Julkaistavaksi hyväksytty versio

Olli-Pekka Kasurinen, Lauri Leinonen ja Sanna Supponen

Modernien tutkimustietokantojen ja tiedonjalostusmenetelmien mahdollisuudet historiantutkimukselle
Tapaus Diplomatarium Fennicum

Historiantutkimusta tehdään lähteiden ehdoilla ja niiden saavutettavuus on välttämättömyys tutkimukselle. Aineistojen saavutettavuuden ja käytettävyyden keskeisiä haasteita on pyritty ratkaisemaan jo vuosisatojen ajan lähde-editioiden systemaattisella julkaisemisella. Nykyään painettujen lähdejulkaisujen rinnalle ovat tulleet editio- ja tutkimustietokannat. Modernit tutkimustietokannat ja tiedonjalostusmenetelmät tuovat merkittäviä uusia mahdollisuuksia historiatieteelle sekä uudenlaisten analyysimenetelmien että lähdevarantojen muodossa. Tässä artikkelissa tarkastelemme, miten nämä menetelmät parantavat aineistojen saavutettavuutta ja käytettävyyttä. Tapausesimerkkinä käytämme Suomen keskiaikaisten asiakirjojen tietokannan, *Diplomatarium Fennicumin* (DF), parissa tehtyä kehitystyötä.[1]

Painetut editiot ja nykyään myös niiden digitoidut versiot ovat historiantutkimuksen perusta. Suomen keskiajantutkimuksen keskeisen asiakirjakorpuksen muodostavat valtionarkistonhoitaja Reinhold Hausenin (1850–1942) julkaisemat *Finlands Medeltidsurkunder I–VIII* (FMU, 1910–1935) ja *Registrum Ecclesiae Aboensis eller Åbo Domkyrkas Svartbok* (REA, 1890). Tutkijat tukeutuvat Hausenin editioihin edelleen, vaikka niiden virheet ja puutteellisuudet ovat yleisessä tiedossa.[2] Hausenin editiot ovat myös DF:n perusta. Ensimmäinen DF-tietokanta, joka kehitettiin vuosituhannen vaihteessa, toi Hausenin FMU-editiosarjan tekstit sekä osan REA-editiojulkaisusta sähköisessä muodossa verkkoon.[3] Tietokannan yksinkertaiset hakuominaisuudet osoittautuivat

[1] Artikkeli keskittyy pääasiassa olemassa olevien lähdevarantojen käytettävyyden parantamiseen, eikä siinä oteta laajemmin kantaa esimerkiksi lähteiden digitointiin tai digitaalisten editioiden tuottamiseen ja niiden haasteisiin. Näistä teemoista, Mats Dahlström, Critical editing and critical digitisation. Teoksessa Wido Peursen (toim.) *Text Comparison and Digital Creativity. The Production of Presence and Meaning in Digital Text Scholarship.* Brill 2010.

[2] "Som urkundsutgivare kan det ibland förefalla som om han satte kvantitet före kvalitet, hans register var inte alltid genomarbetade, hans dateringar ibland vågade och ofta funderar man vad som egentligen motiverat tryckningen av källmaterialet. Hans styrka var att kombinera olika typer av källor med varandra." Lena Huldén, Reinhold Hausen och källutgåvorna. Teoksessa Elisa Orrman (toim.) *Reinhold Hausen (1850–1942). Kansallisen arkiston rakentaja.* Kansallisarkisto 2000, 42.

[3] DF-tietokannan varhaisemmista vaiheista, ks. Seppo Eskola & Lauri Leinonen, Suomen keskiajan asiakirjalähteistä uusi verkkopalvelu. *Historiallinen Aikakauskirja* 114 (2016), 332–336; ks. myös ensimmäisen DF-hankkeen tuottama

nopeasti riittämättömiksi. Pelkkien editiotekstien lisäksi tietokantaan kaivattiin lisätoiminnallisuuksia. Aloite DF-tietokannan jatkokehittämiseen nousikin tutkijayhteisön tarpeista.

DF-tietokanta uudistettiin kokonaisuudessaan Koneen säätiön tukemassa Kansallisarkiston hankkeessa vuosina 2015–2018.[4] Uuden tietokannan runkona käytettiin aiemman tietokannan sisältämiä editiotekstejä, minkä päälle aineistoa laajennettiin ja jalostettiin. Lisäksi tietokanta uudistettiin teknisesti. Tavoitteena oli korvata vanha ’tekstivaranto’ modernilla tutkimustietokannalla, jossa lähdeaineistoon lisätään sitä selittävää metadataa sekä integroidaan tutkimusvälineitä, kuten kartta- ja muita sovelluksia. Hankkeessa sovellettiin tietokoneavusteisia menetelmiä sekä perinteisempää tutkimusta, joilla tuotettiin asiakirjoille niiden käsittelyä ja tutkimusta helpottavaa metadataa. Tietokannan rakenteesta luotiin aiempaa editiohistoriaa ja tutkimusta syntetisoiva: yksittäiseen numeroituun tekstikokonaisuuteen[5] pyrittiin yhdistämään kaikki saatavilla oleva tieto, kuten muut editiot, tutkimuskirjallisuus, lähteiden digitaaliset jäljenteet ja linkitys muihin tietokantoihin. Lisäksi Hausenin editioissaan antamia tietoja korjattiin ja täydennettiin. Tietokantaan myös lisättiin uusia asiakirjoja.[6] Kehitystyön myötä DF-tietokanta on erkaantunut FMU:sta ja kasvanut määrällisesti sitä laajemmaksi siten, että yhtäsuuruusmerkkejä näiden kahden kokonaisuuden väliin ei voi enää vetää. Lopputuloksena syntyi tutkimusinfrastruktuuri, joka tarjoaa uusia tapoja käsitellä, luokitella ja tutkia Suomen keskiaikaisia asiakirjoja.

Tässä artikkelissa tarkastelemme modernin tutkimustietokannan kehitystä ja sen haasteita sekä mahdollisuuksia tutkimukselle ja tuleville tietokantahankkeille. Esimerkit nostamme DF-tietokanta -projektissa tehdystä työstä: mihin kehitystyössä keskityttiin, miksi, ja miten perinteisistä editioista jalostettiin tutkimustietokanta. Samat kysymykset, jotka DF-projekti on joutunut ratkaisemaan,

---

kuvaus *Suomen keskiaikaa koskevien dokumenttien julkaiseminen*, df.narc.fi/Images/Information/DF_artikkeli_historia.pdf (8.8.2018).

[4] Hankkeessa työskentelivät projektipäällikkö FM Seppo Eskola, tutkija FM Lauri Leinonen sekä ICT-suunnittelija MSc Denis Mandrov; Talvella 2017–2018 aineiston läpikäyntiin ja metadatan tuottamiseen osallistuivat FM Olli-Pekka Kasurinen, FM Sanna Supponen ja FT Ville Walta. DF-projektin ohjausryhmään kuuluivat tutkimusjohtaja Päivi Happonen (Kansallisarkisto), dosentti Tuomas Heikkilä (Helsingin yliopisto), dosentti Anu Lahtinen (Turun yliopisto, Helsingin yliopisto), professori Marko Lamberg (Tukholman yliopisto), apulaisprofessori Kirsi Salonen (Turun yliopisto), lehtori Minna Sandelin (Turun yliopisto), lehtori Seija Tiisala (Helsingin yliopisto), kehittämispäällikkö Vili Haukkovaara (Kansallisarkisto), kehittämispäällikkö Anne Wilenius (Kansallisarkisto) ja ylitarkastaja Yrjö Kotivuori (Kansallisarkisto).

[5] FMU sisältää 6714 numeroitua tekstikokonaisuutta. DF:ssa säilytettiin FMU:n mukainen numerointi, joka numeroi asiakirjat 1–6726 (hypäten 12 numeron yli). Lisäksi REA:ssa on oma numerointinsa, joka asiakirjojen jaottelun osalta toisinaan poikkeaa FMU:sta: asiakirja, joka on FMU:ssa yhden numeron alla, voidaan antaa REA:ssa useampana numeroituna kokonaisuutena.

