Kaakkois-Suomen ammattikorkeakoulu

South-Eastern Finland University of Applied Sciences

PLEASE NOTE! THIS IS PARALLEL PUBLISHED VERSION / SELF-ARCHIVED VERSION OF THE OF THE ORIGINAL ARTICLE

This is an electronic reprint of the original article.
This version may differ from the original in pagination and typographic detail.

Author(s): Jääskeläinen, Anssi

Title: Archiving 2018

Version: final draft

Please cite the original version:

Jääskeläinen, A. (2018). Archiving 2018. Faili, 2, 25-28.

HUOM! TÄMÄ ON RINNAKKAISTALLENNE

Rinnakkaistallennettu versio voi erota alkuperäisestä julkaistusta sivunumeroiltaan ja ilmeeltään.

Tekijä(t): Jääskeläinen, Anssi
Otsikko: Archiving 2018
Versio: final draft

Käytä viittauksessa alkuperäistä lähdettä:
Jääskeläinen, A. (2018). Archiving 2018. Faili, 2, 25-28.

Archiving 2018

Anssi Jääskeläinen
TKI-asiantuntija
Digitalia/Xamk

Maanantaina 16.4, klo 6.40 alkoi tämänkertainen reissu. Suuntana oli Washington DC ja kansallisarkiston (NARA) tiloissa järjestettävä Archiving 2018 konferenssi. Pidettävänä oli puheenvuoro avoimen lähdekoodin ohjelmistoilla toteutetuista automatisoidusta OCR luvusta ja sisältöanalyysistä. Molempiin on kehitetty ratkaisua Xamkin Digitalia hankkeessa. Esitysmateriaalini on kokonaisuudessaan kenen tahansa nähtävillä osoitteessa bit.do/jaaskelainen-archiving-2018. Puheenvuoron videointi tulee aikanaan myös IS&T:n Youtube-kanavalle.

Ohjelman mukaan luvassa olisi puheenvuoroja niin erilaisten ääni- ja videomuotojen digitoinnista, taiteen digitoinnista sekä 3D digitoinnista. Pikaisella ennakkokatsauksella reilusti yli puolet puheenvuoroista tulevat keskittymään digitointiin. Varsinaiseen arkistointiin / sähköiseen säilyttämiseen liittyviä puheenvuoroja oli konferenssin nimestä huolimatta valitettavan vähän.

22 tunnin matkustamisen jälkeen uni maistui hyvin. Seuraavan päivän allekirjoittanut saikin käyttää aikaerosta toipumiseen sekä nähtävyyksien kiertelyyn. Jätin suosiolla perinteiset nähtävyydet katsomatta ja suuntasin kävelyretkelle niin monessa elokuvassa näkemälleni Arlingtonin hautausmaalle. Vaikuttava näky kumpuilevan maastonsa, kasvillisuutensa ja kilometrien päähän ulottuvien valkoisten hautakivienriviensä johdosta.

Keskiviikkona alkoi varsinainen konferenssi. Suzanne Grinnan IS&T:stä kiitteli järjestäjiä ja sponsoreita. Seuraavaksi general chair Don Williams avasi virallisesti konferenssin. Kaikkiaan läsnä oli reilut sata osallistujaa.

Festareista arkistoon

Päivän ensimmäinen keynote-puheenvuoro käsitteli Montreaux Jazz festifal tapahtuman livetaltiointien tallentamista ja metatiedottamista. Jo 60 luvulla näiden tapahtumien taltiointeja haluttiin tehdä sekä videona että äänenä mahdollisimman korkeilla resoluutioilla. Alunperin ko. festarit syntyivät tarpeesta saada pienelle paikkakunnalle uutta elinvoimaa. Lisäpontta festivaali sai kun paikallinen kasino paloi ja Deep Purple teki aiheesta yhden suosituimmista lauluistaan “Smoke on the water”. Arkistoinnin suhteen Montreaux Jazz oli edelläkävijöitä: Suurin osa materiaalista on arkistoituna kolmella eri medialla/formaatilla ja HD videointi alkoi jo vuonna 1991. Arkistoissa on tällä hetkellä noin 14 000 pakkaamatonta masternauhaa kahdeksanatoista eri audio/video formaattina. Aikana puhuttaessa videotallennetta on noin 11 000 tuntia ja ääntä noin 6 000 tuntia.

Lisäksi valokuva-arkistossa käsittää noin 100 000 kuvaa, joista reilu puolet on digitoitu. Kuulemamme mukaan, myös analoginen arkisto on tällä hetkellä vielä hyvässä kunnossa ja esim. 45v vanhat tallenteetkin ovat helposti luettavissa.

