Jarmo Saarti

Digitaalinen humanismi pohjoismaissa

Digital humanities in the Nordic countries -konferenssi Oslossa 15–17.3.2016. Suomen tiedekustantajien liitto järjesti jäsenilleen yhteismatkan seminaariin ja allekirjoittanut oli mukana Informaatiotutkimus-lehden edustajana.

Jarmo Saarti, Library Director, Itä-Suomen yliopiston kirjasto, PL 1627, 70211 Kuopio, jarmo.saarti@uef.fi

Oslon yliopistolla järjestettiin keväällä 2016 ensimmäinen Digital humanities in the Nordic countries -seuran konferenssi[1]. Seuran perustamista on puuhattu muutaman vuoden ajan, ja nyt järjestetty konferenssi osoitti seuran ja konferenssin aiheen ajankohtaisuuden: paikalla oli noin 220 osallistujaa eri pohjoismaista ja esityksiä pidettiin noin 80. Järjestäjien mukaan esitysten määrä oli yllättänyt eniten – hyviä esityksiä piti tilojen puutteen takia jättää pois.

Humanistisissa tieteissä on tietotekniikkaa hyödynnetty atk-laitteistojen ja ohjelmistojen käytön alusta asti. Alueen pioneereina ovat olleet kielitieteet ja kirjallisuustiede, joissa digitaalisten korpusten laatimisella ja niiden analysointimenetelmien kehittämisellä on jo pitkä historia. Tietotekniikan hyödyntäminen on levinnyt vähitellen taiteiden tutkimukseen ja muihin ihmistieteisiin. Merkittävänä osa-alueena on ollut digitaalisten kokoelmien ja niihin liittyvien tietojärjestelmien kehittäminen; sekä fyysisten aineistojen digitointi että digitaalisina syntyneiden aineistojen kuratointi.

Tietotekniikan mukaantulo humanistisiin tieteisiin on merkinnyt myös matemaattisten ja tilastollisten menetelmien hyödyntämistä humanistisiin aineistoihin. Sekä Oslon yliopiston rehtori Ole Petter Ottesen (neurotieteilijä) että humanistisen tiedekunnan varadekaani Ellen Rees (kirjallisuustieteilijä) korostivat tätä kehitystä.

Edellinen otti esille erityisesti sen, että tämä antaa mahdollisuuden luonnon- ja ihmistieteiden yhdistämiseen molempia hyödyntävällä tavalla: erityisesti myös sen vuoksi, että geneettisen koodin aukaiseminen on tuonut luonnontieteisiin myös merkitysopillisia piirteitä. Norjan valtio onkin panostanut runsaasti sekä erilaisten aineistojen digitointiin että parhaillaan humanistisen alan strategiseen työhön.

Kirjastot ja digitaalinen humanismi

Konferenssin aikana kävi useassa esityksessä selville, että kirjastotieteen ja informaatiotutkimuksen perusasioiden uudelleen löytäminen on tapahtumassa: tiedon järjestäminen, kokoelmatyö ja kokoelmien kuratointi ja erityisesti sisältöjen semanttisten merkitysten analysointi on nousemassa uuteen arvoon semanttisen webin kehittämisen kohdattua sen haasteen, jonka erityisesti humanististen tekstien monimerkityksellisyys ja –tulkinnallisuus ja merkitysten historiallisuus aiheuttavat. Apulaisprofessori Francesca Tomasi Bolognan yliopistosta toi tämän hyvin esille alustusesitelmässään.

