavuuksia ajojen välillä ei havaittu. DAM-hankkeen kahden ratkaisumallin lisäksi Kansallisarkistossa tehtiin vielä oma syväoppivaan neuroverkkoon pohjautuva ratkaisu suuremmalla koulutusmateriaalilla.

SÄÄNTÖPOHJAINEN RATKAISU

Ensimmäinen osa pilotissamme kehitetystä ratkaisusta pohjautuu Python-ohjelmointikieleen, OpenCV -kirjastoon sekä Tesseract OCR -ohjelmaan. Työnkulussa jokainen löydetty kuvatiedosto käsitellään erikseen omana kokonaisuutenaan ja rinnakkaisia suorituksia on meneillään yhtä monta kuin CPU-ytimiä on käytössä.

Työnkulun aluksi kuvasta rajataan jokaisesta reunasta 45 pikseliä. Tämä siksi, että Kansallisarkiston aineistossa kuvien ympärillä oli lähes poikkeuksetta reunusalue, joka vaikeutti Tesseractin toimintaa. Seuraavaksi rajatulle kuvalle tehdään kohinan poisto ja adaptiivinen mustavalkoiseksi muuttaminen (binärisöinti). Kuva 1 esittää samaisen työnkulun kuvina poislukien rajaamisen, koska kuvassa esitetään vain pieni ote koko alkuperäisestä kuvasta. Lisäksi kuvassa on mukana esimerkin vuoksi tehty CLAHE (Contrast Limiting Adaptive Histogram Equation) -kuva, joka kertoo hyvin läpikuultavien taustojen ongelman. Kuvassa sekä binärisöity että clahe-versio on tehty samasta kuvasta, josta kohina on poistettu.



Kuva 1. OpenCV:n avulla toteutettu kuvanparannustyönkulku

Kuvanparannuksien ja mustavalkoiseksi muuntamisen jälkeen kuva syötetään ensimmäisen kerran Tesseractille, joka tekee kuvalle nopean OSD (Orientation and Script Detection) -tunnistuksen. Jos tämä tunnistus löytää kuvasta jotakin, kuva tulkitaan tietoa sisältäväksi. Jos OSD ei saa tulosta, samainen sivu syötetään uudelleen Tesseractille, joka tekee tällä kertaa syvemmän analysoinnin sivulle. Tämä tunnistus palauttaa lähes sivusta kuin sivusta jotakin, joten Tesseractin tuottamia numeerisia tuloksia on analysoitava lisää ohjelmallisesti, tai lähes kaikki tyhjät sivut tunnistuvat tietoa sisältäviksi.

Tesseractin tuottamista tuloksista hyödynnetään tässä työnkulussa vain tekstiksi tunnistetut osat ja tunnistuksen luotettavuus. Näihin sovelletaan sekä pituuteen, määriin, keskiarvoihin että yksittäisiin tietokenttiin perustuvia jaottelumenetelmiä. Lopputuloksena lähes kaikki oikeasti tyhjät sivut saadaan eroteltua joukosta, mutta virheitäkin luonnollisesti tulee.

Toteutettujen kuvanparannustoimenpiteiden lisäksi kokeiltiin myös viivojen poistoa openCV:n avulla. Tämä kuitenkin heikensi tunnistustarkkuutta, koska poisto ei ole täydellinen ja saattaa poistaa myös osia tekstistä. Lisäksi kokeiltiin myös sivujen automaattista suoristamista, mutta sekään ei parantanut tunnistustarkkuutta. Tesseractissa on ilmeisesti jonkinlainen sisäänrakennettu sivujen suoristaminen, joka toimii yhtä hyvin kuin openCV:n

avulla ohjelmallisesti tehty suoristaminen. Pilotin aikana kokeiltiin myös kuvan syöttämistä Tesseractille jokaisen välivaiheen jälkeen ja vaihtamalla vaiheiden järjestystä. Kokeiluissa parhaaksi osoittautui valitsemamme työnkulku, jossa ensin rajataan, sitten tehdään kohinan poisto ja viimeisenä binärisöinti.

SÄÄNTÖPOHJAISEN RATKAISUN TULOKSIA

Käytettäessä virtuaalipalvelinta ja 32 virtuaali-CPU:ta ajoaika per sivu on noin 0,7 s. Jos CPU-määrä pudotetaan puoleen, ajoaika on noin 1,2 s per sivu. Seuraavaksi työlistallamme onkin tutkia, voiko prosessia tehostaa hyödyntämällä GPU:n CUDA-ytimiä suorituksessa. Tähän liittyen saimme palvelimeemme Nvidia Quadro M4000 -näytönohjaimen, jolla kokeiluissa pääsee hyvin alkuun.

Sääntöpohjaiselle päättelylle erityisen ongelmalliseksi osoittautuivat sellaiset sivut, joissa on ainoastaan hieman käsinkirjoitettua materiaalia, etenkin jos käsiala on epäselvää. Toinen selkeä ongelmakohta Tesseractille ja sääntöpohjaiselle päättelylle ovat sivut, jossa tekstiä on jonkinlaisen värillisen laatikon sisällä. Molemmissa tapauksissa Tesseract saattaa tulkita sisällön niin väärin, että sääntömme, joille Tesseractin tulokset syötetään, tulkitsevat sivun tyhjäksi. Sääntöjä muuttamalla tätä voisi luonnollisesti korjata, mutta sääntöjen korjaaminen toisesta päästä aiheuttaa virheitä toiseen päähän, jolloin enemmän tyhjiä sivuja tunnistuu tietoa sisältäviksi. Toinen vaihtoehto olisi käyttää OpenCV:n segmentointia yhtenä tunnistavana tekijänä Tesseractin tulosten lisäksi, mutta tämä jää kokeiltavaksi tulevissa hankkeissa.

KEINOÄLYRATKAISU

Digitalian ratkaisun toinen osa pohjautuu keinoälyyn. Tesseract toimii itsekin valmiiksi opetetun LSTM (Long Short Term Memory) -neuroverkon päällä, joten periaatteessa jo ensimmäinenkin osa oli keinoälyratkaisu. Toteuttamamme keinoälyratkaisu tutkii Tesseractin antamaa numerotietoa, joka sisältää mm. jokaiselle tunnistetulle sanalle annetun luotettavuusarvon (confidence rate). Se ilmaisee todennäköisyyden, jolla sana on tunnistettu. Karkeasti pelkistäen: iso prosentti indikoi konekirjoitettua tekstiä, mutta todellisuudessa sivukohtaista numerodataa päättelyyn tarvittiin paljon enemmän. Lisäsimme malliimme erilaisia tilastollisia arvoja, joita pyöriteltiin yrityksen ja erehdyksen kanssa Scikit-learn -paketista löytyvien luokittimien avulla. Kehittämämme skripti laskee sivukohtaisia suureita csv-tiedostoon, siten että yksi rivi vastaa yhtä sivua. Alla oleva rivi esittää yhdestä sivusta kerätyt tiedot ja näitä rivejä on siis lopullisessa csv-tiedostossa yhtä monta kuin käsiteltyjä kuvia oli. Vaikka muodostamamme csv-formaatti onkin sekavan näköinen, se mahdollistaa mm. luokkien helpon yhdistelyn ja kerättyjen tietojen ohjelmallisen jatkokäsittelyn.

- 64.41818181818182,1,85,29.09090909090909,16,70.9090909090909,39,96,0,743.877441077441,96,sekaetta,/home/digitalia-tr/DAM/AI/Final-report/Opetusaineisto/7352.KA/327535.KA/350737.KA/3223793.KA/lajiteltu/sekaetta/0121.jpg

Lopulliseen keinoälymalliin valikoitui mukaan seuraavat viisi suuretta:

1. Tunnistettujen sanojen luotettavuuden keskiarvo
2. Tunnistamattomien sanojen määrä
3. Tunnistetut tekstialueet, ei niinkään yksittäisiä sanoja
4. Alle 50 %:n tarkkuudella tunnistettujen sanojen luotettavuuden keskiarvo
5. Alle 50 %:n tarkkuudella tunnistettujen sanojen määrä

Yksittäisiltä sivuilta kerättiin myös seuraavat suureet, joita ei kuitenkaan sisällytetty lopulliseen keinoälymalliin:

6. Yli 50 %:n tarkkuudella tunnistettujen sanojen keskiarvo
7. Yli 50 %:n tarkkuudella tunnistettujen sanojen määrä
8. Luotettavuuksien moodi
9. Minimi luotettavuus
10. Luotettavuuksien varianssi
11. Maksimi luotettavuus

Numerotietojen jälkeen riville liitettiin oikean luokan tyypin lisäksi tiedoston absoluuttinen hakemistopolku, jonka avulla tuloksien tarkastelu helpottui huomattavasti. Ensimmäiset viisi numeroa riviltä ovat siis mallien kannalta merkitseviä. Muunkin numerotiedon käyttöä kokeiltiin, mutta nämä valikoituivat tähän kokeiluun. Numerotietoja on mahdollista laskea monipuolisemminkin, joten tästä on hyvä jatkaa mallien kehittämistä.

Taulukon 1 esittämästä opetusaineistosta käytettiin 80 % mallien muodostamiseen ja loput mallin toiminnan varmistamiseen. Varsinaiset tulokset syntyvät kuitenkin sitten, kun mallit ajetaan testiaineistolla. Käytimme avoimen lähdekoodin Scikit-learn -paketista löytyviä luokittimia, jotka esitetään taulukossa 2.

Taulukko 2. Käytetyt luokittimet

Lyhenne	Nimi	Selite
LR	logistic regression	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
LDA	linear discriminant analysis	https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html
KNN	nearest neighbor	https://scikit-learn.org/stable/modules/neighbors.html
CART	decision tree classifier	https://scikit-learn.org/stable/modules/tree.html
NB	gaussianNB	https://scikit-learn.org/stable/modules/naive_bayes.html
SVM	support vector machine	https://scikit-learn.org/stable/modules/svm.html

TULOKSIA DIGITALIAN KEINOÄLYRATKAISUSTA

Tuloksia tarkastelemme ensi alkuun painotetun kokonaistarkkuuden mukaan, pitäen mielessä alkuperäisen ongelman eli luokittelemme aineiston neljään osaan. Käyttämällä Tesseractin moodeja psm-3 (automaattinen sivun segmentointi ja OCR, ei OSD:tä), psm-6 (olettaa lukevansa isoa tekstiblokkia) ja psm-12 (olettaa lukevansa hajallaan olevaa tekstiä). Näin ollen keinoälymallien kokonaismääräksi muodostui 36.

Taulukko 3. Tarkkuuksien painotetut keskiarvot prosentteina eri luokittimille

Malli / psm-moodi	psm-3	psm-6	psm-12
LR	77	80	79
LDA	75	78	79
KNN	79	83	80
CART	79	76	77
NB	80	77	72
SVM	77	79	72
LR-re	78	80	77
LDA-re	78	78	79
KNN-re	78	81	80
CART-re	76	76	77
NB-re	74	77	72
SVM-re	76	78	73

Yksityiskohtaisempaa tietoa mallien toiminnasta saa kuitenkin tutkimalla tapahtumia luokkakohtaisesti. Tähän tarkoitukseen käytetään luokitteluraporttia, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html. Taulukko 4 esittää Taulukon 2 KNN mallin psm-12 moodin luokitteluraportin.

Taulukko 4. KNN psm-12 tuloksia luokitteluraportin muodossa

Aineisto	precision	recall	f1-score	support
kasin	0.52	0.66	0.58	201
konekirjoitetut	0.67	0.57	0.62	1899
sekaetta	0.80	0.86	0.83	3931
tyhjat	0.97	0.93	0.95	1760
Macro avg	0.74	0.75	0.74	7791
Weighted avg	0.80	0.80	0.80	7791
Total accuracy	0.80			7791

Luvut kertovat sadasosina tarkkuuksia. Näiden termien selitykset löytyvät esim. https://en.wikipedia.org/wiki/Precision_and_recall. Koska virheitä tapahtui, niin seuraavaksi onkin mielenkiintoista tutkailla, mihin luokkiin virheet muodostuivat keinoälyn tuloksissa. Keinoäly ei osaa sanoa "en tiedä", vaan tulkitsee käsiteltävän sivun johonkin luokkaan, käytettävän luokittimen ja sivukohtaisen numerodatan avulla. Tätä havainnollistetaan tyypillisesti sekaannusmatriisin avulla, englanniksi Confusion matrix, https://en.wikipedia.org/wiki/Confusion_matrix.

Taulukko 5. Sekaannusmatriisi KNN psm-12 tuloksista

	kasin	konekirjoitetut	sekaetta	tyhjat
kasin	132	4	61	4
konekirjoitetut	2	1081	798	18
sekaetta	16	503	3388	24
tyhjat	102	23	6	1629

Koska sääntöpohjaisuudella pyrittiin erottelemaan tyhjät sivut aineistosta, onkin mielenkiintoista verrata tuloksia siihen. Edellä olevassa esimerkissä tyhjien tarkkuus oli 97 %. Informaatiota hukkui kuitenkin 4 + 18 + 24 sivun verran. Tyhjien suhteen parhaaksi malliksi informaation katoamisen suhteen osoittautui LDA-re, josta ohessa tuloksia.

