



MAIJA PAAVOLAINEN

Tiedon järjestäminen käytännössä

Kirjastoissa järjestetään aineistoa harjaantuneeseen abstraktioon perustuvilla työkaluilla ja säännöillä, joiden oikeutuksesta käydään keskustelua. Aineiston järjestäminen näin on tiedon löytämisen edellytys. Nykyisissä kirjastojärjestelmissä ihmisen tekemä luokitus ja kuvailu on korvautumassa automaattisella kuvailulla ja suositteuvalgoritmeilla muun verkon tapaan, mutta samalla tiedonhaku muuttuu keskivertokäyttäjälle läpinäkymättömäksi.

Miksi kirjastolainen on kuuluisa täsmällisyydestään? Jos kirja on väärässä paikassa, se on hukassa. Ainoa tapa löytää kirja tuhansien joukosta on tietää, missä sen pitäisi olla. Kirjojen järjestämisen historia kulkee yhdessä kirjoitustaidon kanssa. Assyrian kuningas Assurbanipalin kirjastossa Ninivessä 600-luvulla eaa. yli 30 000 nuolenpääkirjoituksella täytettyä savitaulua oli otsikoitu taulun reunaan, joka näkyi hyllyn päädystä. Taulujen järjestys noudatti todennäköisesti myös aiheenmukaista luokitusta. Myöhemmin yksityiskirjastojen käsin kirjoitetuissa luetteloissa teoksesta listattiin yleensä perustietoja tekijästä, painovuodesta ja kirjan hankintatavasta; tieto saatettiin jaotella sarakkeisiin. Jotta luetteloa ei tarvitsisi jatkuvasti kirjoittaa puhtaaksi kokoelman täydentyessä uusilla teoksilla, kirjastoissa siirryttiin 1900-luvulla niin sanottuun kortistoluetteloon, johon voi lisätä ja josta voi poistaa vaivattomasti tiettyä teosta kuvaavan kortin kokonaisuutta rikkomatta.

Kirjastoluettelo listaa kirjaston kirjat ja kuvailee ne soveltuvin osin. Luettelon avulla pidetään kirjaa kokoelman sisällöstä eli siitä, mitä kirjoja kirjastossa on. Yhtä lailla voidaan erottaa esimerkiksi saman teoksen eri versiot toisistaan. Koska kuvailutietoa käytetään viime kädessä yksilöimään kirjojen eri kappaleet niiden numeroilla, sen avulla valvotaan myös, mitkä kirjat ovat paikalla kirjastossa ja mitkä lainassa.

Digitaalisten aineistojen metadata muistuttaa painettujen aineistojen metatietoja mutta on käytettävyyden kannalta vielä tärkeämpää, koska digitaalisen kirjan tekniset tiedot kertovat tiedostomuodosta ja lukemiseen tarvittavista ohjelmistoista. Lukualustat ominaisuuksineen eivät ole vielä standardisoituja vaan kilpailevat keskenään. Digitaalisen kirjan haettava teksti ja integroidut muistiinpanovälineet ovat hyödyllisiä mutta samalla riippuvaisia päätelaitteiden käytettävyydestä ja yhteensopivuudesta. Lisäksi digitaalisen tekstin käyttöä joudutaan rajoittamaan keinokekoisesti käyttöehdoilla ja suojaus-

järjestelmillä, koska ilman niitä se olisi monistettavissa ja luettavissa ilman kustannuksia.

Metadatan käsite yhdistää tiedon järjestämisen ja uudelleen löytämisen tapoja. Se on tietoa tiedosta ja myös rakenteista eli potentiaalisesti koneluettavaa. Metadatan avulla kuvailtavien objektien piirteitä voidaan suhteuttaa toisiinsa riippumatta niiden olomuodosta eli esimerkiksi siitä, onko julkaisu painettu vai digitaalinen. Se mahdollistaa objektien tunnistamisen ja etsimisen sekä niiden sisällön hallinnan. Metadatan tuottamista ja toisaalta käyttämistä voidaan pitää kirjastolaisten asiantuntemuksen keskeisimpänä osana. Alan sisällä saatetaan käydä intohimoisia keskusteluja vaikkapa luokitusjärjestelmien paremmuudesta, mutta ammatillista identiteettiä luonnehtii yleensä pikemmin stereotyyppinen vaatimattomuus kuin ymmärrys merkittävän luokitteluvallan käytöstä tai harjaantuneesta käsiteanalyysistä.