Hänen mukaansa näyttää siltä, että kokoel-

[1]http://dig-hum-nord.eu/?page_id=352&lang=en

mien laatiminen ja teosten kontekstualisointi on mahdollista tehdä semanttisen webin tekniikoiden avulla. Tämä on kuitenkin tehtävä kontekstia ymmärtäen ja hyödyntäen sekä automaattisia että ihmisen tulkintaa vaativia tekniikoita. Tietotekniikan kielellä: humanistisista kokoelmista muodostuvat datasetit ovat kulttuuririippuvaisia, ja näyttää siltä, että kulttuuriset peruskokoelmat – auktoriteettikokoelmat – vaativat samankaltaista kokoelmatyötä kuin mitä perinteisessä painetussa kokoelmassa tehtiin. Tietotekninen ympäristö antaa tosin aivan uudenlaisia mahdollisuuksia teosten kontekstualisointiin, niiden merkitysten avaamiseen käyttäjille ja intertekstuaalisten viittausten konkreettiseen avaamiseen hyperteksteinä.

Norjan kansalliskirjasto esitteli useassa esitelmässä mittavaa digitointityötä ja sen mahdollisuuksia tutkimukselle. Norjalaiset voivat jo nyt käyttää vuoteen 2000 asti digitoituja kokoelmia[2] – kun kattava digitointi 2000-luvulle asti saadaan vuonna 2017 valmiiksi, on teoksia kaikkiaan käytettävissä ilmaiseksi noin 250.000 kappaletta.

Kirjasto on luonut tutkijoille tähän aineistoon työkaluja, joiden avulla on helppoa tutkia mm. digitoitujen aineistojen sanastoa, sen tilastollisia tunnuslukuja ja sanojen esiintymistiheyksiä. Tutkija voi valita digitoiduista teoksista oman korpuksensa analysoitavaksi. Lisäksi aineiston avulla voidaan analysoida sanojen kontekstuaalista esiintymistä esimerkiksi eri lajityypeissä, teemoittain, aikakausittain tai tekijöittäin.

Tekstit, digitointi ja kokoelmat

Iso osa esityksistä painottuikin vanhojen tekstin digitoinnille: sen aiheuttamille haasteille ja mitä mahdollisuuksia digitoidut kokoelmat antavat tutkimukselle ja sen tekemiselle. Suomen Kansalliskirjaston Kimmo Kettusen esitys käsitteli historiallisen sanomalehtikirjaston massadigitoinnin haasteita. Vanhojen digitoitujen aineistojen kasvu on ollut nopeaa viime vuosikymmenen aikana. Europeanan mukaan vuonna 2012 jo pelkästään Euroopassa oli digitoitu 129 miljoonaa sivua noin 24000 lehtinimekkeestä. Pohjoismaat ovat olleet erittäin aktiivisia vanhojen sanomalehtien digitoinnissa.

Kettusen mukaan suomalaisessa korpuksessa on noin 2,4 miljardia sanaa, jolloin voidaan puhua aidosti isosta datamäärästä ja sen antamista mahdollisuuksista ja haasteista. Vanhojen sanomalehtien digitoinnissa on hänen analyysinsa mukaan päästy Suomessa noin 75 %:n yhdenmukaisuuteen originaalitekstien kanssa. Puutteet johtuvat mm. vanhojen aineistojen automaattisen luennan ongelmista, sanaston variaatioista vanhassa kielessä ja alkuvaiheen digitointitekniikkojen laadusta.

Tutkijan kannalta digitoiduissa korpuksissa on tärkeää niiden luotettavuus erityisesti silloin, kun aletaan käyttää automaattisia datanlouhinnan menetelmiä. Digitoidut kuvat riittävät hyvin kun ihminen on tulkitsemassa aineistoja ja pystyy silloin näkemään alkutekstin, mutta dataksi muutettu teksti tällaisilla suuruusluokilla on mahdotonta tulkita ihmisen lukemana. Kettusen mukaan kokoelmien laadun kehittämisessä voidaan hyödyntää sekä tietoteknisiä välineitä että ihmistyötä. Tosin hän suhtautui jälkimmäisen mahdollisuuksiin kriittisesti: niistä on käytännössä apua vain suppeiden kokonaisuuksien tai yksittäisten tekstien korjailussa.