Taulukko 6. LDA-re psm-12 -luokittimen luokitteluraportti

Aineisto	precision	recall	f1-score	support
kasin	0.17	0.83	0.28	201
konekirjoitetut	0.52	0.83	0.64	1899
sekaetta	0.86	0.57	0.68	3931
tyhjat	0.99	0.65	0.79	1760
Macro avg	0.63	0.72	0.60	7791
Weighted avg	0.79	0.66	0.69	7791
Total accuracy	0.66			7791

Nyt tyhiien sivujen osalta tunnistumisprosentti on jo mukavat 99 %. Sekaannusmatriisi kertoo, mitä on tapahtunut.

Taulukko 7. LDA-re psm-12 luokittimen sekaannusmatriisi

	kasin	konekirjoitetut	sekaetta	tyhjat
kasin	166	1	34	0
konekirjoitetut	33	1583	277	6
sekaetta	259	1436	2234	2
tyhjat	517	37	59	1147

Eli nyt informaatiota ”katosi” (kriittisiä virheitä tuli) 6 + 2 kappaletta, joka laskennallisesti on 99.3 % tarkkuus. Kävi kuitenkin niin, että tyhjiä meni muihin luokkiin 517 + 37 + 59 sivua. Käytetyllä aineistolla sääntöpohjainen ratkaisu osoittautui vielä paremmaksi kriittisten virheiden suhteen.

KANSALLISARKISTON RATKAISU JA TULOKSET

Kansallisarkistolla päädyttiin erilaiseen lähtökohtaan, esikoulutettuun konvoluutioneuroverkkoon. Mallina kokeilussa oli 18-kerroksinen ResNet18-verkko siirretyllä oppimisella ja työkaluna PyTorch-kirjaston päälle rakennettu Fastai-kirjasto. Huomioitavaa on, että Kansallisarkistolla oli opetuskäytössään noin nelinkertainen kuvamäärä verrattuna Digitalian keinoälyratkaisussa käytettyyn opetusmateriaalin määrään. Opetusaineiston määrä keinoälyratkaisussa on verrannollinen mallin tarkkuuteen, joten KA:n ja Digitalian keinoälyratkaisut eivät täten ole keskenään vertailukelpoisia.

Opetusvaiheessa, joka kesti tehokkaalla GPU-koneella noin 5 h, ResNet18-verkko säätää kohdilleen noin 11 miljoonaa parametria ja varsinaisessa tunnistusvaiheessa se käyttää näitä parametreja ns. black box -moodissa. Kansallisarkisto painotti erityisesti tietoa sisältäviä kuvia koulutuksessa, koska aiemmin kuvattuja kriittisiä virheitä ei saisi päästä syntymään. Tämä koulutusaineiston painotus osaltaan selittää sitä, miksi niin moni tyhjä sivu tunnistui aineistoa sisältäväksi. Kuten Digitalian molemmissa menetelmissä, erityisen ongelmallisia sivuja ovat sellaiset, joissa on ainoastaan hieman käsinkirjoitettua materiaalia, etenkin epäselvällä käsialalla.

Digitalian kokeiluista sääntöpohjainen malli teki vähemmän kriittisiä virheitä, joten se pääsi mukaan Kansallisarkiston toteuttamaan testiin. Sekä sääntöpohjainen malli sekä ResNet18-malli ajettiin läpi samoilla aineistoilla ja tulokset esitetään taulukossa 8. Muu-luokkaan kuuluvia tiedostoja testissä oli 17359 ja tyhjä-luokkaan kuuluvia tiedostoja 11928. Taulukossa oikein menneet on esitetty vihreällä ja väärin menneet valkoisella pohjalla. Kriittisten virheiden kannalta paras kokonaistarkkuus on korostettu vaaleansinisellä värillä.

Taulukko 8. Sääntöpohjainen vs. Resnet18-malli

Digitalian sääntöpohjainen malli				
Oikea luokka	**N**	**Muu-luokkaan**	**Tyhjä-luokkaan**	**Tarkkuus**
muu	17359	17169	190	98,905 %
tyhjä	11928	287	11641	97,594 %
Kansallisarkiston ResNet18 malli				
Oikea luokka	**N**	**Muu-luokkaan**	**Tyhjä-luokkaan**	**Tarkkuus**
muu	17359	17338	21	99,879 %
tyhjä	11928	1076	10852	90,979 %

Pilotin tuloksien valossa opetettu ResNet18-neuroverkkopohjainen ratkaisu teki vähiten kriittisiä virheitä. Sääntöpohjainen malli olisi parempi, jos kokonaistarkkuus olisi määrittelevä tekijä. Neuroverkkopohjaista mallia voidaan myös säätää luokittimen antamien confidence-arvojen avulla tarvittaessa, mutta silloin ”hyvien” virheiden osuus kasvaa entisestään.

Muu-luokan virheet sisälsivät pääasiassa kuvia, jotka sisälsivät muutamia sanoja epäselvää käsinkirjoitusta. Samankaltainen kuvien korjaaminen, jota sääntöpohjaisessa päättelyssä tehtiin, saattaisi parantaa tuloksia myös neuroverkkopohjaisessa ratkaisussa. Tämä jää seuraavien hankkeiden tutkittavaksi.

Neuroverkkojen opettamisessa yksi isoimmista haasteista on kerätä tarpeeksi valmiiksi luokiteltua opetusaineistoa. Neuroverkkoja voidaan pitää ns. “Black Box” -malleina, koska ei voida tutkia, mitä sääntöjä se on oppinut opetusvaiheessa. Sääntöpohjaiset menetelmät eivät välttämättä tarvitse paljoa opetusaineistoa, mutta myös yleispätevien sääntöjen kehittäminen on hankalaa. Aineistosta löytyi loppuvaiheessa myös kuvia, jotka eivät sisältäneet tekstiä.

Onkin tapauskohtaisesti ratkaistava, kumpaan suuntaan virheitä voidaan tehdä. Kansallisarkiston tapauksessa tärkein lähtökohta oli, että asiaa sisältäviä kuvia ei saisi tunnistaa tyhjiksi. Toiseen suuntaan tapahtuvat virheet – tyhjien tunnistaminen sisällöksi – on vähemmän haitallista tiedonhallinnan ja sen hyödyntämisen näkökulmasta. Keinoälyratkaisuissa on potentiaalia tyhjien tunnistamisessa ja laajemminkin luokittelussa, mutta aineiston tulisi olla suhteellisen homogeenista, jotta keinoälyratkaisu toimii hyvin ilman uudelleen opetusta. Tyhjien tunnistamisen suhteen sääntöpohjainen malli sen sijaan toimii suhteellisen samoilla tarkkuuksilla tekstiaineistosta riippumatta.

Loppukaneettina todettakoon, että tyhjät ja aineistoa sisältävät saadaan eroteltua toisistaan 97-99 % tarkkuudella ja nopeudella, joka on alle sekunti per sivu. Tämä on jo merkittävä harppaus eteenpäin siitä, että kesäharjoittelija siirtelee tiedostoja kansiosta toiseen käsityönä.

LÄHTEET

Massadigitoinnin suunnitteluprojektin loppuraportti, Kansallisarkisto, saatavilla https://arkisto.fi/uploads/Viranomaisille/Massadigitointi%20alasivujen%20liitteet/Massadigitoinnin_suunnitteluprojektin_loppuraportti.pdf [viitattu 6.5.2021].

KANSALLISKIRJASTO KEHITTÄÄ KÄYTTÄJÄLÄHTÖISIÄ DATA-PALVELUITA

Liisa Näpärä, FT, suunnittelija, Kansalliskirjasto

Perinteiset tutkijapalvelut eivät enää kykene vastaamaan kaikkiin tutkijoiden kysymyksiin: Miten ja millaisia digitaalisia aineistoja ja dataa on saatavilla tutkimuksen käyttöön? Miten niitä voi käyttää, säilyttää ja jakaa? Millaista tukea ja yhteistyötä digitaalisten aineistojen ja datan käytön ympärille voidaan kehittää? Kysymykset selittyvät digitaalisten ihmistieteiden nousulla ja datan käsittelyyn tarvittavien työkalujen kehittämisellä.

Tutkijoilla on suuret odotukset digitaalisten aineistojen ja datan käytön mahdollisuuksista, mutta odotukset kaikille vapaasta ja avoimesta tai edes helposti käytettävästä aineistosta eivät aina vastaa todellisuutta tai ole realistisia. Jotta voisiin kehittää digitaalisten aineistojen käytettävyyttä, tarvitaan tietoa tutkijoiden tarpeista digitaalisiin aineistoihin ja dataan liittyen. Lisäksi tarvitaan uudenlaista osaamista ja yhteistyötä digitaalisten aineistojen käyttäjä- ja kehittäjäyhteisön ylläpitoon. Se tarkoittaa sitä, että digitaalisten aineistojen käyttäjät osallistuvat eri tavoin niiden käyttöön ja kehittämiseen joko suoraan tai epäsuorasti.

Digitaalisiin aineistoihin liittyvien tiedontarpeiden täyttämiseksi Digitaalinen avoin muisti (DAM) -hankkeessa kerättiin käyttäjälähtöistä tietoa tutkijoilta ja haettiin vertaisoppia (*benchmarking*) muilta kansalliskirjastoilta siitä, miten ne ovat toteuttaneet digitaalisten aineistojen ja datan ympärille rakentuneet palvelunsa. Näitä palveluja kutsutaan kansainvälisesti nimellä library lab. Lab-palveluiden toimintaa selvitettiin, ja selvityksen pohjalta luotiin käyttäjälähtöinen Kansalliskirjaston library lab -palvelukonsepti eli datapalvelut.

Hankkeessa keskityttiin tutkijoiden datapalveluiden luomiseen ja sisällön määrittelyyn Kansalliskirjaston Mikkelin toimipisteessä tuotettujen digitaalisten aineistojen ja datan ympärille. Lisäksi hankkeessa kehitettiin digitaalisten aineistojen lataustyökalu sekä luotiin opetuskokonaisuus Finna Luokkahuoneeseen. Digitaalisten aineistojen kehittämistyö hyödyttää tutkijoiden lisäksi myös muita aineistojen käyttäjiä, mutta tutkijat ovat usein etulinjassa luomassa uusia ratkaisuja. Hankkeessa tunnistettiin laajasti erilaisia digitaalisten aineistojen käyttötarpeita ja osaamista sekä luotiin käyttäjille uusia välineitä aineistojen hyödyntämiseen.

KÄYTTÄJÄLÄHTÖINEN TIETO

Käyttäjälähtöinen (*user-driven*) tieto tarkoittaa sitä, että tietoa kerätään suoraan jonkin asian käyttäjiltä. Käyttäjälähtöistä tietoa hyödynnetään uusien asioiden innovointiin ja olemassa olevien palveluiden tai esimerkiksi teknologioiden käytettävyyden kehittämiseen. Metodina käyttäjälähtöisen tiedon kerääminen mahdollistaa käyttäjien kokemuksiin perustuvien kehitystarpeiden kuuntelemisen, toiveisiin vastaamisen sekä joissakin tapauksissa vuoropuhelun palveluiden tai digitaalisten aineistoja kehittävien suunnittelijoiden ja niiden käyttäjien välillä. Usein käyttäjien toiveet ovat kuitenkin keskenään ristiriitaisia ja moninaisia. Kehitystyön prioriteettilistalla korkealle nousevat pääsääntöisesti paljon eri suunnista huomiota saaneet asiat. Toisaalta kehityslinjoja määriteltäessä on toisinaan tarvetta kompromisseille. Vaikka kaikkia toiveita ei voida toteuttaa esimerkiksi resurssien tai strategisten valintojen vuoksi, on hyvä muistaa, että palveluiden kehittäminen tai innovointi tapahtuvat prosesseina, eivätkä hetkellisinä ja äkkinäisinä tapahtumina. (Kankainen ym. 2011; Park ym. 2015)

DAM-hankkeen tapauksessa käyttäjälähtöistä tietoa kerättiin sekä palveluiden kehittämiseen että digitaalisten aineistojen käytettävyyden kehittämiseen. Hankkeen käyttäjälähtöisen tiedonkeruun kiinnostuksen kohteena olivat Kansalliskirjaston digitaalisia aineistoja käyttävät tutkijat. Digitaalisista aineistoista huomio kiinnittyi erityisesti digi.kansalliskirjasto.fi -palvelun tarjoamiin aineistoihin. Aiempien digi.kansalliskirjasto.fi -palvelusta tuotettujen yleisten kyselyjen mukaan digitaalisten aineistojen käyttäjissä on useita erityyppisiä käyttäjiä, jotka voidaan jakaa käyttötarkoituksen mukaan edelleen eri ryhmiin ja käyttötarkoituksiin (Pääkkönen & Lilja 2018). Digitaalisia aineistoja ja dataa käyttävät tutkijat muodostavatkin vain yhden käyttäjäryhmän kaikkien digitaalisia aineistoja käyttävien keskuudessa. Fokusoituminen auttoi hahmottamaan monin eri keinoin yhden käyttäjäryhmän tarpeita datan ja digitaalisten aineistojen käytöstä.

