Kaakkois-Suomen ammattikorkeakoulu

South-Eastern Finland University of Applied Sciences

PLEASE NOTE! THIS IS A PARALLEL PUBLISHED VERSION / SELF-ARCHIVED VERSION 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 & Räisänen, Tuomo

Title: Digitalia kulttuurihistoriaa pelastamassa

Version: Publisher´s PDF

Please cite the original version:

Jääskeläinen, A. & Räisänen, T. (2020). Digitalia kulttuurihistoriaa pelastamassa. Faili 1, 11–12.

HUOM! TÄMÄ ON RINNAKKAISTALLENNE

Rinnakkaistallennettu versio voi erota alkuperäisestä julkaistusta sivunumeroiltaan ja ilmeeltään.

Tekijä(t): Jääskeläinen, Anssi & Räisänen, Tuomo

Otsikko: Digitalia kulttuurihistoriaa pelastamassa

Versio: Publisher´s PDF

Käytä viittauksessa alkuperäistä lähdettä:

Jääskeläinen, A. & Räisänen, T. (2020). Digitalia kulttuurihistoriaa pelastamassa. Faili 1, 11–12.

TEKNIIKKA

Digitalia kulttuurihistoriaa pelastamassa

Anssi Jääskeläinen
Tutkimuspäällikkö
Xamk
Digitalia

Tuomo Räisänen
IT-asiantuntija
Xamk
Digitalia

Kun hätä on suuri, apu on Digitaliassa. Tämä lausahdus kuvaa hyvin tässä jutussa kuvaamaamme ex tempore projektia, jossa pelastimme osan "kulttuurihistoriallisesti" arvokasta Yahoon musiikkiaiheisten keskustelupalstojen materiaalia.

Kaikki alkoi Yahoon ilmoituksesta sulkea "Yahoo Groups" keskustelupalstansa 14.12.2019. Mielenkiintoiseksi asian teki se, että ilmoituksesta sulkemiseen oli aikaa vain muutama kuukausi. Keskustelupalstoja oli mahdollista perustaa vuodesta 2001 alkaen ja viime vuosina ihmiset ovat siirtyneet käyttämään toisenlaisia "ilmaisia" palveluja, kuten Facebook, Twitter jne. Niinpä Yahoo ei pystynyt myymään mainoksia tai käyttäjien dataa ansaintatarkoituksessa. Koska jäljelle jäi kuitenkin ylläpitokustannuksia, oli päätös siltä osin odotettu.

#Digitalian henkilöstöön kuuluva Tuomo Räisänen ajautui sattumalta keskusteluun koskien Digitalian DAM-hanketta, jossa siis yhtenä työpakettina on sosiaalisen median tallennus. Viidakkorumpu oli kiirinyt sen verran, että muutaman kontaktin kautta tuli meille kysely, onnistuisiko tietojen kerääminen joistakin ryhmistä. Näin ollen työpaketista tuli heti ajankohtainen siten, että deadline oli todellinen.

Tutkailemalla asiaa havaittiin, että tarjolla on muutamia avoimeen lähdekoodiin perustuvia työkaluja, joiden avulla ohjelmointitaitoinen pystyy lataamaan ryhmien keskusteluja. Näistä Yahoo Group Archiver (https://github.com/IgnoredAmbience/yahoo-group-archiver) osoittautui suhteellisen helpoksi ja hyvin toimivaksi tavaksi aloittaa tietojen kerääminen.

Ryhmiä Yahoossa oli kahden tyyppisiä, avoimia ja suljettuja. Jotta pystyisit lataamaan ryhmien tiedot, sinun tuli olla ryhmän jäsen. Suljettuihin pääsi vain laittamalla hakemuksen sisään, jonka sitten pääkäyttäjä joko hyväksyi tai hylkäsi. Ajan puutteen vuoksi suljettuihin ryhmiin ei kuitenkaan päästy pyynnöstä huolimatta.

