RAJAPINNOILLA-TEEMANUMERO

Informaatiovuorovaikutus historian tutkimuksessa: tiedonlähteiden käytön ymmärtämisestä käytön tukemiseen

Heikki Keskustalo
Tampereen yliopisto
heikki.keskustalo@tuni.fi
https://orcid.org/0000-0003-1415-373X

Elina Late
Tampereen yliopisto
elina.late@tuni.fi
https://orcid.org/0000-0002-3232-1365

Laura Korkeamäki
Tampereen yliopisto
laura.korkeamaki@tuni.fi
https://orcid.org/0000-0002-0985-1333

Sanna Kumpulainen
Tampereen yliopisto
sanna.kumpulainen@tuni.fi
https://orcid.org/0000-0002-7016-257X

Kimmo Kettunen
Itä-Suomen yliopisto
kimmo.kettunen@uef.fi
https://orcid.org/0000-0003-2747-1382

Pysyvä osoite: https://doi.org/10.23978/inf.107890

In information science, the user-oriented research tradition focuses on information seeking in the context of human behavior, while the system-oriented tradition focuses more narrowly on the information retrieval. These traditions still form largely disparate research fields. As an attempt to bridge these traditions, we present a metasynthesis of studies on information interactions in the historical domain, which is related to an ongoing research project. We approach our research task by focusing on both the cognitive space of human actors with task-specific information needs requiring interpretative close-reading and reasoning, and the document space containing potentially relevant pieces of information for the task at hand. We first study information needs and interactions in real work tasks of historians to understand the desired conceptual access points into relevant information (cognitive space). Then we study historical sources from the point of view of task-specific needs at the level of text (document space). We utilize the task-based information interaction (TBII) model to conceptualize the different types of activities during interaction which are important to consider. We believe that this approach is required to learn to support complex task-based information needs extending beyond topicality.

Asiasanat: tiedonhaku, informaatiovuorovaikutus, historiantutkimus, käyttäjät, järjestelmät, kognitiivinen avaruus, dokumenttiavaruus, tehtäväperusteinen informaatiovuorovaikutus, tiedontarpeet

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Johdanto

Teknologinen kehitys muuttaa ihmisten toimintatapoja ja informaatiovuorovaikutuksen käytäntöjä syvällisesti. Historiantutkimuksen alalla digitoitujen aineistojen määrä kasvaa jatkuvasti ja niiden saavutettavuus avaa uusia tutkimusmahdollisuuksia tutkijoille. Digitaalinen historiantutkimus on suuntaus, joka tutkii mennyttä hyödyntäen uutta teknologiaa ja laskennallisia menetelmiä aineistojen tuottamisessa, analyysissä ja jakamisessa (Salmi, 2021, 7). Tässä yhteydessä tehtävien suoritusprosessi, käytetyt tutkimusmenetelmät ja jopa tutkimuksen tavoitteet voivat muuttua. Muuttuneista käytännöistä on toistaiseksi vielä vähän tutkimustietoa ja kokonaisuutta pitäisi ymmärtää aiempaa paremmin.

Digitaalisia historiallisia kokoelmia kehitetään usein data- ja järjestelmävetoisesti. Aineistoja ja digitaalisia työvälineitä kuitenkin käytetään moninaisilla, jopa yllättävillä tavoilla, joita kokoelman kehittäjät eivät välttämättä ole pystyneet kuvittelemaan. Esimerkiksi yksittäisen aineiston data- tai järjestelmälähtöinen tarkastelu ei paljasta erilaisten aineistojen yhdistämisen ja yhteiskäytön tarpeita. Siksi näkökulma, jossa tarkastellaan aineistojen

roolia tutkimusprosesseissa ja kuinka niitä hyödynnetään erilaisiin tarpeisiin, on tärkeä. Tämä näkökulma auttaa ymmärtämään paremmin käytön konteksteja ja suunnittelemaan teknologiaa näkökulmista, jotka tukevat ihmisiä työtehtävien tavoitteiden saavuttamisessa.

Käyttäjäkeskeinen tiedontarve- ja hankintatutkimus sekä järjestelmäkeskeinen tiedonhaun tutkimus ovat erillisiä tutkimustraditioita, jotka kohtaavat käytännössä harvoin, vaikka kumpikin pyrkii auttamaan ihmisiä tietointensiivisten tavoitteidensa saavuttamisessa. Tässä artikkelissa pyrimme yhdistämään näitä lähestymistapoja tarkastelemalla sekä ihmisten tiedontarpeita (kognitiivisen avaruuden tarkastelu) että dokumenttien sisältämiä informaatiopalasia (dokumenttiavaruuden tarkastelu) tiedontarpeiden näkökulmasta. Esitämme artikkelissa metasynteesin tutkimuksista, jotka toteutettiin Suomen Akatemian rahoittamassa EVOLUZ-tutkimusprojektissa vuosien 2019–2021 aikana. Tarkastelemme ihmisten informaatiovuorovaikutusta monipuolisten digitoitujen historiallisten aineistojen kanssa historiantutkimuksen tekemisen kontekstissa. Pääkysymykset ovat seuraavat:

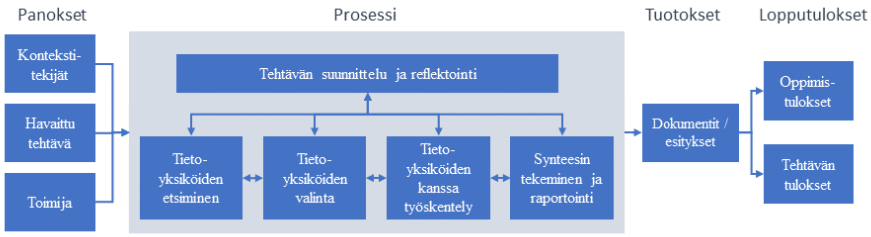
1. Millaista on tehtävälähtöinen informaatiovuorovaikutus historian tutkimuksen työtehtävissä?
2. Millaisia informaatiovuorovaikutuksen esteitä tutkijat kohtaavat pyrkiessään käyttämään digitoituja aineistoja?
3. Kuinka relevantit pääsykohdat aineistoihin ilmenevät käyttäjän näkökulmasta (kognitiivinen avaruus)?
4. Kuinka pääsykohdat ilmenevät aineistoissa tekstin tasolla (dokumenttiavaruus)?
5. Kuinka käyttäjälle tarjottujen informaatiopalasten laatu vaikuttaa käyttäjän kokemukseen aineiston hyödyllisyydestä?

Tutkimusmenetelminä käytettiin osallistavan suunnittelun menetelmiä, tutkijoiden haastatteluita, käyttötapojen demonstraatioita, digitaalisen primääriaineiston sisällönanalyysiä sekä koehenkilöiden suorittamia simuloituja työtehtäviä. Ihmisiä tarkasteltiin tutkimusasetelmassa itsenäisinä toimijoina, joiden tietoresurssien käyttö nähdään osana yksilöiden intentionaalista toimintaa, erotuksena järjestelmäkeskeiselle tarkastelulle, jossa tutkimuskohde redusoidaan järjestelmän käyttöön. Tarkastelemme seuraavassa kysymyksiä yksityiskohtaisemmin.

Tehtäväperusteisen informaatiovuorovaikutuksen evaluointimalli

Informaatiovuorovaikutuksella tarkoitetaan vuorovaikutusta ihmisen ja informaation välillä (Fidel, 2012, s. 7). Sen tutkiminen lisää ymmärrystä inhimillisestä informaation hyödyntämisestä ja auttaa muun muassa kehittämään parempia tiedonhakujärjestelmiä sekä rakentamaan siltaa käyttäjälähtöisen ja data- ja järjestelmälähtöisten tutkimusotteiden välille. Tässä tutkimuksessa käytetään kokoavana näkökulmana *tehtäväperusteisen informaatiovuorovaikutuksen* (task-based information interaction, TBII) evaluointimallia (Järvelin et al., 2015). Malli jäsentää kokonaisvaltaisesti työtehtävien, kuten esimerkiksi oppimistehtävien, informaatiovuorovaikutusta. Malli kiinnittää huomiota vuorovaikutuksen monimuotoisuuteen ja kannustaa pohtimaan kausaalisuhteita prosessin panosten, sen eri vaiheiden sekä lopputulosten välillä.