Paikka hyllyssä

Painetun kirjan fyysistä paikkaa kuvaavaa, joskus kryptistä merkijonoa kirjastoluettelossa kutsutaan *signumiksi*. Kansalliskirjaston vanhimmat käytössä olevat paikanmerkit ovat kolmiosaisia numerosarjoja, joiden historia juontaa Engelin rakennuksen etelä- ja pohjoisaleissa sijainneisiin kirjakaappeihin. Kaapeilla on juoksevat numerot, niiden hyllyillä roomalaiset numerot ja kunkin hyllyn kirjoilla juokseva numero vasemmalta oikealle. Ongelma tällaisessa paikanmerkissä on, että kokoelma pysyy staattisena eikä kirjoja voi lisätä sen hyllyväleihin. Kaappeja voisi toki tuoda lisää. Nykyisin Kansalliskirjaston kirjakellarin painetussa yleiskokoelmassa noudatetaan kirjoille saapumisjärjestyksessä annettavaa juoksevaa numeroa *numerus currensia*, mutta asiakkaiden saatavilla jatkuvasti oleva avokokoelma on käytön helpottamiseksi jaettu aiheenmukaisiin luokkiin. Tämä onkin tapana silloin, kun ihmiset saavat itse etsiä kirjansa.

Kun aloitin työt Helsingin yliopiston Opiskelijakirjastossa viisitoista vuotta sitten, sinne palautettiin

satoja kurssikirjoja päivittäin. Ensimmäisellä viikolla sain käteeni useamman sivun perehdytysmonisteen kurssikirjojen aakkostamisesta ja aloin uusien siviilipalvelusmiesten kanssa harjoitella sitä kärry kerrallaan. Kokeneempi nuoriso kisasi täyden palautuskärryn aakkostus- ja hyllytysnopeudessa. Ohjetta seuraten kirjat aakkostetaan pääsanan mukaan, joka useimmiten on teoksen nimeke eli kirjan nimi tai sen tekijän nimi. Pääasiassa toimitetut teokset aakkostetaan nimekkeen mukaan ja muut ensimmäisen tekijän mukaan. Pääsana on käytännön syistä merkattu kirjan takakanteen. Kurssikirjakokeelma on niin suuri, että monista kirjoista on käytössä useamman eri vuoden painoksia. Ne aakkostetaan vanhin ensin ja uusien viimeiseksi. Jostain syystä oli päädytty siihen, että V:n ja W:n välille ei tehdä eroa, mutta Mc-alkuiset tekijänimet aakkostetaan kohtaan Mac. Ymmärrätte, että perehdytys tarvitaan. Näitä samoja kultajyviä jaoimme asiakkaille tarvittaessa.

Kirjat aakkostetaan tietyn luokan sisällä. Kirjan luokittelu voi perustua määrättyyn käyttötarkoitukseen (kurssikirjat, hakuteokset, bestsellerit) tai se voi olla aiheenmukainen. Luokitusjärjestelmiä kehitettäessä on pyritty yleispätevyyteen, mutta käytännön esteet tulevat usein vastaan. Monesti on käytännöllisempää luokitella pieni kokoelma väljästi.

Helsingin yliopiston kirjaston lukuisat eri tarkoituksia ja laitoksia palvelevat kirjastot yhdistettiin yhdeksi isoksi pääkirjastoksi yliopiston keskustakampuksella vuonna 2010¹. Jokaisella laitoskirjastolla oli oma, tieteenalan sisäisiin jakoihin ja todennäköisesti henkilökunnan mieltymyksiin perustuva luokituksensa. Myös kurssikirjoja lainaavan Opiskelijakirjaston kirjat oli järjestetty kokoelmien käytön ja hyllytilan mukaan numeroituihin luokkiin. Kun uusi kirjastotalo valmistui, muuttoa oli suunniteltu joitakin vuosia. Jokaisessa kirjastossa henkilökunta oli käynyt läpi kokoelmaa tehden poistoja kokoelmien yhdistämistä varten ja samalla luokittanut teoksia yleisen kymmenluokituksen mukaan, jotta uudessa talossa voitaisiin luopua laitoskirjastojen omista luokituskaavoista ja järjestää kaikki aineisto yhdeksi kokoelmaksi. Hanke ei koskaan valmistunut. Edelleen 12 vuotta myöhemmin pääkirjaston eri tieteenalojen hyllyluokat perustuvat vanhojen tiedekunta- ja laitoskirjastojen luokituksiin. Sittemmin poistoja on tehty paljon ja suurin osa uusista nimekkeistä hankitaan sähköisinä.

Kuvaava keskustelu luokituksesta käytiin vuonna 1982 silloisen nuoren ammattilaisen, Sotkamon kunnankirjaston kirjastonhoitajan Heikki Poroilan ja kirjastokoulutuksen kehittäjän, sittemmin filosofian professorin Raili Kaupin välillä *Kirjastotiede ja informatiikka* -lehdessä². Kaupin kirjoitus ”Kirjastotiede ja filosofia” korostaa kirjaston merkitystä yksilön tukena tämän sivistyksen ja älyllisen potentiaalın toteutumisessa. Kauppi pitää luokitusta käytännön kysymyksenä eikä jää pohtimaan sitä käsitteellisesti tai sen mahdollisen normittavuuden ja luokitteluvallan kannalta. Poroila puolestaan kysyy, eikö luokituksen tule heijastaa tieteiden sys-