[6] DF-projektissa lisättyjen uusien asiakirjojen numerointi aloitettiin FMU:n numeroinnin jatkeeksi DF 6727:stä.

koskevat pitkälti myös muita historian alan tutkimustietokantoja ja lähdevarantoja.[7] Aluksi tarkastelemme aineiston ajantasaisuutta ja saavutettavuutta koskevia kokonaisuuksia, sitten aineiston käytettävyyden parantamista metadatan lisäämisellä ja hakutyökaluilla, ja lopuksi pohdimme tietokannan jatkokehitysmahdollisuuksia. Suhteutamme hankkeen tuloksia erityisesti muihin keskiaikaista aineistoa sisältäviin pohjoismaisiin tietokantoihin ja niiden toteutustapoihin sekä ajantasaiseen tutkimukseen.

Aineiston ajantasaisuus ja saavutettavuus

Lähde-editioiden ja niihin perustuvien tietokantojen yhtenä tarkoituksena on tehdä aineistosta helpommin saavutettava kokoamalla yhteen eri arkistoissa ja kokoelmissa sijaitsevia asiakirjoja. Editioiden käyttö säästää alkuperäisaineistoa kulumiselta, mutta auttaa myös varmistamaan aineiston tekstisisältöjen säilymisen niissä tapauksissa, joissa alkuperäislähde katoaa tai tuhoutuu. Myös DF:n aineistoon kuuluu paljon sellaista materiaalia, joka tunnetaan nykyään vain myöhempien kopioiden tai editioiden pohjalta. Vaikka painetut editiot ja niiden sähköiset versiot ovat näin edesauttaneet lähteiden saavutettavuutta, editioiden käytössä on edelleen useita ongelmia.

Samoja asiakirjoja on esimerkiksi editoitu erilaisin perustein eri lähdejulkaisuihin. Kuhunkin editioon ovat vaikuttaneet tutkijan tulkinnat, julkaisustandardit ja editointikäytännöt.[8] Kaikkien eri versioiden hankkiminen ja rinnakkain vertailu on tutkijalle työlästä. Toinen painettujen editioiden ongelma on se, että ne voivat vanhentua sisällöltään tutkimuksen osoittaessa vääräksi esimerkiksi aiempia ajoituksia, paikannuksia ja lukutapoja. Korjaukset tai uusien löydettyjen asiakirjojen lisääminen kokonaisuuksiin vaativat erillisten lisäosien tai uuden korjatun painoksen teettämistä. Editioiden puutteiden ja epätäydellisyyden takia alkuperäislähteiden ja eri editioversioiden lukeminen rinnakkain on usein välttämätöntä. Digitaaliset tutkimusinfrastruktuurit ovat pyrkineet tuomaan näihin ongelmiin parannusta. DF:ssa tutkijoille tarjotaan näkyville useita eri editioversioita

[7] Samanlaisia kysymyksiä on pohdittu muun muassa Ruotsin diplomatoriumin suhteen, esim. Claes Gejrot, Diplomatarium Suecanum in the digital world. *Proceedings from the Symposium "Virtually Medieval"? Mirator* (2005), https://journal.fi/mirator (18.9.2018). Norjan diplomatariumiin liittyviä kysymyksiä on pohdittu Christian-Emil Smith Ore, Datateknologi og gamle brev. Diplomatarium Norvegicum i elektronisk form. Teoksessa Claes Gejrot, Roger Andersson & Kerstin Abukhanfusa (toim.) *Ny väg till medeltidsbreven. Från ett medeltidssymposium i Svenska Riksarkivet 26–28 november 1999*. *Skrifter utgivna av Riksarkivet 18*. Riksarkivet 2002. Ks. myös Mikko Piippo, Uusia keskiajan lähdejulkaisuja – ja lähdejulkaisun uusia tuulia. *Historiallinen Aikakauskirja* 101 (2003), 306–308.
[8] Dahlström 2010, 80–83.

aineiston tulkinnan luotettavuuden parantamiseksi.[9] Samalla tutkimushistoria tulee monipuolisesti näkyväksi.

Moderneissa tutkimustietokannoissa, kuten DF:ssa, tarjotaan alkuperäisestä lähteestä sekä editio että digitaalinen jäljenne rinnakkain, jolloin materiaalin tarkastelu on helpompaa.[10] Digitaalisen jäljenteen ja edition tarkastelu rinnakkain auttaa tutkijaa saamaan parhaan mahdollisen kuvan lähteen sisällöstä ja siitä tehdyistä tulkinnoista. Milloin digitaalista jäljennettä ei ole saatavilla, tieto lähteen arkistosijainnista on keskeinen. Yksi painettujen editioiden ongelmakohdista ovat nimenomaan vanhentuneet tiedot aineiston arkistosijainneista. Tästä johtuen DF-projektissa tarkistettiin ja päivitettiin lähteiden arkistosijainteja[11] sekä yhdistettiin asiakirjoja laajempiin kokonaisuuksiin ja muihin editiosarjoihin. Esimerkiksi suora kytkös *Svenskt Diplomatariumin* editioihin luotiin *Svenskt Diplomatariums Huvudkartotek* (SDHK) -tietokannan kautta.[12] Osaan asiakirjoista lisättiin myös käännöksiä kokoteksteistä ja regestoista.[13] Svenska Kulturfondenin rahoittamassa *Nådendalsdiplomen*-sisarhankkeessa tuotettiin kokonaan uusia, nykyiset kielitieteelliset kriteerit täyttäviä, kriittisiä editioita Hausenin jo aiemmin editoimista teksteistä.[14] Asiakirjakorpusta myös laajennettiin lisäämällä asiakirjoja Vatikaanin paavillisen katumustuomioistuimen eli Penitentiariaatin arkistosta[15] ja aiemmin editoimattomia

[9] DF tietokantaan sisällytettiin editiot mm. seuraavista kokoelmista: A. I. Arwidsson, *Handlingar till upplysning af Finlands häfder.* Norstedt 1846–1857; Edward Grönblad, *Nya källor till Finlands medeltidshistoria.* [E. Grönblad] 1857. Uusimpien editioiden lisäämisen esteenä ovat usein tekijänoikeuskysymykset, minkä vuoksi DF:aan liitetyt editiot ovat pääsääntöisesti varhaisia.

[10] Vrt. Espen S. Oren artikkeli, jossa pohditaan erityyppisten editioiden rinnakkaista käyttöä. Espen S. Ore, Monkey Business – or What is an Edition? *Literary and Linguistic Computing* 19:1 (2004), 35–44.

[11] Hausenin editioiden lähdeviitteet ovat usein epämääräisiä tai vanhentuneita, eivätkä auta tutkijaa löytämään asiakirjaa nykypäivänä. DF-hankkeessa keskityttiin selvittämään Hausenin antamien tietojen pohjalta Suomessa sijaitsevien alkuperäislähteiden nykysijainnit. Lähes kaikki pystyttiin paikallistamaan ja päivittämään tietokantaan. Ulkomailla sijaitsevista lähteistä paikallistettiin mahdollisuuksien mukaan säilyttävä arkisto ja arkistokokoelma, jos yksikön tasolle ei päästy.

[12] Ruotsin SDHK-tietokanta on syntynyt keskiaikaisten asiakirjojen kortistojen digitalisoinnin kautta, johon myöhemmin on yhdistetty *Diplomatarium Suecanumin* editiotekstit. Claes Gejrot, Swedish Charters Online. The Digitization of Diplomatarium Suecanum. Teoksessa Georg Vogeler (toim.) *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*. Böhlau Verlag 2009.