Vuorovaikutusprosessi koostuu viidestä aktiviteetista, jotka mallinnetaan erillisinä, vaikkakin käytännössä ne ovat iteratiivisia ja toisiinsa limittyviä (kuvio 1). Ensimmäinen aktiviteetti on *tehtävän suunnittelu ja reflektointi*. Suunnittelu pitää sisällään esimerkiksi tavoitteiden asettamisen tehtävälle ja sopivien tehtävän suoritustapojen valinnan. Yksilö monitoroi suoritustaan läpi tehtäväprosessin reflektoiden onnistumistaan prosessin eri vaiheissa. Toinen aktiviteetti, *tietoyksiköiden etsiminen*, pitää sisällään tehtävän suorittamisen kannalta relevantin informaation etsinnän. Käytännössä tämä voi merkitä esimerkiksi tiedonhakuja aineistoista ja niiden selailua pyrittäessä vastaamaan erityyppisiin tiedontarpeisiin. Kolmantena on *tietoyksiköiden valinta*, jossa yksilö arvioi informaation relevanssia suhteessa tiedontarpeisiin ja voi tallentaa informaatiota myöhempää käyttöä varten. Neljäs aktiviteetti, *tietoyksiköiden kanssa työskentely*, sisältää monenlaisia toimintoja, kuten lukemisen, tiedon organisoinnin ja analysoinnin. Viidenteen, *synteesin tekemiseen ja raportointiin* lukeutuvat opittujen asioiden yhdistäminen, uuden tiedon luominen sekä tehtävän suorituksen lopputuloksista viestiminen esimerkiksi artikkelina tai lopputulosta kuvaavana esityksenä.



Kuvio 1. Tehtäväperusteisen informaatiovuorovaikutuksen evaluointimalli (Järvelin et al., 2015).

Tässä tutkimuksessa TBII-malli toimii yhteisenä kielenä pyrittäessä yhdistämään historiallisten aineistojen informaatiovuorovaikutuksen tarkastelua käyttäjän näkökulmasta ja järjestelmän näkökulmasta. Malli nostaa etualalle vuorovaikutusprosessin monimuotoisuuden ja haastaa pohtimaan, kuinka eri aktiviteetteja voitaisiin tukea tietojärjestelmän tasolla.

Historiallinen päättely ja primäärilähteet

Historiantutkimuksen tavoitteena on luoda johdonmukaisia kertomuksia, jotka perustuvat historiallisiin aineistoihin ja tietoihin. Menetelmää kutsutaan historialliseksi päättelyksi (Kuhn, Weinstock, & Flaton, 1994). Historiallinen päättely on prosessi, johon kuuluvat historiallisten kysymysten esittäminen, kontekstualisointi, argumenttien rakentaminen esitettyjen väitteiden tueksi, lähteiden käyttö, informaation organisointi historiallisten ilmiöiden selittämiseksi, sekä metodologisten ja historiallisten käsitteiden hyödyntäminen (Van Drie & Van Boxtel, 2008; vrt. Autio, Katajala-Peltomaa, & Vuolanto, 2001). Tutkimuksemme osa-alueita yhdistävä tavoite on oppia tukemaan historiallista päättelyä toimittaessa nykyaikaisessa digitaalisessa hybridi-informaatioympäristössä. Historiallisissa tutkimusprojekteissa käytetään monentyyppisiä tutkimusaineistoja. Tämän monipuolisuuden valottamiseksi kuvailemme moninaisten aineistojen käyttöön liittyviä haasteita historiantutkimuksen todellisten tutkimusprojektien työtehtävissä.

Työtehtävissä pyritään hyödyntämään aineistoja tyypillisesti siten, että ne edistävät käsillä olevan tehtävän etenemistä. Tutkimustehtävissä usein työtehtävien tavoitteena on vastata valitulla menetelmällä tutkimuskysymyksiin sekä niistä johdettuihin osatehtäviin (vrt. Byström & Hansen, 2005; Byström & Kumpulainen, 2020). Etsiessään ja valitessaan tietoyksiköitä ihmiset pyrkivät

hyödyntämään informaatioresursseja tietyistä käsitteellisistä pääsykohdistä käsin, joita voidaan kutsua kognitiivisiksi pääsykohdiksi tietoon (Ingwersen, 1992). Näiden pääsykohtien ymmärtäminen ja mallintaminen informaatioympäristöihin auttaa historiantutkijoita pääsemään vuorovaikutukseen haluamiensa informaatiosisältöjen ja -palasten kanssa.

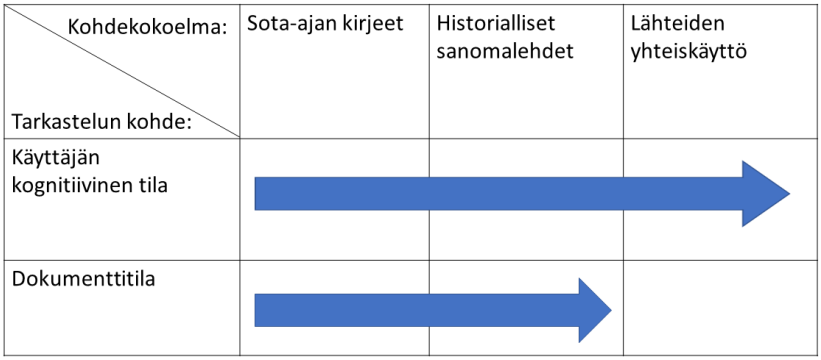
Tässä tarkastelussa erottelimme kaksi keskeistä käsitettä – inhimillisen **kognitiivisen avaruuden** sekä **dokumenttiavaruuden**, joiden välistä kuilua pyritään kuromaan umpeen käsitteellisesti. Kognitiivisessa avaruudessa on tiedontarpeita, joihin etsitään vastauksia. Kognitiiviset pääsykohdat tietoon ovat ne käsitteet, käsitekokonaisuudet tai informaatiopalaset, joiden kautta tiedonhakija haluaa ja olettaa pääsevänsä käsiksi tietoon. Dokumenttiavaruudessa on erilaisia informaatiopalasia, tietoyksiköitä sekä metadataa, jotka sisältävät sekä suoranaista informaatiosisältöä että vihjeitä, joiden avulla vastauksia voidaan päätellä tai rakentaa. Digitaalisen ja laskennallisen historiantutkimuksen menetelmät, kuten optinen tekstintunnistus, nimettyjen entiteettien tunnistus ja sanojen frekvenssianalyysi auttavat osaltaan rakentamaan siltaa käyttäjän kognitiivisen avaruuden ja dokumenttiavaruuden välille. Kuitenkin algoritminen ongelmanratkaisu edelleen vaatii usein rinnalleen manuaalisia työvaiheita, esimerkiksi työskenneltäessä roskaisten, monentyyppisiä virheitä sisältävien aineistojen parissa (Jarlbrink, 2020).

Hakumenetelmät nähdään usein osana prosessia, jonka tavoitteena on välittää lähettäjän tarkoittama viesti vastaanottajalle. Digitaalisten tietoaineistojen käytön näkökulmasta tämä näkemys on kuitenkin liian rajallinen (Kumpulainen et al., 2020). Esimerkiksi historiallisia tekstejä tulkitaan näkökulmista, jotka eivät suoraan vastaa alkuperäisen kirjoittajan tarkoitusperiä (ks. esim. Taskinen, 2021). Laajojen digitointiprojektien ja esimerkiksi käsin kirjoitetun tekstin automaattisten tunnistusmenetelmien edistyessä (Muehlberger et al., 2019) digitoituja historiallisia aineistoja on tutkijoiden saatavilla yhä laajemmin. Perinteiset hakujärjestelmät tukevat dokumenttien hakua vapaita hakusanoja, asiasanastoja ja taksonomioita käyttäen. Monia tutkijan hakutarpeita on kuitenkin vaikea kuvata yksittäisillä hakusanoilla, tai edes määritellä käsitteellisesti (Oberbichler et al., 2021). Käyttäjälle silti olisi tärkeää saada tukea monenlaisiin aktiviteetteihin, joihin voi liittyä aineistojen yhteiskäyttöä, tekstin lähilukua sekä inhimillistä päättelyä ja tulkintaa. Näiden tukemiseksi tulisi ymmärtää käyttäjän tiedontarpeita ja kuinka kognitiivisen avaruuden viitoittamia pääsykohtia voisi havaita dokumenteissa. Viime kädessä pyrkimyksenä on oppia muodostamaan hyödyllisiä ”kahvoja” dokumenttiavaruuteen, esimerkiksi tunnistamalla relevantteja tietoyksiköitä automaattisesti ja annotoimalla niitä sopivalla metadatalla.

Tarkastelemme tässä tutkimuksessa käyttäjien kognitiivista avaruutta ja teemme selkoa historiantutkijan toimintaan vaikuttavista tiedontarpeista. Dokumenttiavaruutta tarkastelemme kahden tekstiaineiston avulla (historialliset kirjeet ja lehtileikkeet) kognitiivisen avaruuden määrittämien pääsykohtien valossa. Historiallisten kirjeiden pääsykohtiin liittyy tarve tunnistaa kirjeenvaihdon osapuolten ominaisuuksia. Historiallisia lehtileikkeitä puolestaan tarkastelemme simuloitujen työtehtävien määrittämän tiedontarpeen valossa.