Käyttäjälähtöisiä tiedontarpeita selvitettiin kyselyllä, haastatteluilla ja yhteistyön havainnoinnilla. Kyselyn 130 vastaajaa, 18 haastateltua ja muutama yhteistyöpilotti muodostavat toisiaan täydentävän aineistokokonaisuuden, jonka perusteella Kansalliskirjaston digitaalisia aineistoja käyttävät voidaan jakaa kolmeen ryhmään heidän teknisten taitojensa ja digitaalisten metodien käyttötapojen mukaan:

1) Digitaalisia aineistoja käyttävät, jotka **eivät hyödynnä digitaalisia metodeja** tai hyödyntävät niitä vain vähäisesti, esimerkiksi taulukoimalla tuloksia. Heidän digitaaliset tai tekniset taitonsa ovat joko vähäisiä tai kiinnostus suuntautunut muuhun kuin digitaalisten aineistojen metodiosaamiseen.
2) Digitaalisia aineistoja käyttävät, jotka **hyödyntävät jonkin verran digitaalisia metodeja**. Heidän digitaaliset tai tekniset taitonsa ovat keskitasoa. He osaavat hyödyntää esimerkiksi analyysiohjelmia, tekevät aineistolle digitaalista luokittelua tai ymmärtävät jonkin verran koodista, rajapinnoista tai algoritmien toimintaperiaatteista. He eivät kuitenkaan itse luo koodia tai tuota algoritmeja, tai tekevät niitä vain vähäisesti.

3) Digitaalisia aineistoja käyttävät, jotka **ovat digitaalisten metodien käytössä edistyneitä**. Heillä on edistyneet digitaaliset tai tekniset taidot. He osaavat luoda itse omia algoritmeja ja työstävät suuria aineistomassoja itse luomillaan työkaluilla.

Vastaavan jaottelun aloittelijoista edistyneisiin digitaalisten metodien käyttäjiin teki myös Sarah Ames (2021) artikkelissaan, mutta DAM-aineiston perusteella luokitteluryhmien rajat ovat häilyvät. Käyttäjät voivat eri tilanteissa sijoittua eri kategorioihin. DAM-aineiston perusteella suurin osa Kansalliskirjaston digitaalisia aineistoja käyttävistä tutkijoista kuuluu edellä esitellyssä luokittelussa joko ryhmään yksi tai kaksi. Lisäksi he hyödyntävät pääasiallisesti digitaalisia historiallisia sanomalehtiä. He eivät käsitelleet aineistoaan digitaalisesti kovinkaan pitkälle, mutta esimerkiksi avainsanojen lisääminen ja aineiston osien tallentaminen sähköisesti kuuluvat heidän digitaalisten aineistojen käyttö- ja käsittelytarpeisiinsa. Metodisesti tutkijat hyödynsivät sekä laadullisia että määrällisiä metodeja, mutta DAM-aineiston kokonaisuudessa käyttäjien metodiosaamisen tarve näyttää edelleen olevan varsin laadullinen ja ei-digitaalisten metodien orientoima. Kolmannen ryhmän tarpeet ovat kuitenkin digitaalisen humanismin suunnannäyttäjiä, ja ne on huomioitava tulevaisuuden suunnittelutyössä.

Käyttäjälähtöisen tiedon perusteella Kansalliskirjaston aineistoilla on hyvin monenlaisia tutkimuksellisia käyttötarkoituksia, ja myös muut kuin varsinaiset akateemisen tutkimuksen käyttötavat heijastuivat aineiston keruuseen. Osa käytti aineistoja aktiivisesti myös vapaa-ajan kiinnostuksen kohteidensa tukemiseen ja jakoi tietoa aineistoista ja niiden mahdollisuuksista esimerkiksi sosiaalisessa mediassa. Tutkijat eivät rajoittuneet mihinkään yksittäiseen tieteenalaan, vaikkakin käyttäjälähtöisen tiedonkeruun painopiste oli humanistisessa tutkimuksessa. Lisäksi raportoitiin jonkin verran taiteellisia aineiston käyttötarkoituksia. Suuri osa käyttäjistä tarvitsi aineistoa Kansalliskirjaston lisäksi myös muista kulttuuriperintöorganisaatioista joko kansallisesti tai kansainvälisesti. Pitkän tähtäimen tavoitteeksi lieneekin tarpeen lisätä erilaisten aineistotyyppien ja formaattien yhteentoimivuus digitaalisten aineistojen käytettävyyden parantamiseksi.

Monet tutkijat operoivat kansainvälisellä tutkimuksen kentällä, tarjoavathan digitaaliset aineistot mahdollisuuden ajasta ja paikasta riippumattomaan työskentelyyn. Monille on tästä huolimatta tärkeää olla samanaikaisesti paikallinen. Alueellinen sidonnaisuus voi näkyä esimerkiksi tutkimusaiheessa tai alueellisten yhteistyökumppanuuksien hyödyntämisessä sekä kotipaikassa.

Esimerkiksi Etelä-Savon alueella toimivat tutkijat ja heidän edustamansa organisaatiot, joissa käytettiin Kansalliskirjaston aineistoja, pitivät Etelä-Savon alueella tapahtuvaa alueellista yhteistyötä tärkeänä. Yhteistyötä haluttiin vahvistaa, ja eri tahojen toiminnan ymmärtäminen koettiin molemmin puolin tärkeänä sekä tutkijoiden että Kansalliskirjaston suunnasta. Haastattelut toimivat monien tutkijoiden mielestä hyvänä vuoropuheluna, jossa pohdittiin

aiemman yhteistyön kokemuksia ja avattiin keskustelua uuden yhteistyön kehittämiselle. Esimerkiksi alueelle sijoittuvista tapahtumista oli hyviä kokemuksia ja niitä toivottaisiin järjestettäväksi useammin. Haastatellut pitivät myös tärkeänä, että heidän äänensä pääsee kuuluville ja he voivat vaikuttaa palveluiden kehittämiseen.

KÄYTTÄJIEN TARPEESEEN VASTAAMINEN

Kaikissa käyttäjälähtöiseen aineistonkeruun vaiheissa keskeisiä toiveita olivat digitaalisten aineistojen löydettävyys, saatavuus ja käytettävyys. Yksi hankkeen keskeisimmistä tuloksista oli Kansalliskirjaston vanhimman digitoidun sanomalehtiaineiston tekstin tunnistuksen uudelleen tekeminen. Se edistää digitaalisten aineistojen löydettävyyttä ja käytettävyyttä kaikkien käyttäjien keskuudessa, koska se parantaa esimerkiksi hakutuloksien tarkkuutta. Työssä hyödynnettiin Euroopan Unionin Horizon-ohjelman NewsEye-projektissa kehitettyä mallia, joka paransi aiempaa tekstintunnistuksen tapaa tekoälyn avulla.

Aineistojen löydettävyydessä panostettiin niiden näkyvyyteen viestimällä digitaalisista aineistoista ja niiden kehittämisestä useissa kanavissa. Esimerkiksi keskustelut matkailu- ja kulttuurialan toimijoiden kanssa sekä aineistojen esittelyt Kaakkois-Suomen ammattikorkeakoulun opiskelijaryhmille toivat aineistoja tunnetuksi heidän keskuudessaan. Lisäksi hankkeen aikana pystyttiin käyttäjälähtöisen tiedon perusteella luomaan yleiskuva siitä, millaisia teknisen kehityksen tarpeita sisältyy varsinaisten digitaalisten aineistojen hakutoimintojen kehittämiseen. Yleisesti myös tämä tulee tulevaisuudessa palvelemaan kaikkia digitaalisten aineistojen käyttäjiä.

Digitaalisten aineistojen saatavuuteen ei tekijänoikeuksien asettamissa rajoissa voitu juuri vaikuttaa, mutta hankkeen aikana pystyttiin selkeyttämään aineistojen uudelleen käytettävyyttä tuottamalla ohjeellisia käyttöoikeustekstejä, joissa oli ohjeellinen vuosiraja siitä, milloin aineisto on käytettävissä sellaisenaan (https://digi.kansalliskirjasto.fi/terms). Uusien ohjeistuksien jälkeen käyttäjälle jää kuitenkin edelleen vastuu siitä, että hän noudattaa yksittäisten aineistojen käyttöoikeuksia tieteellisten käytäntöjen mukaan tutkittavien henkilöiden tietosuojaa ja tekijänoikeuksia kunnioittaen.

Digitaalisten aineistojen käytettävyys parani uuden työkalun kehittämisellä. Digitaalisten aineistojen lataustyökalu palvelee erityisesti teknisiltä taidoiltaan vähemmän edistyneitä käyttäjiä. Se vastaa käyttäjien tarpeeseen käyttää aineistoja uusilla tavoilla, koska se mahdollistaa suurten aineistomassojen lataamisen ilman ohjelmointitaitoja. Lataustyökalun avulla on yksinkertaisen käyttöliittymän avulla mahdollista ladata digi.kansalliskirjasto.fi -palvelusta tekijänoikeudesta vapaat hakutulokset omalle koneelle erilaisissa tekstimuodoissa ja/tai sivukuvina. Työkalun avulla ladattavat aineistot on helppo siirtää erilaisiin analyysiohjelmistoihin, alkaen taulukkolaskentaohjelma Excelistä aina tutkijoiden itse kehittämiin analyysityökaluihin asti. Työkalun tuottamat csv-tiedostot ovat yhteensopivia

monien ohjelmistojen ja työkalujen kanssa, mikä on omiaan lisäämään digitaalisten aineistojen käytettävyyttä eri yhteyksissä.

Käyttäjälähtöinen tiedonhankinta toimi laajemmin mahdollisuutena tutkijoiden ja Kansalliskirjaston välisen vuorovaikutuksen kehittämiskohteiden tunnistamiseen. Tiedonhankinnan aikana pystyttiin tunnistamaan joitakin eroja esimerkiksi siinä, miten datan merkitys ymmärretään eri käyttötarkoituksissa ja miten tutkimusprosessia hahmotetaan eri tavoin. Vaikka ymmärrys näistä kasvoi hankkeen aikana, on tarpeen edelleen kiinnittää huomiota siihen, miten digitaalisista aineistoista, datasta ja niiden käytöstä viestitään.

VERTAISOPPIA LIBRARY LABEISTA

Samaan aikaan käyttäjälähtöisen tiedon kanssa hankkeessa hankittiin vertaisoppia library labeista. Tätä benchmarkkaamiseksikin kutsuttavaa metodia toteutettiin haastattelemalla ja havainnoimalla Euroopassa toimivia datalähtöisiä tutkijapalveluita. Apuna toimi myös aiheeseen liittyvä kirjallisuus (esim. Ames 2021; Candela ym. 2020; Mahey ym. 2019; Snickars 2018) ja laajasti erilaiset tapahtumat, jotka vuoden 2020 koronapandemia vei verkkoon.

Hankkeessa haastateltiin 14 henkilöä seitsemästä maasta. Haastattelut noudattivat puolistrukturoidun teemahaastattelun mallia ja ne toteutettiin englanniksi verkkoyhteydellä (Hirsjärvi & Hurme 2008). Viimeinen kahden hengen haastattelu oli erään teknisen alustan ympärille keskittynyt, eikä heidän haastattelunsa ollut mukana 12 ensimmäiselle haastatellulle lähetyssä ja avoimesti julkaistussa haastatteluraportissa (Näpärä & Liukkonen 2020). Lisäksi tämän jälkeen olimme yhteydessä vielä yhteen kansainväliseen kontaktiin, mutta tiedonvaihtoa tapahtui enemmänkin hankkeesta ulospäin, ja näin ollen 15. kansainvälistä kontaktia ei voi pitää varsinaisen vertaisoppimisaineiston osana.