[13] Regestalla tarkoitetaan lyhyttä referaattia asiakirjan keskeisestä sisällöstä. Mikko Piippo, Diplomatiikka. Teoksessa Marko Lamberg, Anu Lahtinen & Susanna Niiranen (toim.) *Keskiajan avain.* SKS 2009, 408. Käännettyjä regestoja lisättiin REA:n vuoden 1996 faksimilepainoksesta, jossa annetaan englanninkieliset regestat asiakirjoista: *Registrum Ecclesiae Aboensis eller Åbo domkyrkas Svartbok – The Black Book of Abo Cathedral*. National Archives of Finland 1996. Lisäksi projekti laati suomenkieliset regestat sekä uudelleen että ensi kertaa editoituihin asiakirjoihin. Suomenkielisiä regestoja lisättiin myös teoksesta Mikko Juva, Vilho Niitemaa & Päiviö Tommila (toim.) *Suomen historian dokumentteja* 1. Otava 1968. Käännökset asiakirjoista olivat yleisöpalautteessa yleisimmin toivottu lisä tietokantaan. Ne ovatkin erittäin tärkeitä tiedon saattamisessa suuren yleisön saataville sekä opetuskäyttöön.

[14] Nådendalsdiplomen-hankkeessa editioita tuottavat Ft Viveca Rabb (Åbo Akademi), FT Oliver Blomqvist (Uppsala universitet, Riksarkivet) sekä FT Mikko Kauko (Turun yliopisto). Uudet editiot on julkaistu DF-tietokannassa.

[15] Penitentiariaattiarkiston asiakirjat on julkaistu painettuna editioina ja SDHK-tietokannassa. Sara Risberg & Kirsi Salonen, *Auctoritate papae. The church province of Uppsala and the apostolic penitentiary 1410–1526*. Riksarkivet 2008.

pergamenttikirjeitä Tallinnasta[16]. Moderneihin tietovarantoihin onkin mahdollista lisätä uusia editioita ja asiakirjoja suhteellisen helposti. Uusi tieto julkaistaan jatkoksi jo olemassa olevaan resurssiin, eikä erillisenä julkaisuna, mikä parantaa käytettävyyttä.

Digitaaliset resurssit ovat edellä kuvatuista, usein käytettävyyteen ja saavutettavuuteen palautuvista syistä jo syrjäyttämässä vanhoja painettuja editioita. Tuore kyselytutkimus suomalaisten humanistitutkijoiden digitaalisten metodien käytöstä osoittaa, että digitaalisia aineistoja käytetään kattavammin kuin painettuja. Erityisesti arkistoaineistoihin tutustaan yhä useammin digitaalisen version kautta. Lähes kolme neljäsosaa tutkimuksen vastaajista kertoi käyttävänsä digitoituja arkistoaineistoja. Vastaajat kokivat nimenomaan aineistojen saavutettavuuden (*accessibility*) tärkeimmäksi vaatimukseksi digitaalisen ajan tutkimukselle.[17] Tulevaisuudessa tuleekin miettiä, miten uusia lähteitä kannattaa julkaista.[18] Digitaalisten editioiden saavutettavuus ja käytettävyys on lähtökohtaisesti painettua teosta parempi.

Metadata

Aineistojen saavutettavuuden rinnalla toinen keskeinen kehityskohde oli aineiston käyttökelpoisuuden parantaminen, sillä FMU:n kuten monien varhaisten editioiden hakemistot ja asiasanaluettelot ovat varsin usein puutteellisia tai lähinnä viitteellisiä.[19] Ensimmäisessä DF-tietokantaprojektissa oli jo puututtu tähän ongelmaa ja tuotettu metadataa, mutta osa tästä oli korruptoitunutta eikä siksi täysin palvellut tarkoitustaan.[20] Lisäksi FMU:n sisältämä materiaali on hyvin sekalaista – muodon, tyypin ja alkuperän suhteen – ja vaati siksi erilaista käsittelyä kuin

[16] Tallinnan asiakirjat (9 kpl) editoi DF-projektille FT Tapio Salminen (Tampereen yliopisto).
[17] Tutkimukseen vastasi 239 eri vaiheen tutkijaa kaikista kahdeksasta Suomen yliopistosta, joissa on humanistisen alan tohtoriohjelma. Inés Matres, Mila Oiva & Mikko Tolonen, In between research cultures – the state of digital humanities in Finland. *Informaatiotutkimus* 37:2 (2018), 39–42.
[18] Mm. Vujošević et al. pohtivat kannattaisiko painetuista editioista luopua kokonaan ja käyttää kaikki voimavarat suoraan digitaalisten editiotietokantojen tekemiseen. Žarko Vujošević, Nebojša Porčić & Dragić M. Živojinović, Das serbische Kanzleiwesen. Die Herausforderung der digitalen Diplomatik. Teoksessa Antonella Ambrosio, Sébastien Barret & Georg Vogeler (toim.) *Digital Diplomatics. The Computer as a Tool for the Diplomatist? AfD-Beiheft 14*. Böhlau Verlag 2014. Vrt. Dahlström 2010.
[19] Hakemistojen puutteista esim. Piippo 2003, 308; Huldén 2000, 32, 38–39.
[20] Uusi DF-hanke peri edeltävältä hankkeelta tietokantarakenteen (SQL) ja metadatamallin jota projektissa jatkojalostettiin. DF:n metadata on räätälöity projektin tarpeisiin. Metadata on jaoteltu SQL-tietokannassa omiin tauluihinsa (antopaikat, antajatiedot, jne.). Jatkokehityksessä erityisesti muihin tietokantoihin yhdistämisen kannalta tietokanta ja metadata tulisi muuttaa kansainväliseen standardin, esimerkiksi SDHK:n ja Codices Fennicin käyttämän TEI/XML-mallin mukaiseksi. TEI/XML-mallin mahdollisuuksista ks. Camile Desenclos & Vincent Jolivet, Diple, propositions pour la convergence de schémas XML/TEI dédiés à l'édition de sources diplomatiques. Teoksessa Antonella Ambrosio, Sébastien Barret & Georg Vogeler (toim.) *Digital Diplomatics. The Computer as a Tool for the Diplomatist?* AfD-Beiheft 14. Böhlau Verlag 2014.

puhtaasti diplomataarinen asiakirjakokoelma.[21] Kenties olennaisin sisällön ja käytettävyyden kehitystyö tapahtuikin tietokannan metadatan parissa.

Metadatalla tarkoitetaan tässä yhteydessä käytännössä kaikkea itse asiakirjoja kuvailevaa, selittävää, yhdistävää tai tukevaa dataa. Näihin lukeutuvat niin paikka- ja henkilötiedot, ajoitukset, kielitiedot, asiakirjojen kategorisoinnit kuin asiasanatkin. Metadatan kautta massiivinen ja sekalainen aineisto voidaan jäsentää eri tavoin, mikä edesauttaa huomattavasti tutkijoiden työtä. Metadata auttaa muodostamaan aineiston sisältä osajoukkoja yhdistäviä kokonaisuuksia ja kehittää niihin tukeutuvia tutkimustyökaluja. Siinä missä perinteisen editiosarjan käyttö rakentuu pitkälti hakemistojen selailulle ja lähiluennalle, tietokanta mahdollistaa metadatan avulla monia muita tiedon haku- ja käsittelytapoja. Esimerkkinä voidaan mainita käsiteltävän aineiston rajaaminen tietyn aikavälin tai asiakirjan sisällön mukaan.