Tutkimusasetelma

Tutkimme ihmisten informaatiovuorovaikutusta historiallisten aineistojen kanssa kognitiivisen avaruuden ja dokumenttiavaruuden näkökulmista. Kognitiivista avaruutta tarkastelimme tutkimalla informaatiovuorovaikutusta informaatiointensiivisissä työtehtävissä, käytettyjä tiedonlähteitä sekä tiedonlähteiden käyttötapoja, aineistojen hyödyntäjän näkökulmasta (käyttäjälähtöinen tutkimusote). Dokumenttiavaruutta tarkastelimme reflektoimalla kokoelmien piirteitä suhteessa tiedettyihin kognitiivisiin pääsykohtiin (kokoelmalähtöinen tutkimusote). Jäsennys on esitetty kuviossa 2.



Kuvio 2. Tutkimuksen tarkastelun kohteet sekä käytetyt kokoelmat.

Kognitiivista avaruutta tutkittaessa keskiössä on ymmärtää informaatiovuorovaikutusta työprosessien eri vaiheissa ja erilaisten hakutyyppien vaikutusta hyödyllisyyden kokemukseen. Tarkastelimme historian tutkijoiden tehtävälähtöisiä tiedontarpeita ja niiden tukemista sekä informaatiovuorovaikutusta historiallisten aineistojen kanssa. Tunnistimme tässä yhteydessä

sekä aineistojen yhteiskäytön tarpeita että tiedonhaun näkökulmasta haastavia hakuaiheita. Tutkimuksen fokus tarkentui tutkimuksen edetessä siten, että ensimmäisessä vaiheessa saatiin geneeristä ymmärrystä tiedontarpeista ja informaatiovuorovaikutuksesta, toisessa fokusoiduttiin enemmän yhteen aineistoon ja sen kanssa työskentelyyn.

Siirryttäessä tarkastelemaan dokumenttiavaruutta, aiemmista vaiheista valikoitui tutkimuskohteiksi sukupuoleen viittaavien ilmaisujen esiintyminen historiallisissa kirjeissä, sekä se, kuinka koneellisesti tuotettujen tekstien (OCR-)laatu vaikuttaa tiedonhaun onnistumiseen. Taulukko 1 esittää poikkileikkauksen metasynteesimme kohteena olevien osatutkimusten tutkimusongelmista, tutkimusmenetelmistä sekä käytetyistä digitaalisista primääriaineistoista.

Taulukko 1. Metasynteesin kohteena olevat tutkimusongelmat, tutkimusmenetelmät ja digitaaliset primääriaineistot.

Tutkimus-ongelmat	**Tutkimus-menetelmät**	**Digitaaliset primääriaineistot**	**Alkuperäis-tutkimus**
Millaista on informaatio-vuorovaikutus digitaalisten aineistojen kanssa?	Yhteistyötapaamisten havainnointi Tutkijahaastattelut Aineistojen käytön demonstrointi	Historialliset sanomalehdet Sotakirjekokoelma Kaatuneiden tietokanta Sota-ajan valokuvakokoelma	Korkeamäki ja Kumpulainen (2019) Kumpulainen et al. (2020) Late ja Kumpulainen (2021) Kumpulainen ja Late (2021)
Millaisia informaatio-vuorovaikutuksen esteitä tutkijat kohtaavat käyttäessään aineistoja?			
Millaisia ovat kognitiiviset pääsykohdat aineistoihin?			
Miten kognitiiviset pääsykohdat ilmenevät tekstissä?	Primääriaineiston sisällönanalyysi	Sotakirjekokoelma	Keskustalo et al. (2021)
Kuinka tekstin OCR-laadun vaihtelu vaikuttaa dokumentin koettuun hyödyllisyyteen?	Kokeellinen tiedonhaun asetelma	Historialliset sanomalehdet	Kettunen et al. (2021)

Kognitiivinen avaruus

Käyttäjälähtöisellä otteella tutkittiin historiantutkijoiden informaatio-vuorovaikutusta, vuorovaikutuksen esteitä ja tapaa ymmärtää aineistoissa

esiintyviä kognitiivisia pääsykohtia. Tämä toteutettiin yhdistämällä haastatteluita ja havainnointiaineistoja aineistokeruussa ja analyysissä. Historiallisten aineistojen kanssa työskenteleviä tutkijoita havainnoitiin 12 osallistavassa yhteistyötapaamisessa syyskuulta 2017 maaliskuulle 2018. Yhdestä kahteen tuntiin kestävissä tapaamisissa tavoitteena oli pääsyn parantaminen historiallisiin aineistoihin sekä niiden rikastaminen ja analysointi. Yhteistyötapaamisista tehtiin yksityiskohtaiset havainnointimuistiinpanot, jotka kirjoitettiin puhtaaksi. Tapaamisia täydennettiin viidellä haastattelulla.

Aineisto analysoitiin aineistolähtöisesti ja samalla teoriaohjautuneesti. Ensin analysoitiin tehtävälähtöiset aktiviteetit valitun kehyksen mukaisesti. Koko prosessin kattamiseksi käytettiin haastatteluaineistoja holistisen ymmärryksen muodostamiseksi siitä, miten informaatioympäristössä työskenneltiin. Toiseksi analysointiin osallistujien kuvailemat kognitiiviset pääsykohdat perustuen haastattelun aikana tehtyihin demonstraatioihin ja yhteistyötapaamisiin. Historiantutkijat käyttivät omia käsitteellistyksiään kuvaillessaan päättelyprosessejaan. Kolmannessa vaiheessa kognitiiviset pääsykohdat eristettiin muusta aineistosta ja koodattiin.

Haastatteluita jatkettiin fokusoimalla tutkimuskohdetta aineistotyypin osalta historian tutkijoiden informaatiovuorovaikutukseen historiallisten digitoitujen sanomalehtien kanssa. Tutkimusaineistoa kerättiin haastattelemalla digitaalisia sanomalehtiä tutkimusaineistona käyttäviä historian tutkijoita (N=13) vuosina 2018–2021. Haastatteluissa hyödynnettiin ns. kriittisten tapausten menetelmää (Critical Incident Technique, Flanagan 1954) ja haastateltavia pyydettiin kuvailemaan käynnissä olevaa tai päättynyttä tutkimusprojektia, jossa sanomalehtiaineistoa käytettiin. Haastattelukysymysten muodostamisessa hyödynnettiin tehtäväperusteisen informaatiovuorovaikutuksen mallia varmistaen, että kaikki mallin aktiviteetit käytiin läpi haastattelun kuluessa. Haastateltavia pyydettiin lisäksi demonstroimaan aineiston käyttötapoja, jolloin haastateltava pystyi paremmin muistamaan ja kuvailemaan aineistojen käyttöä, haastattelijan saadessa paremman käsityksen aineistojen todellisesta käytöstä.

Haastatteluaineistoja analysoitiin sekä teoria- että aineistolähtöisesti. Aineistoa käyttäen kuvailtiin ensin tehtävälähtöisiä informaatiovuorovaikutuksen tapoja mallin eri aktiviteeteissa, minkä jälkeen analysoitiin tutkijoiden kohtaamia informaatiovuorovaikutuksen esteitä. Kerätty aineisto tarjoaa rikkaan kuvan historian tutkijoiden tavoista hyödyntää digitaalisia historiallisia sanomalehtiä primäärilähteinä.

Dokumenttiavaruus

Kokoelmalähtöisessä tutkimusotteessa tarkastelu fokusoitiin kahteen aineistoon – digitoituihin historiallisiin kirjeisiin ja sanomalehtileikkeisiin. Kirjeiden tutkimukseen liittyy usein tarve ymmärtää kirjeenvaihdon osapuolten keskinäistä suhdetta (Keskustalo et al., 2021; kirjeaineistoista tutkimuksen osana tarkemmin ks. Lahtinen et al., 2011). Syvennyimme dokumenttiavaruuden tutkimuksessa sukupuoleen (gender) assosioituviin merkkijonotason johtolankoihin kirjeiden tervehdyksissä: kuinka johtolankoja voi havaita ja tyypitellä – ja voitaisiinko niitä tunnistaa automaattisesti? Annotoimme manuaalisesti 3094 toisen maailmansodan aikaisen kirjeen alku- ja lopputervehdykset ja haimme kirjeitä hakukriteereinä lähettäjän (esim. mies) sekä vastaanottajan (esim. nainen) sukupuoli. Haku perustui hypoteesiin siitä, että alkutervehdys sisältää viitteitä vastaanottajasta ja lopputervehdys viitteitä lähettäjästä. Sukupuolen tunnistus perustui havaittuihin etunimiin ja Omorfi-sovelluksen tarjoamaan gender-metadataan (Pirinen, 2015). Kirjeiden korrekti lähetyssuunta (esim. “mieheltä naiselle”) oli tiedossa historioitsijan ennalta muodostaman tiedon (ground truth) perusteella. Tutkimuksemme ytimessä oli johtolankojen intellektuaalinen analyysi yksinkertaisen nimiperustaisen haun onnistuessa ja epäonnistuessa. Keskeinen kysymys oli, kuinka kirjeen lähettäjän ja vastaanottajan sukupuoli on pääteltävissä dokumenttiavaruudessa ilmenevien merkkijonotason johtolankojen perusteella.