teemiä ja todellisuutta. Ja eikö systeemiä pitäisi pyrkiä kehittämään kattavammaksi ja paremmaksi? Poroila on myöhemmin tullut tunnetuksi aktiivisena ammattikeskustelijana tekijänoikeus- ja sensuurikysymyksissä. Teoksessa *Luurangot portinvartijan kaapissa* esitetään vaikeita kysymyksiä siitä, mitä yleisten kirjastojen kokoelmiin hankitaan ja miten verkon käyttöä kirjastoissa rajataan sopivaan ja sopimattomaan käyttöön henkilökunnan omien arvostusten mukaan.³ Kirjastolaisten tuntuu olevan vaikea myöntää tekevänsä valintoja ja käyttävänsä valtaa.

Kaupin vastauksessa kuuluu pitkä käytännön kokemus. Hän kirjoittaa luokitusjärjestelmän uudistamisen olevan hidasta ja pitkäjänteistä työtä; valmistuessaan se on todennäköisesti jo vanhentunut. Sanastot ja luokituskaavat jäävät nopeasti jälkeen tieteellisen terminologian kehityksestä. Luovan hengen tuotteina kirjat voivat paeta luokitteluja, mutta silti jokaiselle kirjalle olisi löydettävä paikka. Luokitusongelmat ovat tyypillisiä kirjastotyön käytännön ongelmia. Kauppi kehottaa edistämään tieteiden systematiikan tunnettuutta kirjaston asiakkaiden keskuudessa muulla tavoin. Käytössä olevan luokitusjärjestelmän ja hyllykartan viereen voi kirjastotilaan ripustaa jonkin toisen esityksen tiedon kokonaisuudesta ja taustoittaa sitä kirjanäyttelyllä tai esitelmillä: ”Tallaisiin kokeiluihin jokainen voi ryhtyä *itse* ja *nyt*, luokitusjärjestelmästä riippumatta”⁴.

Mitä kirja käsittelee?

Luokituksen lisäksi kirjojen järjestyksestä ja löytämisestä huolehditaan luetteloinnin ja sisällönkuvaailun avulla. Luettelointi kuvailee kirjaa fyysisenä esineenä, ja sisällönkuvaailun tarkoituksena on kertoa kirjan sisällöstä tiivistetyssä muodossa. Kirjojen kuvailu nojaa standardeihin. Kuvaailutietojen bibliografiset tiedot listaavat tuttuja tietoja tekijästä, ilmestymisvuodesta ja ilmestymispaikasta, mutta myös painetun kirjan fyysisistä ominaisuuksista kuten sivumäärästä tai kansien materiaalista. Asiasanoituksella puolestaan tarkoitetaan muiden kuvaailutietojen mukana kulkevaa tiivistettyä kuvausta kirjan sisällöstä.⁵

Kirjastojen tekemän sisällönkuvaailun tarkoitus on palvella tiedon etsijää. Tieto kirjan summittaisesta sisällöstä täytyy välittää jotenkin, jotta teosta tuntematta voisi arvioida, haluaako lukea sen. Kuvailusanojen valintaa voi ohjata se, että kirjastolainen oikeasti tuntee kirjan sisällön, mutta yleisimmin kuvaailun tekijä ei ole lukenut käsittelemäänsä kirjaa. Hän selaa kirjaa, lukee takakansitekstin ja mahdollisesti aiheita käsittelevän artikkelin hakuteoksesta. Hän tuntee kuvaailussa käytettävät välineet ja noudattaa ohjeita, joiden avulla hän valitsee sopivat sanat kontrolloidusta sanastosta. Kirjastolainen toimii välittäjänä kirjan tekijän ja kirjan etsijän välillä, kun hän muiden puolesta päättää tietyn teoksen käsittelevän tiettyä aiheita.

Asiasanastot (*subject headings, subject terms*) eli tesau-rukset säätelevät kuvaailua määrittelemällä, mikä yleis-

”Kirjastolainen ei enää määrittele teoksen asemaa yksin.”

kielen sana valitaan vakiintuneeksi kuvailusanaksi eli asiasanaksi.

Sanastoilla on hierarkkinen rakenne, jossa asiasanojen keskinäisiä suhteita kuvataan. Esimerkiksi Kansalliskirjaston ylläpitämä Finto-YSO, Yleinen Suomalainen Ontologia, listaa ”pukineet” yläkäsitteeksi ”vaatteille”, ”asusteille” ja ”jalkineille” ja suosittaa sen käyttämistä ”asun” tai ”pukimien” sijasta.⁶ Pukineet liittyvät (assiosiatiivisina käsitteinä) muotiin ja pukeutumiseen, mutta on listattu kuuluviksi ”Kotitalouden” ja ”Vaateusteollisuuden” käsitteiden aiheryhmiin. Termeihin voi myös ehdottaa perusteltuja muutoksia. Pukineet-tietue on luotu vuonna 1992, mutta jokin päivitys on tehty vielä vuonna 2020.