Ensimmäinen metadatan kehityskohde olivat asiakirjojen antopaikkatiedot, jotka ovat tutkimuksen kannalta hyvin keskeisiä. Sen lisäksi, että paikkatiedot auttavat sijoittamaan asiakirjoja sekä niissä mainittuja henkilöitä maantieteellisesti, ne kytkevät asiakirjat laajempiin kokonaisuuksiin ja auttavat rakentamaan lähteiden tulkintakontekstia ja tapahtumanarratiivia. Hausen panostikin paljon asiakirjojen antopaikkojen tunnistamiseen. Tyypillisesti hän antaa ne asiakirjojen otsikkotiedoissa yhdessä ajoituksen kanssa. Uudessa DF-tietokannassa tietoja täydennettiin ja tarkistettiin.[22] Tarkistuksen lisäksi paikannimet geokoodattiin[23] käyttäen sekä koneellista paikantunnistusta[24] että käsin tehtävää koordinaattien hakua. Historiallisten paikkojen etsiminen koneellisesti osoittautui odotetun haasteelliseksi, sillä paikannimet vaihtelevat ja tiedot ovat usein epätarkkoja. Suomen keskiaikaisista paikannimistä ei myöskään ole sähköisessä muodossa saatavilla kattavia rekistereitä tai ontologioita. Ongelmia aiheuttivat erityisesti samannimiset paikat sekä pienten kylien ja tilojen nimet, jotka vaativat tunnistukseen laajempaa tutkimusta. Nykyinen samanniminen paikka ei aina vastaa keskiaikaista paikkaa: esimerkiksi Kokemäen Pilton tila sijaitsi keskiajalla Kokemäenjoen varrella, mutta nykyinen Pilton nimellä tunnettu tila sijaitsee siitä yli viiden kilometrin päässä

[21] Hausen poimi FMU:hun kaikkea Suomea koskevaa materiaalia eikä tehnyt erotusta maininnan alkuperän suhteen. Tämän vuoksi materiaalia on kerätty esimerkiksi kronikoista, esineiden kaiverruksista, ja muista ei-diplomataarisista lähteistä. Toisaalta Hausen on poiminut eriaikaisista kopiokirjoista, koontipöytäkirjoista ja jäljennöksistä mainintoja asiakirjoista. Huldén 2000, 39, 42.
[22] Antopaikoista erotettiin paikan nimi, paikan vaihtoehtoinen nimi (Tallinna/Reval), paikkaan liittyvän paikan nimi (usein pitäjä), paikan tyyppi (linna, luostari), sekä maatieto (pääosin modernien maiden rajojen mukaisesti).
[23] Geokoodaus tarkoittaa tietyn paikan yhdistämistä tiettyyn karttapisteeseen.
[24] Ajo tehtiin Google Mapsin avoimesta rajapinnasta VBA-skriptillä.

keskellä peltoja.[25] Paikkojen tunnistamisessa korostui lopulta tapauskohtainen tutkimustyö. Manuaalisen listan läpikäynnin jälkeen lähes kaikki antopaikat saatiin kuitenkin tunnistettua ja sijoitettua kartalle, ja näin käyttäjä pystyy saamaan kattavan kuvan aineiston maantieteellisestä jakautumisesta. Muita samankaltaisia keskeisiä tietoja asiakirjoissa ovat ajoitus- ja kielitiedot, jotka myös tarkistettiin niissä havaittujen puutteiden ja virheiden vuoksi.[26]

Toinen keskeinen kehityskohde olivat varsin puutteellisina ja toisinaan vääristyneinä säilyneet antajatiedot. Tämä johtui siitä, että Hausen on nostanut asiakirjoissa mainitut suomalaiset usein regestaan, vaikka he olisivat olleet vain asian todistajina. Samalla asiakirjan varsinainen antaja jäi usein nimeämättä. Koska varhaisemmassa DF-versiossa antajatiedot oli poimittu suoraan regestoista, saattoi asiakirjan antajaksi päätyä sivuroolissa ollut henkilö. Antajatiedot tarkistettiin asiakirja kerrallaan ja tietoihin kerättiin antajan nimitiedot ja määritteet, jotka esiintyvät asiakirjassa.[27] Nimi poimittiin omaksi tekstikentäkseen, josta on erotettu henkilöön liittyvät paikanmääreet ja tittelit, sillä samaan henkilöön liitetyt määreet saattavat vaihdella ajan tai kontekstin mukaan (esim. *riddare*/*fogde* tai *hustru*/*änka*). Henkilöille annetut määreet valikoituivat asiakirjoissa esiintyvien termien mukaan, joskin niiden kirjoitusasua nykyaikaistettiin ja yhdenmukaistettiin.[28] Jollei tarkempaa antajaa ilmennyt, merkittiin asiakirja sen antaneen instituution nimiin. Raatien tili- ja pöytäkirjojen katkelmat merkittiin täten raatien nimiin ja kuninkaiden ja paavien kanslioiden dokumentit asianomaisten kanslioiden nimiin. Aineistossa on myös ei-diplomataarista materiaalia sekä lukuisia myöhäisempiä, esimerkiksi tuomiokirjoihin sisältyviä viitteitä kadonneisiin keskiaikaisiin asiakirjoihin, joista antajatietoja ei pystytty poimimaan. Osalla teksteistä ei täten ole selkeää antajaa, jolloin antajaksi merkittiin ”ei määritetty”. Lopputuloksena asiakirjoista poimittiin yhteensä 9541 antajatietoa, jotka voidaan jatkojalostaa kattavaksi henkilörekisteriksi. Lisäksi DF-tietokannan sisältämää dataa olisi mahdollista hyödyntää tutkittaessa näiden henkilöiden välisiä suhteita, jos halutaan tehdä kattava prosopografinen tai verkostotutkimus Suomen keskiajasta.[29]

[25] Kyseinen tapaus ei olisi paljastunut ilman Kokemäen historiaa syvällisesti tuntevan tutkijan apua. Kaisa Kyläkoski, Rauhankadulta Kokemäen kautta Ruotsiin. *Sukututkijan Loppuvuosi* -blogi, sukututkijanloppuvuosi.blogspot.com/2017/11/rauhankadulta-kokemaen-kautta-ruotsiin.html (14.8.2018).
[26] Ajoitus- ja kielitietoja tarkistettiin muun työn yhteydessä sitä mukaa kuin tarkempia tietoja tai korjauksia tuli vastaan. Milloin FMU:ssa oli vain regesta, asiakirjan kielitiedot tarkistettiin muista editioista, kuten *Liv-*, *Esth- und Curländisches Urkundenbuch*ista (1853–1914).
[27] Tiedot poimittiin ensisijaisesti editiosta ja FMU:n lisäksi tietoja kerättiin myös muista editioista, erityisesti SDHK:sta. Jos editiota ei ollut saatavilla, käytettiin regestan tietoja.
[28] Keskiaikaisista ammattinimikkeistä ei ole olemassa “standardirekisteriä”, jota voisi käyttää pohjana. Sen vuoksi hankkeessa toimivat tutkijat tekivät linjaukset yhtenäistämisistä ja ammattinimikkeiden tulkinnasta tapauskohtaisesti.
[29] DF:n koko tietokanta ei ole kokonaisena, rakenteistetussa muodossaan toistaiseksi ladattavissa verkkokäyttöliittymän kautta, vaan aineisto tulee erikseen pyytää toimituskunnalta. Henkilötietoja sekä tietoja henkilöiden rooleista on kerätty