Historiallisten lehtileikkeiden tutkimukseen liittyi halu ymmärtää, kuinka OCR-laatu vaikuttaa siihen, kuinka hyödyllisinä järjestelmän käyttäjät pitävät tarjottuja leikkeitä. Historiallisten leikkeiden laadun vaikutusta tarkasteltiin simuloidussa asetelmassa, jossa haun kohteena olivat Uuden Suomettaren automaattisesti tuotetut digitaaliset lehtileikkeet. Lehteä painettiin 49 vuoden aikana hieman yli 86 000 sivua. Aineistosta on automaattisesti tuotettu 1.46 miljoonaa leikettä PIVAJ-ohjelmistolla (Kettunen et al., 2019a,b). Koehenkilöt hakivat leikkeitä sekä valmiilla kyselyillä että hakuaiheista vapaasti muotoilemillaan omilla kyselyillä. Leiketeksteistä oli hakujärjestelmän tietokannassa kaksi erilaatuista versiota: optisen luvun peruslaatu sekä optisen luvun kohennettu laatu. Hakukoneen haku kohdistui kohennetun laadun indeksiin ja hakujärjestelmä valitsi satunnaisesti käyttäjälle näytettävän leikkeen laadun. Hakijat eivät tienneet leikkeiden OCR-laadun vaihtelevan. Keskeinen kysymys oli, kuinka digitoinnin laatu vaikuttaa lehtileikkeen koettuun hyödyllisyyteen hakijoiden suorittaessa simuloituja työtehtäviä.

Tarkastellut kokoelmat

Haastatellut tutkijat hyödynsivät erilaisia digitaalisia kokoelmia primääriaineistoinaan. Kansalliskirjaston tarjoama digitaalinen historiallinen sanomalehtikokoelma käsittää suomalaisia sanomalehtiä vuodesta 1771 vuoteen 1929. Aineistoon on vapaa pääsy Kansalliskirjaston käyttöliittymän kautta (digi.kansalliskirjasto.fi), jossa aineistoa voi hakea ja mm. koota aineistoa omaan leikekirjaan. Aineisto on ladattavissa tekstimuodossa myös Kielipankista. Kettunen ja Pääkkönen (2018) ovat aiemmin esitelleet sanomalehtikokoelman tarjoamia mahdollisuuksia Informaatiotutkimus-lehdessä. Järvelin et al. (2016) tekee selkoa historiatiedonhaun haasteista käytettäessä suomenkielisiä tekstikokoelmia.

Muita hyödynnettyjä digitaalisia kokoelmia olivat kaatuneiden tietokanta sekä sota-ajan valokuvakokoelma. Ensin mainitusta löytyvät kaikkien sodassa kaatuneiden henkilöiden tiedot (94 673 tietuetta) ja jälkimmäisessä on noin 140 000 kuvaa kuvateksteineen ja kuvailutietoineen.

Historiallisten kirjeiden analyysimme kohteena on toisen maailmansodan (1939–1945) aikainen kirjekokoelma, joka on näyte (N=3094) Tampereen yliopiston Kansanperinteen arkiston laajemmasta digitoitujen sota-ajan kirjeiden kokoelmasta. Aineisto koostuu yksityishenkilöiden kirjeistä koti- ja sotarintaman välillä. Annotoimme kirjeiden alku- ja lopputervehdykset manuaalisesti syksyllä 2018 ja analysoimme kesällä 2020 intellektuaalisesti kirjeiden tervehdysten sisältämiä ilmaisuja. Keskeinen kysymys oli kuinka kirjeen vastaanottajan ja/tai lähettäjän sukupuoli on pääteltävissä tervehdysten sisältämien ilmaisujen perusteella.

Historiallisten lehtileikkeiden tutkimuksen kohdeaineistona oli Uuden Suomettaren (1869–1918) automaattisesti segmentoituja digitoituja lehtileikkeitä (N=1.46 miljoonaa). Testihenkilöt (N=32) hakivat ennalta määriteltyjä hakuaiheita (N=30) ja arvoivat haettujen digitaalisten lehtileikkeiden hyödyllisyyttä perustuen simuloituun työtehtävään (toimintaa taustoittavaan kuvailevaan tarinaan). Hakija käytti jompaakumpaa kahdesta hakujärjestelmäversiosta, jotka poikkesivat toisistaan käyttäjän näkemän lehtileikkeen laadun suhteen (peruslaatu ja kohennettu laatu). Dokumenttiavaruuden tarkastelun kannalta tutkimuksen keskeinen kysymys oli kuinka digitoinnin laatu vaikuttaa käyttäjän kokemaan lehtileikkeen hyödyllisyyteen simuloidussa työtehtävässä.

Tulokset

Kognitiivinen avaruus

Informaatiovuorovaikutus digitaalisten kokoelmien kanssa

Historian tutkijoiden informaatiovuorovaikutuksen tapoja historiallisten digitoitujen aineistojen kanssa tutkittiin sarjassa tutkimuksia (Korkeamäki & Kumpulainen, 2019, Kumpulainen et. al., 2020, Late & Kumpulainen, 2021). Tutkimuksissa keskityttiin erityisesti tehtäviin, joissa digitaalisia aineistoja käytettiin primäärilähteinä. Haastateltavien tutkimusprojektit käsittelivät esimerkiksi sota-aikaa sekä sanomalehtien, kielen ja yhteiskunnallisten ilmiöiden historiallista kehitystä. Haastateltavat hyödynsivät erilaisia digitaalisia kokoelmia niin että osa käytti tutkimuksessa ainoastaan yhtä aineistolähdettä, kun taas osa integroi aineistoa monesta kokoelmasta. Historian tutkimukselle onkin tyypillistä, että aineistoja ei voida luoda, vaan tutkijat käyttävät niitä aineistoja, jotka ovat säilyneet. Osa haastatelluista tutkijoista työskenteli monitieteisissä tutkimusryhmissä, joissa hyödynnettiin laskennallisia tutkimusmenetelmiä. Toiset taas työskentelivät yksin. Haastatteluaineistojen analyysissa informaatiovuorovaikutuksen tapoja tarkasteltiin tehtäväperusteisen informaatiovuorovaikutuksen evaluointimallin eri aktiviteeteissa. Tulokset osoittivat, että informaatiovuorovaikutuksen tavat eri aktiviteeteissa olivat runsaita ja vaihtelivat aktiviteettien välillä.

Kun tarkastelemme tehtävän suunnittelua ja reflektointia, haastatellut tutkijat painottivat, että tutkimusprosessi ei ollut lineaarinen vaan tyypillistä oli hyppiminen aktiviteettien välillä. Suunnitteluvaiheessa lähestymistapa oli usein aineistolähtöinen ja tutkijat selailivat aineistoja löytääkseen uusia kiinnostavia tutkimuskohteita ja tutustuakseen kokoelmiin. Tutkimusryhmien osalta tässä aktiviteetissa pyrittiin löytämään yhteinen tiedonintressi ja jakamaan tehtävät ryhmän kesken. Lisäksi hankittiin pääsy aineistoihin.

Suurin osa tutkijoista muodosti digitaalisista kokoelmista oman aiheenmukaisen kokoelmansa, kun taas osa tutkijoista hyödynsi kokonaisia kokoelmia, jolloin omaa kokoelmaa ei muodostettu. Tutkijoille, jotka muodostivat omaa kokoelmaansa, tietoyksiköiden etsiminen oli usein työläs aktiviteetti, joka saattoi edellyttää viikkojen ja jopa kuukausien työtä. Tutkijat selailivat ja tekivät sanahakuja kokoelmiin käyttäen erilaisia hakutermejä, kuten kirjoittajien, paikkojen ja osastojen nimiä. Toisinaan tutkimusaiheita käsitteleviä aineistoja ei voinut hakea sanahauilla, jolloin haluttu aineisto oli paikallistettava selailemalla. Osa tutkijoista piti kirjaa käyttämistään hakusanoista ja hyödynsi niin sanottuja aputiedostoja tiedonhakujen dokumentointiin.