Aineistoa kuvaillessa edetään systemaattisesti tietystä piirteestä toiseen ja tuloksena on kenttiin jaoteltu sopiva määrä tietoa. Parhailaan käytössä oleva sisällönkuvailun ohje havainnollistaa rakenteisuutta hyvin ja sisältää vuokaavioita kuvailun vaiheittaisesta etenemisestä.⁷ Ensin listataan julkaisun lajityyppi ja mahdollinen kohderyhmä, sitten listataan aiheena olevia henkilöitä, paikkoja ja aikakausia. Lopuksi listataan tutkimusmenetelmä. Ohjeet ovat ihastuttavan tarkat ja pyrkivät minimoimaan harvinaisista syntyviä ongelmia.

Asiasanoja valitessa on toki periaatteessa mahdollista tehdä valintoja omien asenteidensa pohjalta tai jättää joi-takin piirteitä näkymättömiin. Luokitteluvallan käyttö ei kuitenkaan ole ammattikulttuurissa kovin tiedostettu tai toimintaohjeissa julkilausuttu asia. Jo vuonna 2011 ilmestyneessä kirjoituksessaan ammattilehti *Signumissa* joukko alan tutkijoita ja vaikuttajia kirjoittaa sisällönkuvailun olevan kriisissä⁸. Paineet ”intellektuaalisen indeksoinnin” perinteisiä työtapoja kohtaan ovat moninaiset. Julkaisumäärän kasvaessa kirjastolaiset eivät enää ehdi tehdä kuvauksia käsityönä, mutta toisaalta valmista kuvailutietoa on yhä enemmän saatavilla. Sähköisten aineis-

tojen automaattisella kuvailulla voidaan rikastaa teoksen tietoja liittämällä siihen vaikkapa sisällysluettelo, abstrakti ja kansikuva, mutta vuorovaikutteiseen verkkoon tottuneet käyttäjät odottavat jo voitavansa itse jättää teok-sista omia kommenttejaan yhteiseen tietokantaan. Kirjas-tolainen ei enää määrittele teoksen asemaa yksin.

Automaattinen sisällönkuvailu ei myöskään ole enää scifiä, vaan aitoa resurssipulaa ratkotaan Kansalliskirjas-tossa tekoälyyn perustuvalla työkalulla. Nimellä Finto.ai tunnettu ohjelmisto osaa tekstikatkelman perusteella eh-dottaa asiasanoja käytössä olevista sanastoista, ja jokainen voi kokeilla sitä omilla teksteillään. En pidä tätä kovin ongelmallisena, koska sekä lähtöteksti että sanasto ovat työkalun käyttäjällä saatavilla ja tuloksia voi vertailla.

Matthew Crawford kuvaa kirjassaan *Elämän korjaajat* eloisasti pakkotahtista kuvailutyötä, jota hän käyttää esimerkkinä pahimmanlaatuisesta vieraannuttavasta tietotyöstä⁹. Crawford kertoo työskentelystään 1990-lu-vulla yrityksessä, jossa tuotetaan abstrakteja tieteellisistä artikkeleista haettavaksi sen aikaisissa tietokannoissa. Tekstit ovat pitkälle erikoistunutta tutkimusta, joita kuvailijoiden on parhaila lahjoillaankaan mahdoton todella ymmärtää ilman alan opintoja ja kontekstia. Ly-hennelmät täytyy laatia ja avainsanat valita mekaanisesti muodon ja toiston perusteella. Päiväkohtaiset tulosta-voitteet ovat niin kovat, että niiden täyttäminen vaatimattomammallakaan työetiikalla ja itsekunnioituksella on mahdotonta. Kuulostaa siltä, että tämäntyyppiseen kuvailutyöhön tekoäly soveltuu hyvin!

Tutkimuksen lähteet

Kirjastoammattilaiset eivät osaa suositella tiedonlähteitä siksi, että he tietäisivät tai olisivat lukeneet kaiken, vaan he ovat harjaantuneita tiedon etsimisessä. Yleensä kirjas-

”Ammattitaito on sekä tietoa julkaisemisen tavoista ja julkaisukanavien kirjosta että tietoa siitä, miten tieto on järjestetty.”

tolainen toki tuntee tietyn tieteenalan lähteitä kuten alan tutkijatkin. Tieto siitä, mitkä lehdet painottavat mitäkin tutkimussuuntausta ja mitkä ovat arvostettuja alallaan helpottaa niiden selailua ja suosittelua. Seuratessaan tiettyä tieteenalaa jonkin aikaa oppii myös, mitkä kirjat ovat tiettyinä vuosina olleet tärkeitä avauksia ja mitkä muodostuneet klassikoiksi. Osa tiedonhakutaidosta on kuitenkin teknisempää systemaattista ja analyttistä taitoa, jonka peruskäsitteet ja menetelmät ovat suhteellisen yksinkertaisia. Menetelmät perustuvat pitkälti kirjastossa tapahtuvaan tiedon järjestämiseen ja kuvailuun. Ammattitaito on silloin sekä tietoa julkaisemisen tavoista ja julkaisukanavien kirjosta että tietoa siitä, miten tieto on järjestetty.