Kolmas projektissa toteutettu metadatakokonaisuus oli kattava asiasanoitus, jollaista ei ole juurikaan aiemmin tehty pohjoismaisissa diplomataareissa. Mallia haettiin muun muassa *Diplomatarium Danicumista*. Siinä tehty asiasanoitus todettiin kuitenkin riittämättömäksi ja liian hajanaiseksi DF:n tarpeisiin.[30] DF:n asiasanat muodostettiin aineiston perusteella mahdollisimman selkeiksi erillisiksi käsitteiksi. Asiasanoitusta tehtiin sekä antajien että asiakirjan tyypin mukaan.[31] DF:n asiasanat eivät ole toisiaan poissulkevia kategorioita, vaan asiakirjalla voi olla useampi päällekkäinen määrite.[32] Käyttökelpoisuuden vuoksi asiasanottamisessa jouduttiin kuitenkin tekemään suhteellisen yleisiä kategorioita ja yksittäistapauksien linjavedot olivat toisinaan haastavia. Siinä missä varsinaisiin antajatietoihin merkittiin tarkkaan asiakirjan tai regestan antama tieto nimestä ja tittelistä, asiasanoitus vaati antajan mukaan enemmän tulkintaa siitä, missä roolissa antaja on toiminut asiakirjaa antaessaan. Sama henkilö on voinut esimerkiksi valtaneuvoksena allekirjoittaa keskushallinnon asiakirjan, antaa tuomion paikallishallinnon edustajana tai antaa huomenlahjan tulevalle vaimolleen yksityishenkilönä. Yksittäisellä asiakirjalla saattaa olla myös useita antajia, esimerkiksi kuningas ja valtaneuvoston jäsenet, jolloin asiakirja on merkitty useampaan rinnakkaiseen kategoriaan. Asiakirjan tyypin mukainen asiasanoitus vaati alkuun asiakirjojen jakamista kahteen luokkaan, ei-diplomataarisiin ja diplomataarisiin, sillä kuten edellä mainitaan DF sisältää muutakin kuin yksiselitteisiä asiakirjoja. Ei-diplomataariseen aineistoon sisältyy muun muassa kronikkamerkintöjä, hagiografista materiaalia sekä erilaisia esineissä olevia kirjoituksia, kuten hautakivien muistokirjoituksia tai kirjojen omistusmerkintöjä. Toisaalta myös varsinaiseen diplomataariseen aineistoon sisältyy erityyppisiä kokonaisuuksia, sillä asiakirjojen ja kirjeiden lisäksi Hausen on kerännyt mukaan erilaisista rekistereistä tili- ja pöytäkirjamerkintöjä, joissa usein vain referoidaan alkuperäisen asiakirjan sisältö lyhyesti. Tyypillinen esimerkki ovat Tukholman raastuvanoikeuden pöytäkirjat ja maakirjat, joista on poimittu kaikki maininnat suomalaisista, jopa silloinkin, kun he ovat vain todistajia tai heidät mainitaan naapurikiinteistön omistajina. Projektissa toteutetun asiasanoituksen avulla tuodaan käyttäjälle esiin yksinkertaisessa muodossa sekä yksittäisen asiakirjan sisältö ja konteksti että koko tietokannan sisältämä

myös Codices Fennici -hankkeessa. Näiden kahden tietokannan metadata olisi linkitettävissä, vaikka ne eroavat toteutustavoiltaan.

[30] Diplomatarium Danicumin asiasanoitus käsittää vain kuusi aihetta (*emner*), joilla voi rajata aineistoa: *afladsbreve*, *arveret*, *kirkeret*, *mageskifte*, *regnskab*, *pantebrev*, *skøder*, *testamenter*. Asiasanoitus ei siis kata kuin osan asiakirjojen tyypeistä ja antajatahoista. Asiasanoitus ei myöskään koske koko aineistoa: 5685 asiakirjasta ainoastaan 706:lla on asiasana.

[31] Vaikka asiasanoitusta tehtiin kahden eri kategorian mukaan, ne on tietokannassa annettu samassa valikoissa, jotta käyttöliittymä pysyisi mahdollisimman yksinkertaisena. Ainoastaan ”Diplomataariset” ja “Muut lähteet” (ei-diplomataariset) on erotettu toisistaan erillisiin valikoihin, sillä ei-diplomataarisen aineiston osalta asiasanoitusta voitiin tehdä vain asiakirjan tyypin mukaan.

[32] Diplomataareissa asiakirjoja on historian saatossa luokiteltu usein eri tavoin, ks. esim. Piippo 2009, 395–397.

materiaalien kirjo. Lisäksi asiasanoitus mahdollistaa aineiston käsittelyn osajoukkoina tyyppien ja asiasisällön mukaan.

Tekstinlouhinta ja haut

Metadatalla, kuten asiasanoilla, pystytään edesauttamaan hakuja ja ryhmittelemään asiakirjoja suuremmiksi kokonaisuuksiksi, mutta asiakirjojen yksityiskohtainen tekstisisältö, kuten tekstissä mainitut henkilöt ja paikat, jäävät kuitenkin suurilta osin näiden koko asiakirjaa määrittävien, kattotason tietojen ulkopuolelle. Vaikka digitaalisessa muodossa olevat editiot mahdollistavat haluttujen sanojen hakemisen suoraan tekstihaulla, hakemista vaikeuttavat tietokannan kielijakauma ja ortografian vaihtelu.[33] Samojen henkilön- ja paikannimien kirjoitusasun vaihtelu asiakirjasta toiseen on huomattavaa, esimerkiksi Henrik voidaan kirjoittaa myös muodoilla Heinrich, Hinrik, Hinrich, Henricus.[34] Tästä johtuen yksittäisten nimettyjen entiteettien, keskeisten termien ja sanojen löytäminen asiakirjoista on erittäin työläs prosessi, joka vaatii tutkijalta suunnattoman paljon aikaa vievää dokumenttien läpikäymistä. Paikat ja henkilöt ovat kuitenkin tutkijoiden kiinnostuksen kannalta keskeisimpiä hakukohteita, tiettyjen sisältöä määrittävien asiasanojen lisäksi.
Vastauksena näihin haasteisiin DF-tietokannassa kehitettiin kolme eri työkalua, jotka tukevat toisiaan.[35] Ensinnäkin hakumoottori vaihdettiin sumeat haut mahdollistavaksi. Käyttäjä voi antaa termille tarkkuusarvon ja hakumoottori hakee kaikki tarkkuusarvon sisään mahtuvan osuman sisältämät asiakirjat.[36] Toisena työkaluna tietokantaan toteutettiin konkordanssihaku. Käyttäjä voi hakea tiettyä sanaa, käyttäen edellä mainittua sumean haun operaattoria. Hakutulos on kaikki haetun sanan esiintymät korpuksessa tekstirivinä, jossa sana näkyy keskimmäisenä.[37] Näin sanojen esiintymät on suoraan nähtävissä kontekstissaan, ja tutkija saa hakutuloksista luettua itseään

[33] Teksteistä valtaosa on joko latinaksi (n. 1600 kpl), keskiaikaiseksi ruotsiksi (n. 3000 kpl) tai keskialasaksaksi (n. 1200 kpl), mutta asiakirjoissa esiintyy myös laajalti sekakielisyyttä ja muitakin kieliä (islanti, venäjä jne.). Lisäksi keskiajalla kirjoitetun kielen ortografia ei ollut vakiintunut kuin latinassa, pois lukien alkujaan kansankielisten henkilön- ja paikannimien latinalaiset muodot. Esimerkiksi suomenkieliset nimet ovat usein taipuneet lähes tunnistamattomiksi ruotsinkielisen kirjurin kirjoittaessa ne latinalaistettuun muotoon. Ortografia, kieli ja nimien kirjoitusmuodot ovat haasteita myös perinteisessä indeksoinnissa, vrt. Kate Mertes, Holding hands with the past. Indexing historical documents. *The Indexer* 31:3 (2013), 96, 98–99.
[34] Yksittäisellä nimellä voi olla tietokannassa useita kymmeniä eri kirjoitusasuja.
[35] Kehitetyt työkalut valikoituivat tutkijayhteisön toiveiden mukaisesti. Käyttäjätestausta suoritettiin alusta asti siten, että ohjelmoija kävi läpi tietokannan käyttöä konkreettisessa tutkimustilanteessa tutkijoiden kanssa. Tutkijoilta kerättiin palautetta, kehitystoiveita sekä referenssejä valmiisiin toteutustapoihin, joista DF:aan valikoitiin resurssien puitteissa toteutuskelpoisimmat.
[36] Sumean haun tarkkuusarvo annetaan välillä 0.0–1.0. Arvo 1 edellyttää täsmälleen samaa muotoa, ja mitä pienempi hakuarvo, sitä enemmän hakutulokseen sisällytetään varianssia. Esim. hakutermillä ”Erik~0.4” haetaan sanaa ”Erik” tarkkuusarvolla 0,4, jolloin hakutuloksiin sisältyvät myös esim. “Eryk” ja “Erick”.
[37] Esim. hakemalla konkordanssihaulla ”sockn”, yksi tulosrivi on ”att thet sijn iagh ting hullt medh almogen i Poijo sockn anno Domini 1456 opo sancti Hendrici affton i erlige” asiakirjassa DF 3003.

kiinnostavan sanan tai sen variantin esiintymät. Sumea haku ja konkordanssihaku mahdollistavat erilaisia hakutapoja, ja palvelevat historian tutkimuksen lisäksi esimerkiksi filologeja. Konkordanssihaun toteuttamiselle haettiin mallia muun muassa Språkbanken Korpista.[38]