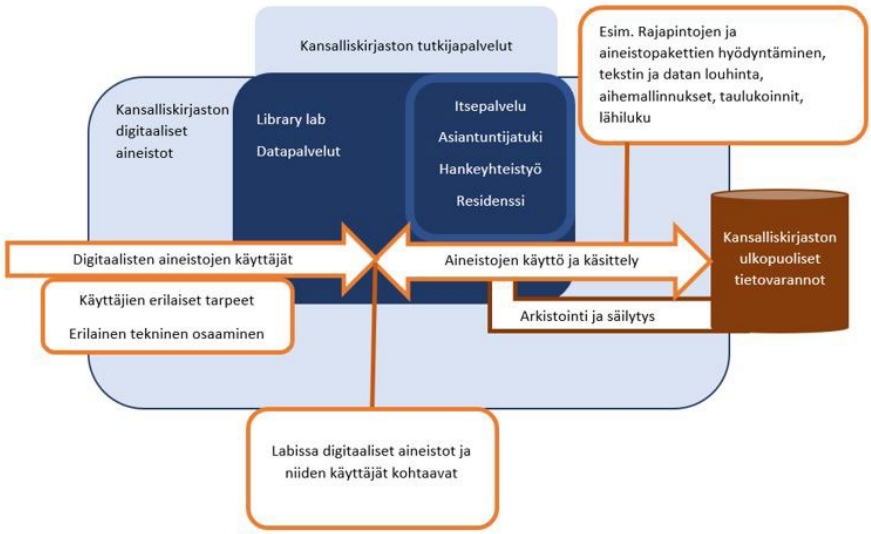
Aineistonkeruun perusteella voidaan määritellä, että library lab on paikka, jossa tutkimus ja kirjastoaineistot kohtaavat – useimmiten datana ja datalähtöisenä tutkimuksena. Datalla tarkoitetaan tässä koneluettavaa digitaalista aineistoa, josta voidaan tuottaa tutkimusaineistoa tai sen avulla kouluttaa koneoppimiseen perustuvia työkaluja käsittelemään sitä. Labeissa työstetään, kehitetään ja luodaan uusia työkaluja datan käyttöön ja saadaan selville tutkimuksen avulla jotain sellaista, jota ei aiemmin tiedetty. Useimmiten datan käsittelyssä hyödynnetään nimenomaan digitaalisten ihmistieteiden menetelmiä, joita voi kutsua myös tietokoneavusteisiksi tai laskennallisiksi menetelmiksi. Uudenlaiset ja vielä kehittyvät digitaaliset menetelmät avaavat uusia mahdollisuuksia aineistojen käsittelyyn tutkimusperusteisesti, mutta edelleen on huomioitava esimerkiksi aineistoon, digitointiin ja algoritmeihin liittyvät ominaispiirteet, kuten digitaalisen otoksen luotettavuus, tai vääristyminen, kuten historiallisen aineiston säilymiseen ja digitointiprosesseihin liittyvät aineistojen valinnat. Tutkimusprojekteille tämä tarkoittaa, että datan metatietojen pitää olla kunnossa, ja datan syntyhistoriaan liittyvän tiedon ymmärtämiseen tarvitaan kirjastosta tulevaa asiantuntemusta.

Aineistonkeruun mukaan labien toiminnassa on tarkoitus innovoida ja kehittää sekä myös epäonnistua ja oppia yhdessä. Tällainen ei onnistu ilman ihmisten välistä vuorovaikutusta, joka jää pelkkään dataan keskittyvillä internet-sivuilla toisinaan vajavaiseksi. Laajemmin tarkasteltuna taas labien toimintakulttuurin täytyy aineiston analyysin mukaan sallia erilaisen toiminnan kokeiluluonteisuus, ja organisaation pitää olla valmis joustamaan esimerkiksi sallimalla ajoittaiset takapakit, joiden kuitenkin ajatellaan lopulta johtavan toimintakulttuurin edistämiseen ja uusien asioiden luovaan synnyttämiseen.

Jonkin verran labien kokeilukulttuuria ja samalla eräänlaista keskeneräisyyttä ja kehittyvyyttä oli näkyvissä hankkeessa tehdyissä havainnoissa. Esimerkiksi joidenkin labeissa kehitettyjen työkalujen dokumentaatiossa havaittiin puutteita. Kaikkea ei jaettu julkisesti tai tieto oli hieman hankalasti saatavissa. Syitä selvitettäessä havaittiin, että erilaisten työkalujen jatkokäyttöä ei pidetty todennäköisenä tai sitä ei edes tavoiteltu. Kokonaiskuvassa ja verrattuna labien yleisiin avoimuuden periaatteisiin tämä antaa hieman ristiriitaisen kuvan, mutta asia ei ole täysin yksinkertainen. Laajemmin projekteissa syntyvien datojen ja työkalujen säilytys- ja jatkokäyttö on herättänyt keskustelua, eikä se tapahdu ilman suunnittelua ja eri tahojen välistä yhteistyötä (Näpärä 2021). Asiaan on myös kiinnitetty labeissa huomioita esimerkiksi suunnittelemalla projektien ylläpitoa rahoituksen päättymisen jälkeen. Aiheen ympärillä riittää kuitenkin vielä työtä. Esimerkiksi nopeasti kehittyvät teknologiat, päivittyvät standardit ja muuttuvat formaatit aiheuttavat omat haasteensa projekteissa syntyvien tuotosten kestävälle ja oikeasuhtaiselle ylläpidolle.

LIBRARY LAB TUTKIJOIDEN DATAPALVELUNA

Kansalliskirjaston datapalvelut kehittyvät osaksi laajempien, erilaisiin aineistotyyppeihin perustuvien tutkija- ja aineistopalveluiden kokonaisuutta. DAM-hankkeen aikana uutta library lab -muotoista datapalvelua määriteltiin suhteessa olemassa oleviin palveluihin, ja täten voitiin kehittää prosessi datalähtöisten tutkimusyhteistyöprojektien hakuvaiheeseen. Digitaalisten aineistojen ja datan tutkimuskäyttö läpäisee erilaiset Kansalliskirjaston tarjoamat tutkijoille tarjottavat datan itsenäiset käyttömahdollisuudet, räätälöityjen datapakettien luomisen ja dataan liittyvän tutkimusyhteistyön. Datapalvelut muodostavat kuitenkin oman lab-kokonaisuutensa digitaalisia aineistoja ja dataa hyödyntäville käyttäjille. Se huomioi käyttäjien erilaiset teknisen osaamisen tasot ja tarjoaa erilaisille käyttäjille asiantuntevaa palvelua ja tukea varsinaisen datan ja sen käsittelyyn tarvittavien työkalujen lisäksi. Näitä on havainnollistettu kuvassa 1.



Kuva 1. Datapalvelut (library lab) osana kokonaisuutta

Kansalliskirjastossa aivan uudenlaisena datan käytön mahdollistamistapana suunniteltiin tutkijaresidenssiä ja sen pilotointia. Residenssin ideana on tarjota digitaalisten aineistojen käyttöön uusia mahdollisuuksia silloin, kun ne ovat tekijänoikeuden alaisia ja niiden käyttö on vaikeaa tai jopa mahdotonta kirjaston tilojen ulkopuolella. Esimerkkeinä olevien library labien mukaista digitaalisten aineistojen käyttöön tarkoitettua tutkijaresidenssiä pilotoidaan DAM-hankkeessa saadun informaation perusteella omana hankkeenaan vuodesta 2021 alkaen.

LOPUKSI

Digitaalisten aineistojen ja datan ympärille pystytään tuottamaan aivan uudenlaisia käytettävyyteen liittyviä parannuksia sekä uudenlainen palvelukonsepti. Library lab -muotoisina datapalvelut kehittyvät jatkuvasti vastaamaan käyttäjiensä tarpeeseen ja sopeutuvat tarvittaessa muuttuviin digitaalisten aineistojen käyttöratkaisuihin. Kehitystä tarvitaan, sillä osa artikkelin alussa esitetyistä kysymyksistä vaatii vielä jatkotyöstöä. Esimerkiksi kysymykset datan ja työkalujen kestävästä säilyttämisestä tutkimushankkeiden jälkeen eivät saa täydellistä vastausta, vaan vaativat lisää kansallista ja jopa kansainvälistä yhteistyötä eri toimijoiden välillä. Lisää tietoa tarvitaan myös aineistojen jalostajista ja uuden tiedon synnyttämiseen liittyvistä prosesseista aineistojen hyödyntämiseen liittyvän profiloinnin ohella. Haasteeksi jäävät edelleen kuitenkin sellaiset digitaaliset aineistot, joiden tekijänoikeus on voimassa ja sopijakumppania niiden käytöstä ei ole.

LÄHTEET

Ames, S. 2021. Transparency, provenance and collections as data: the National Library of Scotland's Data Foundry. LIBER Quarterly, 31(1), pp.1–13.

Candela, G., Dolores Sáez, M., Escobar Esteban, M. & Marco-Such, M. 2020. Reusing digital collections from GLAM institutions. Journal of Information Science. 46:5, 1–17, DOI: https://doi.org/10.1177/0165551520950246.

Digi.kansalliskirjasto.fi käyttöehdot https://digi.kansalliskirjasto.fi/terms [viitattu 10.5.2021].

Hirsjärvi, S. & Hurme, H. 2008. Tutkimushaastattelu: teemahaastattelun teoria ja käytäntö. Gaudeamus.

Kankainen, A., Vaajakallio, K., Kantola, V. & Mattelmäki, T. 2012. Storytelling Group – a co-design method for service design. Behaviour & Information Technology, 31:3, 221-230, DOI: https://doi.org/10.1080/0144929X.2011.563794

Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., Derven, C., Dobreva-McPherson, M., Gasser, K., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S-C. and Wilms, L., with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P. and Papaioannou, G. 2019. Open a GLAMLab. Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23–27 September 2019. https://glamlabs.pubpub.org/ [viitattu 10.5.2021].

Näpärä, L. & Liukkonen E. 2020. Report on the benchmarking interviews in the Digital Open Memory project. Zenodo. DOI: http://doi.org/10.5281/zenodo.4285836.

Näpärä, L. 2021. Tutkimusdatan säilyttämiseen tarvitaan kansallista yhteistyötä. Signum, 52(4), 30–31. DOI: https://doi.org/10.25033/sig.101390.

Park, H. Y., Il-Hyung Cho, Jung, S., & Main, D. 2015. Information and communication technology and user knowledge-driven innovation in services. Cogent Business & Management, 2(1). DOI: http://dx.doi.org/10.1080/23311975.2015.1078869

Pääkkönen, T. & Lilja, J. 2018. Hieno palvelu, mutta sisältöä lisää – Kansalliskirjasto kyseli Digi-palvelun käyttökokemuksia. Tietolinja, 2018(2). http://urn.fi/URN:NBN:-fi-fe2018092336401.

Snickars, P. 2018. Datalabb på KB. https://pellesnickars.se/2018/10/datalab-kb-se-a-report-for-the-national-library-of-sweden/ [viitattu 10.5.2021].

DIGITOIDUISTA HISTORIAN LEHDISTÄ MUOKATTIIN OPPIMATERIAALIA

Marja-Leena Hynynen, HuK, tietoasiantuntija, Kansalliskirjasto
Liisa Näpärä, FT, suunnittelija, Kansalliskirjasto

Kansalliskirjaston digi.kansalliskirjasto.fi -palvelu (Digi) sisältää tällä hetkellä yli 20 miljoonaa sivua aineistoja. Kokoelmiin kuuluu muun muassa lehtiä, kirjoja, pienpainatteita, karttoja ja nuotteja. Käytetyimpiä ja laajimpia aineistoja ovat sanoma- ja aikakauslehdet. Kaikki Suomessa vuoden 1939 loppuun mennessä ilmestyneet sanomalehdet ja yleiset aikakauslehdet ovat Digissä vapaasti yleisön selattavissa ja luettavissa. Vanhimmat näistä ovat vapaita tekijänoikeuksista, ja tuoreemmista aineistoista on solmittu käyttöoikeussopimus Kopioston kanssa. Historialliset lehdet ovat suosittua tutkimusmateriaalia sukututkijoille ja historioitsijoille. Kuka tahansa kansalainen voi löytää Digistä esimerkiksi kotiseutunsa historiaan liittyvää aineistoa.

Ei ole kuitenkaan itsestään selvää, että kansalaiset löytävät tiensä Digin sisältöjen äärelle. Aineistojen käytettävyyden ja tunnettuuden parantamiseksi on tehtävä työtä, jotta historian aarteet eivät hukkuisi digitaaliseen tietotulvaan. Digitaalinen avoin muisti (DAM) -hankkeen tavoitteena oli parantaa mahdollisuuksia hyödyntää laajoja digitaalisia aineistoja. Osana hanketta Kansalliskirjastossa tuotettiin digitaalisia opetuspaketteja, joilla pyrittiin tutustuttamaan opettajia ja oppilaita Digin lehtiaineistoihin. Opetussuunnitelmaan sidotut aineistopaketit suunnattiin yläkoulun ja lukion opetukseen, ja ne julkaistiin Finna Luokkahuone -sivustolla toukokuussa 2021.

MUISTIORGANISAATIOT TUOTTAVAT OPPIMATERIAALEJA VERKKOON

Kirjastojen, arkistojen ja museoiden aineistoja on mahdollista hyödyntää opetuksessa monin tavoin, entistä useammin myös digitaalisessa muodossa. Eri organisaatioiden kulttuuri- ja tiedeaineistoja voi Suomessa hakea keskitetysti Finna.fi-palvelusta. Erityisesti opettajille suunnattu osa tätä palvelua on Finna Luokkahuone, joka otettiin käyttöön syksyllä 2019. Finna Luokkahuone sisältää valmiita aineistopaketteja ja avoimia oppimateriaaleja oppitunneilla käytettäväksi. Museot ja muut muistiorganisaatiot ovat yhteistyössä opettajien kanssa räätälöineet aineistopaketteja eri oppiaineiden kuten historian, yhteiskuntaopin, kuvataiteen, kotitalouden sekä äidinkielen ja kirjallisuuden oppitunneille. Paketit sisältävät runsaasti kuva-aineistoja ja jonkin verran myös ääni-, kartta- ja tekstiaineistoja.