Aiheenmukaisella haulla etsitään teoksia, joiden olemassaolosta ei vielä tiedetä, mutta niiden toivotaan käsittelevän esimerkiksi tutkimuksen tai kirjallisuuskatsauksen kohteena olevaa aihetta. Aiheenmukainen haku kirjastoluettelosta edellyttää aiheen analysointia sen käsitteellisiin osiin. Hakuun tarvitaan aihetta kuvaava sanajono eli hakulause. Se on tyypillisimmillään kahden tai useamman käsitteen yhdistelmä (vaatetus AND työolot). Joskus yksi käsitteistä on kuvattu useammalla synonyymillä, jolloin lauseke on monimutkaisempi ((vaatteet OR asusteet OR jalkineet) AND työolot). Tarkoitus on verrata hakulausekkeen kirjainjonoja kirjastoluettelon kuvailutietoihin, kuten nimikkeisiin ja asiansanoihin ja kerätä talteen ne, joissa kirjainjonot esiintyvät. Yhdistämiseen on perinteisesti käytetty Boolean logiikan operaattoreita AND, OR ja NOT tai tarkoitukseen rakennettuja tietokannan käyttöliittymän kenttiä.

Boolean logiikka on yksinkertaista siinä mielessä, että se ei millään lailla ymmärrä haettua aihetta tai anna apua suosittelemalla jotakin asiaan liittyvää. Jos hakija ei vaihda keksimään vaihtoehtoisia sanoja kuvaamaan ai-

heensa tiettyä käsitteellistä osaa, hakutulosten määrä voi jäädä pieneksi. Toisaalta jos vaihtoehtoja annetaan liikaa, tulos kasvaa niin suureksi ja menee sisällöltään ohi halutusta aiheesta, ettei sitä kannata järkevästi käydä läpi. Hakulausekkeiden rakentaminen on opiskelijalle kevyt aivojumppa, joka opettaa ottamaan vastuuta lähteiden etsimisestä ja havainnollistaa, miten oma näkökulma tutkimuskysymykseen täytyy suhteuttaa aiheesta käytävään tieteelliseen keskusteluun. On kirjaimellisesti löydettävänä sanat, joilla aiheesta on aikaisemmin kirjoitettu, jotta voi perustaa oman tutkimuksensa yhteiseen, olemassa olevaan asiayhteyteen.

Systemaattisella tiedonhaulla tähdätään tieteelliseen kirjallisuuskatsaukseen. Systemaattinen tiedonhaku tarkoittaa yksinkertaisimmillaan hakulausekkeen ja haun eri versioiden sekä saatujen tulosten dokumentoimista, jotta kirjallisuuskatsauksen kattavuutta voi tarvittaessa arvioida ja haut toistaa. Tietyillä tieteenaloilla kirjallisuuskatsaus luo puitteet empiiristen tutkimusasetelmien suunnittelulle tai tavoitteena on tuottaa näyttöön perustuvia käytännön ratkaisuja (*evidence based practice*). Tällöin on tärkeää, että katsaus on tehty huolella ja hakutulokset kattavat aiheesta käydyt tieteellisen keskustelun riittävän hyvin. Systemaattista tiedonhakua opetetaan osana maisteriopintoja esimerkiksi lääketieteen koulutusohjelmassa. Sitä tehdään edelleen maksusta tieteellisissä kirjastoissa, vaikka järjestelmien olisi tarkoitus olla niin helppoja käyttää, ettei ammattimaista tiedonhankintaa enää tarvita.

Opettaessani tiedonhankintaa humanistiopiskelijoille Helsingissä lähdin ajatuksesta, että kirjaston tietokantaa selaamalla kirjaston painetusta kokoelmasta voi saada kattavan käsityksen tietystä aiheesta, jos hakee muutamalla eri tavalla ja vertailee hakutuloksia. Tutkielman tekijä voi silloin sanoa, että hän on tietoinen aiheitaan

käsittelevästä keskeisestä keskustelusta. Jos hän valitseekin löytämistään teoksista lähteiksi vain osan, hän osaa tarvittaessa myös perustella, miksi tuli rajanneeksi lähdekirjallisuuden käytön tietyllä tavalla.

Tilanne on erilainen käytettäessä suuria monitieteisiä artikkelitietokantoja ja Google Scholaria. Rajatusta tietokannasta voi ajatella, että tehdessään siinä useampia hakuja ja vertaillaessaan eri hakutulosten määrää, relevanssia ja osumia, on kohtuullisella tavalla tietoinen käydystä keskustelusta. Sikäli se muistuttaa toiminnallisuuksiltaan kirjastoluetteloa. Yleensä tuloksia on kuitenkin niin paljon, että kaikkia ei koskaan tule käytyä läpi.