Kolmantena hakuja helpottavana uudistuksena tietokannassa tunnistettiin nimettyjä entiteettejä sekä luotiin entiteettihaku.[39] Mallia entiteettihaun toteutukseen ja käyttöliittymään otettiin muun muassa Svenska Litteratursällskapetin *Zacharias Topelius Skrifter* -hankkeen tuottamista sähköisistä editioista.[40] Keskiaikaisten tekstien eri tavoin kirjoitetut nimivariantit ovat niin merkittävästi toisistaan poikkeavia, että jopa sumean haun kanssa niitä on vaikea löytää ilman huomattavaa määrää virheosumia.[41] Tietokannan tekstisisältöihin kohdistettiin sekä koneellista että käsin tehtävää entiteettien ja niihin liittyvien määreiden poimintaa. Eri varianttikirjoitusasuja tunnistettiin ja yhdistettiin moderniin standardimuotoon.[42] Lopuksi tietokantaan toteutettiin entiteettihaku, jolla saa haettua kuhunkin standardimuotoon yhdistettyjä varianttimuotoja sisältävät asiakirjat. Lisäksi käyttäjä voi asiakirjanäkymässä korostaa asiakirjasta tunnistetut paikannimet, henkilönimet sekä näihin liittyvät määreet, ja hakea tunnistetut varianttimuodot kunkin standardimuodon takaa. Erilaisia entiteettien varianttimuotoja tunnistettiin yli 23 000 ja ne esiintyvät teksteissä yli 200 000 kertaa. Entiteettien tunnistus kohdistettiin kuitenkin vain yleisimmin esiintyviin sanoihin.[43] Harvoin

[38] https://spraakbanken.gu.se/korp (18.8.2019).

[39] Nimetyt entiteetit (*Named entites*) on termi, jota käytetään tekstissä nimetyistä toimijoista, paikoista, instituutioista ja muista nimetyistä kokonaisuuksista, kuten teoksista, tuotemerkeistä jne. Entiteettien tunnistamisesta ohjelmallisesti käytetään termiä Named Entity Recognition (NER). Tekstinlouhinnan työvälineeksi valikoitui Python-ohjelmointikieli, josta entiteettien tunnistamiseen käytettiin erityisesti Natural Language Toolkit (NLTK) -kirjastoa. Entiteettien varianttimuotoja poimittiin Pythonilla koneellisesti monin eri räätälöidyin hauin. Esimerkiksi pitäjien nimiä tunnistettiin aineistosta etsimällä sanan ”socken” eri varianttien kanssa esiintyviä isolla alkukirjaimella alkavia sanoja (Hausen on editioissaan standardisoinut erisnimet alkamaan isolla alkukirjaimella). Lisäksi variantteja poimittiin teksteistä käsin. Poimittujen muotojen yhdistäminen ja standardisointi suoritettiin sekä käsin että OpenRefine-ohjelman sisältämiä sanojen klusterointimetodeja ja sanojenyhdistämisalgoritmeja käyttäen (mm. Fingerprint, N-gram fingerprint, Phonetic fingerprint, Levensteihn distance, PPM). Standardimuotoja ja varianttimuotoja vertailtiin myös mm. Sveriges medeltida personnamn (SMP) -tietokantaan sekä muihin nimiaineistoa sisältäviin resursseihin (mm. Kotuksen verkkoaineistot). Niiden käyttöä kuitenkin rajoittivat DF:n varianttien paljous ja SMP:ssä tehty suomenkielisten nimien poissulkeminen. Lisäksi nimiaineistoresursseja ei ole juuri kerätty keskiaikaisesta aineistosta. Hakujen perustana olevat tunnistetut standardimuodot ovat omina tauluinaan SQL-tietokannassa. Entiteettien varianttimuodot teksteissä on yhdistetty standardimuotoihinsa tekstin sisäisin XML-tägein, jotka ajettiin tunnistusaineiston perusteella takaisin tekstiaineistoon. Tägäysajoon syötettiin myös erinäisiä sääntöjä esimerkiksi homonyymisten henkilön- ja paikannimien erottamiseksi. Pitkälle viedyn automatisoinnin ansiosta prosessi on kokonaisuudessaan toistettavissa, mikäli esimerkiksi tunnistusaineistoa jalostetaan ja laajennetaan edelleen, tai mikäli XML-tägien skeemaa halutaan tulevaisuudessa vaihtaa.

[40] http://www.topelius.fi (18.8.2019).

[41] Esim. hakemalla ”Erik” riittävän sumealla haulla saa myös osumat sanoista ”rike” tai ”kaere”.

[42] Esim. ”Erich” ja ”Ericus” yhdistettiin standardimuotoon ”Erik”.

[43] Entiteettipoiminta suoritettiin sana-aineistossa, jotka esiintyvät tietokannassa vähintään 3 kertaa (25% uniikista sana-aineistosta, 99% kaikista tietokannan sanoista). DF-tietokannan 1,45 miljoona sanaa koostuvat noin 165 000 uniikista sanasta; näistä uniikeista vain kerran esiintyviä on 98 000 (lähes 60%) ja kaksi kertaa esiintyviä sanoja yli 23 000 (n. 14%). 1 tai 2 kertaa esiintyvät sanat kattavat siis lähes 74% kaikista uniikeista sanoista, mutta toisaalta vain noin yhden prosentin tietokannan kaikista sanoista, noin 145 000 esiintymällään. Sekakielisyys ja kielitietojen puutteellisuus tarkoittivat lisäksi sitä, että aineistoa ei voitu käsitellä kielikohtaisesti eri kriteerein. Tämä tarkoittaa sitä, että osa

esiintyvien entiteettien kohdalla lista ei siis ole täydellinen. Ratkaisu on tehty tietokannan käytettävyyttä ajatellen. Yksittäiset sanat ovat löydettävissä sekä normaaleilla konkordanssi- että sanahauilla, eikä niiden standardisointi tai modernisointi olisi usein kovinkaan suoraviivaista.[44] Aikaa myöten aineistoa ja entiteetintunnistusta voidaan parantaa ja saada kaikista teksteistä poimittua merkittävät tunnisteet. Tällaisenaan ne toimivat lähinnä asiakirjojen löytämisen apuna, mutta jatkossa yhdistettynä metadataan ne voitaneen jalostaa ja yhdistää kattavaksi paikka- ja henkilörekisteriksi.

Tulevaisuus ja kehitys

Kansalliset editiosarjat, kuten FMU ja *Diplomatarium Suecanum*, ovat osa oman aikansa kansallisen historian rakentamista. Sarjoihin on poimittu asiakirjoja useista eri arkistoista ja arkistosarjoista, ja samalla usein asiakirjan alkuperäinen konteksti on kadonnut. Kun aineisto on koottu anakronistisin perustein kansallisiin editiosarjoihin, se on myös pirstaloitunut.[45] Esimerkkinä FMU:hun on koottu suomalaisia koskevia mainintoja kaikkialta Itämeren alueen arkistoista, mutta nämä maininnat jäävät usein hyvin irrallisiksi vailla kontekstualisoivaa tietoa. Toisaalta Suomea koskevaa keskiaikaista aineistoa on koottu eri instituutioiden toimesta myös muihin tietokantoihin aineiston tyypin mukaan: Suomalaisen Kirjallisuuden seuran ylläpitämä *Codices Fennici*[46] kokoaa yhdeksi digitaaliseksi kokoelmaksi keskiaikaiset käsikirjoitukset ja Kansalliskirjaston *Fragmenta membranea* -tietokanta[47] sisältää tiedot fragmentteina säilyneistä käsikirjoituksista. Syntesoiva ote asiakirjakokoelmiin tarjoisi mahdollisuuden yhdistää irrallaan olevaa ja aiemmin erotettua. Tulevaisuudessa tietojen yhdistely eri verkkotietokantojen välillä saattaakin olla mahdollista, mutta se vaatii yhteistyötä tietokantoja ylläpitävien toimijoiden kesken.[48] Syitä tämänkaltaisen holistisen

entiteeteistä on virheellisesti tunnistettu. Virhetunnistusten laskennallinen osuus on kuitenkin varsin pieni. Aineisto haluttiin julkaista keskeneräisessä ja epätäydellisessä muodossaankin hakujen tueksi.