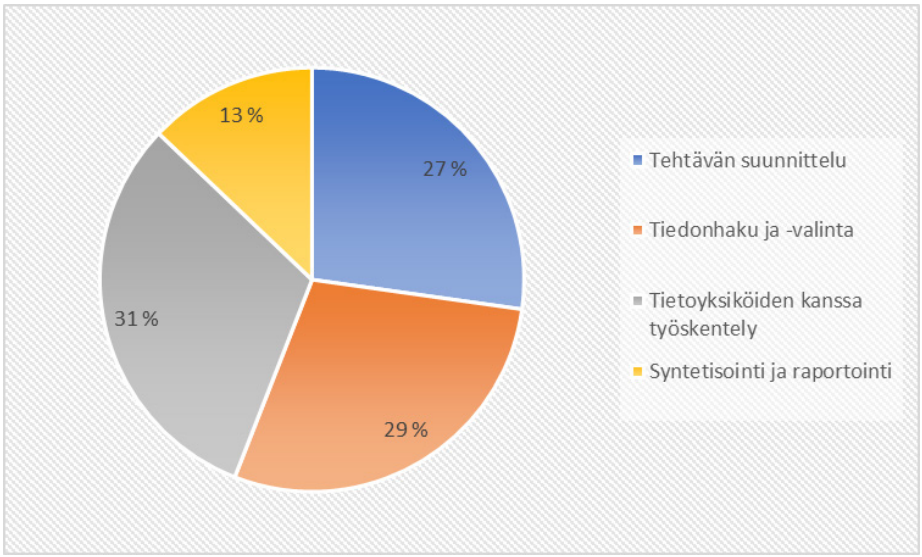
Löytämistään hakutuloksista tutkijat valikoivat omaan kokoelmaansa sopivat aineistot niiden relevanssin perusteella esimerkiksi taulukkolaskentaohjelmistoon. Tiedonhaun ja valikoinnin ohella tutkijat analysoivat aineistoja tehden muistiinpanoja ja kooten esimerkiksi tilastotietoja sanomalehtiartikkeleiden sisällöistä. Aineistojen valinta useimmiten edellytti aineiston silmäilyä ja tarkkaakin lukemista. Tutkijoille myös löydetyn tekstin kontekstin hahmottaminen oli tärkeää. Esimerkiksi sanomalehtien PDF-kuvista tutkijat näkivät artikkelin kokonaisuudessaan ja siihen liittyvät kuvat, mutta koneluetusta tekstistä konteksti ei ollut samaan tapaan nähtävissä. Yksityiskohtainen ja hidas työskentelytapa auttoi historiantutkijoita tutustumaan primääriaineistoihin syvällisesti, eikä muilla keinoilla heidän mukaansa tähän voitu päästä. Tutkimusprosessin laatu oli sen nopeutta tärkeämpää. Toisinaan aineiston valinta oli työlästä prosessien ollessa semidigitaalisia: ne sisälsivät digitaalisia materiaaleja, mutta eivät aina kunnollisia työkaluja, jotka olisivat tukeneet aineistoista hakemista, informaatioyksiköiden valintaa ja tallentamista.

Työskennellessään löydettyjen tietoyksiköiden kanssa tutkijat hyödynsivät erilaisia laadullisia ja määrällisiä menetelmiä. Tämä aktiviteetti kietoutui vahvasti synteesin tekemiseen ja kirjoittamiseen. Lähiluku oli läsnä kaikissa projekteissa, myös niissä, joissa tutkimusote oli laskennallinen. Tiedonhakuja kohdistettiin omaan valikoituun aineistoon. Laskennallisia menetelmiä käytettäessä korostui aineistojen siivoaminen eli muokkaaminen sellaiseen muotoon, jota voitiin käsitellä laskennallisin menetelmin. Esimerkiksi korpuslingvistiikan menetelmät olivat yleisesti käytössä tutkittaessa erilaisten sanojen esiintymistä aineistossa. Tällöin aineistoa perusmuotoistettiin pyrkien huomioimaan analyysissä OCR-virheitä. Lisäksi aineistoja annotoitiin manuaalisesti sekä koneoppimista hyödyntäen. Osa tutkijoista myös hyödynsi analyysissa aineistojen metadataa.

Synteesin ja raportoinnin vaiheessa tutkijat esittelivät tuloksia suullisesti ja kirjallisesti. Yleisimmin tulokset julkaistiin tieteellisissä lehdissä, jolloin keskeistä oli mm. historiallisiin faktoihin perustuvan hyvän tarinan kertominen. Tässä aktiviteetissa tietoyksiköiden hallinta oli keskeistä, jotta tietoihin pystyttiin palaamaan ja viittaamaan tekstissä oikeisiin tietoyksiköihin. Tulosten visualisointi oli useille tutkijoille tärkeää, esimerkiksi digitaalisia karttoja käytettiin visualisoimaan uutisten leviämistä. Suunnitellessaan tulosten raportointia yhteistyössä tutkijat hyödynsivät erilaisia digitaalisia työkaluja. Aineistojen jakaminen voidaan nähdä TBII-mallissa yhtenä tehtävän tuotoksista. Tutkijat suhtautuivat aineistojensa avoimeen jakamiseen positiivisesti, mutta toistaiseksi siihen ei ollut vakiintuneita käytäntöjä.

Informaatiovuorovaikutuksen esteet

Kumpulainen ja Late (2021) analysoivat historian tutkijoiden kohtaamia informaatiovuorovaikutuksen esteitä historiallista digitoitua sanomalehtikokoelmaa hyödynnettäessä. Haastatteluaineistossa artikuloitiin yhteensä 202 estettä, jotka esiintyivät informaatiovuorovaikutuksen eri aktiviteeteissä (Kuvio 3).



Kuvio 3. Informaatiovuorovaikutuksen esteiden jakautuminen eri aktiviteetteihin.

Yli neljännes esteistä ilmeni tehtävän suunnittelun ja reflektoinnin aktiviteetissa, jossa lähes puolet esteistä liittyi itse tehtävään, tutkimusaiheen määrittelyyn, tutkimusprojektin organisointiin ja projektin työntekijöihin. Esimerkiksi monitieteisissä projekteissa tutkimusryhmä käytti paljon aikaa yhteisen tutkimusongelman muotoiluun ja eri tieteenalojen tutkijoiden intressien yhteensovittamiseen. Tutkijat kamppailivat puuttuvan infrastruktuurin ja aineistoihin pääsyn kanssa.

Lähes kolmannes esteistä liittyi tietoyksiköiden etsimiseen ja valintaan. Esteet liittyivät pääosin kokoelman sisältöihin, metadataan ja kokoelman informaatioarkkitehtuuriin. Lähes jokaisessa haastattelussa ongelmana mainittiin OCR:n laatu, joka vaihteli merkittävästi kokoelman sisällä. Joissakin tapauksissa lähes joka toinen merkki aineistossa oli luettu väärin, jolloin aineisto nähtiin lähes lukukelvottomana. Kokoelman monikielisyys ja tutkimusaiheiden luonne myös asettivat haasteita aineistojen hakemiselle.

Lisäksi työkalut, kuten vaikeakäyttöiset käyttöliittymät, aiheuttivat tutkijoille ongelmia.

Noin kolmannes esteistä liittyi tietoyksiköiden kanssa työskentelyyn. Myös tässä aktiviteetissa pääosassa olivat kokoelmaan liittyvät esteet. Merkittävänä pidettiin sitä, että kokoelmasta ja sen muodostumisesta oli huonosti tietoa saatavilla. Tutkijat pohtivat esimerkiksi, mikä oli se "Suomi", jota kokoelma koski. Lisäksi OCR-aineistojen heikko laatu vaikeutti aineistojen kanssa työskentelyä, kuten myös kokoelman rakenne. Sanomalehtikokoelma on muodostettu skannaamalla sanomalehtisivuja kokonaisuudessaan, minkä vuoksi tutkijat joutuivat tekemään paljon manuaalista työtä tunnistaakseen yksittäiset artikkelit aineistosta. Lisäksi metadatan puutteellisuus koettiin esteenä. Työkalujen puute tai vaikeakäyttöisyys aiheuttivat haasteita.

Pienehkö osa havaituista esteistä (13 %) liittyi prosessiaktiviteettiin synteetisointi ja raportointi. Tällöin esteitä tuottivat itse tehtävä sekä sosio-organisationaaliset tekijät. Historia-alan vakiintunut julkaisukulttuuri ei aina tukenut monitieteisiä laskennallisia menetelmiä hyödyntäviä tutkimuksia. Tulosten ja menetelmäkuvausten sovittaminen artikkelimuotoon nähtiin haasteena. Humanistisella alalla tarjolla oleva tutkimusaineistojen ja työkalujen jakamiseen tarkoitettu tutkimusinfrastruktuuri nähtiin puutteellisena ja jopa avoimuutta estävänä.

Kognitiiviset pääsykohdat aineistoihin

Kumpulainen et al. (2020) analysoivat kognitiivisia pääsykohtia historiallisiin digitaalisiin aineistoihin. Kognitiivisilla pääsykohdilla tarkoitetaan käyttäjän tehtävälähtöisiä tiedontarpeita ja haluttuja tapoja päästä käsiksi aineistoihin. Kognitiiviset pääsykohdat heijastavat historian tutkijoiden käsityksiä siitä, kuinka primaarilähteet voivat auttaa tutkimusongelmien ratkaisemisessa. Tunnistamalla kognitiivisia pääsykohtia voidaan tukea historiallista päättelyä dokumenttiavaruudessa ja ehdottaa pääsykohtia tarvittuun tietoon. Tutkimukseen osallistuneet historian tutkijat käyttivät primaarilähteinä sota-ajan kirjeitä, kaatuneitten tietokantaa sekä sota-ajan valokuva-arkistoa.