Vuonna 2007 alkoi kuitenkin digitointiprojekti, jotta materiaalit säilyvät siinäkin tapauksessa, että analoginen arkisto tuhoutuu. Vuonna 2010 projektille saatiin ensimmäinen yksityinen rahoittaja, jonka turvin perustettiin osaamiskeskus EPFL:ään (Lausanne Polytechnique Ecole). Digitoitujen materiaalin tekijänoikeudet on hoidettu niin, että tullessaan festareille soittamaan, artistit allekirjoittavat sopimuksen, joka käsittää materiaalin esittämisen TV esityksinä, optisilla medioilla sekä Montreaux Jazz kahviloissa. Se ei kuitenkaan anna oikeutta julkiseen näyttämiseen, joten netin kautta pääsyä materiaaliin ei ole. Digitointiprojektissa EPFL hoiti kaiken muun toiminnan paitsi varsinaista digitointia, koska siihen tarvittiin medioiden ja formaattien monimuotoisuuden takia digitoinnin ammattiosaajia.

Varsinainen tallennus on hoidettu sekä LTO4 että LTO6 nauhoille. Broadcast ympäristönä toimii levytallennusjärjestelmä. Tällä hetkellä käytössä on kolme erillistä 4.7 petatavun järjestelmää, joista yksi sijaitsee eri kaupungissa. Materiaalin indeksointiin on käytetty asiasta kiinnostuneita EPFL:n opiskelijoita. Tallennetun materiaalin pohjalta on tehty monia innovatiivisia sovelluksia, mm. Sound Dots by Hidacs,

National Archives, Washington DC. Kuva: Anssi Jääskeläinen.

Remix the archive, Open Mic, 360 video, jne. joiden avulla menneitä konsertteja voi kokea uudelleen mm. Montreax Jazz kahviloissa. Tulevaisuuden suunnitelmissa on pyrkiä tunnistamaan artisti, tunteet yms. asioita soitosta automaattisesti. Toisena aikeena on pyrkiä avaamaan arkistot digitaalisen humanismin tutkijoille. Futuristisena kokeiluna ovat Washingtonin yliopiston kanssa tallentaneet konsertteja myös synteettiseen DNA:han, on kuulemma vielä turha kallista massatuotantoon.

Uusia tekniikoita arkistointiin

Ensimmäisessä sessiossa keynote puheenvuoron jäleen Benoit Sequin edellämainitusta EPFL:stä kertoi uusista tekniikoista taiteen digitoimiseksi. Heidän arkistonsa on pääpiirteittäin hyvin järjestettyä, mutta luonnollisesti löytyy myös epämääräistäistä "arkistoitua materiaalia". Analogisessa arkistossa on yli miljoona valokuvaa, joista noin 350 000 on standardinmukaisesti pahvilla metatietoineen. Digitointia varten he olivat kehittäneet pyörivän levyillä varustetun skannerin, jonka avulla he digitoivat noin 1500 valokuvaa päivässä. Skannista luettiin semanttisen segmentoinnin ja syväoppimisalgoritmien avulla varsinainen kuva ja metatiedot. Opettaminen 120:llä mallikappaleella oli kuulemmamme mukaan riittävä hyvän tarkkuuden aikaansaamiseksi. Syväoppimisalgoritmin ja rakenneanalyysin käyttöön he olivat päätyneet törmätessään ongelmiin mm. käpristyneiden pahvitaustojen ja muuttuvien metatietojen sijoittelun kanssa.

Seuraavat kolme puheenvuoroa käsittelivät myös digitointia. Ensimmäisenä kerrottiin siitä, kuinka photometristä stereokuvausta käyttäen teksturoituja kolmiulotteisia objekteja voidaan digitoida realistisemmin. Hyödyt nähtiin erityisesti siinä tapauksessa, että varsinainen objekti on vaarassa tuhoutua ja siitä halutaan säilyttää mahdollisimman paljon tietoa tuleville tutkijoille. Sinänsä mielenkiinoista tutkimusta, mutta koskettanee vain hyvin pientä osaa tutkijoista. Lopuksi Arnold Chevea totesi, että menetelmä toimii hyvin kohtalaisen tasaisille pinnoille, joilla on suhteellisen tasainen heijastavuus, mutta monimutkaisemmat objektit ja objekti, jotka heijastavat valoa epätasaisesti vaativat myös muita digitointimenetelmiä tuekseen.