Finna Luokkahuone helpottaa opettajien työtä tarjoamalla opetussuunnitelman mukaisia, laadukkaita sisältöjä ja valmiita tehtäväideoita oppitunneilla käytettäviksi. Luokkahuone on toimiva palvelu myös muistiorganisaatioille, sillä se tarjoaa valmiin alustan aineistopakettien julkaisemiseen. DAM-hankkeen opetuspaketin julkaisemiseen käytettiin tätä olemassa olevaa palvelua, jonka kautta oli mahdollista tavoittaa laaja joukko opettajia.

Kansalliskirjastossa ei ole aiemmin tuotettu oppimateriaalia Digin lehtiaineistoista. Projektin alkuvaiheessa haettiin ideoita ja vertailupohjaa oppimateriaaleista, joita eri toimijat ovat julkaisseet verkossa. Esimerkiksi Kongressin kirjasto Yhdysvalloissa ja Ranskan kansalliskirjasto ovat laatineet kouluja varten monipuolisia aineistokokonaisuuksia. Suomalaisia esikuvia olivat Museoviraston ylläpitämä Opi aineeton kulttuuriperintö -sivusto sekä tietenkin Finna Luokkahuoneessa jo julkaistut aineistopaketit.

AINEISTON HAUN SUUNTAVIIVAT JA HAASTEET

Projektissa tarkasteltiin perusopetuksen ja lukion opetussuunnitelmien perusteita ja todettiin, että digitoidut lehtiaineistot sopivat hyvin yhteen monien laaja-alaisen osaamisen tavoitteiden kanssa. Laaja-alainen osaaminen käsittää muun muassa monipuolisia oppimisen taitoja. Oppilaita tulee rohkaista hankkimaan tietoa eri lähteistä, kohtaamaan ristiriitaistakin tietoa, arvioimaan lähteiden luotettavuutta ja kehittämään omaa ajatteluaan. Elinympäristön kulttuuriperintöön tutustuminen ja kulttuurien moninaisuuden havainnointi auttavat oppilaita näkemään kulttuurin merkitystä. Monilukutaito sekä tieto- ja viestintäteknologinen osaaminen varustavat nuoria toimimaan yhteiskunnassa ja rakentamaan tulevaisuutta. (Opetushallitus 2014; Opetushallitus 2019)

Oppiaineet, joissa luontevimmin voisi hyödyntää Digin sisältöjä, ovat historia sekä äidinkieli ja kirjallisuus – unohtamatta kuitenkaan monialaista tutustumista erilaisiin ilmiöihin ja suurempiin oppiainerajat ylittäviin kokonaisuuksiin. Äidinkielen sisältöalueissa kiinnitti huomiota erityisesti tekstien tulkinta ja historian sisältöalueissa arkipäivän historia, elinkeinorakenteen muutos, maatalous, teollistuminen, lähteiden ja etenkin lähiympäristön lähteiden käyttö. (Opetushallitus 2014; Opetushallitus 2019)

Näiden opetussuunnitelmista löytyneiden suuntaviivojen avulla hahmoteltiin muutamia teemoja, joista ryhdyttiin hakemaan aineistoja. Haut rajattiin maantieteellisesti Etelä-Savoon liittyviin sanoma- ja aikakauslehtisisältöihin. Tavoitteena oli kuitenkin löytää aineistoja, joita on mahdollista hyödyntää kouluissa ympäri Suomen. Lehdistä etsittiin artikkeleita, joiden aiheet olivat tuttuja tai vertailukelpoisia maaseudulla ja pikkukaupungeissa eri puolilla maata, tai joissa oli selkeä kytkös Suomen historian suuriin tapahtumiin.

Ajallinen rajaus tehtiin tekijänoikeuksien pohjalta. Tekijänoikeus on voimassa 70 vuotta tekijän kuolinvuoden päättymisestä. Jos tekijä ei ole tiedossa, rajana on 70 vuotta teoksen

julkistamisvuodesta. Tekijänoikeuden alaista aineistoa ei saa esimerkiksi kopioida ja julkaista uudelleen, vaikka se olisikin Kopioston kanssa tehdyn käyttöoikeussopimuksen perusteella luettavissa Digissä. Projektissa päätettiin käyttää vuoden 1918 loppuun mennessä ilmestyneitä sanoma- ja aikakauslehtiä, jotta aineisto olisi mahdollisimman suurelta osin vapaata tekijänoikeuksista ja riski oikeuksien rikkomisesta olisi pieni.

Sopivien sisältöjen hakeminen valtavasta aineistomassasta vaati paljon työtä. Aluksi aineistoa haravoitiin laajasti, jotta nähtiin, millaisia aiheita esimerkiksi tiettyjen paikkakuntien yhteydessä nousi esiin. Vanhimpien aineistojen haravoinnissa käytettiin apuna 1800-luvun lopulla koottua artikkelihakemistoa ja etenkin sen Topographica-osiota, josta lehtiartikkeleita voi hakea paikkakunnittain. Hakemisto oli alun perin käsin kirjoitettu kortisto, ja nykyisessä digitaalisessa muodossaan se sisältää linkit artikkeleihin (Kokko 2017).

Tarkempia hakuja tehtiin esimerkiksi yhdistämällä paikannimiä ja valittuun aihepiiriin liittyviä hakusanoja. Digin hakutoiminnot mahdollistavat muun muassa Boolen operaattoreiden ja erilaisten rajaimien käytön, mutta tiedonhaku ei silti ollut ongelmatonta. Alkuperäisen aineiston vaihteleva laatu ja optisen luvun virheet toivat omat haasteensa, ja vuosien varrella tapahtuneet sanaston ja nimistön muutokset vaikeuttivat oikeiden hakusanojen löytämistä.

Alussa hahmoteltuja teemoja muokattiin projektin edetessä sen mukaan, kuinka aineistolöydöt painottuivat. Elinkeinot olivat teema, johon löydettiin hyvin erityyppisiä, osin vaativia aineistoja. Lopulta teemaa käsiteltiin sanomalehtien ilmoitusaineiston avulla. Joitakin teemoja jätettiin pois, kun hakutuloksista ei muotoutunut toimivaa kokonaisuutta. Toisaalta esimerkiksi Olavinlinna nousi yhden aineistopaketin aiheeksi, koska siihen liittyviä artikkeleita löytyi runsaasti.

LEHTIAINEISTON ERITYISPIIRTEITÄ

Lehtiaineistoja ei voitu käyttää opetuspaketeissa täysin saman kaavan mukaan kuin kuva-aineistoja. Paketteja ei esimerkiksi voitu kääntää sellaisinaan suomesta ruotsiksi. Jotta pystyttiin julkaisemaan Finna Luokkahuoneessa sekä suomen- että ruotsinkieliset paketit, oli haettava erikseen aineistoa ruotsinkielisistä lehdistä ja laadittava löydettyihin aineistoihin sopivat tehtävät. Aihepiireiltään lehtiaineistot olivat samankaltaisia, vaikka aineisto ja tehtävät erosivat toisistaan.

Oman haasteensa työhön toi se, että sanoma- ja aikakauslehdet ovat löydettävissä Finnasta vain nimekkeiden tasolla, eikä yksittäisiin artikkeleihin ainakaan toistaiseksi ole suoraa reittiä. Siirtyminen Finnasta Digiin oli tehtävä aineistopaketeissa mahdollisimman selkeäksi, jotta oppitunneilla ei tuhlautuisi aikaa artikkelien etsimiseen. Asia ratkaistiin tarjoamalla suorat linkit artikkelisivuille, mutta edelleen haasteeksi jäi kokonaisuudessa tässä kohdassa

ylimääräisiksi tulkittavat linkit lehtinimekkeisiin. Navigoinnin helpottamiseksi tarvittiin selkeitä ohjeita ja huomion kiinnittämistä siihen, että oikeasta kohdasta painamalla pääsee suoraan aineistona olevan artikkelin sivulle. Vaikka nimekkeiden näkyminen ja niiden käyttäminen varsinaisten artikkelien etsintään palvelisivat laajasti ajateltuna tiedonhakua historiallisesta tietokannasta, tämän arveltiin käytettävyyden tasolla olevan haastavaa peruskoululaiselle ja vievän liikaa aikaa varsinaiselta oppimistehtävältä. Toisaalta opettajan on mahdollista opettaa tiedonhakua nimekkeistä esimerkiksi lisätehtävien muodossa.

Projektissa pohdittiin myös sitä, että kouluopetuksessa käytettävien tekstien on oltava oppilaiden ikätasolle sopivia. Kohderyhmä päädyttiin rajaamaan lukiolaisiin ja yläkoululaisiin, koska heillä on alakoululaisia paremmat edellytykset kyetä lukemaan ja ymmärtämään vanhoja lehtitekstejä. Suomen ja ruotsin kieli ovat muuttuneet aikojen kuluessa sekä sanastoltaan että kirjoitusasultaan, ja yli sata vuotta sitten kirjoitetut tekstit kuulostavat ilmaisutyyliltään vanhahtavilta.

Oppimateriaalina toimii parhaiten sellainen teksti, jossa ulkoiset seikat eivät häiritse sisällön ymmärtämistä ja luettavuutta. Jotta sanojen ja pitkien virkkeiden merkityksiä voisi edes yrittää tulkita, on ensiksi saatava selvää kirjaimista. 1800-luvulla ja 1900-luvun alussa Suomen lehdistö käytti kirjasinlajina pääasiassa fraktuuraa. Nykyisin fraktuuraa näkee harvoin muualla kuin Sisu-askin kannessa. Jos koululaisille annettaisiin historian tunnilla luettavaksi fraktuuralla painettu lehtiartikkeli, tunti kuluisi todennäköisesti kirjainten opetteluun, eikä tekstin sisällöstä ehdittäisi keskustella lainkaan. Aineistona oli siis viisainta käyttää antiikvalla painettuja tekstejä. Antiikvan käyttöön siirryttiin vähitellen ensin aikakauslehdissä ja ruotsinkielisissä sanomalehdissä ja lopulta myös suomenkielisissä sanomalehdissä. Tästä syystä suomenkielisiin aineistopaketteihin valikoitui aikakauslehtien artikkeleita ja sanomalehti-ilmoituksia.

Luettavuuteen vaikuttaa myös alkuperäisen lehden laatu. Kovin himmeä tai sotkuinen painojälki ei näytä selkeältä digitoitunakaan. Digitoidun tekstin laatu vastaa käytännössä lähes täysin alkuperäistä lehteä, ja oikein säilytettynäkään yli sadan vuoden takainen lehden laatu ei aivan vastaa tämän päivän HD-tason standardeja, joihin käyttäjät ovat tottuneet. Joitakin kiinnostavia lehtiartikkeleita jouduttiin karsimaan pois siksi, että tekstistä oli vaikea saada selvää.

Tekstiaineiston erityispiirteistä tuli lopulta oppimateriaalien aineistojen valintaa ohjaavia tekijöitä enemmän kuin oli ajateltu. Toisaalta ne rajasivat positiivisella tavalla valtavaa aineistomassaa pienemmäksi otokseksi, mutta edelleen valinnanvaraa oppimateriaalien aineistoksi oli runsaasti. Digin sanoma- ja aikakauslehdet osoittautuivat todellakin aarreaitaksi.

PAKETTIEN SOVELTUVUUS OPETUKSEEN

Opetuspakettien suunnittelussa ja teossa huomioitiin se, miten ne soveltuvat opetukseen sekä opettajien taitoihin sisällöllisesti, pedagogisesti ja teknologisesti (Mishra & Koehler 2006). Vaikka hankkeessa ei lähestytty oppimateriaalien tekoa varsinaisten opettajien taitojen näkökulmasta, opettajien on osattava hyödyntää oppimateriaaleja mahdollisimman helposti opetuksessaan pedagogisesti ja sisällöllisesti. Sen lisäksi teknologisen sisällön käytettävyyden pitää olla tarpeeksi helppoa, jotta opettajat osaavat vaihtelevallakin teknologiaosaamisellaan mahdollisimman helposti käyttää historiallisen sanomalehtiaineiston opetuspaketteja (Näpärä 2019).

Opetuspakettien sisällöllinen sopivuus opetukseen suhteutuu perusopetuksen ja lukion voimassa oleviin opetussuunnitelmiin. Opetuspakettien pedagoginen soveltuvuus testattiin siinä, kun ne kävivät yläkoulun ja lukion opettajien kommentoitavina. Yksi opettajista myös testasi niitä varsinaisessa opetuksessa. Teknologinen opetukseen soveltuvuus tulee Finna Luokkahuoneen valmiista ratkaisusta. Melko uutena konseptina havaittiin kuitenkin joitain haasteita siinä, miten se soveltuu historiallisen tekstiaineiston opetuskäyttöön. Ratkaisut tekstimuotoisen aineiston esittämiseen löytyivät hyvässä yhteistyössä, ja kokonaisuudesta yritettiin tehdä mahdollisimman käyttäjäystävällinen.

PALAUTETTA OPETTAJILTA JA OPPILAILTA

Opetuspaketeissa DAM-hankkeen käyttäjälähtöisyyden tavoitetta toteutettiin tekemällä yhteistyötä muutamien historian ja äidinkielen opettajien kanssa. Opettajat tutustuivat aineistopakettien luonnoksiin ja antoivat niistä palautetta. Sekä yläkoulun että lukion opettajat olivat ilahtuneita siitä, että tämäntyyppistä oppimateriaalia tuotetaan. He antoivat useita hyödyllisiä kehittämisehdotuksia, joiden pohjalta paketteja muokattiin edelleen.