Avoimiin ryhmiin liittymällä tuli mahdolliseksi viestiketjujen ohjelmallinen lataaminen. Meille tullut pyyntö sisälsi kahden avoimen, musiikkiin liittyvien keskustelupalstojen (ooooo ja phinnweb) tallennuksen. Näissä oli yhteensä yli tuhat jäsentä ja kymmeniä tuhansia viestejä, jotka saatiin kaikki tallennettua. Aikaa viestien lataamisessa kului muutamia tunteja.

Toinen vaihtoehto olisi ollut käyttää Yahoon omaa viestiketjun tallennuspalvelua, mutta pyyntöjen käsittelyajaksi arvioitiin reilua kahta viikkoa, joten käytimme edellä kuvattua työkalua ajan vähyyden vuoksi.

Kun data oli saatu otettua talteen, alkoi tietorakenteiden perkaaminen. Kansioiden nimeämisessä Yahoo oli onneksi ollut yhtenäinen ja kaikkien käsiteltävien viestiryhmien alta löytyi toisiaan vastaava kansiorakenne. Ratkaisu, jonka toteutimme, toimii siis kaikkiin olemassa oleviin viestiryhmiin, jotka on ladattu yllä mainitulla Yahoo group archiver työkalulla.

Emme ole saaneet testattavaksemme yhtäkään Yahoon omalla palvelulla tuotettua pakettia, joten toimivuus tämän kanssa on kysymysmerkki, johon todennäköisin vastaus on "ei toimi ilman lisäkoodausta" tai ainakaan ilmaiseksi. Jos jollakulla lukijalla tällainen Yahoon omalla palvelulla tuotettu export paketti löytyy, kokeilemme mielellämme ratkaisumme toimintaa myös sen kanssa.

Tietorakenteista merkittävimmäksi paljastui emails kansion sisältö, joka pitää sisällään sekä viestit kahdessa eri muodossa että viestien mahdolliset liitetiedostot omissa kansioissaan. Molemmat viestiä esittävät tiedostot ovat JSON (JavaScript Object Notation), joka on hyvin samankaltaista kuin XML. Rakenteen läpikäyminen oli siis hyvin yksinkertaista.

Rakenteellisten ratkaisujen selvittyä oli aika kääriä hihat ja kaivaa Python naftaliinista. Digitalian toteuttama Python koodi lukee viestit yksi kerrallaan, generoi niistä suhteellisen paljon alkuperäistä viestiä muistuttavan HTML-esitysmuodon ja tallentaa sen PDF/A-3b -muotoon. Lopuksi vielä JSON-tiedostosta löytyvät metatiedot tallennetaan luodun PDF-tiedoston sisään.

Lopputuloksena jokaisesta JSON-muotoisesta viestistä ja mahdollisista liitetiedostoista muodostuu siis täysin standardin mukainen arkistokelpoinen PDF/A-3b-tiedosto, jossa näkyy viestin ID, aihe, lähettäjä, aika sekä itse viesti. Saman sisältöinen HTML-tiedosto ja toki alkuperäinen JSON-tiedostokin säästetään mahdollista jatkokäsittely varten.

Lopuksi alkuperäiset ja luodut tiedostot vielä jaotellaan sopivan kokoisiin kansioihin, koska havaitsimme, että esim. 150 000 (alkuperäinen + HTML + PDF) tiedostoa yhden kansion sisällä hidastaa merkittävästi jopa Linuxin tiedostojärjestelmän toimintaa.

Lopuksi koko muodostettu kansiorakenne kaikkine sisältöineen vielä paketoidaan siistiin zip-pakettiin siirrettävyyden helpottamiseksi. Luonnollisesti, kaikki edellä kuvattu tapahtuu täysin automaattisesti – kuten digitalialla on mottona "Jos jokin pitää tehdä useammin kuin kerran se pitää automatisoida".

Tästä pikaprojektista oppina muillekin sähköisen aineiston kanssa toimiville olkoon seuraava. Arkistointia tulisi miettiä etukäteen, eikä vasta sitten, kun on liian myöhäistä kuten tässä tapauksessa. Ihmiset siirtyvät sutjakkaasti somesta tai palvelusta toiseen ja harvemman kaupallisen toimijan intresseihin kuuluu vanhan (heille tuottamattoman) tiedon tallentelu.