[44] Esimerkiksi yksittäisen, kerran esiintyvän paikannimen standardisoiminen edellyttäisi käytännössä paikan tunnistamista ja yksilöimistä, jotta modernisoitu kirjoitusasu voitaisiin antaa oikein.

[45] Suomen kansallisen keskiaikaisen historian luonnista ja arkistokokoelmien editoinnista esim. Jussi Nuorteva & Päivi Happonen, *Suomen Arkistolaitos 200 vuotta. Arkivverket i Finland 200 år.* Kansallisarkisto 2016, 44–48, 79–85.

[46] https://www.codicesfennici.fi/ (18.8.2019).

[47] https://fragmenta.kansalliskirjasto.fi/ (18.8.2019). Fragmenta-hankkeesta, ks. Liisa Savolainen, Pergamenteista tietokannaksi. Digitaalisen sisältötuotannon haasteet. *Signum* 5 (2012), 20–23. Fragmenta-hankkeen tuloksia ja yleisemmin Suomen keskiajan kirjallisia lähteitä käsittelee Seppo Eskola et al., *Kirjallinen Kulttuuri Keskiajan Suomessa*. SKS 2010. Ruotsin vastaava käsikirjoitusfragmenttitietokanta on Medeltida pergamentomslag: https://sok.riksarkivet.se/MPO (18.8.2019). MPO-hankkeesta, ks. Jan Brunius, *From Manuscripts to Wrappers. Medieval Book Fragments in the Swedish National Archives: Archival Guide*. Riksarkivet 2013.

[48] Katja Fält näkee erilaisten aineistojen toisiinsa linkittämisen olevan keskeisessä asemassa humanistisen tutkimuksen tulevaisuuden näkymissä. Linkittäminen avaisi ja täydentäisi eri kulttuuriperintölaitosten kokoelmien välisiä suhteita ja tehostaisi aineistojen käyttöä. Hän visioi myös, että tutkijoiden arkistoaineistojen pohjalta tuottamat jatkojalostetut materiaalit voisivat olla omana kokonaisuutenaan, joka olisi linkitetty arkistoaineistoihin. Katja Fält, Tutkijoiden

lähtökohdan esittämisen puolesta on useita, joista keskeisimpinä on mainittava ekonomisuus ja aineiston yhteneväisyys, jotka ovat saman kolikon kaksi puolta. Esimerkiksi Itämeren alueen kansallisten diplomataarien aineisto jakaa provenienssinsa, minkä vuoksi on resurssien tuhlausta ja hajanaisuutta lisäävää käsitellä asiakirjoja monin eri kriteerein erillisissä kokoelmissa. Asiakirjoissa mainitut henkilöt, paikat ja instituutiot jakautuvat nykyisten kansallisten rajojen yli. Mahdollisuudet korpusten yhdistämiseen niin metadatan kuin asiakirjojen itsensäkin tasolla ovat ilmeiset.[49] Itsestään selvä lähtökohta olisi ensin kartoittaa, kuinka suuri osa asiakirjoista on jo julkaistu erinäisissä muissa tietokannoissa, ja koota näistä syntesoiva verkkotietokanta, jonka puitteissa nykyiset kansalliset kokoelmat voitaisiin esittää erillisinä virtuaalisina kokoelmina.

Toisaalta modernit digitaaliset menetelmät tarjoavat uusia keinoja käyttää ja yhdistää tietoa, joka oli aikaisemmin hankalasti kerättävissä.[50] Rikastettua metadataa ja digitaalisia tekstisisältöjä voidaan käyttää suoraan pohjana hyvinkin erilaisissa tutkimuksissa.[51] Geokoodatut paikkatiedot antavat mahdollisuuden tarkastella aineistoa ja sen sisältöä maantieteellisessä jakaumassa. Ajoitukset tuottavat ajallisen sarjan aineistosta. Asiasanat, henkilötiedot ja henkilöitä määrittävä metadata yhdistettyinä esimerkiksi aika- ja paikkatietoihin mahdollistavat jo hyvin laajoja uudenlaisia tutkimuksia. Mahdollista olisi tarkastella esimerkiksi Suomen keskiaikaiseen paikallishallintoon osallistuneiden ihmisten taustoja ja yhteyksiä ajallisessa ja paikallisessa sarjassa.[52]

Metadata esimerkiksi paikoista ja henkilöistä on jalostettavissa kattaviksi rekistereiksi, lisäämällä kutakin nimettyä entiteettiä itseään koskevaa metatietoa järjestelmään. Rekistereissä voidaan käyttää pohjana sekä tunnistettuja, standardisoituja antaja- ja antopaikkatietoja, mutta myös

digitaaliset tutkimusaineistot ja datan avoin jakaminen digitaalisen humanismin kontekstissa. Teoksessa Kimmo Elo (toim.) *Digitaalinen humanismi ja historiatieteet*. Turun historiallinen yhdistys 2016, 65.

[49] Yhdistäminen edellyttäisi samojen metadatastandardien käyttöä. Christian-Emil Ore, New digital assets. How to integrate them? Teoksessa Georg Vogeler (toim.) *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*. Böhlau Verlag 2009, 238–239.

[50] Kvantitatiivisen ja kvalitatiivisen, tai digitaalisen ja perinteisen humanistisen tutkimuksen suhde on suuren muutoksen alla. Tietokoneavusteisten menetelmien käytöstä ja merkityksestä tutkimukselle esim. Helle Porsdam, Digital Humanities. On Finding the Proper Balance between Qualitative and Quantitative Ways of Doing Research in the Humanities. *Digital Humanities Quarterly* 7:3 (2013).

[51] Digitaalisin menetelmin on suoritettu muun muassa latinankielisten asiakirjojen ajoittamista metadataan ja luonnollisen kielen prosessointiin (Natural Language Processing, NLP) perustuen, sekä kirjurien tunnistamista stylometrisin perustein. Michael Gervers & Michael Margolin, Managing Meta Data in a Research Collection of Medieval Latin Charters. Teoksessa Georg Vogeler (toim.) *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*. Böhlau Verlag 2009; Sidsel Boldsen & Patrizia Paggio, Automatic Dating of Medieval Charters from Denmark. *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries. Copenhagen, March 6–8 2019*, https://cst.dk/DHN2019Pro/DHN2019Proceedings.pdf (13.5.2019).

[52] Useissa hankkeissa on jo suoritettu verkostoanalyysiä, esimerkiksi keskiaikaisten asiakirjojen antajien ja vastaanottajien kesken tai teoksissa mainittujen henkilöiden kesken. Esim. Robert Gramsch et al., Medieval Historical, Hagiographical and Biographical Networks. Teoksessa Ralph Kenna, Máirín MacCarron & Pádraig MacCarron (toim.) *Maths Meets Myths. Quantitative Approaches to Ancient Narratives*. Springer 2017.

tekstilouhittuja entiteettejä. Tunnistettujen entiteettien jatkojalostamisen kautta voitaisiin rakentaa esimerkiksi henkilörekisteri, johon saataisiin listattua kunkin mainitun henkilön jokainen esiintymä asiakirjoissa, ei vain antajan roolissa.[53] Toisaalta korpuslingvistiikan keinoin voidaan aineiston vaihtelevan ortografian asettamat haasteet selättää: aineiston teksti voidaan sana sanalta palauttaa perusmuotoihinsa eli lemmatisoida, mikä mahdollistaa koko tekstikorpusta koskevat standardisoidut tekstihaut.[54]