Opettajien kommenteissa korostui se, että koulussa ei ehditä käsitellä kovin suuria tekstimääriä. Oppitunnit ovat lyhyitä, ja osalla yläkoululaisista on vaikeuksia lukemisessa. Luonnoksissa oli liian paljon ja liian pitkiä tekstejä. Konkreettinen parannusehdotus oli esimerkiksi se, että tehtävissä viitattaisiin tarkemmin tiettyyn kohtaan tekstissä, jolloin vastaukset löytyisivät nopeammin. Palautteissa tuli ilmi myös, että luonnosten tehtäväideat eivät vastanneet yläkoulun tarpeita. Opettajat toivoivat eritasoisia tehtäviä, jotka ryhmiteltäisiin vaikeustason mukaan perus- ja syventäviin tehtäviin.

Yksi opettajista testasi pakettiluonnosta seitsemännen luokan äidinkielen tunnilla ja pyysi oppilaita kirjoittamaan ensivaikutelmiaan. Oppilaat kommentoivat muun muassa näin: ”Hieman liian pitkää tekstiä mutta muuten hyvää asiaa.” ”Uutiset olivat vanhoja ja tylsiä. Niitä oli vaikea tulkita ja lukea.” ”Katsottiin vanhaa kirjaa, jossa kerrottiin Runebergistä. Lopussa oli myös Suomen paikkojen nimi luettelo. Materiaali oli mielenkiintoista.

Löydettiin mielenkiintoisia juttuja.” Oppilaiden kommenteissa toistui sama, mitä opettajat olivat sanoneet tekstien pituudesta. Tuli myös hyvin ilmi, että materiaalia oli muokattava selkeämmäksi. Kun Olavinlinnaa käsittelevien lehtiartikkelien sijaan oli päädytty katselemaan kirjaa Runebergistä, ohjeissa oli selvästikin parantamisen varaa.

Tekstiaineistoja karsittiin, ja osa paketeista kohdennettiin vain yläkoululle tai lukiolle. Tehtäviin lisättiin luetun ymmärtämiseen keskittyviä perustehtäviä, ja pohdintaa ja analysointia vaativat kysymykset sijoitettiin syventäviin tehtäviin. Opettajat toivoivat myös mallivastauksia. Tätä toivetta ei toteutettu, sillä mallivastaukset eivät kuulu Finna Luokkahuoneen käytäntöihin. Sen sijaan toive aineistopakettien tehtävien ja ohjeistuksien selkeyttämisestä otettiin huomioon, kun paketteja työstettiin lopulliseen muotoonsa.

Lopputuloksena oli neljän aineistopaketin muodostama suomenkielinen kokonaisuus, jolle annettiin nimeksi Lehtien kertomaa Savosta, ja vastaava ruotsinkielinen kokonaisuus, joka otsikoitiin Tidningarna berättar om Savolax. Julkistaminen tapahtui 6.5.2021, ja asiasta tiedotettiin Kansalliskirjaston verkkosivuilla ja sosiaalisen median kanavilla sekä erityisesti Etelä-Savon alueen medioille ja kouluille. Lisäksi päätettiin tehdä viestintää uudelleen syyslukukauden alussa esimerkiksi lähettämällä sähköpostia opetustoimelle Etelä-Savon kunnissa.

LOPUKSI

Opetusaineiston kuratointi lehtiaineistoista oli Kansalliskirjastolle tuore avaus, ja hankkeen aikana opittiin paljon uutta. Opetukseen soveltuvien aineistojen hakeminen oli aikaa vievää ja haasteellista, mutta työn tuloksena syntyi toimiva aineistokokonaisuus. Yhteistyö opettajien kanssa osoittautui hedelmälliseksi: kirjaston aineistotuntemuksen rinnalle saatiin käytännön pedagogista näkökulmaa koulumaailmasta.

DAM-hankkeessa tuotetut Finna Luokkahuoneen opetuspaketit tutustuttavat peruskoululaiset ja lukiolaiset digitaalisten aineistojen käyttöön ja opettavat opetussuunnitelman mukaista sisältöä arjen historiasta kannustaen samalla opetussuunnitelman mukaiseen lähdekriittisyyteen. Ne parantavat nuorten koululaisten ja opettajien tietoisuutta digitaalisista aineistoista ja niiden käytöstä myös opetus- ja oppimistilanteiden ulkopuolella.

LÄHTEET

Kokko, H. 2017. Digitaalisten aineistojen sanomalehtihakemiston historiasta. Verkkojulkaisu. https://blogs.helsinki.fi/scriptaselecta/2017/07/13/digitaalisten-aineistojen-artikkelihakemiston-historiasta/ [viitattu 22.2.2021]

Mishra, P. & Koehler, M. 2006. Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. Teachers College Record 108:1017–1054 [viitattu 13.4.2021].

Näpärä, L. 2019. Mikä ihmeen digiloikka? Opettajuuden rakentuminen digiloikkadiskurssissa. Itä-Suomen yliopisto. Väitöskirja. PDF-dokumentti. http://urn.fi/URN:ISBN:978-952-61-3212-9 [viitattu 13.4.2021].

Opetushallitus 2014. Perusopetuksen opetussuunnitelman perusteet. PDF-dokumentti. https://www.oph.fi/fi/koulutus-ja-tutkinnot/perusopetuksen-opetussuunnitelman-perusteet#34746527 [viitattu 19.2.2021].

Opetushallitus 2019. Lukion opetussuunnitelman perusteet. PDF-dokumentti. https://www.oph.fi/fi/koulutus-ja-tutkinnot/lukion-opetussuunnitelmien-perusteet#34746527 [viitattu 19.2.2021].

KULTTUURIELÄMYKSIÄ COVID19-AIKANA DIGITALIAN DYNAAMISELLA VIRTUAALI-MUSEOLLA

Anssi Jääskeläinen, TkT, tutkimuspäällikkö, Xamk

Euroopan parlamentti on päätöslauselmassaan korostanut, että kulttuuriperinnön säilyttäminen on tärkeää monestakin eri näkökulmasta (European Parliament resolution 2014/2149). Tässä julkaisussa esitelty työ keskittyy kulttuurielämyksien tarjoamiseen, mutta ratkaisumme toimisi hyvin myös päätöslauselmassa mainitussa opetuksessa. Kulttuurielämykset ovat osalle väestöstä erittäin tärkeitä, mutta Covid-19 -pandemian vuoksi museot, arkistot ja galleriat joutuivat pitämään ovensa kiinni ja aineistot ovat olleet kävijöiden ulottumattomissa.

Oikein toteutettuna virtuaalimuseo voi tuoda kulttuurielämyksen museokävijän omaan kotiin ja näin ollen osaltaan ylläpitää henkistä hyvinvointia. Koronapandemia ei kuitenkaan ollut laukaisevana tekijä kehitystyössä, vaan aloimme tutkia tätä ongelmaa jo aiemmin - ajankohta vain sattui osumaan erityisen hyvin kohdalleen. Työskentely aiheen parissa alkoi jo syyskuussa 2019 ja ensimmäisiä pilotointeja kehitetyn museon kanssa tehtiin alkuvuodesta 2020.

Varsinaisesti ajatus virtuaalimuseosta lähti liikkeelle lomamatkastani Las Palmasiin, jossa sijaitsee El Museo Canario. Tämä museo on kiinnostava jo siksi, että se kertoo kyseisen saaren esihistoriallisesta ajasta ja kansoista. Museossa on myös paljon kalloja, luita ja muita entisaikojen elämään liittyvää materiaalia. Lisäksi museossa oli virtuaalilaseilla toteutettu elämys, jossa pääsi osaksi kyseisen aikakauden kylän elämää. Tällaisia toivoisi näkevänsä / kokevansa enemmänkin GLAM (Galleries, Libraries, Archives, Museums) -sektorilla, mutta ongelmaksi nousevat usein

1. teknisen tietotaidon puute
2. resurssit
3. digitaalisessa muodossa olevan materiaalin vähyys.

ONGELMAKENTTÄ

Yhteenkään yllä mainituista ongelmista ei ole ollut helppoa ratkaisua ennen Digitalian DAM (Digitaalinen avoin muisti) -hankkeen virtuaalimuseota, joka ratkaisee kaksi ensimmäistä ongelmaa täysin. Teimme hankkeen aikana myös pilotoinnin, Suomen Elinkeinoelämän Keskusarkiston yhteydessä toimivan Designarkiston kanssa. Digitoimme hankkeelle ostetulla Shining EinScan-SP skannerilla yhdeksän Designarkiston esinettä, jotka lisättiin mukaan kiinteänä osana Digitalian virtuaalimuseota. Välillisesti ratkaisumme kattaa siis myös kolmannen ongelman, koska olemme testanneet työnkulkua esineestä 3D-mallin kautta museoon. Tämä työnkulku on suhteellisen pienellä investoinnilla monistettavissa myös muihin museoihin.

Mainituista ongelmista huolimatta maailmalta löytyy myös monia virtuaalimuseototeutuksia, mm. Viking VR (Schofield ym. 2018). Vaikka takana olisi suuriakin tekijöitä, kuten National Museum of Ireland[1] tai Natural History Museum[2], niillä on kuitenkin usein teknisiä ongelmia, joista tarkemmin seuraavissa kappaleissa. Onneksi myös parempia esimerkkejä löytyy, kuten Smithsonian 3D Digitization[3], mutta tässäkin ratkaisussa on omat haasteensa.

Aloitamme teknisten ongelmien purkamisen Irlannin kansallismuseon toteutuksesta. Jo ennen yhdenkään saatavilla olevan näyttelyn aloittamista kävi selväksi, että toteutukset on tehty Matterport-kameroilla samaiselle alustalle, joita on käytetty menestyksekkäästi mm. arkeologisten kaivausten mallintamisessa (Shults ym. 2019). Alustan taustalla on 3D-pisteverkko tilasta ja verkko on teksturoitu kameran ottamilla valokuvilla. Lopuksi malliin on määritetty hotspotteja, joissa esineistä ja asioista saa lisätietoa. Lopputulos on näennäisesti hyvä ja replikoi täydellisesti tilan sillä hetkellä, kun se oli kuvattaessa, mutta tila on siis täysin staattinen ja aina samanlainen.

Samankaltaiseen lopputulokseen päästäisiin myös hyödyntämällä perus 360-kameraa ilman 3D-ominaisuutta. Ratkaisu on helppo, jos tila halutaan toistaa samanlaisena kuin se oli. Tässä toteutuksessa päivitettävyys tai uusien esineiden lisääminen on kuitenkin mahdotonta tai vaatii vähintäänkin uuden kuvauksen ja kohteiden yhdistämisen. Myös liikkuminen tilassa ennalta määrättyjen hotspottien avulla on ainakin Digitalian näkökulmasta turhan rajoittavaa. Viimeisenä ja ehkä tärkeimpänä ongelmana näissä Matterport / 360 -toteutuksissa on interaktio esineiden kanssa, jota ei siis ole lainkaan. Esineitä voi ainoastaan katsella tietyistä kameralla kuvatuista kulmista ja zoomata kuvaa lähemmäksi tiettyyn pisteeseen asti, mutta esineitä ei voi esimerkiksi vapaasti pyöritellä, zoomata tai muutella niiden valaistusolosuhteita.

---

[1] https://www.museum.ie/en-IE/Museums/Natural-History/Visitor-Information/3D-Virtual-Visit

[2] https://sketchfab.com/3d-models/hintze-hall-nhm-london-surface-model-b2f3e84112d04b-f1844e7ac2c4423566

[3] https://3d.si.edu/collections

Kun kohdistamme kriittisen silmämme kohti toista toteutusta, huomaamme heti, että olemme jälleen tekemisissä oikean 3D-mallin kanssa, joka ensi katsomalta näyttää hyvältä. Jos kuitenkin siirrämme museovierailijaa lähemmäksi kohdetta, karu totuus paljastuu. Tämä on esitetty kuvassa 1. Vertailun vuoksi kuvan oikeassa laidassa on valokuva kohteesta.

Kuva 1. Photogrammetrian ongelmat paljastuvat yksityiskohdista

Tämä kuvassa selkeästi havaittavissa oleva ”mössöytyminen” johtuu siitä, että museo on toteutettu photogrammetrian avulla, jossa suuri määrä eri kulmista otettuja valokuvia yhdistetään laskennallisesti 3D-malliksi. Menetelmä toimii hyvin, jos halutaan kokonaiskuva suuresta alueesta, mutta kuvia ei mitenkään ole mahdollista ottaa tai prosessoida niin montaa, että yksityiskohdista tulisi selkeitä. Huonon tarkkuuden lisäksi myös tämä virtuaalimuseototeutus kärsii samoista liikkumiseen, zoomaamiseen ja objektien pyörittelyyn liittyvistä ongelmista kuin edellinenkin toteutus.