Sen sijaan Google Scholarin tulosten järjestys selailtavalla sivulla on heidän liikesalaisuutensa. Todennäköisesti tuloksissa painottuvat suositut ja uudet nimekkeet, mutta tulosten järjestys on aina käyttäjälle läpinäkymätön. Hakukoneiden toiminnallisuudet ohjaavat myös tieteellisiä kustantajia tuottamaan julkaisujen abstraktit ja avainsanat jo julkaisuvaiheessa niin, että ne saisivat heti mahdollisimman paljon osumia. Vaikka Google Scholar on läpinäkymätön, se ei tuota samanlaista nopeasti henkilön hakuhistorian perusteella suodattuvaa kuplaa kuin Googlen arkiversio¹⁰. Kärkeen nousseiden tulosten viittaussuureudet eivät puolestaan kuvaa niiden relevanssia suhteessa tutkimuskysymykseen tai hakulausekkeeseen lainkaan. (On mahdollista, että algoritmi nostaa runsaasti viitattuja julkaisuja tuloslistalla ylempäs relevanssin kustannuksella.)

Algoritmin läpinäkymättömyyden vuoksi Google Scholar sopii tutkimuksen tekemiseen tietyin varauksin. Sille ei voi pohjata systemaattista kirjallisuuskatsausta¹¹, mutta tietyn artikkelin etsimiseen, uutuusseurantaan tai uuteen aiheeseen tutustumiseen se sopii hyvin. Scholarin indeksoima aineisto keskittyy kansainvälisiin tieteellisiin lehtiin, joiden julkaisukieli on englanti. Kotimainen keskustelu ja muunkielinen aineisto kannattaa etsiä muualta.

Algoritmit muuttavat hakemista

Verkossa syntyvän metadatan kulttuurinen ja taloudellinen merkitys on kasvanut voimakkaasti. Ihmisten vapaaehtoisesti eri alustoille luovuttamia henkilötietoja ja tieto käyttäjien liikkeistä verkossa on jo jonkin aikaa ollut merkittävä hyödyke, jota verkkoalustat myyvät mainostajille. Juuri voimaan astunut EU:n tietosuojasetus on askel tällaisen tiedonkeruun rajoittamiseksi. Se on kuitenkin vasta ensimmäinen askel, kun otetaan huomioon, kuinka pitkään dataa on kerätty ja miten monimuotoiselle liiketoiminnalle se luo pohjan. MyData-kansalaisliike on pyrkinyt tekemään näkyväksi tiedonkeruun laajuutta ja vaatii ihmisille oikeuksia saada päättää kaikesta heistä kerätystä tiedosta¹². Näkökulman muutos nykytilanteeseen olisi valtava.

Samalla kun metatiedon määrä kasvaa, sen tulkinta ja yhdistely ihmisavoilla vaikeutuu. Viimeisen 15 vuoden aikana myös kirjastojen pitkään läpinäkyvinä pysyneet tiedon järjestämisen käytännöt ovat digitalisoitumisen myötä ottaneet askelia sattumanvaraisempaan suuntaan.

Kun tiedon käsittely on nopeampaa ja tehokkaampaa, myös kirjastojärjestelmän metatieto voi olla runsaampaa ja alkuperältään vaihtelevaa. Osa tiedosta tuotetaan edelleen käsin, mutta osa haravoidaan automaattisesti esimerkiksi julkaisujen sisällysluetteloista ja abstrakteista. Kirjakokoelman sisältö heijastaa osaltaan uusinta tutkimusta, mutta myös digitaalisen kirjabisneksen lainalaisuuksia. E-kirjoja ostetaan tarpeen mukaan yksittäin, mutta myös paketteina, jotka kustantaja on koostanut oman etunsa mukaan. Paketeissa voidaan myydä uudelleen esimerkiksi 1990-luvulla ilmestyneiden nimekkeiden verkkoversioita.

Digitaalista kokoelmaa selataan esimerkiksi Helsingin yliopiston kirjastossa kirjaston omalla käyttöliittymällä, jonka hakutuloksissa painettua ja digitaalista aineistoa ei lähtökohtaisesti erotella. Liittymä tarjoaa hyvät vaihtoehdot haun rajaamiselle ikään kuin relevanttien tulosten määrä olisi aina suuri, mutta systemaattista aiheenmukaista hakua on vaikea saada toimimaan. Aineiston määrä on niin suuri, että tuloslistalta ei voi selata kaikkia osumia. Täsmällisen merkkijonojen toistamisen sijasta käyttöliittymän toiminta muistuttaa muualta verkosta tuttuja arvauksia ja suositteluja.

