Digitoitujen lehtien vallankumous

(julkaistu Turun Sanomissa 25.11.2020)

Digitoitujen sanoma- ja aikakauslehtien hyödyntämisestä on tullut Suomessa arkipäivää. Kansalliskirjasto avasi jo vuonna 2001 portaalin digitoiduille historiallisille lehdille. Silloin esillä oli 36 000 sivua. Tällä hetkellä kuka tahansa voi lukea Suomessa julkaistuja lehtiä, ensimmäisestä sanomalehdestä *Tidningar utgifne af et Sällskap i Åbo* (1771) vuoteen 1929 asti. Aineistoa on huikeat 7,4 miljoonaa sivua. Kattavuudessaan kokoelma on ainutlaatuinen, sillä tavoitteena on ollut digitoida kaikki painetut lehtien numerot. Kansalliskirjasto on digitoinut myös vuoden 1929 jälkeistä aineistoa, mutta se ei tekijänoikeussyistä ole avoimesti luettavissa.

Digitoidut kulttuuriperintöaineistot ovat muuttaneet sitä tapaa, jolla voimme ammentaa tietoa ja mielikuvia menneisyydestä. Historiallisia lehtiä voivat käyttää taustatutkimusta tekevät kirjailijat, biografisia tietoja etsivät sukututkijat, ympäristön muutoksista kiinnostuneet luonnontutkijat, kotikaupunkinsa menneisyydestä innostuneet kaupunkilaiset – ketkä tahansa, joilla on tarve päästä alkuperäisaineistojen äärelle.

Kirjailija Edward Bulwer-Lytton kuvasi 1800-luvulla sanomalehdistöä kulttuurin kronikaksi, ”yhteiseksi altaaksi, johon jokainen virta laskee vetensä ja jonka partaalle kuka tahansa voi asettua ammentamaan juotavaa”. Lehdistö oli oman aikansa big dataa, jossa kohtasivat uutiset ja kannanotot, mainokset ja kaskut, torihinnastot ja säätiedot.

Lehdistön digitoinnin taustana on pyrkimys säilyttää hauraat sivut tuleville polville. Vielä 1990-luvulla ajateltiin, että mikrofilmi on varmin tallennusmenetelmä, kunnes tietokoneistuminen muutti kaiken: digitaaliset tallenteet tulivat jäädäkseen. Jos alussa ajateltiin säilyttämistä, vuosituhannen vaihteessa aineiston avoimuus ja erilaiset hakumenetelmät nousivat keskiöön.

Valtaosaltaan digitoidut lehdet perustuvat alkuperäisten painettujen sivujen sijasta mikrofilmeihin, joiden skannaaminen on ollut myös nopeaa ja taloudellista. Digitointi lähti liikkeelle mikrofilmien hyödyntämisestä monessa muussakin maassa, kuten Australiassa, Iso-Britanniassa ja Yhdysvalloissa. Tällä on sivuvaikutuksensa: mustavalkoisessa mikrofilmissä kuva ei aina ole tarkka tai sävykäs, eikä filmausta tehtäessä ole osattu ajatella, että tulevaisuudessa aineistoon tehtäisiin sanahakuja.

Digitoituja lehtiä voi lukea sivu ja numero kerrallaan, mutta valtaosa hakijoista lähestyy miljoonien sivujen kokoelmaa juuri sanahakujen avulla. Tämä perustuu optiseen tekstin tunnistukseen. Lehtien sivuista otetut kuvat on käsitelty ohjelmistolla, joka tunnistaa tekstin. Todellisuudessa haku ei kohdistu lehden sivuun tai siitä otettuun kuvaan, jonka käyttäjä näkee, vaan taustalla olevaan tekstitiedostoon, joka pohjautuu kuvan koneelliseen tulkintaan. Käyttäjän on hyvä muistaa, että nämä tekstinnökset sisältävät myös tunnistusvirheitä, minkä vuoksi osa avainsanoista jää löytymättä. Haaviin tarttuu myös virheitä. Kokeilin kerran, viitattiinko suomalaisessa lehdistössä 1800-luvulla ”aikamatkaan”. Haku tuotti paljon tuloksia, mutta tarkemmin tutkittaessa paljastui, että suurin osa viittasikin kävelyretkiin, ”jalkamatkoihin”. Ohjelmisto oli tulkinnut l-kirjaimen i:ksi ja erottanut j-kirjaimen muusta sanasta.

Aina virheet eivät ole ongelma. Olemme Svenska Litteratursällskapet i Finlandin rahoittamassa ja Turun, Helsingin, Uumajan ja Örebron yliopiston yhteishankkeessa tutkimassa ruotsalaisen ja suomenruotsalaisen lehdistön välistä kopiointia ja tekstien uudelleenkäyttöä. Olemme tuoneet digitoidut lehdet yhteen ja analysoimme seuraavaksi, kuinka paljon sisältöjä jaettiin Pohjanlahden yli. Laskennallisin menetelmin tekstien samanlaisuus voidaan paikantaa silloinkin, kun tekstintunnistuksessa on paljon virheitä.

Tällä hetkellä digitoidut lähteet vaikuttavat voimakkaasti historiakuvaamme ja siihen, mitä ylipäätään tutkimme. Omalle tietokoneelle saatavat aineistot voivat tuntua helpoilta käyttää. Tämä herättää kuitenkin kysymyksiä. Kuinka paljon lähteiden saatavuus ohjaa tutkimuksen kohteita? Tällä hetkellä yliopistojen opettajat ja tutkijat voivat työssään käyttää myös tekijänoikeuden alaista aineistoa, mutta 1900-luvun kohdalla digitointiaste on vielä matala. Ne harvat lehdet, jotka ovat saatavissa, painottuvat kohtuuttomasti. Tässä tilanteessa digitointipolitiikkaa pitäisi huolellisesti miettiä, samoin sitä, miten toisen maailmansodan jälkeisen aineiston digitointi tapahtuu ja miten voitaisiin taata resurssit työn jatkamiseen.

Hannu Salmi