Historioitsijat käyttivät useita erilaisia käsitteellisiä rakennelmia kognitiivisina pääsykohtina tietoon (Taulukko 2). Tunnistettuja pääsykohtia olivat henkilöt ja heidän roolinsa, sukupuoli ja henkilöiden väliset suhteet sekä henkilöiden ja paikkojen väliset suhteet; sodassa kaatuneet ja kaatuneiden tyypit; organisaatioiden ja paikkojen nimet; sekä ajalliset ilmaisut, kuten päivämäärät, vuodenajat ja vuorokaudenajat.

Taulukko 2. Kognitiiviset pääsykohdat ja esimerkkejä dokumenttiavaruudessa.

Kognitiivinen pääsykohta	**Esimerkkejä dokumenttiavaruudessa**
Henkilöt ja heidän roolinsa	Kuvaaja, lähettäjä/vastaanottaja, sotilasarvo
Henkilöiden väliset suhteet	Etunimi, sukunimi, siviilisääty
Sukupuoli ja sukupuolten väliset erot	Sukupuolta ilmaisevat sanat: etunimi, hellittelysanat, sukulaissuhteet
Sodassa kaatuneet ja kaatuneiden tyypit	Kuollut taistelussa, haavoittunut
Organisaatiot	Kenttäsairaala, kenttäposti
Paikat, paikannimet	Alue, rintama, kotirintama, kuolinpaikka
Ajalliset ilmaisut	Päivämäärä, syntymäaika, vuodenaika

Mahdollinen keino tukea tehtävien suorittamista olisi lisätä aineistoihin rakennetta, joka tekee näkyväksi niissä esiintyviä kognitiivisia pääsykohtia. Esimerkiksi kuva-aineistojen kuvailutekstejä ja kirjeitä voitaisiin rikastaa lisäämällä tekstiin automaattisesti relevanttia metadataa, kuten päivämäärään perustuvaa tietämystä merkittävistä sota-ajan tapahtumista (esim. ”katastrofaalinen tapahtuma”). Kirjeiden alku- ja lopputervehdyksiin sisältyviä vinkkejä kirjeenvaihdon osapuolten rooleista ja heidän välisestä suhteestaan voitaisiin ilmaista automaattisesti lisätyn metadatan avulla (esim. ”avioparin välinen kirje”). Mahdollisten kognitiivisten pääsykohtien ehdottaminen voisi tukea historioitsijaa relevanttien tietoyksiköiden etsinnässä ja valinnassa. Niitä voitaisiin hyödyntää myös aineistojen visualisoinnissa aineistojen valinnan tueksi (esim. sisältääkö aineisto relevanttiin katastrofaaliseen tapahtumaan liittyviä pääsykohtia). Käytännön haasteena onkin historiallisten resurssien rakentaminen tarvittujen pääsykohtien tukemiseksi, esim. tiedonlouhintaa hyödyntäen. Seuraavassa siirrymme tarkastelemaan lähdeaineistoja huomioiden käyttäjien näkökulman teksteihin.

Dokumenttiavaruus

Sukupuolivihjeiden löytäminen dokumenttiavaruudessa

Kognitiivinen avaruus sisältää inhimillisiä informaatioon kohdistuvia toiveita ja tavoitteita. Dokumenttiavaruutta puolestaan määrittää se, millaisia pääsykohtia relevanttiin informaatioon dokumenteissa on tarjolla niiden sisältämien ”informaatiopalasten” kautta. Esimerkkejä on esitetty Taulukossa 2. Edellä kuvasimme historiantutkijoiden kognitiivisia pääsykohtia todellisten

työtehtävien yhteydessä. Seuraavassa tarkastelemme pääsykohtia dokumenttiavaruuden tasolla (Keskustalo et al., 2021). Rajauduimme tarpeeseen tunnistaa kirjeen vastaanottajan ja lähettäjän sukupuoli. Hakukriteerinä käytettiin kirjeen vastaanottajan ja lähettäjän sukupuolten kombinaatiota (F=nainen, M=mies) suunnissa FF (N=340), FM (N=1976), MF (N=575), ja MM (N=161). Lähetetyn kirjeen korrekti suunta oli tiedossa 3052 kirjeelle (42 tapauksessa informaatio oli epätäydellistä). Hakumenetelmä perustui vastaanottajan ja lähettäjän etunimien tunnistukseen kirjeiden alku- ja lopputervehdyksissä hyödyntäen Omorfi-sovellusta, joka tunnistaa nimiin liittyviä sukupuolia. Intellektuaalinen analyysi keskittyi sukupuolen johtolankoihin merkkijonotasolla tarkastelemalla erityisesti, kuinka lähettäjän/vastaanottajan sukupuoli olisi pääteltävissä myös etunimiperustaisen haun epäonnistuessa. Hakujen saanti vaihteli välillä 18.5 % – 30.2 % ja tarkkuus välillä 38.7 – 96.1 %. Hakujen epäonnistumisen tyypillisiä syitä olivat etunimen puuttuminen alku- tai lopputervehdyksestä (tai molemmista), nimen sukupuolen tunnistuksen epäonnistuminen, ja tervehdyksen puuttuminen kirjeestä. Etunimien lisäksi intellektuaalinen analyysi paljasti kirjeissä seuraavia sukupuolen johtolankatyyppejä:

- Sukulaisuustermit (esim. äiti, isä, sisko, veli)
- Lempinimet (esim. Väiski, Mappesi)
- Tittelit (esim. herra, neiti) ja sotilasarvot (alokas, vääpeli)
- Puolisoon viittaavat termit (vaimo, eukko, ukkoni, miehesi)
- Hellittelysanat (Irmeliinus, Mamma)
- Muut sukupuoleen viittaavat termit (sotasisareni, voimamies)

Hakujen onnistuessa sukupuolet tyypillisesti oli tunnistettu korrektisti alku- ja lopputervehdysten sisältämien etunimien avulla, mutta osa relevanteista dokumenteista oli haettu väärästä syystä. Lisäksi hakumenetelmä haki myös epärelevantteja dokumentteja. Virheellisiä hakutuloksia aiheuttivat useampia kuin yhden nimen sisältävät tervehdykset, virheellisesti tunnistetut homografiset ilmaisut sekä OCR-virheet. Esimerkiksi lopputervehdys ”Monin terveisin …” tulkittiin naisen lähettämäksi, tervehdyksen sisältämän nimen (Mona) vuoksi. Sukupuolineutraalit tervehdysilmaisut ”Monin”, ”Toivon”, ”Armas” ja ”Kallein” tulkittiin automaattisessa haussa virheellisesti nimiksi. Lisäksi kirjeissä esiintyvät OCR-virheet aiheuttivat virheitä hakutuloksissa (esim. reservin kersantin hälyisen lyhenteen ”Rea. kers.” tulkittiin sisältävän nimen Rea). (Keskustalo et al., 2021.)

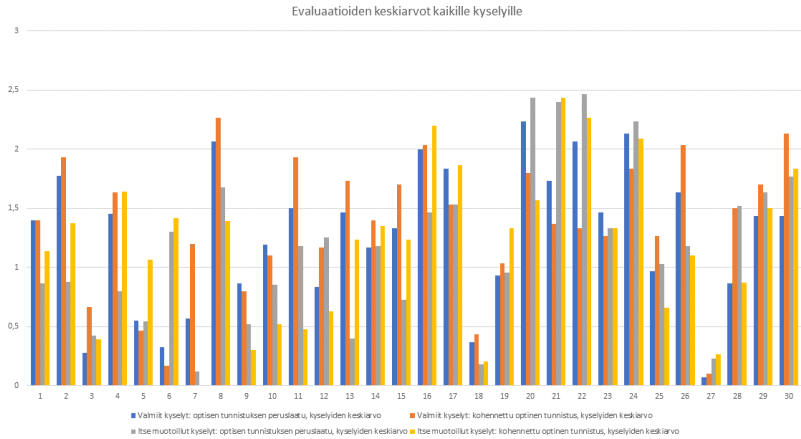
Ihminen kykeni päättelemään kirjeen vastaanottajan ja lähettäjän sukupuolen tyypillisesti assosioimalla tervehdysten sanoja (“vaimoni”, “Eevalle”,

"isä", jne.) yllä kuvattuihin johtolankatyyppeihin, myös OCR-hälystä huolimatta. Toisinaan sananloput kuten possessiivisuffiksi ja allatiivin pääte (-ni ja -lle edellä) antoivat lisävinkkejä henkilöiden välisen suhteen läheisyydestä ja/tai kirjoitussuunnasta. Johtolankojen hakua voitaisiin pyrkiä tukemaan imitoimalla automaattisesti esimerkiksi sukupuolen praktista päättelyä vinkkisanojen avulla (päätelmätyypeistä ja argumentoinnista ks. esim. Kalela, 2020). Kirjeiden sanoja voitaisiin täsmäyttää johtolankatyyppien mukaisesti muodostettuihin leksikkoihin, joihin on sisällytetty relevanttia metadataa. Homografiset ilmaisut voitaisiin huomioida erillisissä sulkusanastoissa. (Keskustalo et al., 2021.)