Digitointi on saanut aikaan myös mielenkiintoisen historiallisen käänteen: painetun paperin standardoidusta ja originaalitekstiin luottavasta ajasta olemme siirtymässä digitaalisten kopioiden variaatioiden maailmaan. Se alkaa lähennellä jo käsikirjoitusajan tekstien kopiointia luostareissa – hermeneutiikalle avautuukin tästä aivan uusi teoreettinen tutkimuskohde: miten tulkita eri aikoina digitoituja tekstejä, niiden variantteja ja versioita!

Historian tutkimukselle digitointi avaa uusia mahdollisuuksia myös vanhojen kielten ja uniikkien asiakirjojen kohdalla. Ensinnäkin digitaaliset kopiot voidaan jakaa laajalle yleisölle pelkäämättä niiden katoamista tai kulumista. Toiseksi vanhojen tekstien massadigitointi antaa tutkimuksellisia mahdollisuuksia dokumenttien analysointiin.

Lasse Mårtenson Gävlen yliopistosta esitteli projektia, jossa vanhoja ruotsalaisia asiakirjoja oli analysoitu tietoteknisin menetelmin. Tavoitteena on ajoittaa, tunnistaa tekijät ja määrittää kirjoituspaikat. Esityksessä oli mielenkiintoista

[2]http://www.nb.no/Tilbud/Samlingen/Samlingen/Boeker/Bokhylla.no

se, että tekstianalyysin lisäksi mukaan oli otettu kuvan tunnistamiseen liittyviä tekniikoita: kirjoittajia pyrittiin tunnistamaan tekstaamiseen liittyvien parametrien avulla (mm. musteterän kaltevuus paperilla). Mårtenson korosti, että uusilla menetelmillä on helppoa saada runsaasti dataa, mutta sen merkityksen analysointi vaatii vielä ihmisen tulkintaa erityisesti historiallisissa aineistoissa. Hän mainitsi myös tarpeesta kouluttaa humanistisia aloja ymmärtäviä atk-alan taitajia (ja päinvastoin).

Historiallisten aineistojen digitointi ja näiden aineistojen analysointi näyttääkin olevan yksi keskeisistä digitaalisen humanismin kohteista tällä hetkelle. Konferenssissa esiteltiin erilaisia digitoinnin tekniikoita: mm. käsikirjoitusten sanelu tietokoneohjelmalle, joka muuttaa sanelun tekstiksi. Uppsalan yliopiston kirjasto on tällä tavoin digitoinut kirjeitä varsin hyvin tuloksin. Ludvig Zeevaert Islannin yliopistosta korosti puolestaan omassa esityksessään jo edellä mainittua eri tekstiversioiden luomista. Tietotekniikka mahdollistaa vanhojen tekstien kopioinnin kuvana, niiden muuttamisen digitoiduksi transkriptioksi ja transkription normalisoinnin nykylukijalle ymmärrettäväksi tekstiksi. Tanskan kuninkaallinen kirjasto – Jonas Nielsen – puolestaan esitteli sanakirjojen linkittämistä normalisoinnin vaihtoehtona: lukija ohjataan heidän kokoelmissaan suoralla linkillä vanhan kielen sanakirjoihin nykykielelle harvinaisemmissa termeissä ja käsitteissä.

Tekstien louhinta ja sen haasteet

Tekstin louhinnan tekniikoita käytetään nykyisin ensinnäkin kokoelmien ja kokonaisuuksien hahmottamisessa. Becke Stegman Kööpenhaminan yliopistosta esitteli hanketta, jossa tietokoneen avulla tunnistettiin digitoiduista, hajallaan olevista arkistokäsikirjoituksista yhteneviä piirteitä ja muodostettiin siten tutkijoille ja lukijoille korpuksia. Hänen mukaansa nykyisillä tekniikoilla päästään jo 80 prosentin tarkkuuteen tekstien yhdistämisessä kokoelmiksi. Lisäksi näitä tekniikoita on voitu hyödyntää korruptoituneiden dokumenttien rekonstruoinnissa.