Toisessa puheenvuorossa käsiteltiin 35mm diojen digitointia ja sitä miten käsinkirjoitetut "metatiedot" saadaan talteen. Ongelmia oli tullut mm. pölyhiukkasten ja erilaisten valaistusmenetelmien kanssa. Myös kuvan orientaation kanssa oli jouduttu näkemään hieman vaivaa.

Kolmas puheenvuoro käsitteli pistekirjoitusmusiikin digitointia (OBR) Optical Braille Recognition työkalujen avulla. Allekirjoittanut ei edes

tiennyt, että tällaista musiikkia on olemassa ja jotenkin pisti miettimään että kyseessä on niin marginaaliryhmä joten kannattaisiko tutkimusta edes tehdä. Vielä kun huomioidaan se fakta, että pistekirjoitus vie tilaa paljon enemmän kuin normaali kirjoitus ja jo yhdenkin pisteen muuttaminen suuresta joukosta muuttaa koko äänen kyseenalaistin mielessäni ko. tutkimuksen vieläkin enemmän. Lopuksi, kun OBR luettujen materiaalien oikolukemiseen tarvittiin vielä pistekirjoitusmusiikin asiantuntija, oli johtopäätökseni lopullinen. Marginaaliryhmän kannalta kiinnostava tutkimus, arvo suurelle yleisölle hyvin vähäinen. Ongelmiakin koko OBR prosessissa oli ollut, suurimpana ehkä se, että ohjelmiston aikanaan kehittänyt firma ei ole enää toiminnassa, joten ei ole ketään, joka osaisi tehdä edes pienimuotoisia virheiden korjauksia. Tässä yksi hyvä syy pyrkiä käyttämään avoimen lähdekoodin ohjelmistoja.

Iltapäivän keynote puheenvuoro oli jälleen kiinnostava myös arkistointinäkökulmasta. Alonso, C Addison kertoi 3D tekniikan kehittymisestä viimeisen 30-vuoden aikana. Hän sivusi puheessaan niin laserkeilausta, paikannusta kuin mallinnustakin. Hänen puheensa löyhä otsikko liittyi askelmerkkeihin katoavan kulttuuriperinnön arkistoinnissa. Uusi teknologia sinänsä mahdollistaisi moniakin erilaisia lähestymistapoja niin digitointiin kuin 3D mallintamiseenkin. Varsinaisiin ratkaisuihin asti tässä puheenvuorossa ei kuitenkaan päästy, mutta hän herätteli yleisöä muutamilla mielenkiintoisilla kysymyksillä kuten: Jos kulttuurihistoriallisesti tärkeästä kohteesta tehdään 3D malli, pitäisikö se tehdä alkuperäisestä (tai oletuksesta millainen alkuperäinen on ollut), vaiko moneen kertaan vuosien saatossa korjatusta nykyisestä versiosta, vaiko eri paikkaan tehdystä kopiosta, vaiko kenties pelkästään vanhojen valokuvien ja piirroksien pohjalta? Esimerkkinä hän esitti Pyhän Markuksen kellotornin, jonka alkuperäinen kappale tuhoutui 1900-luvun alkupuolella. Olisiko tässä tapauksessa oikein mallintaa samalle paikalle rakennettu “kopio”, vaiko kenties joku monista muualle rakennetuista kopioista. Jatkoksi edelliseen hän kysyi; miten 3D mallista voidaan luotettavasti sanoa, onko se tehty aidosta esineestä vai ei? Onko sillä lopultakaan merkitystä, muuten kuin syvälle menevän historian tutkinnan kannalta. Seuraavaksi hän tiedusteli sitä, kenen pitäisi vastata uusien mallintamismenetelmien tuottamista petatavujen tietomassoista, pistepilvistä, yms. Samaan yhteyteen liittyen hän totesi, että usein ei edes tiedetä, jos jotain on jo digitoitu/mallinnettu, koska tietoa ei jaeta tai sen on tuottanut kaupallinen toimija. Esim. Kotikaupungistani Mikkelistä on tehty useita 3D malleja, mutta suurin osa kaupallisten toimijoiden toteuttamina. Näin ollen uudelleenkäyttö ilman korvausta on mahdotonta.