Valtaosa digitaalisista kirjastoaineistoista on haettavissa metadatan lisäksi etsimällä merkkijonoja kokotekstistä. Kokotekstihauksessa hakulauseketta vastaavia sanoja ei siis haeta pelkästään kuvailutiedosta vaan myös teoksen sisästä. Tällöin esimerkiksi niiden esiintymistiheys tekstissä voidaan ottaa huomioon arvioitaessa, miten hyvin hakuosuma vastaa haettua aihetta. Kokotekstihaku tuottaa suuremman määrän hakutuloksia kuin käsin luotuihin avainsanoihin perustuva metadatahaku. Pitkän hakutulostilan järjestämiseen tarvitaan muita kriteerejä. Järjestys perustuu silloin usein relevanssiin, joka tietokannasta riippuen ottaa huomioon, kuinka tiheästi hakusanat toistuvat kohdetekstissä. Toki hakutuloksia voi järjestää tutuin kriteerein vaikkapa ilmestymisvuoden mukaan. Osa tieteellisiä tekstejä tarjoavista kaupallisista tietokannoista käyttää myös ”Etsi samanlaisia” -hakuja. Niissä tuloksiin luodaan relevanssia vertaamalla joko pelkkää metadataa tai tietyn dokumentin kokotekstiä koko tietokannan sisältöön.

Onko tiedonhaku välttämätöntä ymmärtää? Onko suosittelu- ja läheisyysalgoritmien käyttö ongelma vakavasti otettavan tutkimuksen tekemiselle? Kuvasin aiemmin tiedonhakuja, joka perustui vastaavuuksien etsintään aineistojen kuvailutiedoista, kuten nimekkeistä ja asiasanoista. Kirjastossa on siis yleensä itse tuotettu se tieto, jota hakemiseen käytetään. Tiedonhaun kuvaaminen näin auttaa ymmärtämään metadatan merkityksen tiedon järjestämisessä. Uusien järjestelmien myötä tämä on kuitenkin muuttunut ideaalimalliksi kirjastotyön käsitteellisistä piirteistä ja ”sisällönkuvaailun informaationsiirtomallista”¹³. Välittömän käytettävyyden kannalta voi tuntua, että kirjastoluettelon oli korkea aika muuttua joustavammaksi. Silti jos tieteellisen tiedonhankinnan tarkoituksena ei ole löytää vain jotain tarkoitukseen sopivaa vaan kaikki aiheeseen liittyvä ja lopulta

vielä arvioida sitä itsenäisesti, tulosten sattumanvaraisuudesta voi muodostua ongelma.

Ero on periaatteellinen, vaikka kirjaston kokoelma olisi digitaalisen aineiston suhteellisen osuuden kasvettua muuttunut laajemmaksi tai vaikeammin selailtavaksi. Algoritmien ei sinänsä tarvitse olla läpinäkymättömiä – hyvässä digitaalisiin aineistoihin perustuvassa tutkimuksessa data-analyysialgoritmit ovat osa tutkimuksen menetelmiä ja niitä kehitetään ja arvioidaan suhteessa tutkimuksen tuloksiin. Tämä edellyttää kuitenkin ymmärrystä tilastollisista menetelmistä ja analysoitavana olevan datan laadusta. Kirjastolaiset tarvitsevat siis parempaa tietoteknistä ymmärrystä siitä, miten heidän tuottamaansa kuvailutietoa nykyisissä järjestelmissä käsitellään.

Samalla kun tieteellisen tiedon hakeminen on tekniikkansa puolesta muuttunut läpinäkymättömämmäksi ja automatisoitunut, vastauksia haetaan valtavasta ja rajoiltaan epäselvästä julkaisujen joukosta. Lähteitä ei enää haeta tietyn talon seinien sisäpuolelta, eikä edes tietyn yliopiston lisensoimista digitaalisista kokoelmista. Kansainvälisen avoimen tiedejulkaisemisen myötä hyvää tutkimusta voi löytyä vapaasti ympäri maailmaa. Tiedon käyttäjät tarvitsevat kontekstuaalista ymmärrystä julkaisemisesta.

Jos käsitteellinen analyysi ja tiedon järjestäminen ovat kirjastotyön ytimessä, mutta tarve painettujen kokoelmien järjestämiselle ja hoidolle vähenee, mitä asiantunteumuksella voi tehdä? Kirjastolaiset voivat välittää ymmärrystään muuttuvista julkaisukanavista: miten esimerkiksi yliopistojen tutkimustietojärjestelmät ja avoimet julkaisuarkistot täydentävät kuvaa tutkijan vertaisarvioituista julkaisuista? Toisaalta tarvitaan ymmärrystä tieteellisen kustantamisen bisnesmalleista ja yrityksistä muuttaa niitä julkisen talouden kannalta kestävämpään suuntaan, kuten monissa maissa yleistyvistä yliopistojen omista avoimista julkaisusarjoista. Kirjastossa tulisi olla riittävästi tietoa vähintään omassa yliopistossa tehtävän tutkimuksen sisällöistä ja myös sen poliittisesta kontekstista. Yhtä lailla ymmärrys lainsäädännön asettamista puitteista kuten tekijänoikeusasioista, lisensoimisesta ja yksityisyysensuojasta kuuluu kirjastoon. Tieto tieteellisen julkaisemisen kanavista ja käytännöistä palvelee tutkijoita edelleen, vaikka julkaisuja ei osteta paperilla. Ymmärrys tiedon järjestämisestä on digitaalisessa ympäristössä vähintään yhtä tärkeää kuin materiaalisessa maailmassa.