Andrew Salway Bergenin Uni Research -tutkimusryhmästä puolestaan korosti sitä täysin uutta mahdollisuutta, jonka suuret digitoidut tekstimassat mahdollistavat: tietotekniikan avulla on mahdollista löytää sellaista, mihin ihmisen kyvyt ja ominaisuudet tekstien analysoinnissa ei riitä. Tätä hän kutsui datalähtöiseksi tekstianalyysiksi.

Heidän ryhmänsä oli lähestynyt kaunokirjallisia tekstejä tällä tekniikalla ja pyrkinyt löytämään datanlouhinnan avulla tekstien sisäisiä,

Kokoelmatyö ja dokumentaatio	**Datan digitointi ja kuratointi**	**Tutkimusmenetelmien kehittäminen**
-luotettavien digitaalisten kokoelmien luominen eri aineistotyypeistä -metadatatyö -käyttöliittymät ja niiden kehittäminen	-standardointityö -eri versioiden – painettujen ja digitaalisten hallinta ja dokumentointi -pitkäaikaissäilytys -laadun varmistaminen	-tiedonhaun ja hallinnan tekniikoiden kehittäminen -kokoelmien, dokumenttien ja niiden rakenteiden analysointi -kuvantaminen ja mallintaminen -valmiiden tutkimusmenetelmäsovellusten kehittäminen

Taulukko 1. Digitaalinen humanismi ja informaatiotutkimus

tilastollisesti toistuvia rakenteita. Hänenkin mukaansa tällaiset tekniikat ovat vasta kehittymässä. Tällä hetkellä näyttää siltä, että tällaisella analyysilla voidaan löytää rakenteita, joita voidaan tarkastella sitten perinteisen, inhimillisen tulkinnan välinein. Esimerkkinä hän käytti kaunokirjallisten perusrakenteiden (henkilöt, paikat, teemat) analysointia ja niiden ympärille kasautuvia, tilastollisesti merkittäviä termiluokkia, joiden avulla voidaan sitten jatkossa analysoida kirjoittajan tekstejä.

Humanistisissa aineistoissa näyttää vielä vahvasti siltä, että datanlouhintaa voidaan tehokkaasti käyttää niissä tilastollisissa ja tekstillisissä rakenteissa, jotka ovat helposti kuvattavissa ja analysoitavissa matemaattisesti. Tietokoneiden siirtyminen tekstien varsinaiseen tulkintaan näyttää edelleen olevan kesken.

Digitaalinen humanismi, sen mahdollisuudet ja haasteet informaatiotutkimukselle

Konferenssi painottui pitkälti vanhoihin teksteihin ja niiden digitointiin ja mahdollisuuksiin tutkimukselle. Postereissa oli esimakua myös tulevasta: nykyaikaisen kulttuurisen aineiston analysoinnista uusilla tekniikoilla ja dokumenttien monimuotoisuuden tunnistamista. Digitaalinen humanismi, sen tutkimuskohteet ja käytettävät menetelmät ovat siis kehittymässä nopeasti.

Kirjastoille ja informaatiotutkimukselle alalla on paljon tutkittavaa ja yhteistyömahdollisuuksia. Näitä on avattuna taulukossa 1. Kirjastotieteen perusasiat tulevat uudelleen muotiin: digitaalisten kokoelmien järjestäminen, dokumentaatio ja kokoelmatyö kaipaavat selvästi alamme osaamista – toki päivitettynä tietotekniikan tarjoamiin uusiin mahdollisuuksiin.

Informaatiotutkimuksen alan koulutus voisi olla myös mukana luomassa alalle maisteriohjelmia, joissa yhdistettäisiin vahva humanistinen osaaminen tietotekniseen ja informaatiotutkimuksen osaamiseen - tämäkään ei ole mitään uutta alallamme. Suurimman haasteen – ja monitieteisen – tarjoaa suurten datamassojen louhinnan menetelmien kehittäminen. Tätä kautta on selvästi saatavissa aivan uusia tutkimustuloksia ihmistieteisiin ja siitä on tulossa myös kirjasto- ja informaatioalan perusosaamista.