Viimeisenä esimerkkinä tarkastellaan Smithsonianin 3D-skannauksia. Näistä yksittäisten objektien skannauksista ei suoraan käy ilmi, millä laitteistoilla objektit on skannattu ja millaisella teknisellä alustalla ne näytetään. Sivun syvempi tarkastelu kertoo, että kyseessä on Smithsonian-instituutin itse rakentama Voyager[4] -työkalu, joka on saatavilla Github-palvelussa. Sen lisäksi, että työkalua voi käyttää verkossa, sen voi myös asentaa itselleen ja näyttää siinä omia objektejaan. Voyagerissa mallia voidaan tarkastella ja zoomata vapaasti erilaisissa valaistuksissa. Lisäksi

Kuva 2. Neil Armstrongin avaruuspuku Voyager-ohjelmassa

4 https://smithsonian.github.io/dpo-voyager/

objekteja voidaan tarkastella mm. XRay- ja Wireframe-moodeissa. XRay-moodissa oleva Neil Armstrongin avaruuspuku on esitetty kuvassa 2. Pelkästään objektien tarkastelun kannalta tämä on parempi tekninen ratkaisu kuin DAM-hankkeessa toteutettu. Voyagerin parhaita paloja voisikin poimia myös DAM:n jatkokehityslistalle. Ongelmat Voyagerin toteutuksessa liittyvät kuitenkin siihen, että objektit ovat yksittäisiä esineitä ja siirtymät objektien välillä vievät aikaa. Varsinaista museokokemusta ei siis pääse syntymään. Epäselväksi jää myös se, onko objektien metatiedot kirjoitettu Voyager-järjestelmän avulla vai onko ne mahdollista ladata esinekohtaisesti jostakin ulkoisesta erillisestä palvelusta. Positiivista on kuitenkin se, että objektit on luokiteltu järkeviin kokonaisuuksiin web-käyttöliittymässä.

Toinen Voyageria vastaava toteutus on Sketchfab-palvelu, joka puolestaan loistaa Final Render -moodinsa kanssa. Malli osaa hyödyntää tekstuurilla mahdollisesti olevia karttoja (normal, bumb, metalness, roughness, specular ja displacement), joiden avulla pinnasta saadaan todenmukainen ja näyttävä. Tästäkin on toki huomautettava, että Sketchfabin Final Render -moodi ei ole kokonaan reaaliaikaisilla valaistuksilla toteutettu, vaan osittain ennalta laskettu malli, joka vain näytetään käyttäjälle. Tämän havaitsee esimerkiksi siitä, että objektin liikuttelu ja pyörittely eivät vaikuta objektin pinnalla näkyviin varjoihin mitenkään. Valon heijastuminen pinnoilta sen sijaan näyttäisi toimivan reaaliaikaisesti.

Yhteenvetona voidaankin todeta, että olemassa olevien virtuaalimuseoiden ongelmat liittyvät huonoon tarkkuuteen / yksityiskohtien toistoon, interaktion puutteeseen tai siihen, että kokemus on joka kerralla samanlainen. Toinen mahdollinen ongelma virtuaalimuseonäkökulmasta katsottuna on objektien esittäminen yksittäin, jolloin varsinainen museokokemus jää syntymättä.

DAM-HANKKEEN VIRTUAALIMUSEO

Kehittämämme virtuaalimuseo ratkaisee yllä mainittuja ongelmia yksinkertaisella mallilla. Koko museoympäristö luodaan dynaamisesti Unity-pelimoottorin ja C#-koodin avulla. Aluksi museon pohjapiirros luodaan annettujen määreiden perusteella satunnaisuutta hyödyntäen. Näin ollen todennäköisyys sille, että käyttäjä vierailisi kaksi kertaa täysin identtisessä museossa, on olematon. Itse 3D-museon rakentamisessa käytetään Unityn valmiita primitiivejä, kuten tasoja ja kuutioita, jotka teksturoidaan sattumanvaraisilla hyvin toteutetuilla tekstuureilla. Tällä tavalla valojen taittumiset ja heijastumiset pinnoilta saadaan myös varioitua ja toteutus on tehokas, koska primitiivien laskenta on nopeaa.

Valaistuksen laskennassa ja muodostamisessa hyödynnetään Unityn URP (Universal Render Pipeline) -moodia, jolla saavutetaan hyvän laadun lisäksi myös yhteensopivuus useimpien laiteympäristöjen kanssa. Myös museoesineiden sijoittelut lasketaan muodostetun pohjapiirroksen avulla, joten museo on objektien sijoittelujenkin suhteen aina erilainen. Liikkuminen museossa on toteutettu Unityn FPS (First Person Controller) -toiminnolla, joten

liikkuminen vastaa normaalia 3D-peleistä totuttua kaavaa. Hiirellä katsellaan ja hoidetaan osa interaktiosta ja näppäimistön wasd-/nuolinäppäimien avulla liikutaan maailmassa.

Lisäksi museoon on toteutettu muutamia pikanäppäinkomentoja helpottamaan käyttäjän toimia museoympäristössä. Kaikki toiminnot esitellään käyttäjälle selkeässä Info-näkymässä, johon käyttäjä voi siirtyä ollessaan museoiden valintaruudulla. Periaatepäätöksenä museosta ei ole tehty selainympäristössä toimivaa, koska se rajoittaisi jo tietoturvasyistäkin mm. objektien lataamista käyttäjien koneelta. Lisäksi objektien lataaminen verkkosijainneista voisi hitaiden verkkoyhteyksien takia tehdä käyttökokemuksesta todella huonon.

Yksi suurista ongelmista museosektorilla on osaamisen puute digitaalisten työkalujen suhteen. Tässä DAM-hankkeessa toteutettu virtuaalimuseo tekee myös uusien museoiden ja uusien objektien lisäämisestä todella helppoa. Ladattavan museopaketin mukana tulee config.txt -tekstitiedosto, jossa kaikki saatavilla olevat museot määritellään. Lisäämällä tähän tiedostoon museon nimen ja hakemistopolun, josta museoesineet löytyvät, kaikki muu hoituu automaattisesti. Jokaiseen annettuun museon hakemistopolkuun on myös mahdollista lisätä toinen tekstitiedosto, jossa on joko linkit esineiden Yksa-arkiston metatietoihin tai suoraan esineiden metatiedot määritetyssä rakenteessa.

DESIGN-ARKISTON PILOTTI

Kuten jo aiemmin todettiin, teimme myös pilotoinnin Designarkiston kanssa. Saimme digitoitavaksi kymmenen Designarkiston esinettä, joista onnistuimme digitoimaan kunnolla yhdeksän. Kymmenes esine, trukin aisa, oli skannerillemme mahdoton pala purtavaksi muotonsa ja materiaaliensa takia. Skannerin ohjelmisto skannaa isommista objekteista pienen osan kerrallaan ja pyrkii sitten laskennallisesti yhdistämään skannatut osat toisiinsa. Tämä trukin aisa on niin symmetrinen ja materiaaliltaan sellainen, että valo ei heijastu kunnolla, joten jokainen skannaus saa vain pienen osan pinnasta kaapattua. Näiden pienien osien yhdistäminen oli automatiikalle mahdotonta. Käsin yhdistämistäkin kokeiltiin, mutta tällöinkin aisasta saatiin kasaan alle puolet. Muiden esineiden kohdalla automatiikka ja kääntöpöytä toimivat todella hyvin ja tuloksista tuli jopa liian tarkkoja.

Skannauksen jälkeen käsissämme oli 30-800Mt:n kokoisia 3D-malleja, jotka sellaisenaan olivat liian raskaita pelimoottorille. Ennen pelimoottoriin tai ulkoiseen latauskansioon vientiä käsittelimmekin mallit Blender-ohjelman Decimate modifierin avulla. Tämä työkalu vähentää mallien vertexien/pintojen määrää poistamalla sellaisia osia, joilla on mahdollisimman vähän merkitystä pinnan muotoon. Käytimme modifierin arvona 0.1, joka pienentää mallin noin 1/10 alkuperäisestä koostaan. Lähelle objektia zoomatessa aiheutunut tarkkuuden lasku on havaittavissa, mutta esineestä saa kuitenkin erittäin hyvän kokonaiskuvan. Kuva 3 esittää joitakin ruudunkaappauksia pilottikohteestamme, jonka sisältävän virtuaalimuseon voit ladata artikkelin lopussa olevasta linkistä.

Kuva 3. Kuvakaappauksia DAM-hankkeen virtuaalimuseosta

Sen lisäksi, että DAM-hankkeen virtuaalimuseo osaa näyttää käyttäjälle 3D-malleja, se hyväksyy myös kuvatiedostoja esineiksi. Kuvatiedostojen tapauksessa automaattinen työnkulku rakentaa kuvista kaksipuoliset 3D-objektit, joita voidaan tarkastella täsmälleen samalla tavalla kuin ”oikeitakin” 3D-malleja. Tätä toiminnallisuutta on tarkoitus hankkeen aikana vielä pilotoida Mikkelissä sijaitsevan Jalkaväkimuseon kuvamateriaalien kanssa.

YHTEENVETO

Yhteenvetona DAM-hankkeessa toteutetusta virtuaalimuseosta voidaankin todeta, että se tekee museoiden rakentamisesta hyvin helppoa ilman teknistä tietotaitoa ja osaamista, sillä ainoat muutokset tai lisäykset on tehtävä tekstitiedostoihin ja tiedostokansioiden sisältöihin. Ennen DAM-hankkeen päättymistä on tarkoitus rakentaa museoon vielä esineiden metatietojen lataaminen Yksan lisäksi myös Finna.fi -palvelusta. Lisäksi tutkimme myös mahdollisuutta lisätä videotiedostoja museoon. Mahdollisissa jatkohankkeissa toimintoja on tarkoitus laajentaa lisätyn todellisuuden eli AR-maailman (Augmented Reality) puolelle. Kerromme mielellämme nykyisestä ratkaisusta lisää ja esittelemme toimintaa tarvittaessa kasvotusten tai erilaisten neuvottelukanavien kautta. Jutussa kuvatun virtuaalimuseon demopaketin voi ladata itselleen kokeiltavaksi ja ihmeteltäväksi täältä: http://digitalia.xamk.fi/demos/VirtualMuseum/virtualmuseum.zip

LÄHTEET

European Parliament 2014. Towards an integrated approach to cultural heritage for Europe, European Parliament resolution of 8 September 2015, towards an integrated approach to cultural heritage for Europe (2014/2149(INI)).

Schofield, G., Beale, N., Beale, G., Fell, M., Hadley, D. M., Hook, J., Murphy, D., Richards, J.D. & Thresh, L. 2018. Viking VR: Designing a Virtual Reality Experience for a Museum.

Shults, R., Levin, E., Habibi, R., Shenoy, S., Honcheruk, O., Hart, T. & An, Z. 2019. Capability of matterport 3d camera for industrial archaeology sites inventory. International archives of the photogrammetry, remote sensing and spatial information sciences. Vol. XLII-2-W11. 1059-1064.

SOSIAALISTA MEDIAA TALTEEN: TAPAUS TWITTER

Tuomo Räisänen, FT, ohjelmistosuunnittelija, Xamk
Miia Kosonen, KTT, tki-asiantuntija, Xamk

Digitaalinen avoin muisti -hankkeessa toteutettiin Twitter-tileiltä jaettujen tietojen keräys rajatulle määrälle tilejä automaattisesti siten, että olennainen tieto kerätään tietokantoihin ja mediat omiin kansioihin. Kerättäviä tilejä ja niihin liittyviä tietokantoja voidaan helposti lisätä ja poistaa. Twiittien sisältö medioineen voidaan palauttaa vaivattomasti, mikä voi mahdollistaa pitkäaikaistallennuksen ja tutkimuskäytön. Tässä artikkelissa kuvaamme tehdyn toteutuksen tarkemmin.

Digitaalisen viestinnän ja mahdollisuuksien mukaan myös sosiaalisen median arkistointi tulee ajankohtaiseksi ennemmin tai myöhemmin. Joissakin tapauksissa arkistoala on jo myöhässä. Tästä esimerkkinä ovat Yahoo! -sähköpostiryhmät, joissa käytiin vuosituhannen vaihteessa vilkasta keskustelua. Erilaisia ryhmiä, sekä avoimia että suljettuja, löytyi liki kaikilta yhteiskunnan aloilta. Kun ylläpitäjä ilmoitti, että kaikki ryhmät katoavat vuoden 2020 alussa, saimme Digitaliassa pyynnön Musiikkiarkistolta pelastaa muutamien ryhmien viestinvaihdot. Työ onnistui avointen ryhmien osalta hyvin.

Toivottavasti tämä esimerkki havainnollistaa hyvin sitä, ettei mikään säily internetissä ikuisesti. Viestintään ja tiedon jakamiseen käytettävien palveluiden tarjoajista huomattava osa on voittoa tavoittelevia globaaleja yritysjättejä. Jos datasta ei ole yrityksille enää mitään taloudellista hyötyä (öljy on loppunut), ei heillä ole intressiä sitä säilyttää. Tutkimuskäytössä aineistojen arvo puolestaan saattaa konkretisoitua vasta vuosia tai vuosikymmeniä myöhemmin. Tätä kuilua Digitalia voi osaltaan olla kuromassa umpeen.