Tietokanta tarjoaa uudenlaisia mahdollisuuksia, mutta myös vastuun siitä, että sitä kehitetään ja ylläpidetään jatkuvasti.[55] Uudet sisällöt eivät synny itsessään, tiede menee eteenpäin ja verkkotietokannat vanhentuvat käyttöliittymältään usein alle vuosikymmenessä, ellei niitä päivitetä ja pidetä ajan tasalla.[56] Internet on pullollaan eri projektien tuottamia tietokantoja, joiden tavoitteet ovat olleet kunnianhimoisia, mutta päivitys on loppunut samaan päivään kuin projektin rahoitus ja vanhentuneina ne eivät enää kunnolla palvele tarkoitustaan.[57] DF-projektissa aineiston säilyvyyttä on pohdittu jo projektin aikana ja siksi tietokannan kolmikerrosmallissa tietokantakerros, logiikkakerros ja käyttöliittymäkerros on erotettu toisistaan. Näin käyttöliittymä on suhteellisen helposti muutettavissa, mutta data eli tietokantakerros säilyy koskemattomana. Aineiston pitkäaikaissäilytyksestä ja tietokannan ylläpidosta vastaa Kansallisarkisto. Lisäksi jatkokehitystä varten on perustettu toimituskunta, joka vastaa hankkeen tulevaisuuden suunnittelusta.[58] Sisällönkehittämisen osalta yksi mahdollisuus on osallistaa tutkijayhteisö kehitystyöhön, jolloin käyttäjien asiantuntemusta saataisiin hyödynnettyä tutkimustietokannan kehitystyössä.[59] Uusien

[53] Esimerkiksi The Making of Charlemagne's Europe -hankeessa on poimittu Kaarle Suuren ajan asiakirjoista kaikki niissä esiintyvät henkilö- ja paikkatiedot kattaviksi rekistereiksi, jotka mahdollistavat mm. prosopografisen tutkimuksen, http://www.charlemagneseurope.ac.uk/ (18.8.2019).
[54] Esimerkiksi Språkbanken Korp: https://spraakbanken.gu.se/korp/ (18.8.2019).
[55] Miguel Escobar Varela, The Archive as Repertoire. Transscience and Sustainability in Digital Archives. *Digital Humanities Quarterly* 10:4 (2016).
[56] Fält katsoo, että metatietojen standardisointi helpottaa niiden käsittelyä ja digitaalinen humanismi tarvitsee kukoistaakseen datan ja tutkimusaineistojen huoltamista. Fält 2016, 62–63. Yhtenä esimerkkinä digitaalisen ympäristön nopeasta muutosvauhdista on Paolo Buonoran artikkeli 10 vuoden takaa, jossa hän argumentoi CD-ROMien eduista sähköisessä pitkäaikaissäilytyksessä. Paolo Buonora, Long lasting digital charters. Storage, formats, interoperability. Teoksessa Georg Vogeler (toim.) *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*. Böhlau Verlag 2009.
[57] Digitaalisen datan hallinnasta ja säilyvyydestä käydään tällä hetkellä runsaasti keskustelua tutkimuksessa, esim. Jessica Parland-von Essen et al., Supporting FAIR data. Categorization of research data as a tool in data management. *Informaatiotutkimus* 37/4 (2018).
[58] Toimituskuntaan kuuluvat pj Dos. Tuomas Heikkilä (Helsingin yliopisto), vpj FT Tapio Salminen (Tampereen yliopisto), sihteeri FM Lauri Leinonen (Helsingin yliopisto) sekä jäseninä FM Seppo Eskola (Helsingin yliopisto), FM Maria Kallio (Turun yliopisto), Prof. Anu Lahtinen (Helsingin yliopisto), Prof. Harry Lönnroth (Jyväskylän yliopisto), Prof. Kirsi Salonen (Turun yliopisto) ja FL Minna Sandelin (Turun yliopisto).
[59] Digitaalisten työkalujen kehittämisessä käyttäjien osallistaminen jo varhaisessa vaiheessa on ensiarvoisen tärkeää. Fred Gibbs & Trevor Owens, Building Better Digital Humanities Tools. Toward broader audiences and user-centered designs. *Digital Humanities Quarterly* 6:2 (2012).

tutkimusten tuottamaa tietoa voitaisiin lisätä DF:iin tai jopa joukkoistaa sisällön tuottamista ja muokkaamista, mutta tällöin haasteena on tiedon oikeellisuuden varmistaminen.[60] Uusien sisältöjen lisääminen on kuitenkin digitaalisessa tutkimustietokannassa teknisesti helppoa, koska pyrkimyksenä ei ole tuottaa yhtä ”täydellistä editiota” vaan tarjota laajasti olemassa olevat aineistot tutkijoiden käyttöön ja rikastaa niitä, jolloin varhaisemmat versiot ja vaihtoehtoiset tutkimustulokset voidaan esittää rinnakkain.

Johtopäätökset

Modernit tutkimusinfrastruktuurit ja tiedonjalostusmenetelmät mahdollistavat uusia tapoja esitellä ja käsitellä historiallista aineistoa. Editioiden esittäminen rinnakkain yhdessä lähteen digitaalisen jäljennöksen kanssa mahdollistaa erilaisten luenta- ja tulkintatapojen tarkastelun yhdellä alustalla. Tämä tekee näkyväksi aineiston tutkimushistoriaa, ja on merkittävää paitsi tulevalle aineistoa koskevalle tutkimukselle myös historiografisesti. Eri aineistoja, lähteitä, editioita, tietokantoja, metadataa ja tutkimuskirjallisuutta yhdistävällä tutkimustietokannalla tuotetaan uuden tietokerroksen sijaan syntesoiva rakenne, joka kokoaa hajallaan olevat tiedot yhteen. Tutkimustietokantojen etu verrattuna yksittäisiin editioihin onkin kiistattomasti juuri tiedon koostamisessa yhteen paikkaan.

Tiedon yhteen tuomisen lisäksi tietoa voidaan tutkimustietokannassa jalostaa keskitetysti. Tiedon löydettävyyttä ja analyysiä helpottavan metadatan tuottaminen on tästä paras esimerkki. Modernit menetelmät antavat myös mahdollisuuden yhdistää uudella tavalla aineistoja vanhoja kansallisia editiosarjoja suuremmiksi kokonaisuuksiksi. Seuraava looginen askel on hakea järjestelmällisempää ylikansallista yhteistyötä kansallisten lähdevarantojen laajempaan käytettäväksi saattamiseen ja metatietojen ja asiakirjojen yhdistämiseen.

Huomionarvoista on, että tutkimustietokannan tietosisällön jalostaminen, syntetisointi sekä metadatatyö ovat arvokasta perustutkimusta, joka vaatii aineiston tuntemusta ja asiantuntijoiden työpanosta. Editioiden ja lähteiden jatkojalostaminen voidaan nähdä nyky-tutkijasukupolven

[60] Joukkoistamisen haasteena on ennen kaikkea käyttäjien lisäämän tiedon todenmukaisuuden varmistaminen, mikä on tieteelliselle tutkimustietokannalle toimintaedellytys, mutta myös resurssikysymys. Hyviä kokemuksia joukkoistamisesta on saavutettu niin Suomessa kuin kansainvälisestikin erityisesti kirjastomaailmassa. Christina Manzo et al., ”By the People, For the People”. Assessing the Value of Crowdsourced, User-Generated Metadata. *Digital Humanities Quarterly* 9:1 (2015). Digitaalisten editioiden haasteista ja mahdollisuuksista yleisemmin, ks. Dahlström 2010, 80–86; David J. Birnbaum, Sheila Bonde & Mike Kestemont, The Digital Middle Ages. An Introduction. *Speculum* 92:1 (2017), 5, 12–13.

editiotyönä, edeltäjien työn parantamisena ja nostamisena uudelle tasolle, ja sikäli se on koko historiantutkimuksen kentälle arvokas päämäärä.

FM, väitöskirjatutkija, Olli-Pekka Kasurinen, Suomen ja Pohjoismaiden historia, Helsingin yliopisto, olli.kasurinen@helsinki.fi
FM, väitöskirjatutkija, Lauri Leinonen, yleinen historia, Helsingin yliopisto, lauri.i.leinonen@helsinki.fi
FM, väitöskirjatutkija, Sanna Supponen, yleinen historia, Helsingin yliopisto, sanna.supponen@helsinki.fi