Viitteet

- 1 Ks. Sinikara 2022.
- 2 Poroila 1982; Kauppi 2000. Lehden nimi on vuodesta 1995 ollut *Informaatiotutkimus*.
- 3 Ks. Poroila 2007. Luokitteluvallasta, ks. myös Bowker ym. 2000.
- 4 Kauppi 1982.
- 5 Moderni kuvailustandardi listaa, mitä tietoja kirjoista tallennetaan. Kuvailustandardin osia kutsutaan kentiksi. Bibliografisen datan MARC-standardi syntyi 1960-luvulla Yhdysvalloissa ja sitä ylläpidetään kansainvälisissä kirjastojärjestöissä. Monta muutosta kokoneena se pitää pintansa, koska niin suuri osa maailman kirjastojen tiedoista noudattaa sen muotoa.
- 6 Finto-YSO aiheesta ”Pukineet”: yso.fi/onto/yso/p4731 Ontologialla tarkoitetaan tässä yhteydessä sitä, että sanasto nojaa niin sanotusti pysyviin tunnisteisiin eli teknisiin viitepisteisiin, joihin voidaan liittää esimerkiksi suomen- ja ruotsinkieliset versiot samasta käsitteestä.
- 7 Sisällönkuvailun asiantuntijaryhmän sisällönkuvailuopas suomalaisia kirjoja varten: wiki.helsinki.fi/x/uaLkFw
- 8 Saarti ym. 2011.
- 9 Crawford 2012.
- 10 Verkkohakujen itseviittaavuudesta ja kuplista kirjoitti jo Eli Pariser teoksessa *Filter Bubble* (2011).
- 11 Ks. esim. Bramer ym. 2013.
- 12 Liikkeen manifesti on vapaasti luettavissa verkossa www.mydata.org/participate/declaration/
- 13 Ks. Saarti ym. 2011.

Kirjallisuus

- Bowker, Geoffrey C. & Star, Susan Leigh. *Sorting Things Out. Classification and its consequences*. MIT Press, 2000.
- Bramer, Wichor, Giustini, Dean, Kramer, Bianca & Anderson, P. F., The Comparative Recall of Google Scholar versus PubMed in Identical Searches for Biomedical Systematic Reviews. A Review of Searches Used in Systematic Reviews. *Systematic Reviews*. Vol. 2, No. 115, 2013. Verkossa: systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046-4053-2-115
- Crawford, Matthew, *Elämän korjaajat. Kädentaitojen ja käytännöllisen ammattityön ylistys* (Shop Class as Soulcraft, 2010). Suom. Johan L. Pii & Tuukka Tomperi. niin & näin -kirjat, Tampere 2012.
- Kansalliskirjasto, Sisällönkuvailun asiantuntijaryhmän sisällönkuvailuopas suomalaisia kirjoja varten. wiki.helsinki.fi/x/uaLkFw
- Kauppi, Raili, Kirjastotiede ja filosofia. Teoksessa *Raili Kaupin kirjoitukset 2. Kirjasto, sivistys, kasvatus*. Toim. Ismo Koskinen & Jari Palomäki. Tampere University Press, Tampere 2000, 135–143.
- Kauppi, Raili, Edellisen Johdosta. *Kirjastotiede ja informatiikka* 4/1982, 101. Verkossa: journal.fi/inf/article/view/2322.
- Pariser, Eli, *The Filter Bubble. What the Internet is Hiding from You*. Penguin, London 2011.
- Poroila, Heikki, Filosofia luokitusta auttakoon! Kommentteja Raili Kaupin kirjoitukseen ”Kirjastotiede ja filosofia”. *Kirjastotiede ja informatiikka* 4/1982, 100–101. Verkossa: journal.fi/inf/article/view/2321
- Poroila, Heikki, *Luurangot portinvartijan kaapissa*. BTJ Kirjastopalvelu, Helsinki 2007.
- Saarti, Jarmo, Nykyri, Susanna, Hypén, Kaisa, Tuominen, Kimmo & Kulmala, Seija, Monikulttuurisuus, monimerkityksisyys ja teknologioiden kehitys sisällönkuvailun tulevaisuuden haasteina. *Signum* 4/2011. Verkossa: journal.fi/signum/article/view/3910.
- Sinikara, Kaisa, *Tiedeyhteisön kumppanina. Laitoskirjastoista Helsingin yliopiston kirjastoksi 1828-2020*. Helsingin yliopiston kirjasto, Helsinki 2022. Verkossa: doi.org/10.31885/9789515150462