Digitoinnin laadun vaikutus hyödyllisyyden kokemukseen sanomalehtikokoelmassa

Tarkastelimme historiallisten lehtileikkeiden optisen tunnistuksen laadun vaikutusta koettuun hyödyllisyyteen simuloidussa asetelmassa, jossa hakukohteena olivat Uuden Suomettaren koko historiasta (1869–1918) automaattisesti tuotetut leikkeet (Kettunen et al., 2021). Aineistosta oli automaattisesti tuotettu 1,46 miljoonaa leikettä (käsittäen runsaat 300 miljoonaa sanaa) PIVAJ-ohjelmistolla (Kettunen et al., 2019a,b). Hakutestiä varten oli luotu 30 hakuaiheen ja 30 valmiin kyselyn testikokoelma. Kokeen osallistujat (N=32) hakivat digitoituja leikkeitä sekä valmiilla kyselyillä että hakuaiheista vapaasti itse muotoilemillaan kyselyillä. Leiketeksteistä oli hakujärjestelmän tietokannassa kaksi erilaatuista versiota: optisen luvun peruslaatu sekä optisen luvun kohennettu laatu. Hakukoneen haku kohdistui kohennetun laadun indeksiin ja hakujärjestelmä valitsi satunnaisesti käyttäjälle näytettävän leikkeen laadun. Hakijat eivät tienneet leikkeiden OCR-laadun vaihtelevan. Keskeinen kysymys oli, vaikuttaako digitoinnin laatu käyttäjän kokemaan lehtileikkeen hyödyllisyyteen simuloidussa työtehtävässä.

Arvioitaessa optisen tunnistuksen laadun vaikutusta käyttäjille näytettiin hakutulosten kymmenen relevanteinta leikettä ja pyydettiin arvioimaan leikkeen hyötyrelevanssia asteikolla 0–3. Evaluointi ohjeistettiin simuloidun työtehtävän mallin mukaisesti hakijan ajatellessa hakevansa leikkeitä artikkelin kirjoittamista varten. Kuvio 4 ja taulukko 3 tiivistävät päätulokset. Kuvio 4 esittää hakutulosten keskiarvot hakuaiheittain kumuloituvan hyödyn keskiarvona kymmenen relevanteimman leikkeen joukossa kysely- ja OCR-tyyppien kombinaatioina. Taulukko 3 esittää vastaavat tulokset 30 hakuaiheen keskiarvona.



Kuvio 4. Yksittäisten kyselyiden evaluaatiot kyselytyypin ja OCR-laadun kombinaatioille (kumuloituvan hyödyn keskiarvo kymmenen relevanteimman haetun leikkeen joukossa).

Taulukko 3. Kaikkien kyselyiden evaluaatioiden keskiarvot kyselytyypin ja OCR-laadun kombinaatioittain relevanssiasteikolla 0-3.

Valmiit kyselyt Optisen tunnistuksen peruslaatu, kaikkien kyselyiden evaluaatioiden keskiarvo	**Valmiit kyselyt** Optisen tunnistuksen kohennettu laatu, kaikkien kyselyiden evaluaatioiden keskiarvo	**Itse muotoillut kyselyt** Optisen tunnistuksen peruslaatu, kaikkien kyselyiden evaluaatioiden keskiarvo	**Itse muotoillut kyselyt** Optisen tunnistuksen kohennettu laatu, kaikkien kyselyiden evaluaatioiden keskiarvo
1,26	1,36	1,17	1,19

Kuvasta 4 ja taulukosta 3 nähdään, että erityisesti valmiiksi muotoillut kyselyt hyötyivät kohennetusta tekstien optisesta tunnistuksesta. Valmiiden kyselyiden evaluaatioiden keskiarvo (1,36) on 7,9 % korkeampi kohennetulla optisella tunnistuksella kuin peruslaadulla (1,26). Itse muotoilluissa kyselyissä ero jää 1,7 prosenttiin. Simuloidussa asetelmassa siis havaittiin, että kohennettu optisen tunnistuksen laatu vaikutti selkeän myönteisesti käyttäjän kokemaan hyödyllisyyteen.

Keskustelu ja johtopäätökset

Tämä tutkimus pyrki yhdistämään käyttäjäkeskeistä ja järjestelmäkeskeistä tapaa lähestyä historiallisten aineistojen käyttöä. Yhdistävänä mallina käytettiin tehtäväperusteisen informaatiovuorovaikutuksen mallia, joka nostaa etualalle vuorovaikutusprosessin monimuotoisuuden ja monipuoliset aktiviteetit, joita tulisi ymmärtää ja tukea. Tarkastelimme aineistojen käyttöä kognitiivisen avaruuden ja dokumenttiavaruuden näkökulmasta informaatiointensiivisissä työtehtävissä. Ensin mainittua tutkimme käyttäjälähtöisesti ja jälkimmäistä kokoelmalähtöisesti, kuitenkin reflektoimalla kokoelmien piirteitä suhteessa käyttäjänäkökulmasta nouseviin kognitiivisiin pääsykohtiin.

Informaatiovuorovaikutuksen tapoja historiantutkimuksen työtehtävissä tarkasteltiin haastatellen ja havainnoiden digitaalisia kokoelmia primääriaineistoina hyödyntäviä tutkijoita. Tulokset osoittivat vuorovaikutuksen tapojen olevan monipuolisia ja vaihtelevan eri aktiviteettien kesken. Tyypillistä oli historian tutkimusprosessin epälineaarisuus ja aineistolähtöisyys. Tutkijat muodostivat omia kokoelmia tekemällä tekstihakuja ja selailemalla digitaalisia aineistoja. Aineistojen valikointi selaamalla ja lukemalla koettiin työlääksi, mutta se auttoi tutustumaan kokoelmiin ja hahmottamaan aineistojen kontekstia. Aineistoja analysoitiin laadullisin ja määrällisin menetelmin, lähilukemisen ollessa aina mukana. Analyyseissä hyödynnettiin tiedonhakua käsiteltäessä aineistoja laskennallisin menetelmin. Tutkijat kirjoittivat artikkeleita niin historian kuin teknillisten alojen tieteellisiin julkaisuihin sekä esittelivät tutkimustuloksia suullisesti. Aineistojen jakaminen kiinnosti, mutta siihen ei toistaiseksi ollut olemassa vakiintuneita käytäntöjä.

Informaatiovuorovaikutuksen esteitä tarkasteltiin haastatellen historiallisia sanomalehtiaineistoja käyttäviä tutkijoita. Tulokset osoittavat, että vuorovaikutuksen esteitä esiintyi kaikissa vuorovaikutuksen aktiviteeteissa. Yleisemmin esteet liittyivät käytettyyn kokoelmaan, jolloin OCR-virheet, kokoelman rakenne ja puutteelliset kuvailutiedot vaikeuttivat tutkijoiden työtä. Lisäksi monitieteiseen yhteistyöhön liittyvät sosiokulttuuriset tekijät aiheuttivat esteitä, kun tutkijoiden tuli löytää yhteinen tiedonintressi ja julkaista uusin menetelmin tuotettuja tuloksia eri tieteenalojen julkaisukanavissa.

Kognitiivisia pääsykohtia tietoon eli tehtävälähtöisiä tiedontarpeita ja tapoja päästä käsiksi informaatioon tutkittiin haastattelemalla ja havainnoimalla historian tutkijoita, jotka hyödynsivät tutkimuksissaan eri aineistolajeja. Lisäksi havainnoitiin yhteistyötapaamisia. Tulokset osoittivat, että tarvittavat kognitiiviset pääsykohdat olivat usein erilaisia käsitteellisiä rakennelmia, joiden havaitsemiseen liittyy käytännössä lähilukua ja päättelyä.

Esimerkkejä pääsykohdista ovat henkilöiden roolit ja suhteet, sukupuoli, henkilöiden ja paikkojen väliset suhteet, sekä ajalliset ilmaisut. Haasteena onkin oppia tukemaan tällaisia pääsykohtia hakujärjestelmän tasolla.

Kognitiivisten pääsykohtien ilmenemistä dokumenttiavaruudessa tarkasteltiin kirjeen vastaanottajan ja lähettäjän sukupuolen tunnistamisen näkökulmasta analysoiden 3094 kirjeen alku- ja lopputervehdykset. Tulokset osoittivat, että etunimiperustaisen sukupuolentunnistuksen tulokset olivat vaatimattomat, mutta tervehdyssanojen analyysi paljasti sukupuolisidonnaisten käsitteiden typologian (esim. sukulaisuustermit, tittelit, puolisoon viittaavat nimitykset), jonka perusteella voitaisiin kehittää metadatasanastoja ja hyödyntää niitä kognitiivisten pääsykohtien merkkaamiseen teksteissä loppukäyttäjän tukemiseksi.