Asiakaslähtöistä digitointia

Suurta arkistokansaa kiinnostava puheenvuoro oli myös Amsterdamin kaupungin arkistosta tulleiden Marc Holtmanin ja Nelleke van Zeelandin puheenvuoro siitä, kuinka arkistojen digitoinnissa on päästy huomattavan suuriin nopeuksiin hyödyntäen teollisuudesta tuttuja prinsiippejä. Kaikki toiminta on asiakaslähtöistä. Mikä tahansa asiakkaan pyytämä digitoitava asiakirja, kartta yms. nousee välittömästi digitointivuon ensimmäiseksi ja on sen jälkeen kaikkien saatavilla online arkistossa. Lähtötilanteessa kesti noin 3vk pyynnöstä, että asiakirja oli saatavilla, tällä hetkellä pyynnöstä online tilaan kestää noin viikon. Tavoitteena on 48 tuntia. Vaadittu digitointivauhti on noin 20 000 skannia päivässä. Tähän lukemaan on päästy mahdollisimman vakaalla ja auditoidulla digitointiprosessilla. Koska on mahdotonta tarkistaa 20 000 skannia päivässä ihmisvoimin, koko digitointiprosessin auditointiin kiinnitetään erityistä huomiota. Myös prosessin toteutumista valvotaan mahdollisimman automatisoiduilla tarkistuksilla sekä pistokoemaisesti tehtävillä tarkistuksella. Tulokset ovat puhujien mukaan hyvin lupaavia, mutta prosessipohjainen lähestymistapa vaatii arkistoväeltä ajatusmallin muutosta. Meidän on vain luotettava siihen, että auditoitu digitointiprosessi ja otannat takaavat koko massan laadukkaan digitoinnin. Ja entäpä vaikka virhe sattuisikin. Ei ole mielestäni maailmaa kaatava asia, jos arkistossa on muutama sumea/suttuinen kappale koska se voidaan kuitenkin havaittaessa korvata uudelleen tehtävällä skannauksella.

Yksi suuri uutinen, joka Michael Horsleyn puheenvuorossa kävi ilmi, on että 31.12.2022 jälkeen NARA ei enää ota vastaan asiakirjoja analogisessa formaatissa. Käytännössä tämä tarkoittaa sitä, että kaikki tallentavat tahot ovat velvoitettuja käyttämään NARA:n hyväksymiä ja suosittelemia aineistoformaatteja ja että digitoinnin on täytettävä ISO 15489-1:2016 standardin vaatimukset. Lyhykäisesti tämän voi ilmoittaa siten että digitoinnin on täytettävä FADGI (Federal Agencies Digital Guidelines Initiative) 2-star vaatimukset, jonka tarkemmat vaateet voi tarkistaa sivulta 46 vuonna 2016 hyväksytystä dokumentista .

Torstaina aamun avasi Dorothy Williams, joka puhui digitointimahdollisuuden tuonnista yhteisöjen luokse. Jokin heidän käyttäjänsä oli kuulemma sanonut, että vastaavan digitoin-

tityön hinta kaupallisella toimijalla olisi ollut $5,5/kuva ja olisi yhteensä maksanut noin $16 000. Veikkaisinpa vahvasti, että tällaista toimintaa tekemällä joutuisi Suomessa hyvin pian markkinaoikeuteen.

Digitalian ratkaisu

Keynoten jälkeen olikin kahvitauko jonka jälkeen allekirjoittanut pääsi kertomaan yleisölle Digitaliassa toteutetuista ratkaisuista. Ensimmäisenä selityksen alle pääsi Tesseract 4 ohjelman ympärille rakennettu automatisoitu OCR toiminnallisuus. Se lukee minkä tahansa kuva- tai PDF tiedoston ja OCR:ää sen jollei siitä löydy valmiiksi OCR tietoa. Valmis tuotos tallennetaan OCR tiedon sisältäväksi PDF tai PDF/A tiedostoksi. Luonnollisesti on myös mahdollista tallentaa OCR tieto standardin mukaiseksi XML tiedostoksi. Sanoisin että yksinkertaista ja kohtalaisen tehokasta; 91-sivun diplomityö 1960-luvulta kääntyy OCR tiedon sisältäväksi PDF tiedostoksi noin neljässä minuutissa. Toisena puheenaiheenani oli automaattinen sisällön analysointi ja metatiedottaminen. Ratkaisu lukee OCR tiedon sisältävän PDF tiedoston, analysoi sen sisällön useaa eri työkalua hyödyntäen ja lopuksi generoi ihmisluettavan raportin löydöksistään. Tunnistetut avainsanat, metatiedot, yms. Tärkeät sisältöä kuvaavat sanat myös kirjotetaan PDF tiedostoon, jolloin ne ovat arkistoon viedessä suoraan käytettävissä. Jälleen helppoa ja tehokasta, yhden A4 sivun prosessointi kestää noin 5-15 sekuntia sanamäärästä riippuen. Tätä ratkaisua voi testata Digitalian julkisella demo-sivustolla osoitteessa https://digitalia.xamk.fi/content-analyser