TAVOITTEET

Tavoitteena oli selvittää ja toteuttaa sosiaalisen median sisältöjen tallennusta käyttäen esimerkkinä Suomen kuntien virallisia Twitter-tilejä. Käyttämällä automaattista twiittien keräämistä haluttiin luoda tekninen toteutus sekä tilien hallintaan että kerätyn datan mahdollisimman monipuoliseen jatkokäyttöön, mukaan lukien tutkimustarkoitukset. Ratkaisun tulee myös toimia avoimen lähdekoodin ympäristössä.

Näissä valinnoissa taustalla oli luonnollisesti se, että yksittäisiltä kunnilta yleensä puuttuvat sekä tekniset että taloudelliset resurssit oman sosiaalisen median sisällön talteenottoon. Jos tallennusta voidaan tehdä keskitetysti, automaattisesti ja käyttäjän näkökulmasta mahdollisimman yksinkertaisella ratkaisulla, tarkoittaa se pienempiä kustannuksia ja vähemmän ylläpitotyötä. Varsinaisen digitaalisen arkistoinnin toteutus jää kuitenkin kuntien itsensä ratkaistavaksi.

ONGELMAT

Ensimmäinen ongelma sosiaalisen median talteenotossa ja arkistoinnissa on juridinen. Verkossa julkaistua kulttuuriperintöaineistoa voi Suomessa tallentaa Kansalliskirjasto, siltä osin kuin tehtävä on laissa heille määrätty, kts. Laki kulttuuriaineistojen tallettamisesta ja säilyttämisestä (1433/2007, 9 §) sekä Kansalliskirjaston julkaisema Verkkoaineistojen keräyssuunnitelma (2021). Aineistoa kerätään muun muassa säännöllisillä sisällön keräyksillä eri lähteistä (kohteena keskeiset, laajasti näkyvillä olevat toimijat ja niiden avoimesti saatavilla olevat julkaisut) sekä teemakeräyksin, jotka liittyvät ajankohtaisiin ja yhteiskunnallisesti merkittäviin aiheisiin. Väistämättä syntyy myös katvealueita, jossa arvokasta aineistoa jää tallentamatta ja tuhoutuu sosiaalisen median palveluiden ja tilien lakatessa olemasta.

Viranhaltijoiden ja muiden yksittäisten henkilöiden Twitter-tilien sisältöä ei lähdetty tallentamaan Digitalian toimesta. Sosiaalisen median arkistoinnin kehittämisessä on huomioitava ihmisten jo tekemä tietoinen valinta näkyä ja toimia julkisissa kanavissa, mutta samalla myös GDPR:n mukainen oikeus tulla unohdetuksi. Yksityisyyttä kunnioittava linjaus onkin antaa ihmisille mahdollisuus jättäytyä halutessaan kokoelmien ulkopuolelle (Kirmo & Kosonen, 2017).

Rajasimme Digitaalinen avoin muisti -hankkeessa virallisiin organisaatiotileihin ja niiden talteenoton tekniseen toteuttamiseen. Tekninen haaste liittyy nimenomaan datan keräämiseen. Varsinaista arkistointia ei ole vielä tehty, sillä sen reunaehdot jäävät viranomaisten ja arkistotoimijoiden juridiikan asiantuntijoiden linjattaviksi. Digitalian lähestymistavan mukaisesti ensin tulee konkreettisesti osoittaa, mitä on mahdollista ja ylipäätään tarpeellista tehdä, jotta tarvittavat linjaukset saadaan aikaan. Viivyttelyyn ja vastuun siirtelyyn ei internetin syövereissä ole aikaa.

Lähtötilanteessa yksittäisellä kaupungilla on tyypillisesti useita Twitter-tilejä, joten näitä tulisi pystyä lisäämään tai poistamaan tallennukseen joustavasti. Tutkimuskäyttöä varten tietokantojen käyttö olisi suotavaa.

TOTEUTETTU RATKAISU

Twitter tarjoaa API:n eli rajapinnan, jolla pääsemme yksittäiseen twiittiin kiinni ohjelmallisesti. Käytimme tähän tarkoitukseen Python-ohjelmointikieltä. Loimme ilmaisen ns. Developer accountin, jolla yksittäisen tilin twiittejä voi kerätä 3200 kappaletta taaksepäin. Keväästä 2021 alkaen on mahdollista saada akateemisen tutkimuksen tarpeisiin rajoittamaton määrä twiittejä tietyin ehdoin ja tapauskohtaisen hakemuksen perusteella. Näin ei ollut vielä Digitaalinen avoin muisti -hankkeen alkuvaiheissa, eikä ole varmaa, soveltuisiko tämä lähestymistapa projektimme tavoitteisiin.

Rajoitusten vuoksi toteutus ei ole sinällään riittävä. Tilin omistaja voi kuitenkin ladata omat twiittinsä kokonaisuudessaan json-formaatissa, jolloin yksi mahdollinen ratkaisu on, että omistaja toimittaa twiittaushistoriansa Digitalialle, jolloin voidaan suorittaa migraatio tietokantaan. Tätä kokeilimme manuaalisesti yhden tilin osalta. Tekemämme sovellus on luonteeltaan enemmänkin Proof of Concept (PoC) kuin valmis tuotantokäyttöön soveltuva ohjelmisto. Periaate on kuitenkin selkeä: tilinomistaja luovuttaa twiittihistoriansa yhden kerran ja sen jälkeen tallennus tapahtuu automaattisesti. Toinen vaihtoehto on hankkia Developer account, jolloin koko historia saadaan kyllä kerättyä.

Twiittejä ei alun perinkään yritetty kerätä alkuperäisessä muodossaan. Kerättiin vain oleellinen data eli uniikki tunniste, aikaleima, twiitin sisältö, hashtagit, mediasisältö ja tärkeimmät metriikat (uudelleentwiittaukset, tykkäykset, maininnat). Lisäksi vastausketjut ja uudelleentwiittaukset rajattiin juridisista syistä jo lähtötilanteessa keräyksen ulkopuolelle. Kohteena on vain viranomaistililtä jaettu originaali sisältö, ei muiden käyttäjien luomia twiittejä. Jokainen Twitterin käyttäjäksi rekisteröitynyt huolehtii siitä, että tililtä jaetaan ainoastaan sellaista aineistoa, jota tilin haltijalla on oikeus käyttää.

Jo tällainen data on mielestämme riittävää sekä pitkäaikaistallennuksen että tutkimuskäytön tarpeisiin. Kerätyn datan avulla voidaan esimerkiksi tarkastella, milloin ja miten Covid-19 -pandemia alkoi näkyä kuntien virallisessa Twitter-viestinnässä eri alueilla, tai vertailla eri kaupunkien aktiivisuutta ja sävyä kansalaisille tiedottamisessa sosiaalisen median kautta.

Tilien hallinta on tehty joustavaksi. Kuntia ja niiden tilejä voidaan lisätä tallennukseen helposti. Mikäli kuntaa ei ole vielä listassa, tehdään sille oma tietokanta. Kunnalle voidaan lisätä rajoittamaton määrä tilejä, jotka lisätään tietokantaan. Koska tietokantaan on vain yksi tallentaja, tietokannaksi valikoitui Sqlite. Näin ollen sovellus käynnistyy lukemalla tekstitiedosto, jonka formaatti on seuraava.

```
City1, account11,account12,..
City2. account21,account22, ..
.
.
.
```

Tässä City1 on kaupungin nimi ja samalla myös tietokannan nimi. Rivin loppuosassa määritellään kyseiseen tietokantaan kerättävät tilit. Editoimalla tätä tekstitiedostoa voidaan tietokantoja ja tilejä lisätä vapaasti. Myös tiedoston editointiin on kehitteillä käyttöliittymä.

Twiittiin sisältyvä media on tallennettu kuntakohtaisesti omiin kansioihinsa. Näihin lisättiin twiitin ID-tunnus, joka on uniikki. Näin sekä twiitin sisältö että media voidaan helposti palauttaa. Yksittäisestä twiitistä kerätään seuraavat tiedot.

Taulukko 1. Kerätyt twiittikohtaiset tiedot

Muuttuja	Kuvaus
id	Twiitin tunnus
time	Viestin lähettämisen UTC-aika
message	Viestin sisältö
likes	Tykkäysten määrä
re	Uudelleentwiittausten (retweet) määrä
hasht	Hashtagit
mentions	Muut mainitut tilit
media	Kuvat, animaatiot

Koska olennainen data on nyt tietokannoissa, tietoihin voidaan kohdistaa hakuja, jotka voivat osoittautua arvokkaiksi tutkijoille. Metriikat luonnollisesti elävät muutaman päivän ajan twiitin lähettämisen jälkeen - useimmiten twiitteihin reagoidaan varsin tuoreeltaan. Metriikoilla tarkoitetaan tykkäysten ja retwiittien lukumäärää. Tätä varten toteutettiin ratkaisu siten, että twiittejä haetaan n päivää taaksepäin ja mikäli varsinainen twiitti on jo tietokannassa, päivitetään ainoastaan metriikat.

TULOKSET

Olemme yhdellä sovelluksella ja perus-PC:tä käyttäen keränneet kevääseen 2021 mennessä noin 250 000 twiittiä kaikkiaan yli sadalta kunnalta. Yksittäisten tilien lukumäärä on noin 150. Tekemämme sovellus pystyy teoriassa tallentamaan kymmeniätuhansia twiittejä päivittäin. Tämä raja ei kuitenkaan ole Suomen kuntien kohdalla vielä lähelläkään.

Twiittejä voi lukea erilaisilla viewer-ohjelmilla. Näissä on myös SQL-tuki, jolloin yksittäiseen tietokantaan voidaan tehdä esimerkiksi sanahakuja. Toki on mahdollista toteuttaa tiedonkeruuta ohjelmallisestikin siten, että käydään kaikki tietokannat läpi kerralla samalla haulla, tyyliin ”Laske ’korona’ -termin lukumäärä twiiteissä tietyllä aikavälillä”. Oma jatkokehitysaiheensa ovat yksinkertaiset analyysit ja visualisoinnit organisaation oman Twitter-aineiston sisällöstä.

LOPUKSI

Toteutukselle asettamamme tavoitteet täyttyivät. Kehitetty ratkaisu on yleiskäyttöinen, mahdollistaen erityyppisten organisaatioiden sisäisen tietojen koonnin helposti, ja tekniseltä kannalta myös arkistoinnin.

Twitter julkaisi omasta API:staan uuden version vuoden 2020 lopussa, joten se ei ehtinyt varsinaiseen toteutukseemme mukaan. Olisi kuitenkin luontevaa siirtyä käyttämään tätä rajapintaa. Tämä osaltaan muistuttaa myöskin siitä, että sosiaalinen media muuntuu tutkimus- ja kehittämistyön, tallennuksen ja sähköisen arkistoinnin näkökulmasta varsin vikkelää vauhtia.

Twitterin sisältämä tieto on kuitenkin jo tällä hetkellä käsiemme ulottuvilla tallennettavaksi jälkipolvia varten. Moniin muihin sosiaalisen median kanaviin verrattuna sen etuna on muuntumattomuus: sisältö ei jatkuvasti ”elä”, koska twiittien sisältöä ei voi muokata. Twiitin voi ainoastaan poistaa. Talteenoton tärkeydestä kertoo sekin, että merkittävien organisaatioiden ja henkilöiden tilit houkuttelevat hakkereita. Näin kävi muun muassa Applelle, Jeff Bezosille, Elon Muskille, Barack Obamalle ja Joe Bidenille ns. bitcoin-huijauksen yhteydessä 2020.

Tulevaisuutta ajatellen sosiaalisen median talteenotossa on luontevaa tehdä yhteistyötä Kansalliskirjaston verkkokeräyksen ja Kuntaliiton kanssa siten, että aineiston hallinta ja käytettävyys paranevat. Tutkimuskäyttöä voidaan edistää muun muassa digitaalisten ihmistieteiden parissa toimivan Rajapinta ry:n verkostojen kautta. Kehittämistyötä on hyvä jatkaa myös loppukäyttäjien toiveita kuunnellen. Otamme mielellään yhteisen pöydän ääreen niin tutkijat, julkisten organisaatioiden asiakirjahallinnosta vastaavat asiantuntijat kuin niiden viestintäammattilaisetkin – kannattaa siis nykäistä meitä hihasta. Digitalialle on tulevissakin hankkeissa luvassa mielenkiintoista kehitettävää digitaalisen viestinnän aineistojen parissa.

LÄHTEET

Kansalliskirjasto 2021. Verkkoaineistojen keräyssuunnitelma 2021-2024. https://www.kansalliskirjasto.fi/fi/verkkoaineistojen-kerayssuunnitelma-2021-2024 [viitattu 12.5.2021].

Kirmo, O. & Kosonen, M. 2017. Kuinka arkistoida sosiaalista mediaa? Faili 4/2017.

Laki kulttuuriaineistojen tallettamisesta ja säilyttämisestä, 1433/2007, 9 §.

XAMK
KEHITTÄÄ