Viimeiseksi tarkastelimme historiallisten sanomalehtien digitaalisten leikkeiden OCR-laatua. Käyttäjähaastatteluissa OCR-virheet olivat näyttäytyneet merkittävänä käytön esteenä. Vaikka digitoinnin laatua tarkastellaan tyypillisesti datalähtöisesti, pyrimme tarkastelemaan laadun vaikutusta käyttäjän näkökulmasta. Toteutimme tutkimuksen kokeellisessa asetelmassa, jossa koehenkilöt suorittivat simuloituja hakutehtäviä, näytettyjen digitaalisten lehtileikkeiden kuuluessa kahteen eri laatuluokkaan. Käyttäjätesti osoitti digitoinnin laadun vaikuttavan käyttäjien tekemään arvioon heidän näkemiensä lehtileikkeiden hyödyllisyydestä suhteessa simuloituun työtehtävään.

Tutkimme inhimillistä toimintaa historiantutkimuksen työtehtävissä ja toiminnan tukemista työskenneltäessä moninaisten tietolähteiden parissa. Havaitsimme, että tietolähteiden käyttö on monimenetelmällistä, ulottuen erilaisiin tietoresursseihin, joita integroidaan työprosessin eri aktiviteeteissa niille tunnusomaisilla tavoilla, aineistojen ja välineiden asettamissa rajoissa. Tarkennettaessa asetelmaa yksittäisen lähteen kanssa toimimiseen, toimintatavoista sekä tehtäväprosessien ja vuorovaikutuksen ongelmista saatiin tarkempaa tietoa, mutta samalla menetettiin moninaisten tietolähteiden yhdistämisen näkökulma. Mikäli tukea historioitsijalle halutaan kuitenkin tarjota, voi olla tarpeen valita yksityiskohtaiseen tarkasteluun yksittäisiä tutkimusongelmia. Yksittäisenkään tiedontarpeen ja tietoresurssin vuorovaikutuksen tukeminen ei ole yksinkertaista. Siitä huolimatta opimme tässä tutkimuksessa, miten muodostaa uudenlaisia laadullisia vaatimusmäärittelyitä järjestelmille vastakohtana puhtaasti aineistolähtöisille lähestymistavoille. Tiedonhakujärjestelmien evaluoinnissa tulisi pohjimmiltaan mitata sitä, kuinka hyvin ne auttavat inhimillisen toiminnan tavoitteiden saavuttamisessa.

Tarkastelimme tässä tutkimuksessa informaatiovuorovaikutusta historian tutkimuksen prosesseissa käyttäjien ja aineistojen näkökulmista. Kokonaisuus

on kompleksinen ja sillan rakentaminen kognitiivisen avaruuden ja dokumenttiavaruuden välille on haastavaa, erityisesti kognitiivisten pääsykohtien tunnistamisen perustuessa dokumenttien lähiluvulle ja inhimilliselle päättelylle (Salmi, 2021, 30–33). Aineistojen roolia tutkimusprosessissa ja niiden todellista käyttöä erilaisiin tarpeisiin tulisi ymmärtää ja teknologiaa suunnitella tästä näkökulmasta. Kehitettäessä digitaalisia työkaluja tieteidenvälisenä yhteistyönä on tärkeää tulla tietoiseksi haasteista, joita tieteenalojen keskinäiset erot ja niihin liittyvät osapuolten erilaiset motivaatiot saattavat aiheuttaa (ks. Oberbichler et al., 2021). Kehitystyö edellyttää näkemyksemme mukaan käyttäjä- ja kokoelmalähtöisten tutkimusotteiden yhdistämistä. Tämä johtaisi käyttäjien todellisia tarpeita paremmin tukevien hakujärjestelmien kehittämiseen.

Tutkimus on saanut rahoitusta Suomen Akatemian projektilta numero 326616.

Lähteet

Autio, S., Katajala-Peltomaa, S., & Vuolanto, V. (2001). Ongelma, oivallus ja onnistumisen ilo. Teoksessa S. Autio, S. Katajala-Peltomaa, V. Vuolanto (toim.), *Historioitsijan arki ja tutkimuksen prosessi* (s. 145–155). Vastapaino.

Byström, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the Association for Information Science and Technology, 56*(10), 1050–1061. https://doi.org/10.1002/asi.20197

Byström, K., & Kumpulainen, S. (2020). Vertical and horizontal relationships amongst task-based information needs. *Information Processing & Management, 57*(2), 102065. https://doi.org/10.1016/j.ipm.2019.102065

Fidel, R. (2012). *Human information interaction: An ecological approach to information behavior*. MIT press.

Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin, 51*(4), 327. https://doi.org/10.1037/h0061470

Ingwersen, P., & Järvelin, K. (2006). *The turn: integration of information seeking and retrieval in context*. Springer. https://doi.org/10.1007/1-4020-3851-8

Jarlbrink, J. (2020). All the work that makes it work: Digital methods and manual labour. Teoksessa M. Fridlund, M. Oiva & P. Paju (toim.), *Digital histories: Emergent approaches within the new digital history* (s. 113–126). Helsinki University Press. https://doi.org/10.33134/HUP-5-7

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., & Kettunen, K. (2016). Information retrieval from historical newspaper collections in highly inflectional languages: a query expansion approach. *Journal of the Association for Information Science and Technology, 67*(12), 2928–2946. https://doi.org/10.1002/asi.23379

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., . . . Sormunen, E. (2015). Task-based information interaction evaluation: The viewpoint of program theory. *ACM Transactions on Information Systems (TOIS), 33*(1), Article 3. https://doi.org/10.1145/2699660

Kalela, J. (2020). *Historiantutkimus ja historia*. Gaudeamus.

Keskustalo H., Korkeamäki L., Vanamo S., Kettunen K., & Kumpulainen S. (2021). Analyzing Gender Clues in War Time Letters. Arvioitavana.

Kettunen, K., Ruokolainen, T., Liukkonen, E., Tranouez, P., Anthelme, D., & Paquet, T. (2019a). Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771–1929: Early Results Using the PIVAJ Software. DATeCH 2019.

Kettunen, K., Pääkkönen, T., Liukkonen, E. (2019b). Clipping the Page – Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Historical Journalistic Collection. Teoksessa A. Doucet et al. (toim.), *Digital Libraries for Open Knowledge. TPDL 2019* (s. 356–360). Springer. https://doi.org/10.1007/978-3-030-30760-8_33

Kettunen, K., Keskustalo, H., Kumpulainen, S., Rautiainen, J., & Pääkkönen, T. (2021). OCR quality affects perceived usefulness of historical newspaper clippings. Arvioitavana.

Korkeamäki, L., & Kumpulainen, S. (2019). Interacting with Digital Documents: A Real Life Study of Historians' Task Processes, Actions and Goals. Teoksessa *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)* (s. 35–43). Association for Computing Machinery. https://doi.org/10.1145/3295750.3298931

Kuhn, D., Weinstock, M., & Flaton, R. (1994). Historical reasoning as theory-evidence coordination. Teoksessa M. Carretero & J. F. Voss (toim.), *Cognitive and instructional processes in history and the social sciences* (s. 377–401). Erlbaum.

Kumpulainen, S., Keskustalo, H., Zhang, B., & Stefanidis, K. (2020). Historical reasoning in authentic research tasks: Mapping cognitive and document spaces. *Journal of the Association for Information Science and Technology, 71*(2), 230–241. https://doi.org/10.1002/asi.24216

Kumpulainen, S., & Late, E. (2021). Struggling with Digitized Historical Newspapers: Contextual, Activity-Related and Collaborative Barriers to Information Interaction in History Research Tasks. Arvioitavana.

Lahtinen, A., Leskelä-Kärki, M., Vainio-Korhonen, K., & Vehkalahti, K. (2011). Kirjeiden uusi tuleminen. Teoksessa M. Leskelä-Kärki, A. Lahtinen ja K. Vainio-Korhonen (toim.), *Kirjeet ja historiantutkimus* (s. 9–27). Suomalaisen Kirjallisuuden Seura.

Late, E., & Kumpulainen, S. (2021). Interacting with digitised historical newspapers: Understanding the use of digital surrogates as primary sources. *Journal of Documentation*. https://doi.org/10.1108/JD-04-2021-0078

Muehlberger, G., Seaward, L., Teras, M., Oliveira, S. A., Bosch, V., Bryan, M., . . . Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation, 75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Oberbichler, S., Boros, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., . . . Tolonen, M. (2021). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, 2021: 1–15. https://doi.org/10.1002/asi.24565

Pirinen, T. A. (2015). Development and use of computational morphology of Finnish in the open source and open science era: notes on experiences with Omorfi Development. *SKY Journal of Linguistics, 28*, 381–393. http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Pirinen.pdf (viitattu 4.8.2021).

Salmi, H. (2021). *What is digital history?* Polity Press.

Taskinen, I. (2021). *Social lives in letters: Finnish soldiers' epistolary relationships, intimate practices, and emotionality in World War II*. Väitöskirja. Tampereen yliopisto.

van Drie, J. & van Boxtel, C. (2008) Historical reasoning: towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review, 20*(2), 87–110. https://doi.org/10.1007/s10648-007-9056-1