Samaisen session kahdessa muussa puheenvuorossa käsiteltiin korean kielisen puheen automaattista tekstitystä. Tulokset tästä jäivät epäselviksi. Toisessa puheessa käsiteltiin IBRelight ohjelmistoa, jolla voidaan luoda realistisia 3D objekteja pelkästään valokuvien perusteella. Menetelmä taklaa perinteisten RTI menetelmien ongelmia, mutta toisaalta vaatii tällä hetkellä toimiakseen Agisoftin PhotoScan ohjelman tuottaman "pohjamallin". Lisäksi tarvitaan vähintään sata resoluutioltaan maksimissaan 2048x2048 kuvaa kohteesta jotka ovat kaikki saman kokoisia sekä kohtalaisen järeä työasema kuvien pyörittämiseen ja katseluun.

Torstain puheenvuoroissa käsiteltiin lisäksi mm. sitä kuinka hienoa olisi, jos digikamerat ja skannerit noudattaisivat metamorfoze / fadgi määreitä, jotka ovat yhdistyneet ISO 19262:2015 standardiksi. Kaikkia digitointityötä aloittavia ja jo tekeviä tahoja saattaisi myös kiinnostaa osoitteesta https://intranda.github.io/dos-and-donts/ löytyvät opas, jossa kerrotaan mitä pitäisi tehdä ja mitä ei pitäisi tehdä digitointiprojektien aikana. Pääkohtia ovat esim. automatisoi niin paljon kuin voit ja jaa työnkulku mahdollisimman pieniin osiin, jotta virheiden jäljittäminen on helpompaa. Osoitteesta löytyvää käsikirjaa voi kuka tahansa päivittää normaaleilla git-menetelmillä. Myös visuaalisesta digitointityön tarkastelusta ja sen ongelmista puhuttiin. Kenen silmiin tarkastuksissa pitäisi luottaa, entä ympäristötekijät kuten hajavalo? Voiko automatisoituihin tarkastuksiin luottaa, jne. Vastauksia näihin kysymyksiin ei saatu, mutta automaattisesta teknisestä tarkastuksesta täydet pisteet saava skanni saattaa olla sisällöltään täysin originaalista poikkeava.

Perjantain keynote puheenvuorossa Sam Brylawski puhui äänestä ja sen arkistoinnista. Kuka mm. vastaa mastereista stream aikakaudella? Onko sellaista asiaa kuin master välttämättä edes olemassa, jos jokin konsertti tai tapahtuma ainoastaan live streemataan? Hän kysyi myös oikeutetun kysymyksen, miksi kaikki pitäisi säilyttää vain säilyttämisen takia? Myös allekirjoittaneen mielestä säilyttämisellä tulisi olla muutakin tarkoitusta kuin "kulttuurihistorian" säilyttäminen. Hyvä esimerkki tällaisesta käyttötarkoituksesta on aiemmin kuvattu Montreaux Jazz arkiston monimuotoinen uudelleenkäyttö. Tämä on asia joita myös monissa suomalaisissa arkistoissa kannattaa miettiä. Brylawski myös totesi, että julkisia tuloksia tuottavien digitointiprojektien on huomattavasti helpompi saada rahoitusta kuin sellaisten joissa digitoidaan vain säilyttämisen vuoksi. Usein tietysti kansalliset tekijänoikeuslait ovat esteenä ja koko digitointiprojekti jää siksi ilman rahoitusta. Hyvänä esimerkkinä yhteisövetoisesta digitointiprojektista hän mainitsi "the Great 78 projektin", https://great78.archive.org/, jonka tarkoituksena on digitoida ja saattaa käyttöön kyseisen vuoden vinyylilevyjä. Kyllä, tämä projekti polkee rankasti tekijänoikeuksia, mutta toistaiseksi yksikään taho ei ole projektia estänyt. Eli asioita voi tehdä myös käyttäjälähtöisesti takertumatta turhaan byrokratiaan, jos on riittävästi uskallusta ja tekijöitä kuten tällä projektilla selkeästi on. Kannatan ehdottomasti tällaista lähestymistapaa esim. 15-20 vuotta vanhemmalle materiaalille, jolla ei ole enää juurikaan kaupallista arvoa.

Loppupäivän puheenvuorot oli jätettävä väliin lennolle ehtimisen vuoksi. Seuraava Archiving konferenssi järjestetään mahdollisesti Lissabonissa, toukokuun 2019 alkupuolella, mutta tästä ei ole vielä varmuutta.